Methods, Limitations and Applications of Reversible Phase Demodulation Based Order-Tracking of Rotating Machines

Author:
Coats, Michael

Publication Date:
2015

DOI:
https://doi.org/10.26190/unsworks/18151

PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: Coats

First name: Michael | Other name/s: David

Abbreviation for degree as given in the University calendar: PhD

School: Mechanical & Manufacturing | Faculty: Engineering

Title: Methods, Limitations and Applications of Reversible Phase Demodulation Based Order-Tracking of Rotating Machines

Abstract 350 words maximum: (PLEASE TYPE)

Order-tracking is a method to remove speed fluctuations from a varying frequency vibration signal allowing constant frequency based Machine Condition Monitoring (MCM) analysis techniques to be employed.

This thesis describes the generalised Phase Demodulation based Order-Tracking (PDOT) methodology whereby the "mapping" between rotation angle and time is obtained by phase demodulation of a reference signal.

A variety of reference signal types can be used, including a tachometer, shaft encoder, or an extracted reference signal from the response data signal itself. Each has different properties, which have to be taken into account. The primary advantage of PDOT, as opposed to methods based in the time domain, is that in principle it gives samples of the true relationship between rotation angle and time. However, there has to be no aliasing, in the sense of overlap of sidebands of the order to be demodulated with those of higher orders in the frequency domain. This thesis defines the conditions for which this is valid. The PDOT method can be employed using a single stage, or the result improved by using progressive iterations in a multi-stage approach, even with large speed variations. For iterations to give an improvement there must be no aliasing at the first stage.

A summary of the basic PDOT method is presented, highlighting the maximum speed variations of approximately ±30% which can be compensated for. Then the generalised PDOT method is presented, which can be employed in a modular fashion, using a variety of reference signal types and numbers of stages. One new development is using a segmented approach for very large speed variations, and another is the ability to reverse an order-tracked signal back to the time domain.

Multiple experimental examples are presented for different applications of PDOT, highlighting the suitability of PDOT for many variable-frequency applications. Examples include bearing diagnostic and gear diagnostic applications, in the presence of small to large speed variations, the latter using a segmented approach to order-track a run-up signal. Another uses PDOT in a reversible fashion, combined with cepstrum editing techniques, to pre-process a signal for subsequent Operational Modal Analysis (OMA).

Methods, Limitations and Applications of Reversible Phase Demodulation Based Order-Tracking of Rotating Machines

Michael David Coats

B.E.

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Mechanical & Manufacturing Engineering

Faculty of Engineering

The University of New South Wales

April 2015

Originality Statement

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed ………………………………………………..............

Date ………………………………………………..............

Copyright Statement

Authenticity Statement

‘I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.’

Signed ……………………………………………….............

Date ……………………………………………….............

Abstract

Order-tracking is a method to remove speed fluctuations from a varying frequency vibration signal allowing constant frequency based Machine Condition Monitoring (MCM) analysis techniques to be employed.

This thesis describes the generalised Phase Demodulation based Order-Tracking (PDOT) methodology whereby the “mapping” between rotation angle and time is obtained by phase demodulation of a reference signal.

A variety of reference signal types can be used, including a tachometer, shaft encoder, or an extracted reference signal from the response data signal itself. Each has different properties, which have to be taken into account. The primary advantage of PDOT, as opposed to methods based in the time domain, is that in principle it gives samples of the true relationship between rotation angle and time. However, there has to be no aliasing, in the sense of overlap of sidebands of the order to be demodulated with those of higher orders in the frequency domain. This thesis defines the conditions for which this is valid. The PDOT method can be employed using a single stage, or the result improved by using progressive iterations in a multi-stage approach, even with large speed variations. For iterations to give an improvement there must be no aliasing at the first stage.

A summary of the basic PDOT method is presented, highlighting the maximum speed variations of approximately ±30% which can be compensated for. Then the generalised PDOT method is presented, which can be employed in a modular fashion, using a variety of reference signal types and numbers of stages. One new development

is using a segmented approach for very large speed variations, and another is the ability to reverse an order-tracked signal back to the time domain.

Multiple experimental examples are presented for different applications of PDOT, highlighting the suitability of PDOT for many variable-frequency applications. Examples include bearing diagnostic and gear diagnostic applications, in the presence of small to large speed variations, the latter using a segmented approach to order-track a run-up signal. Another uses PDOT in a reversible fashion, combined with cepstrum editing techniques, to pre-process a signal for subsequent Operational Modal Analysis (OMA).

Acknowledgements

First and foremost, I would like to thank my supervisor Emeritus Professor Robert Bond Randall of the University of New South Wales, Australia, for his assistance, guidance and patience during the course of this research project. Your help has been invaluable.

I would like to thank Dr. Nader Sawalhi and Dr Wade Smith, who were both Research Associates at UNSW during my PhD candidature. Your input, advice, and assistance on multiple research topics, from throughout my candidature, is greatly appreciated.

With many thanks, I acknowledge the financial support from the ARC (Australian Research Council), through Discovery Grant DP0878266, during the initial part of my candidature.

Additionally, I would like to thank the DSTO (Defence Science and Technology Organisation), through the Centre of Expertise for Helicopter Structures and Diagnostics at UNSW, for their financial support during the latter part of my candidature, and for the supply of some experimental signals used in this work.

I would also like to thank the NSW division of the Australian Acoustical Society for their financial support, which allowed me to attend multiple conferences to present my research work.

I would especially like to thank Mr. Russell Overhall for his assistance with laboratory testing, and additionally for sharing his additional insights into vibration theory from a perspective founded on a practical approach. This was markedly different from the perspective in common sources, based on a theoretical approach, and having the opportunity to learn the concepts from both perspectives has given me an understanding much greater than would have been possible with either approach alone.

I would also like to thank Mr. Steven Kuhle and Mr. Allan Coats for their assistance with laboratory testing, in particular assisting with electrical based modifications and repair to the UNSW gear testrig, allowing experimental signals used in this work to be recorded.

In addition, I would like to thank my colleagues, fellow PhD students Lav Deshpande and Karel Ruber, and additionally Tim Philipson, for their assistance and support through many facets of the research work which has culminated in this thesis.

Last but not least, I would like to thank my parents for their patience, understanding and support through my entire university career.

Table of Contents

List of Publications

Peer Reviewed Journal Articles

[1] Coats, M.D. and R.B. Randall, Single and multi-stage phase demodulation based order-tracking. Mechanical Systems and Signal Processing, 2014. 44(1–2): p. 86-117.

[2] Randall, R.B., N. Sawalhi, and M.D. Coats, A comparison of methods for separation of deterministic and random signals. The International Journal of Condition Monitoring, 2011. 1(1): p. 11-19.

Peer Reviewed Conference Papers

[3] Randall, R.B., N. Sawalhi, and M. Coats, Separation of Gear and Bearing Fault Signals from a Wind Turbine Transmission under Varying Speed and Load, in Proceedings of CMMNO'2012 – Condition Monitoring of Machinery in Non-Stationary Operations 2012. p. 3-12.

[4] Coats, M.D. and R.B. Randall, Order-Tracking with and without a tacho signal for gear fault diagnostics, in Proceedings of Acoustics 2012 Fremantle, Fremantle, Australia, 21–23 November. 2012.

[5] Coats, M.D. and R.B. Randall, Compensating for speed variation by order tracking with and without a tacho signal, in In Proceedings of VIRM10 – Tenth International conference on Vibrations in Rotating Machinery, London, United Kingdom, 11-13 September. 2012.

[6] Coats, M.D. and R.B. Randall, Single Record Order-Tracking, in In Proceedings of CM 2010 and MFPT 2010, Ettington Chase, Stratford-upon-Avon, England, 22–24 June. 2010.

[7] Coats, M.D., N. Sawalhi, and R.B. Randall, Extraction of tacho information from a vibration signal for improved synchronous averaging, in Proceedings of Acoustics 2009, Adelaide, Australia, 23–25 November. 2009.

<u>Peer Reviewed Conference Papers, accepted for presentation and publication</u>

[8] Coats, M.D. and R.B. Randall, Order tracking over a wide speed range such as run-up or run-down, in the 9th IFToMM International conference on Rotor Dynamics, Milan, Italy, 22-25 September. 2014.

List of Figures

List of Tables

Chapter 1 Introduction, Literature review, and Theory

Chapter Overview

This chapter firstly gives an introduction to the work presented in this thesis, on phase demodulation based order tracking (PDOT).

A review of the literature on prior works is then presented, highlighting how the work presented in this thesis fits with prior works in the field of order-tracking.

1.1 Introduction

This section provides an overview of the concepts of order-tracking, and introduces the Phase-Demodulation based Order-Tracking (PDOT) methods developed in this thesis.

Firstly, the basics of frequency analysis of vibration response signals are discussed, and the reasons for often requiring the use of Order-Tracking with frequency analysis are explained. The basic principle of resample based order-tracking, or Computed Order-Tracking, is then presented. A summary of the basic single record order-tracking method which I developed for my undergraduate BE thesis in 2006 [9] is given.

The aim of this PhD thesis is then presented, which is to expand upon the work presented in my undergraduate thesis. An introductory summary is then given for the work presented in this thesis, and the layout of this thesis work is given.

1.1.1 Rationale for use of order-tracking

Vibration analysis is a method which can be used to determine the condition of rotating machinery – that is whether a machine has become worn, or is damaged. Vibration analysis used in this manner falls under the field of Machine Condition Monitoring (MCM). Vibration analysis is generally done by monitoring the vibration signals of a machine over time, and seeing if there is an increase in the level of vibration signal components.

Simple analysis in the time domain is limited by the fact that a vibration signal is the sum of many vibration components from different parts of a machine, each of which can increase in magnitude separately depending on which part has become damaged. These vibration components are often of widely different magnitudes. As increased fault severity results in a multiplicative increase in vibration levels from a specific component, if a vibration component which increases due to a fault is not already the largest in magnitude, its increase can be masked when looking at the overall vibration level, and damage or wear to the machine can go unnoticed, if only the latter is being monitored.

A method to overcome this is frequency analysis, where a vibration signal is transformed into the frequency domain. In the frequency domain, the vibration signal is separated into separate frequency components of differing magnitudes. It is then possible to monitor each frequency component to determine if there has been an increase in vibration magnitude at that frequency.

In order to monitor the vibration from each mechanical component properly in this manner, the fundamental vibration from each mechanical component should resolve into a single discrete frequency component. If a vibration component does not resolve into a discrete frequency component but rather has a varying frequency, it is not possible to easily determine its magnitude from the frequency spectrum, as the component is spread over multiple frequency values, giving an overall reduction in magnitude, and a lack of precision.

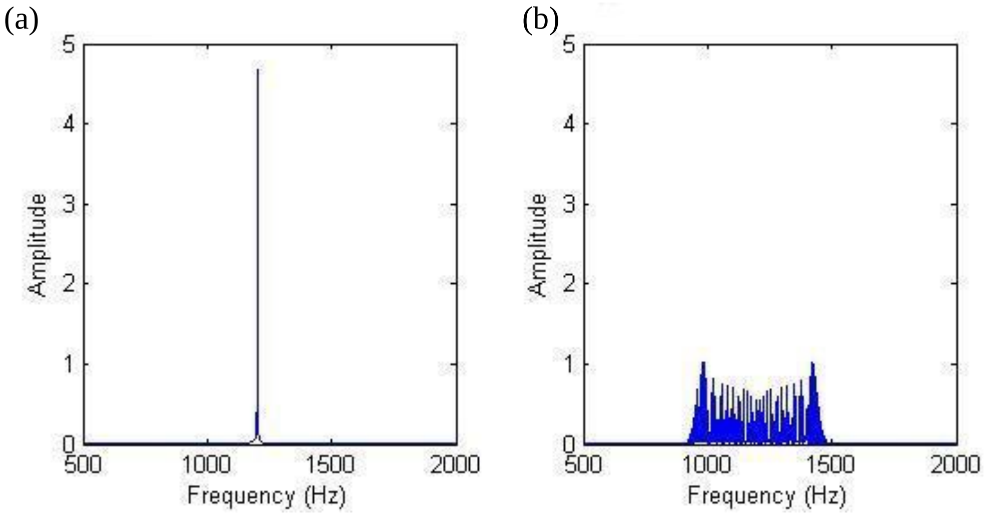


Figure 1.1 – Frequency spectra for vibration components
(a) constant frequency (b) 20% varying frequency

Figure 1.1 shows the frequency spectra for two vibration components. Both these components have exactly the same magnitude, so correspond to the exact same machine condition. However, one of these vibration components occurs at a constant frequency, the other at varying frequency. For the constant frequency case in Figure 1.1 (a), the component has resolved into a discrete value and the magnitude, or amplitude, of the component can be read easily from the diagram. For the varying frequency case Figure 1.1 (b), the vibration component has not resolved into a discrete value, and the overall magnitude of the component cannot be simply determined from the diagram.

In rotating machinery, many vibration components of interest are a function of shaft speed, or frequency. If shaft speed remains constant, the corresponding vibration component will be at a discrete frequency. Discrete frequency values will also result for every other component which is a function of shaft speed.

If shaft speed varies, as with variable speed machines, the resulting vibration components will not be at discrete frequencies and frequency analysis often cannot be directly used for meaningful analysis.

Most machines have a small amount of speed variation with variations in load even at nominally constant speed, which inhibits subsequent frequency analysis when fine analysis is to be carried out, such as time synchronous averaging (TSA) for gear diagnostics. Some machines, such as wind turbines, have much more widely varying operating speeds, which can prevent the direct application of frequency analysis altogether.

A way to gain discrete frequency values from variable speed machines is to change the axis of the vibration signal from equally sampled time increments to equally sampled rotation angle, or phase, increments. A machine shaft will always rotate 2π of phase in one revolution, regardless of shaft speed, giving a constant period along the axis for each rotation, and hence constant "frequency" in the angular domain.

Once a vibration signal has been (re)-sampled at equal phase increments, with units of revolutions of a reference shaft, the reference shaft vibration component will always have a discrete frequency value of exactly one. All other vibration components which are a function of the reference shaft speed will also now have discrete frequency values, and traditional frequency analysis of the resampled signal will be able to give meaningful results.

The name order-tracking is used to describe a method which gives a vibration signal sampled with constant angular sampling (or sampling rate in the angular domain). The name order-tracking comes from the fact that the spectrum of a vibration signal, with constant angular sampling, is expressed in terms of reference component harmonics (shaft harmonics), known as shaft orders. The constant angular sampled vibration signal is said to have been order-tracked, and the corresponding frequency spectrum is called the order spectrum. The term “order” was originally applied to integer harmonics but has been extended to fractional orders so as to apply to non-integer orders, such as bearing frequencies, and other shafts connected through a non-integer gear ratio.

From the order-spectrum, a vibration signal from a machine with varying speed can be subsequently analysed with traditional frequency analysis based MCM techniques, as if the vibration signal were captured from a constant speed machine.

A vibration signal from a machine with varying speed can be directly captured with constant angular sampling, however it is now more common to capture a vibration signal with equally spaced time samples (constant temporal sampling), and then subsequently resample the signal to have constant angular sampling, so as not to lose the information about timing

A method to order-track a signal by resampling, from an initial constant time sampled vibration signal, is called resample based order-tracking, or alternatively Computed Order-Tracking.

1.1.2 Basic principles of resample based order-tracking

In order to resample a vibration signal from equal time to phase increments, a map of phase (rotation angle) vs. time is used. The map of phase vs. time can alternatively be called the phase-time relationship, or phase-time map.

This phase-time relationship is constructed from information in a reference signal. The reference signal must be synchronously recorded with the vibration signal to be order-tracked, and contain an event which occurs with equal increments of phase. A commonly used reference signal is a once-per-revolution tacho pulse signal, but other types can also be used.

Time values corresponding to equal increments of phase are found using interpolation methods with the phase-time relationship. The vibration signal is then re-sampled at these calculated time values, again using interpolation methods. The resulting re-sampled vibration signal, sampled at equal increments of phase, can then be transformed into the "frequency" domain to produce an order-spectrum suitable for use with subsequent constant frequency based frequency analysis methods.

1.1.3 Outline of Single Record Order-Tracking Concept

For my undergraduate BE thesis from 2006 [9] I developed a computed order-tracking method based on phase demodulation. The developed method was titled 'Single Record Order-Tracking', and was built upon work by Bonnardot et al. [10].

In [10] Bonnardot et al. developed an order-tracking method using phase demodulation to conduct order-tracking of a signal, which used information extracted from a vibration signal using phase demodulation of a speed related component as a reference signal to conduct order-tracking, thus not requiring a separately recorded reference signal. The described method [10] utilised high gearmesh harmonics from the response vibration signal, which limited the permissible speed range to nominally constant speed (to avoid smearing of close components), and so only very small speed fluctuations could be compensated for. The work by Bonnardot et al. [10] is summarised in more detail in section 1.2.3.

The single record order-tracking method [9], based on the first harmonic of a synchronously recorded tacho signal, allowed order-tracking of signals with significant speed variations of up to approximately ±30%.

The single record order-tracking method from [9] presented results from simulated testing with single sinusoids, which indicated that highly accurate order-tracking was possible. However no testing with real signals was presented in [9].

1.1.4 Aims of PhD

The basic aim of this doctoral research work was to expand greatly upon the single record order tracking method I developed for my BE thesis [9], as described above in section 1.1.3.

The very first goal was to conduct experimental verification of the single record order tracking method, which had only been tested with simulated signals.

The second goal was to expand the basic single record order-tracking method to a more generalised method applicable to a variety of reference signal types, and to expand the underlying theory to fully encompass these reference types. Where the work and corresponding theory by Bonnardot et al. [10] was focused on using a response signal as a reference, and my own work in my BE thesis [9] was focused on using a once per revolution tachometer as a reference, the goal for this work was to develop a generalised method which could encompass both these reference signal cases and more.

The third goal was to then tailor the developed order-tracking method for a variety of MCM analysis applications. These applications would include gear and bearing analysis, run up signal analysis, and methods to facilitate subsequent Operational Modal Analysis. These analysis applications are relevant to a large number of industrial situations where non-stationary operating conditions are present, including the emerging area of wind turbine diagnostics.

1.1.5 Thesis overview and layout

This thesis describes a generalised methodology whereby order-tracking can be conducted using a Phase Demodulation based Order-Tracking (PDOT) approach for vibration signals, even in the case of large speed variations of up to approximately ±30% for a single continuous record.

This work was first presented by myself in the form of a generalised methodology in 2014 [1]. Some elements of the methodology have previously been presented in isolation in various formats [4-7, 9]. The generalised methodology as presented here includes some previously unpublished methods for PDOT.

The PDOT methodology can be employed with an external reference signal, or alternatively solely using the response vibration signal, allowing flexibility in implementation.

The PDOT methodology can also be employed with either one or multiple iterations, which are each suitable for use with a wide range of different subsequent diagnostic techniques, and result in accurate (to very high frequency) order-tracked signals unconstrained by resolution limitations, such as with time-frequency order-tracking approaches.

The PDOT methodology can also be employed with a segmented approach, allowing speed variations even greater than ±30%, such as found in run-up analysis, to be order tracked.

Finally, the PDOT methodology is also reversible, allowing for MCM methods which require operations to be conducted in the angle domain before transformation back to the original time domain.

Chapter 1 firstly presents this introduction to the thesis work. A summary of the prior literature in the field of order-tracking is given. This summary includes publications from this thesis work, highlighting how this thesis work fits within the body of work in the field of order-tracking.

Chapter 2, a theory section, is then presented, which contains mathematical formulae and methods which will be used within latter chapters of this work.

In Chapter 3, firstly the basic single pass PDOT method is presented. The single pass PDOT method forms the kernel of the generalised methodology, forming the basis upon which subsequent application specific variations are built. The operating limitations of the method are then defined, giving the maximum allowable speed variation and frequency of speed variation for various degrees of aliasing, which determines the extent to which the results can be improved by further iteration. Some of these limitation definitions based on harmonic proportionality have been previously published by myself in [1, 6, 9]. Limitation definitions developed without the assumption of proportionality of harmonics have not been previously published. Simulated results are then presented illustrating these limits of the PDOT method in the absence of noise and contaminating components. Initial experimental testing providing validation of the basic single pass PDOT method is also presented.

In Chapter 4, methods used to extend the basic PDOT method, from Chapter 3, to a generalised methodology applicable to a variety of applications are discussed.

The differences between different types of reference signal which can be used are first discussed. The difference between tacho signals (typically once-per-rev pulses) and shaft encoder signals (multiple pulses per rev) is discussed, in particular with respect to the difference between the signal from an N per rev encoder, and the Nth harmonic of a one per rev tacho. Tacho and shaft encoder signals are typically not amplitude modulated, whereas vibration signals often are, in particular over wide speed ranges, so the effects of this are discussed.

The concept of iterating the PDOT method in multiple passes to improve order-tracking is presented, and the advantages and disadvantages of either a single or multi-pass approach discussed.

The concept of segmenting a signal with the PDOT method is presented, allowing order-tracking of much higher speed variations, the limit of which is constrained only by available processing capability. The concept of reversing the order-tracking process with the PDOT method is also presented. Both the segmentation and reversible PDOT methods are previously unpublished.

In Chapter 5, all of the applications of order-tracking to the angular domain are illustrated using a range of different experimental results. Signals were captured from a gear test rig with either a gear fault, or a bearing fault, over a number of speed ranges up to ±25%. These experimental results are order-tracked using the PDOT method in multiple configurations, and subsequently analysed with appropriate diagnostic techniques resulting in correct fault identification, showing that this was possible in the presence of such large speed variations. At the same time the different requirements for gear and bearing diagnostics are illustrated.

Previously unpublished results of using a shaker test on the gearbox casing from the testrig for a run-up speed range are also presented, showing the success of using a segmented approach to PDOT. In addition to the results from the gear test rig, some results are also given for signals from a gas turbine engine, and a wind turbine, where the multiple stage approach was shown to give extremely high accuracy in determining the relative frequencies of the different components of the machine.

In Chapter 6, results are presented for a shaker test on the gearbox casing, which is used for Operational Modal Analysis. In addition to a stationary random excitation, the casing was excited by a large number of harmonics of a fundamental frequency varying by ± 15%, as with a typical variable speed machine. For this analysis, the PDOT method is used to transform the signal to the order domain where order components are filtered from the signal, which is then reversed back to the time domain to allow subsequent operational modal analysis in the absence of contaminating order components.

Chapter 7 presents the conclusions of the work presented in this thesis, and discusses future work.

1.2 Review of Order-Tracking Methods

In this section, a review of order-tracking methods is presented.

Firstly, a broad definition for order-tracking is given, which encompasses all methods within the field. Following this, a review is given for some of the different order-tracking methods within the field. This review primarily focuses on computed order-tracking methods, of which the methods presented in this work is one. However, a brief summary is also given for the three most common other main order-tracking methods currently developed. Please note that additional approaches to order-tracking exist, beyond the three most common ones for which an overview is presented here.

Firstly, a brief summary is given of hardware based order-tracking, the first order-tracking method.

Then, a review is given of computed order-tracking. As part of this review, summaries of my own publications which form part of this PhD thesis [1, 2, 4-7, 9] are included, highlighting how my own work fitted in to the current body of work at the time of publication, and how some of my work presented here has since been built upon by subsequent authors following its publication.

Included in the review of computed order-tracking methods are some methods developed for Instantaneous Angular Speed (IAS) measurement, due to their commonality with computed order-tracking.

A summary of Vold Kalman filtering based order-tracking methods is then given.

Finally, a summary is presented of the Gabor expansion based method for order-tracking.

1.2.1 Definition of Order-Tracking

The field of order-tracking is rather difficult to define, as term ‘order-tracking’ has been used to describe research from many and different fields.

For this thesis, the term order-tracking refers to the field of signal processing, but even within this field, the term is still difficult to define, since many authors have used significantly different definitions. To complicate matters, many authors use the alternative name of angular resampling to refer to order-tracking work.

To give a broad definition of order-tracking, covering most of the definitions which have been used in the field of signal processing, firstly a definition of an order must be given as:-

An order is defined as a reference frequency measure, based upon the first harmonic order of a reference component, such as the speed of a rotating shaft. The frequency of the order component typically varies with time.

Order-tracking is then broadly defined as any method which allows the frequency of a signal, or part of a signal, to be expressed as a multiple of the order frequency reference.

This is a much broader definition than that typically used for computed order-tracking, such as given in section 1.1.1 on page 3, however it is required to encompass all the works in the general field of order-tracking.

1.2.2 Hardware based order-tracking

The first order-tracking methods simply directly sampled a signal with equal phase increments. This was done by using a varying frequency digital sampler, as a hardware implementation. The digital sampling rate used to record the signal would be varied so the signal is sampled a constant number of times per shaft revolution.

For the most common hardware implementation, this is accomplished by using a counter which measures the time between consecutive tacho or encoder pulses. The time distance between the pulses is then used to estimate the speed of the shaft, and hence determine the digital sampling rate.

A number of other alternative methods to determine the digital sampling rate are used with different hardware implementations, for example using a phase locked loop to produce the required digital sampling rate.

One commonality between all hardware implementations is the digital sampling rate is calculated in real-time. Due to the time delay associated with real-time calculations, the calculated sampling rate is typically not applied until the next shaft rotation period, so the digital sampling rate lags behind the actual speed changes by two periods of shaft rotation.

K.M. Bossley et al [11], who provide a summary of hardware based order-tracking, state that this type of sampling is quite accurate for slowly varying speeds, but due to the delay in adapting the sampling rate this method is not accurate for quickly varying speeds.

1.2.3 Computed Order-Tracking

Computed order tracking is a method where, rather than a signal being directly sampled in the angular domain as with hardware based methods, a signal is first sampled in the time domain and then subsequently resampled from equal time increments to equal phase increments by post processing. This results in a signal that is approximately periodic in terms of rotation angle, allowing subsequent constant frequency MCM techniques to be employed. The equal phase spacing used to resample the signal is determined from a map of phase (rotation angle) vs. time. The phase-time map is determined from a synchronously recorded reference signal.

Computer post processing has the advantage that the physical data capture can be conducted with a comparatively inexpensive data acquisition system, removing the necessity of the more expensive hardware based order-tracking equipment. It also has the considerable advantage of not losing time information so that the reverse process back to the time domain can subsequently be carried out.

The first computed order-tracking methods developed construct the phase-time map directly from information in the time domain. This is most commonly based on the timings from the leading edge of the pulses from a once-per-rev tachometer reference, but other reference signals such as encoders can also be used. Resampling is then generally conducted using a form of polynomial interpolation. Examples include Fyfe and Munck's method from 1997 [12] which was evaluated with both a once-per-rev and 3-pulse-per-rev tacho with comparable results, and which used quadratic interpolation to construct the phase-time map and then evaluated different order polynomial interpolation methods for subsequent resampling, with best results from cubic-spline interpolation. Another example is the 1999 study by Bossley et al, [11], which compared computed order-tracking to older hardware based angular sampling methods and a hybrid hardware/computed order-tracking method. For computed order-tracking multiple interpolation methods were evaluated, with the best results using cubic spline interpolation for both constructing the phase-time map and resampling the vibration signal. In general, time domain based order-tracking methods have been limited to low values of speed variation and frequency of speed variation. Time domain approaches generally only accurately order-track a limited number of low-frequency orders, with errors becoming progressively worse as the order increases, which limits the applicability of the methods. An additional example, which compares computed order-tracking to older hardware based angular sampling methods was presented in 2010 by André et al. [13].

In 2005 Groover et al. [14] presented an extension to time domain based order-tracking, where the order-tracking method is used in a reversible fashion. In a process titled ‘double resampling’, a signal is resampled to the order-domain, edited to remove order-based components, and then the edited signal is resampled back to the time domain for subsequent analysis. For the presented method, frequency bin editing was used to remove order-based components from the spectrum of the order-tracked signal. The initial concept of reversing an order-tracking process, and transform processes more generally, was presented earlier, in 1996 by Lembregts et al. [15-18] (note all four publications are almost word for word identical, and contain no practical differences). However, while Lembregts et al. present some results for reversing an order-tracking process, there is no clarification as to what type of order-tracking process was used and then reversed, and only a cursory methodology for reversing a transform is described.

In 2004 Bonnardot et al. [19] presented results from using a new approach to order-tracking, which was first outlined (in French) by El Badaoui et al. in 2002 [20]. The same new approach was presented in much more detail in 2005 by Bonnardot et al. [10]. Rather than constructing a phase-time relationship directly in the time domain from a tacho signal, as with prior works, the new approach used phase demodulation to directly extract the (continuous) phase-time relationship. Using phase-demodulation for order-tracking is inherently more accurate than time domain polynomial interpolation based methods such as [11, 12]. This is because phase-demodulation effectively performs (almost) “ideal” interpolation in transforming from the frequency domain back to the time domain (equivalent to sinc(x) interpolation in the time domain, but the latter must be truncated, giving indeterminate errors). This thus generates (almost) ideal phase values at the signal sampling rate which can be used for subsequent polynomial interpolation, whereas time-based methods are limited to interpolating between much more widely spaced phase values from a reference signal, with no phase information between them. The described order-tracking approach [10] utilised high gearmesh harmonics from the response vibration signal itself, which successfully allowed order-tracking without a separate reference signal. However, the use of a high harmonic of shaft speed, such as gearmesh, limited the allowable speed range to nominally constant speed, and so only very small fluctuations, or jitter, could be compensated for. The analytical limits to the allowable speed variations for the gearmesh demodulation based method are presented in [10], but it should be noted that the presented limits contain simplifications which make them only suitable for gearmesh and other high order harmonics and not more general demodulation based order-tracking.

In 2007 Combet and Gelman [21] extended the work by Bonnardot et al [10] to develop an automated version of the order-tracking method. This work focused on automating the parameter selection used in the order-tracking process, such as which gearmesh harmonic is most suitable, and on identifying steps which could be replaced with less computationally intensive substitutes with the goal of allowing pseudo real-time analysis. This work did not add anything fundamentally different to the underlying order-tracking process, as described in [10].

In 2006 Coats [9] presented a variant of the phase demodulation based order-tracking method of Bonnardot et al [10]. Rather than using a gearmesh harmonic from the signal itself, the first harmonic of a tacho signal was used, which allowed order-tracking of signals with speed varying up to approximately ±30%. The order-tracking method was titled ‘Single Record Order-Tracking’, and presented results of simulated testing with single sinusoids which indicated that highly accurate order-tracking was possible.

In 2010 Coats and Randall [6] published the methods described in [9], with confirmation from basic experimental testing. Results showed that order-tracking using the first harmonic produced accurate order-tracking up to the first harmonic, with higher orders containing progressively larger residual speed variations comparable to those seen when employing time-based order-tracking methods.

However, in 2009 Coats et al. [7], as part of a study on order-tracking without a directly coupled tachometer, had presented a multi-iteration approach to phase demodulation based order-tracking. The described method was implemented using the response signal itself, as with Bonnardot et al. [10], however the method was also implemented with an external reference signal. Rather than using a gearmesh harmonic as with [10], or the first harmonic as with [6, 9], order-tracking was conducted iteratively with multiple harmonics. The lowest separable harmonic was used as a first stage to allow for larger speed variations, and then successively higher harmonics were used to improve the order-tracking result. A summary of this paper's methodology was presented as part of Randall et al. [2] in 2011.

In 2011 Urbanek et al. [22] compared the results of using two demodulation based order-tracking methods to analyse wind turbine signals. The first method was the multiple-iteration phase demodulation based order-tracking method as described by Coats et al. [7]. The second method constructed a pseudo-encoder signal, by simply inverse transforming the (one-sided) frequency band that was phase demodulated in the first method, and taking the real part of the resulting analytic signal. The zero crossings of this signal (with phase increments of π radians) were then used as phase markers like an encoder signal, to conduct order-tracking as per traditional computed order-tracking methods similar to [11, 12]. Both methods were iterated twice to improve the result, similarly to [7]. The second method was used by Randall et al. in 2012 [3] to analyse wind turbine data, successfully detecting gear and bearing faults, though with modest speed variation and only one stage of order-tracking.

In 2012 Coats and Randall [5] presented results showing the use of a single stage of order-tracking, as described by [6, 9], to successfully diagnose a bearing fault on a gear test-rig with large speed variations of ± 10% and ± 25%. Order-tracking was conducted both with a tachometer, similar to [9], and also using the signal itself, similar to [10]. Later in 2012, Coats and Randall [4] used a similar approach to [5], but with multiple iterations [7], to successfully diagnose a gear tooth root crack on the same test-rig.

In 2014 Coats and Randall [1] presented a generalised methodology whereby order-tracking could be conducted using a Phase Demodulation based Order-Tracking

(PDOT) approach for vibration signals, even in the case of large speed variations of up to approximately ±30%. This generalised methodology included some previous elements which had previously been presented in isolation in various formats [4-7, 9].The PDOT methodology could be employed with an external reference signal, or alternatively solely using the response vibration signal, allowing flexibility in implementation. The PDOT methodology could also be employed with either one or multiple iterations, which are each suitable for use with a wide range of different subsequent diagnostic techniques, and resulted in accurate (to very high frequency) order-tracked signals unconstrained by resolution limitations, such as with time-frequency order-tracking approaches. This thesis work is a further continuation to the generalised methodology first presented in [1].

1.2.3.1 Instantaneous Angular Speed

Some recent research has been done on calculating the instantaneous angular speed (IAS) of machines. This is comparable to methods of computed order-tracking, since both order-tracking and IAS methods determine the machine rotation angle (or its derivative, rotational speed) as functions of time, and so some IAS methods could feasibly also be used for computed order-tracking. The rotational speed can in fact itself be order tracked and expressed as a function of rotation angle.

In 2009 Combet and Zimroz [23] developed a new method to determine IAS directly from a vibration signal. The method estimates an instantaneous time-scaling factor, corresponding to the relative speed gap between segments of the vibration signal, effectively giving IAS. This differs to other IAS methods in that it is based on all harmonics of a speed varying component of interest rather than a specific one. To evaluate the method, it was used to order-track a vibration signal, demonstrating the connection. The new method was compared to a traditional computed order-tracking process similar to [11, 12], and was found to give a similar but slightly poorer order-tracked result.

In 2011 Zimroz et al. [24] presented a method to determine IAS from a vibration signal by using a time-frequency Short Time Fourier Transform (STFT) based approach. Similarly to [23], the method is based on multiple harmonics, however only a limited number are used rather than every harmonic as with [23]. Experimental results are presented showing accurate calculation of IAS when compared to a tachometer. While this method could be applied to order-tracking, subsequent analysis would be significantly limited as a result of the resolution limitations resulting from the STFT method.

In 2013 Urbanek et al. [25] presented a two-step method for determining IAS from a vibration signal. The authors identified that a STFT based approach, similar to that presented in [24], allows IAS to be extracted from a vibration signal for cases with large speed variations where the varying signal harmonics overlap, which could not be order-tracked using phase demodulation based methods. To overcome the resolution limitations inherent in the STFT based approach, the method employed two steps to calculate IAS. For the first step the STFT is employed to coarsely order-track the vibration signal, and then as a second step a phase demodulation based method is used to extract refined IAS information, which is then reverse-order-tracked back to the time domain giving the IAS signal. The phase demodulation step is similar to that from phase demodulation based order-tracking methods, in that machine frequency information is extracted from the spectrum, however the information is not subsequently used for further order-tracking. It should be emphasised that the two-step method in [25] is not equivalent to the multi-iteration approach in [7]. The method in [25], while consisting of two steps, only employs one single stage of phase demodulation, similar to [6, 9, 10, 19, 20], whereas the method in [7], and in this work, implements a phase demodulation based approach multiple times iteratively. While the method in [25] is used to determine IAS, it would be directly applicable to order-tracking, due to the commonality with phase based order-tracking methods.

In 2010 Renaudin et al. [26] presented a novel approach to IAS. Rather than determining the IAS from a vibration signal, the IAS is obtained solely from an encoder-type device, and the IAS is directly used as a fault detection parameter. The IAS signal is directly sampled in the angle domain through a hardware process based on pulse timing. The hardware-based direct angular sampling is comparable to angular sampling of vibration signals in traditional hardware based procedures, such as in [11]. Presented results show accurate fault detection in multiple experimental cases including with variable speed. In 2011 André et al. [27] presented initial results from employing the same methodology as [26] in a long-term wind-turbine study, as well as identifying some specific limitations inherent with the method. One limitation identified was that the angular sampled IAS signals could not be low-pass-filtered and suffered aliasing issues, which were minimised by down-sampling the IAS signal. Presented results showed accurate fault detection on components directly coupled to the encoder, and also on components indirectly coupled through a gear-train. In relation to order-tracking, the approach to IAS presented in [18] and [19] is comparable to traditional hardware-based order-tracking, as described in section 1.2.2, which has been superseded by computed order-tracking methods. As such this IAS method does not currently present any processes which could be applied to computed order-tracking of vibration signals.

1.2.4 Vold-Kalman filtering based Order-Tracking

A good overview of the Vold-Kalman Order Tracking Filter is given by Herlufsen et al. [28].

The Vold-Kalman order-tracking filter was first presented in 1993 by Vold and Leuridan [29]. This filter allows for a markedly different approach to order-tracking to be employed, when compared to the resample based methods discussed in section 1.2.3. As a result, this method is not directly comparable to resample based methods.

The Vold-Kalman order-tracking filter can simply extract out a specific order-based component from a time signal. This gives the waveform corresponding to the specific order in the time domain. The filtered order can then be displayed in the Time-Frequency domain, using a Short Time Fourier Transform (STFT). This allows an order component to be evaluated without resampling the original signal.

The Vold-Kalman order-tracking filter can be employed to simultaneously extract multiple orders from a harmonically related series, e.g. different harmonic orders of a drive shaft. By extracting multiple orders from a signal, it is possible to build up a picture of the total order content of a signal.

Note that as the Vold-Kalman order-tracking filter is a filtering method, it does not extract the true order components, but rather an approximation of the order components.

The main advantage of the Vold-Kalman order-tracking filter is that the maximum allowable speed variations and rate of speed variations, with which the method can be employed, is typically much larger than with resample based order-tracking methods.

The largest limitation is that closely spaced order information cannot be successfully separated and extracted, so it is not possible to obtain order information with the same resolution as with resample based methods, with order information typically being limited to widely spaced order values, e.g. integer order values only.

1.2.5 Gabor expansion based Order-Tracking

A good overview for the Gabor expansion based order-tracking approach is given by Qian [30].

The Gabor expansion based order-tracking method is very similar to the Vold-Kalman order-tracking filter discussed in the previous section. As with the Vold-Kalman based methods, Gabor based order-tracking is not directly comparable to resample based methods, as a markedly different result is produced.

Gabor order-tracking can be thought of as a type of joint time-frequency based filter to extract orders from a signal, with the same basic application as the Vold-Kalman filter.

However, rather than being an actual filter, as with the Vold-Kalman filter, Gabor order-tracking makes use of mathematical transforms (Gabor expansion, discrete Gabor transform etc). As a result, rather than extracting an approximation of an order component, as with Vold-Kalman filter based order-tracking, the Gabor expansion based order-tracking is able to extract the actual order component.

The Gabor order-tracking approach produces similar results to the Vold-Kalman filter based methods, but is typically more accurate, and produces a smoother extracted order with less distortion. The Gabor order-tracking approach is also less computationally expensive than a comparable Vold-Kalman filter based approach.

Chapter 2 Theory

Chapter Overview

This chapter contains mathematical formulae and methods which will be used within latter chapters of this work to develop the PDOT method.

Each of the following sections contains a summary on a topic, for example on frequency analysis and modulation/demodulation, which forms the basis for the work presented in this thesis, but is more clearly and concisely presented in isolation rather than in the main body of this work.

Many of the following topics are considered common knowledge within signal processing, and are available from many reference sources. Some topics have been compiled from multiple sources, where no single general reference was found which covered all facets relevant to the order-tracking methods developed within this thesis. Some topics have been derived by myself, or greatly expanded upon, for example the work on general modulation, where no suitable reference was located.

This chapter can either be read now in order to gain a background into the methods and mathematics to be employed in this thesis, or can be referred to while reading latter chapters of this thesis, depending on the reader's preference.

Note that some theory which is relevant to specific applications, but not used to develop the PDOT method itself, is presented with the applications themselves in Chapter 5 and Chapter 6.

2.1 Frequency Analysis

The concepts in this section have been developed based on R.B. Randall [31, 32], and M. Norton and D. Karczub [33].

The basic principle of frequency analysis is to represent a signal as the sum of separate sinusoidal components of differing frequencies. The signal can then be expressed in the frequency domain as a frequency spectrum, which is a plot of the average value of each sinusoidal component versus its frequency value.

As previously mentioned in section 1.1.1 on page 3, frequency analysis is regularly used in the field of MCM, with the simplest method being the evaluation of the vibration from individual machinery components from a vibration signal without masking from other components with different frequencies. More generally, frequency analysis also forms the basis for a large number of different analysis techniques in MCM, and is one of the fundamental building blocks used for analysis of vibration signals.

Frequency analysis is most commonly conducted by using Fourier analysis. Fourier analysis is a set of mathematical operations used to decompose a signal into a set of sinusoids.

The first method developed for Fourier analysis is what is now known as Fourier series.

For any periodic signal $g(t)$ of period T for which

$$g(t) = g(t + nT) \quad \text{...} \ (2.1)$$

where n is any integer, it can be derived that

$$g(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(k\omega_0 t) + \sum_{k=1}^{\infty} b_k \sin(k\omega_0 t) \quad \text{........................ (2.2)}$$

where a_0 is the zero frequency (DC) coefficient (twice the mean value), k is any integer, t is time, a_k are the cosine coefficients and b_k are the sine coefficients. w_0 is the fundamental angular frequency in rad/s, by which:

$$\omega_0 = \frac{2\pi}{T} = 2\pi f_0 \quad \text{.. (2.3)}$$

where T is the fundamental period and f_0 is the corresponding frequency in Hz. The coefficients of the cosine and sine terms a_k and b_k can be obtained by correlating equation (2.2) with equation (2.1), giving:

$$a_k = \frac{2}{T} \int_{-T/2}^{T/2} g(t) \cos(k\omega_0 t) dt \quad \text{....................................... (2.4)}$$

$$b_k = \frac{2}{T} \int_{-T/2}^{T/2} g(t) \sin(k\omega_0 t) dt \quad \text{.. (2.5)}$$

The value for the DC coefficient a_0 is obtained as

$$a_0 = \frac{2}{T} \int_{-T/2}^{T/2} g(t) dt \quad \text{... (2.6)}$$

Note that $a_0 / 2$ is the mean of $g(t)$, otherwise known as the DC value of $g(t)$.

So, the total component for one sinusoid at frequency ω_k (or $k\omega_0$) is given by

$$a_k \cos(\omega_k t) + b_k \sin(\omega_k t) \quad \text{...} (2.7)$$

By using the values:

$$C_k = \sqrt{a_k^2 + b_k^2} \quad \text{...} (2.8)$$

$$\phi_k = \tan^{-1}\left(\frac{b_k}{a_k}\right) \quad \text{...} (2.9)$$

expression (2.7) can alternatively be written as:

$$C_k \cos(\omega_k t + \phi_k) \quad \text{...} (2.10)$$

Expression (2.10) helps to show that the sinusoidal (frequency) component at value k has a constant (average) amplitude, and a phase angle which is the initial phase ϕ_k at what is arbitrarily chosen to be the time $t = 0$ value. A differently chosen value for initial time t_0 would only affect the initial phase ϕ_k and have no other effect on the sinusoidal component.

An alternative way to write expression (2.10) is as:

$$\frac{C_k}{2}\{\exp[j(\omega_k t + \phi_k)] + \exp[-j(\omega_k t + \phi_k)]\} \quad \text{.........................} (2.11)$$

Expression (2.11) represents the sinusoidal component at value k, which can be summed for all values of k to give the original signal. The individual sinusoidal component from expression (2.11) can be interpreted as being the equivalent of two rotating vectors. Each of these rotating vectors would have identical length $C_k / 2$, and one vector would be rotating at angular frequency ω_k with initial phase ϕ_k, while the second vector would be rotating at angular frequency $-\omega_k$ with initial phase $-\phi_k$ (i.e. the second vector is rotating with the same magnitude of angular frequency as the first, but in the opposite direction).

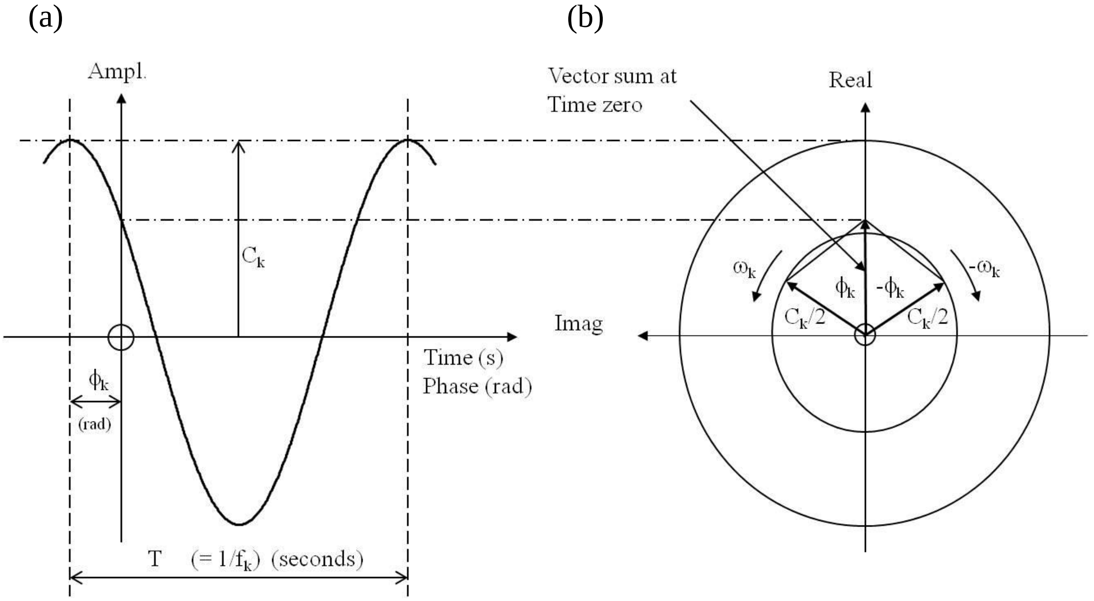


Figure 2.1 – Alternative representation of sinusoid as two counter-rotating vectors (modified, based on [31])

(a) Sinusoid (b) two counter rotating vectors which sum to give Sinusoid

It should be noted that for a real sinusoid, consisting only of non-complex values, the sum of the two rotating vectors must always give a real result. The only way the two vectors (represented as complex numbers) can sum to a real result is if they are complex conjugates, with equal amplitudes and opposite phase. Any real signal consists of real sinusoids, and hence is represented by the sums of counter rotating vector pairs of equal amplitude.

This concept of a sinusoid being alternatively represented as the sum of two counter rotating vectors is useful in understanding the frequency spectrum gained when using Fourier analysis, which will be explained later in this section.

In order to conduct Fourier analysis on a computer, a form of Fourier analysis which has both the time domain and frequency domain as discrete values is necessary as continuous data in either domain cannot be implemented on a computer.

For computer analysis, the Discrete Fourier Transform (DFT) method is employed to decompose a signal into sinusoids for representation in the frequency spectrum. Specifically, the Fast Fourier Transform (FFT) is typically used by most modern software packages, however the FFT is simply a variant of the DFT with optimisation to improve processing speed, and both methods are fundamentally equivalent.

The DFT effectively models a signal as being periodic, based on a finite record length N (samples) corresponding to a time signal, which is repeated for all positive and negative time. This allows for both the time domain and frequency domain representations to be discrete, allowing for analysis on a computer.

The DFT effectively consists of two transforms. The first transform is from the time domain to the frequency domain, giving a resulting frequency spectrum, and is called by the base name, the DFT. The second transform goes from the frequency domain back to the time domain, and is called the Inverse Discrete Fourier Transform (IDFT).

Similarly, the FFT consists of both a forward and an inverse transform, the FFT and IFFT respectively.

The DFT, the forward transform to the frequency values $G(k)$, from the original time data $g(t)$ which is sampled at discrete locations n to give $g(n)$, is given by:

$$G(k) = (1/N)\sum_{n=0}^{N-1} g(n)\exp(-j2\pi kn/N) \quad \text{............................ (2.12)}$$

where n is the discrete time indices and k is the discrete frequency indices, and N is the total number of samples in n and k.

The IDFT, the inverse transform from frequency values back to the discrete time domain values $g(n)$, from the frequency values $G(k)$, is given by:

$$g(n) = \sum_{k=0}^{N-1} G(k)\exp(j2\pi kn/N) \quad \text{.................................. (2.13)}$$

The FFT effectively also makes use of equation (2.12), and similarly the IFFT effectively makes use of equation (2.13), with both giving identical results to the DFT, and IDFT, respectively.

The frequency spectrum formed from $G(k)$ consists of N discrete frequency values, the same number as the periodic record length in the time domain. These discrete frequency values span from $-f_s/2$ to $+f_s/2$, where f_s is the time domain sampling rate in Hz, and where $f_s/2$ is known as the Nyquist frequency. As with the time domain information, the frequency spectrum is also periodic. As a result, the same frequency information from the range $-f_s/2$ to $+f_s/2$ is repeated from $-\infty$ to $+\infty$. So, from this frequency range, it can be said that the frequency spectrum spans both positive and negative frequencies.

For a real signal in the time domain, one which does not contain any complex values such as any recorded real-world signal, the positive and negative frequency components in the spectrum are complex conjugates, with symmetric amplitudes and real parts, and antisymmetric phases and imaginary parts (On an amplitude plot of the frequency spectrum, the positive and negative frequency components are hence simply a mirror image of each other around zero frequency). When considering the representation of a sinusoid as two counter rotating vectors, as introduced earlier in this section, the positive side of a frequency spectrum effectively consists of the rotating vectors with positive frequency for all the separate sinusoids, and similarly the negative side of a frequency spectrum consists of the rotating vectors with negative frequency.

The frequency spectrum information obtained from conducting a DFT is typically complex. As directly plotting complex information is difficult and generally does not aid in subsequent analysis, typically only one component of the frequency spectrum is displayed. Commonly the amplitude values of the spectrum are displayed, from a polar representation where the complex data is expressed with amplitude and phase values. These amplitude values can alternatively be squared to plot what is termed the Power Spectrum, or converted to a decibel scale with the use of a suitable reference level. For some MCM applications phase is also visually analysed from the polar data. Alternatively the complex data can be represented in a Cartesian fashion and either real or imaginary components can be analysed for different MCM analysis techniques.

There are three general layout forms for plotting the frequency axis of the frequency spectrum.

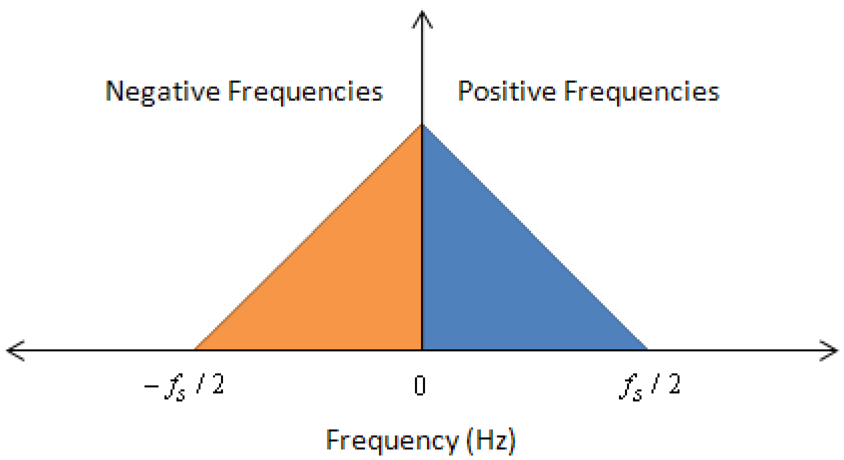


Figure 2.2 – Frequency spectrum, first general form

The first general form for a frequency spectrum can be seen in Figure 2.2. This form has a frequency axis centred on zero frequency, with the negative frequencies to the left and the positive to the right covering the range from $-f_S/2$ to $f_S/2$. This form is typically the easiest to visualise both positive and negative frequencies, and how they mirror each other around zero frequency for a real signal.

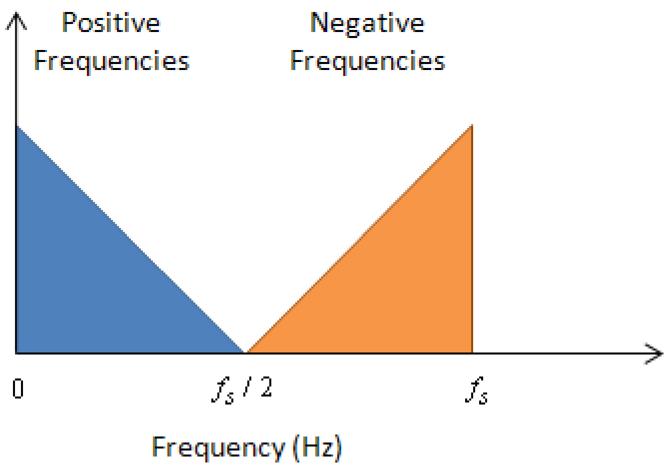


Figure 2.3 – Frequency spectrum, second general form

Figure 2.3 shows the second general form for displaying a frequency spectrum. This form makes use of the fact that the frequency spectrum is periodic, and rather than displaying the negative frequencies from $-f_S/2$ to zero frequency, the negative frequencies are instead shown from the first repetition of the periodic frequency range for the values $f_s / 2$ to f_s (minus one sample, as f_s is the first repetition of the zero frequency value). The negative frequency values are identical to those from the first general form shown in Figure 2.2, just shown in a different location. This form of a frequency spectrum is that most commonly produced in computer software such as Matlab. This is one of the two general forms of frequency spectrum which will be shown in the figures in this work.

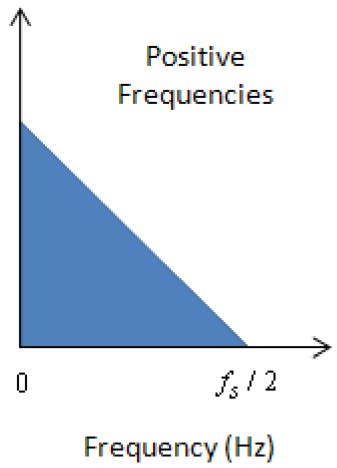


Figure 2.4 – Frequency spectrum, third general form

Figure 2.4 shows the third general form for displaying a frequency spectrum. This form makes use of the fact that for a real signal, the negative frequency data is a complex conjugate of the positive (hence mirror on an amplitude plot) and doesn't add any extra information. In this form of a frequency spectrum the negative frequencies are not plotted, and so frequency values are only shown for the range from 0 to $f_s / 2$. This form is generally used to show the positive frequencies in more detail, without the inclusion of redundant data, and is the second form that frequency spectra will be plotted with in this work.

With this form care must be taken to remember than negative frequencies are still present in a frequency spectrum, and have just been omitted due to redundancy.

Care must also be taken to make sure that a signal is real, as it would not be correct to omit the negative frequency values for a complex signal as they would no longer be mirrored as with the real case.

If any changes or editing is to be done in the frequency domain, as is common with many techniques based on frequency analysis, care must be taken to edit both the positive and negative frequencies in an identical fashion. If the positive and negative frequencies are edited differently, then the time domain signal recovered by an inverse transform (IFFT) would no longer be real valued.

An easier alternative to correctly changing both positive and negative sides of the frequency spectrum identically is to make use of another vector equivalency. Just as a sinusoid can be represented by two counter rotating vectors as shown in Figure 2.1 on page 37, it can be equivalently represented by the projection on the real axis of a single (positive) vector of twice the amplitude (C_k).

(a) (b)

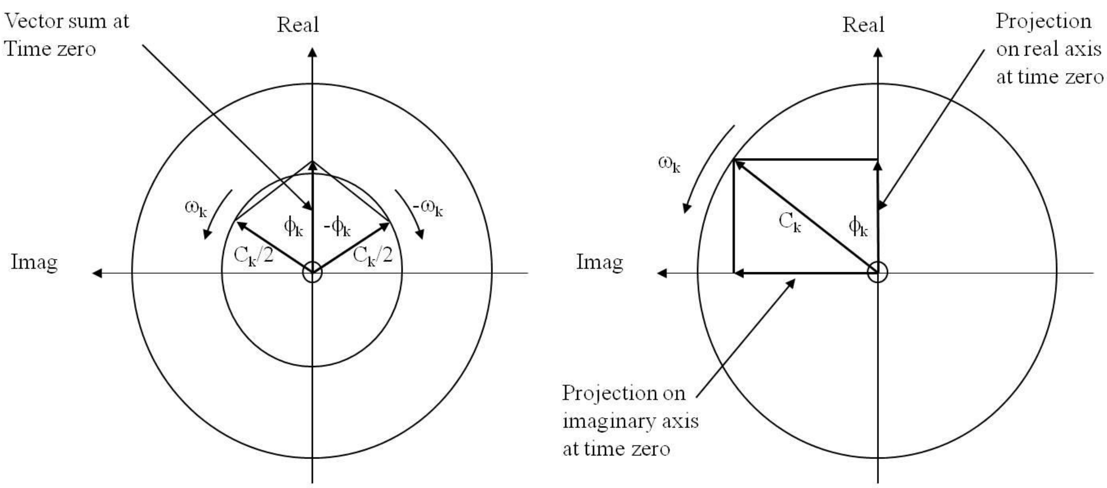


Figure 2.5 – Equivalence of two equal counter rotating vectors with real component of one vector with twice magnitude (modified, based on [31])

Figure 2.5 graphically shows the equivalence of the sum of two equal counter-rotating vectors with the real projection of a single rotating vector with double the magnitude. It should be noted that this equivalency is sometimes erroneously called the ‘Hilbert Transform’, as the Hilbert Transform is often illustrated with the identical equivalency diagram in reference books to Figure 2.5. The Hilbert Transform specifically refers to the relationship between the real and imaginary parts of an analytic signal (with a one sided spectrum, thus containing positive frequencies only, like the single vector shown in the right side of Figure 2.5), where the projection of the single vector on the imaginary axis (imaginary component) is the Hilbert Transform of the projection of the single vector on the real axis (real component).

By making use of the equivalency of two counter rotating vectors with the real component of a single vector with twice the magnitude, it is possible to make changes to the positive frequencies of a frequency spectrum only, discard the negative frequency content and replace with zeroes, transform this positive only spectrum back to the time domain using an IFFT, before finally taking the real component of the complex result. This real component is identical to the real signal which would have been produced, had both positive and negative frequencies of the spectrum been identically changed, before being transformed back to the time domain via an IFFT. This process can help simplify any processes involving the modification of information in the frequency domain.

The optimisations made for the FFT and IFFT result in the best performance boost where the record length N is a power of two, and so practically it is ideal to utilise record lengths of powers of two where possible.

2.2 Aliasing

The concepts in this section have been developed based on M. Norton and D. Karczub [33].

Aliasing is an issue which arises with Frequency analysis, as a result of the periodicity of the frequency spectrum from a FFT (DFT).

As given in 2.1, using the FFT method results in a spectrum giving frequency values from $-f_s/2$ to $f_s/2$. The spectrum is periodic, so the same frequency values are repeated from $-\infty$ to $+\infty$. However, the same restriction on frequency range is not inherently present in the time domain signal $g(t)$. When frequencies outside the range of $-f_s/2$ to $f_s/2$ are present in $g(t)$, these are not correctly shown in the frequency spectrum $G(k)$. Rather than simply being omitted, these out of range frequencies are incorrectly shown as frequencies within the $-f_s/2$ to $f_s/2$ range, which makes the resulting spectrum unsuitable for further analysis. This phenomenon is called aliasing, and a spectrum which contains these incorrect frequencies is said to be aliased.

Once aliasing has occurred, there is no way to remove it from a signal, making any subsequent analysis impossible. As such aliasing is a significant problem if not addressed.

To illustrate this phenomenon, the following figures are for a single sinusoid of 100 Hz. For this sinusoid to be shown correctly, the frequency values (positive and negative) must be within the correct frequency range, so $0 < 100\text{ Hz} < f_s/2$ for the positive frequency component, and $-f_s/2 < -100\text{ Hz} < 0\text{ Hz}$ for the negative frequency component. So in this case, a sampling rate $f_s >$ 200 Hz is required for the signal not to be aliased.

For a first example case, a sampling rate of f_s = 260 Hz is used, which should correctly not give aliasing.

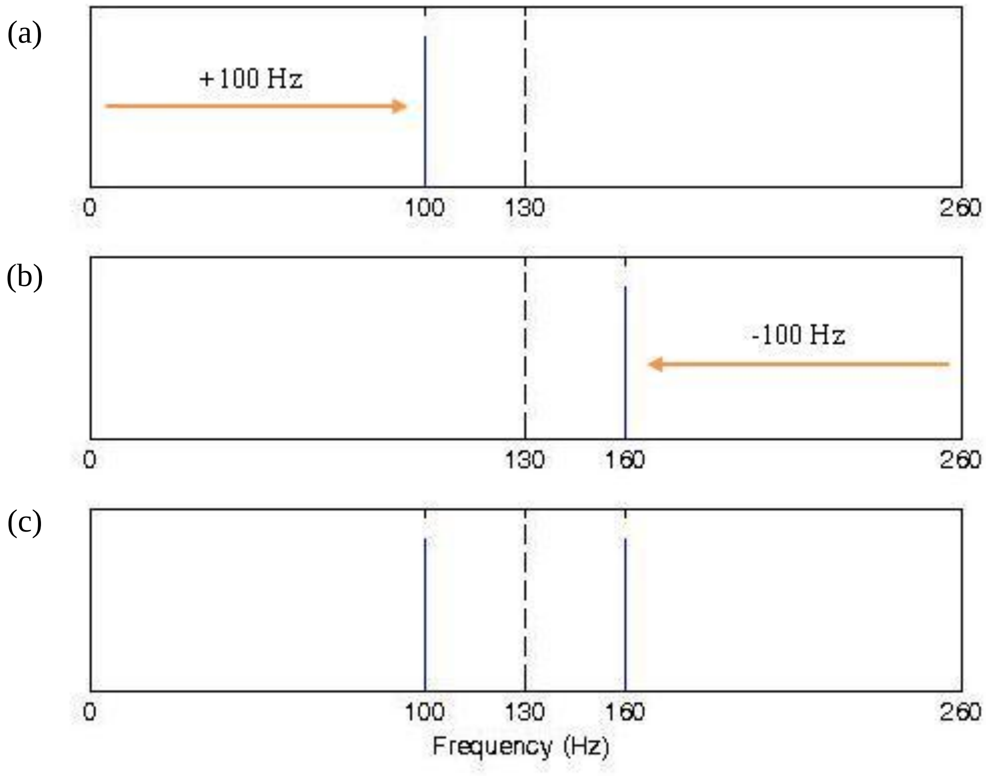


Figure 2.6 – Frequency spectrum for 100 Hz sinusoid with 260 Hz sampling rate
(a) Positive frequency vector only (b) Negative frequency vector only
(c) full spectrum from both vectors combined to give sinusoid

Figure 2.6 shows frequency spectra corresponding to the case with a 100 Hz sinusoid with 260 Hz sampling rate. The dashed line at 130 Hz indicates the Nyquist frequency at $f_s / 2$ indicating the limit that frequencies can be correctly displayed. These plots are shown in the general layout form 2 for a frequency plot, so positive frequency is shown on the left and negative on the right of the plots. Figure 2.6 (a) shows the positive vector only from a 100 Hz sinusoid, and the vector can be seen to be

100 Hz to the right of zero frequency (left side of the plot), and correctly to the left of the Nyquist frequency on the positive side of the spectrum. Figure 2.6 (b) similarly shows the negative vector only, which is 100 Hz to the left of 260 Hz (right side of the plot), and correctly on the negative side of the spectrum. Figure 2.6 (c) shows the full spectrum from the 100 Hz sinusoid, from both positive and negative vectors, which is the correct frequency plot for a single 100 Hz sinusoid.

(a)

(b)

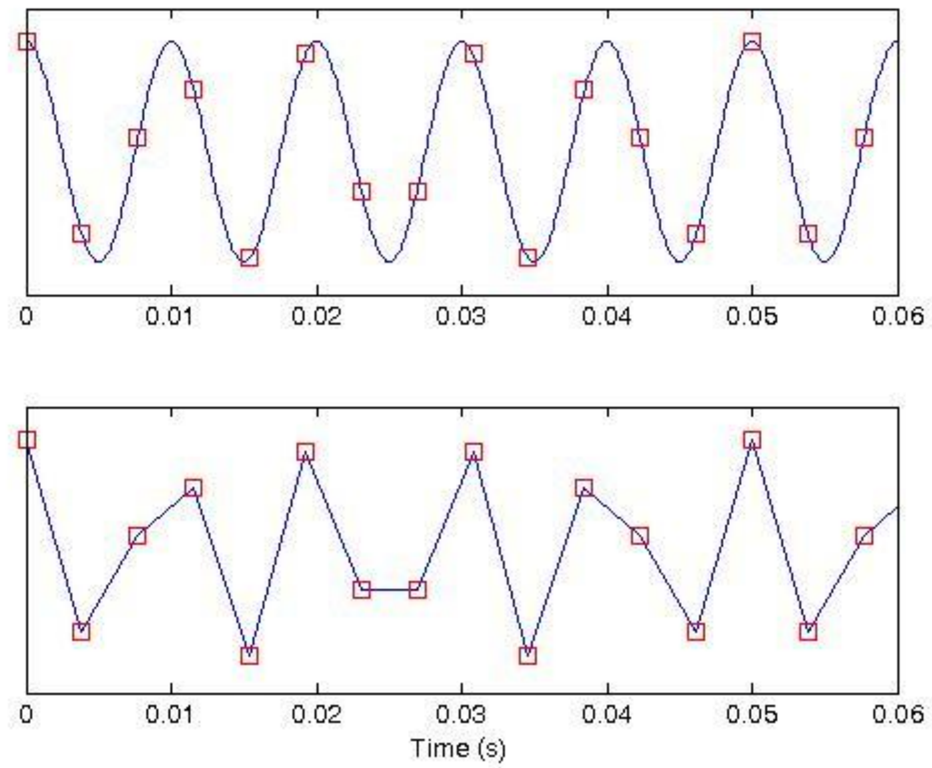


Figure 2.7 – Time domain for 100 Hz sinusoid with 260 Hz sampling rate
(a) Original continuous signal showing samples at 260 Hz
(b) Linearly interpolated signal fitted to discrete samples at 260 Hz

Figure 2.7 shows the corresponding time domain plots for the case shown in Figure 2.6. Figure 2.7 (a) shows the original continuous sinusoid at 100 Hz, with the discrete sampling locations indicated from a sampling rate of f_s = 260 Hz. Figure 2.7 (b) shows the same sampling locations as from (a), and the linearly interpolated signal obtained by joining these samples. The signal in Figure 2.7 (b) appears to roughly have the same frequency as the continuous signal in (a), which can be determined by counting number of peaks in both signals, which is equal, indicating they have the same period and hence frequency. The signal in (b) however cannot be clearly seen as being identical to the continuous signal in (a). However, as was seen from the spectrum in Figure 2.6 (c), this signal is not aliased, and both signals contain identical spectral information and so are effectively equal. As such, the differences in this case are purely visual, and a result of the discrete sampling. That both signals contain the same information can be illustrated by oversampling the discrete signal in (b).

Oversampling can be conducted by transforming to the spectrum with an FFT, padding the spectrum with zeros around the Nyquist frequency, and then returning to the time domain with an IFFT. This oversampling has the effect of raising the sampling rate of the signal, without making any changes to the spectral information.

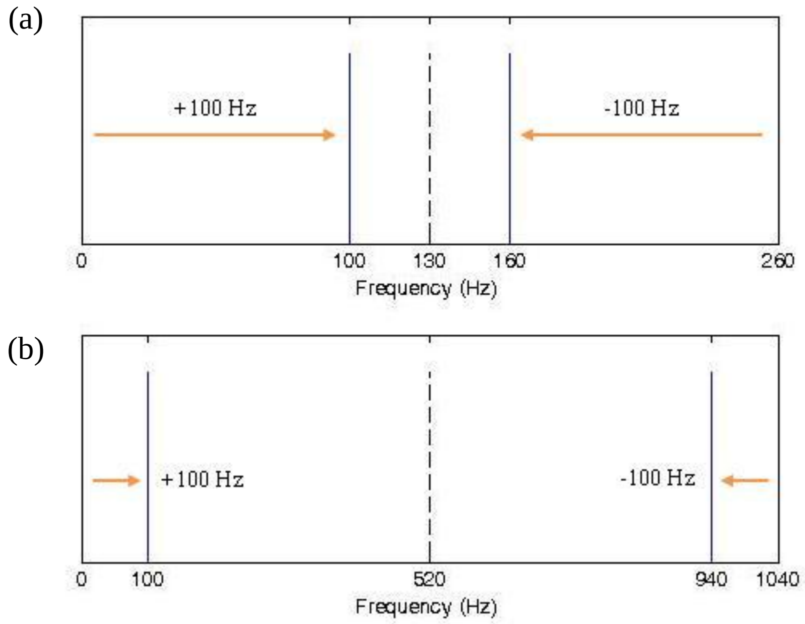


Figure 2.8 – Padding of spectrum with zeros to oversample signal by a factor of four (a) Original spectrum (b) Oversampled spectrum

Figure 2.8 shows the zero padding of the spectrum for this signal, to increase the sampling rate by a factor of four. Figure 2.8(a) shows the original spectrum from Figure 2.6 (c). Figure 2.8 (b) shows the padded spectrum, from which it can be seen that the sampling rate has increased by a factor of four, and the total number of samples has increased by this amount. It can also be seen that the spectral information of the signal has remained identical to that before resampling, with the sinusoid still correctly located at +100 Hz and -100 Hz (1040 – 100 = 940 Hz), and so the oversampling has made no fundamental changes to the signal.

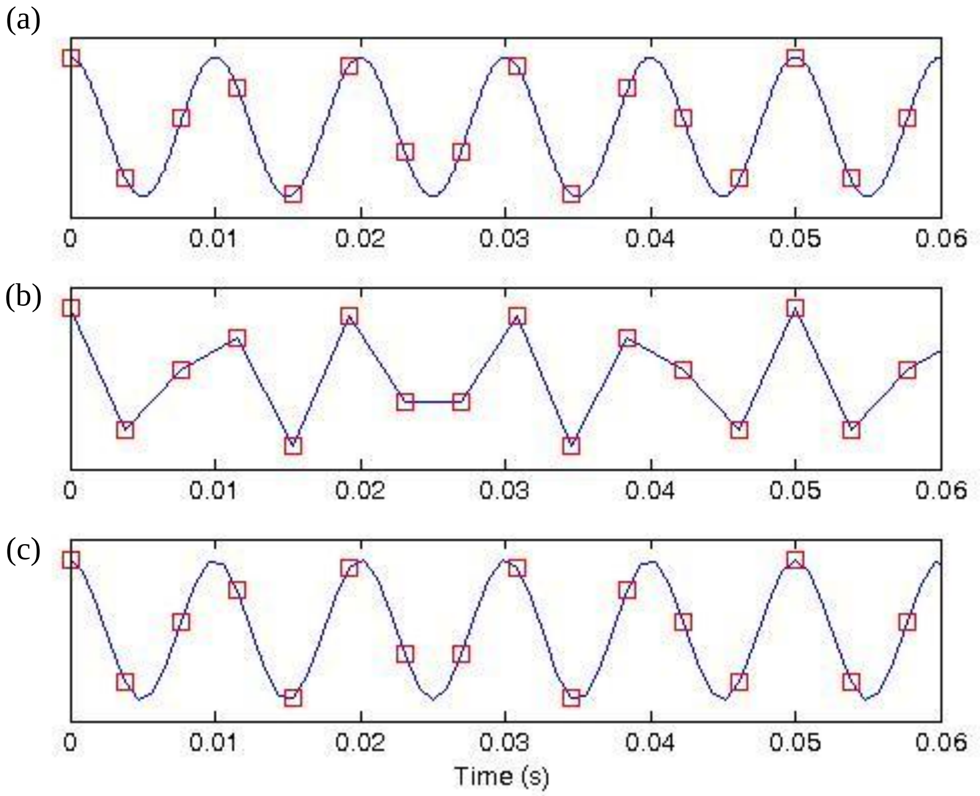


Figure 2.9 – Time domain for 100 Hz sinusoid with 260 Hz sampling rate
(a) Original continuous signal showing samples at 260 Hz
(b) Linearly interpolated signal fitted to discrete samples at 260 Hz
(c) Linearly interpolated signal fitted to discrete samples oversampled by factor of five to 1040 Hz sampling rate, showing fit to original discrete samples at 260 Hz

Figure 2.9 (a) and (b) are a repeat of Figure 2.7, for comparison. Figure 2.9 (c) shows the linearly interpolated signal after oversampling by a factor of five. It can now be clearly seen that this signal is much closer to the original continuous signal, and that no aliasing has occurred. The resampling can be done to any required degree of resolution with no error.

The purely visual difference which makes Figure 2.9 (b) not resemble (a) is referred to as Visual Aliasing, and occurs where the signal is not actually aliased, but just appears not to be an identical representation due to a low sampling rate.

For a second example, the same 100 Hz sinusoid is sampled at f_s = 140 Hz. As this sampling rate is < 200 Hz, this case should give aliasing.

(a)

(b)

(c)

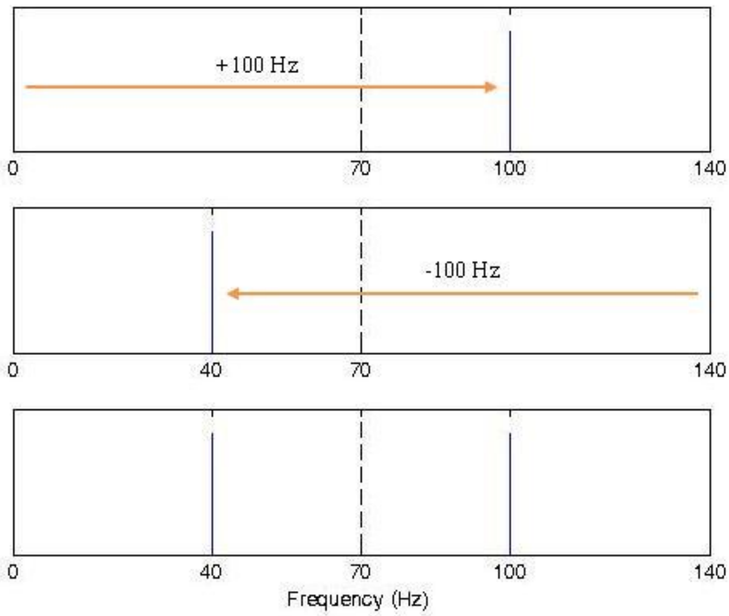


Figure 2.10 – Frequency spectrum for 100 Hz sinusoid with 140 Hz sampling rate
(a) Positive frequency vector only (b) Negative frequency vector only
(c) full spectrum from both vectors combined to give sinusoid

Figure 2.10 shows frequency spectra corresponding to the case with a 100 Hz sinusoid with 140 Hz sampling rate. Figure 2.10 (a) shows the positive frequency vector only from a 100 Hz sinusoid, and the vector can again be seen to be 100 Hz to the right of zero frequency (left side of the plot). In this case it can be seen that the positive frequency vector is now to the right of the Nyquist frequency, and incorrectly on the negative frequency side of the spectrum. Figure 2.10 (b) similarly shows the negative vector only, which is 100 Hz to the left of the sampling frequency 140 Hz (right side of the plot). Similarly, the negative vector is now to the left of the Nyquist frequency, and incorrectly on the positive frequency side of the spectrum. Figure 2.10 (c) shows the combined spectrum for the 100 Hz sinusoid, showing both positive and negative vectors. As can be seen, the complete spectrum is now incorrectly that of a 40 Hz sinusoid, rather than the actual 100 Hz sinusoid. In this case the signal has become aliased, with the wrong frequency information shown, due to the sinusoid not being within the correct frequency range based on sampling rate.

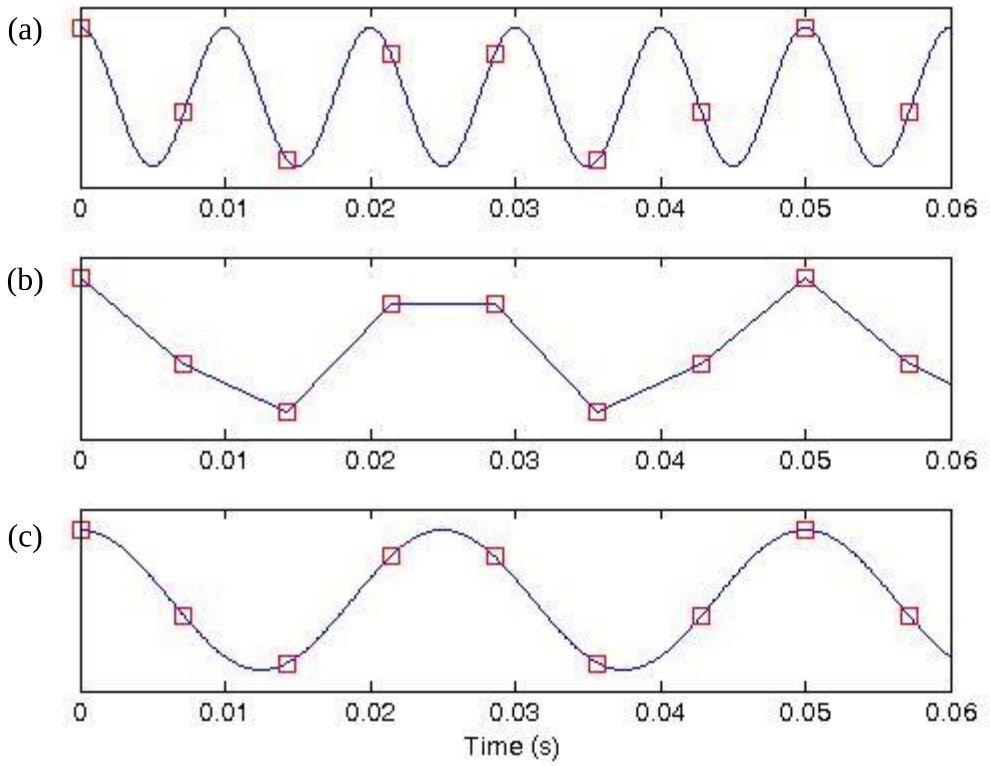


Figure 2.11 – Time domain for 100 Hz sinusoid with 140 Hz sampling rate
(a) Original continuous signal showing samples at 140 Hz
(b) Linearly interpolated signal fitted to discrete samples at 140 Hz
(c) Linearly interpolated signal fitted to discrete samples oversampled by factor of five to 700 Hz sampling rate, showing fit to original samples at 140 Hz

Figure 2.11 shows the corresponding time domain plots for the case shown in Figure 2.10. Figure 2.11 (a) shows the original continuous sinusoid at 100 Hz, with the discrete sampling locations indicated from a sampling rate of f_s = 140 Hz. Figure 2.11 (b) shows the linearly interpolated signal fitted to these samples. It can be seen that these signals do not have the same frequency. Figure 2.11 (c) shows the signal when oversampled by a factor of five to 700 Hz, which was done using zero padding in an identical fashion to that used for the first example. From Figure 2.11 (c) it can be seen that the new signal, which still fits the original samples at 140 Hz, is not now equivalent to the original sinusoid. In this case the discrete signal has been aliased, due to the sampling rate being too low for the signal. Oversampling cannot correct for genuine aliasing as it can for "visual aliasing".

So, to prevent aliasing of a signal when conducting frequency analysis using the FFT (or DFT), it is necessary for all frequency information in a signal to be less than half the digital sampling rate f_S. This rule is commonly referred to as the Nyquist-Shannon Sampling Theorem, or the Nyquist Sampling Theorem.

Vibration signals are typically recorded with continuous analogue sensors, and the result is digitally sampled before use in a computer. Continuous signals have no theoretical limit on frequency, so practically recorded vibrations signals will generally contain frequencies higher than half the sampling rate.

In order to prevent aliasing, a low pass filter must be employed before a signal is digitally sampled, to remove frequency content above half the sampling rate $f_S / 2$. Practically, an analogue low pass filter does not have a perfect brick-wall filter characteristic, and so the filter effect begins below $f_S / 2$. For analogue low pass filters used in signal analysers, while the exact filter characteristics will be different for each analyser, a good rule of thumb is to assume the filter characteristic modifies the signal above $0.4 \times f_S$ and signal information should not be used above this frequency. For a traditional low pass filter, a correctly low pass filtered vibration signal will contain a noticeable filtering effect around the Nyquist frequency, with a clearly visible reduction in signal amplitude.

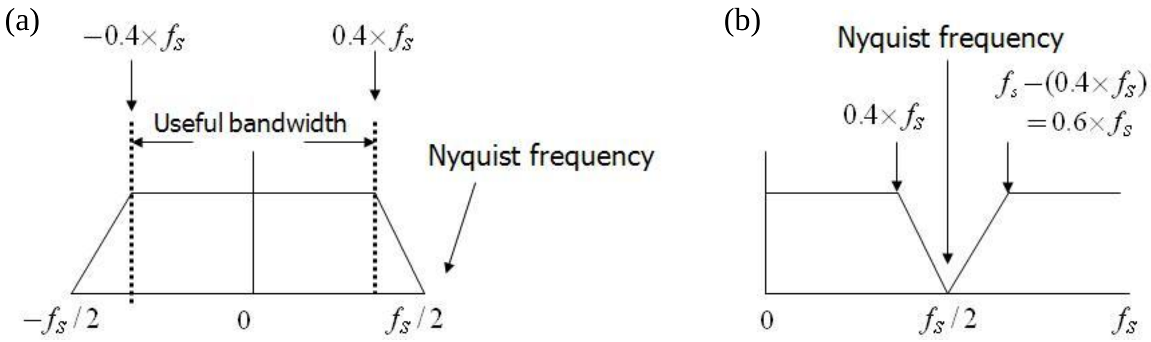


Figure 2.12 – Typical filter effect for an analogue low pass filter
(a) Shown in spectrum of first kind (b) Shown in spectrum of second kind

Figure 2.12 shows the typical pass band for a traditional analogue low pass filter, showing the characteristic dip around the Nyquist frequency typically seen in correctly low pass filtered signals.

Recently, modern signal analysers have begun using Sigma-Delta AD converters, with different low pass filter characteristics when compared to traditional analogue low pass filters. Sigma-Delta AD converters give a low pass filter effect which begins at the Nyquist frequency and extends to $0.6\times f_S$ for the positive frequency vectors, with a corresponding mirror for the negative frequency vectors. This has the result of causing aliasing around the Nyquist frequency, as any positive frequency vectors, which are just above the Nyquist frequency appear on the negative side of the spectrum, and mirror into the positive side of the spectrum. As such it is recommended that the same frequency band be discounted as with an analogue low pass filter, where no frequencies over $0.4\times f_S$ are treated as valid. When signals are to be processed further, it is sometimes advisable to remove the aliased section before doing this.

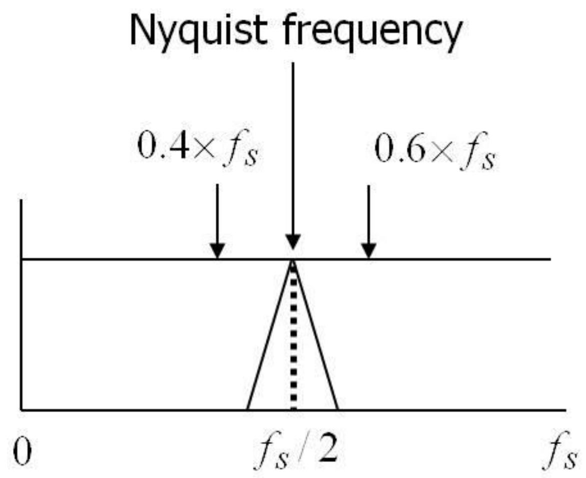


Figure 2.13 – Low pass filter characteristic for Sigma Delta AD conversion, showing aliasing

Figure 2.13 shows the low pass filter characteristic for a Sigma-Delta AD converter, showing the aliasing of some components around the Nyquist frequency, warranting the continuing practice of not using frequency information beyond $0.4\times f_s$. Note that often the Sigma-Delta AD converter is used only for the highest frequency range of the analyser, and further down sampling to lower frequency ranges uses conventional digital low pass filters cutting off below the Nyquist frequency.

All of the signals used in this thesis work have been recorded using signal analysers making use of a Sigma-Delta AD converter, and thus may not show a characteristic low pass filter drop-off, as would be present with the use of a conventional analogue low pass filter.

Aliasing is however not an issue purely limited to initial digital sampling of a signal. Aliasing can also be caused during any process involving resampling. Care must be taken to ensure correct digital low pass filters are applied during any lowering of the digital sampling rate, in particular during any resampling processes. This is specifically relevant to resample-based order tracking, where a signal is resampled from a constant time sampling to a constant angular sampling, as presented in this work.

2.3 Convolution and the Convolution Theorem

The following material has been developed based on Randall [31].

The convolution theorem gives a useful property of Fourier analysis, one use of which is to significantly aid in visualising an expected spectrum based on time domain information.

The convolution theorem states that a Fourier transform (e.g. DFT, FFT) converts a product in one domain to a convolution in the other domain.

So, if two components are a product in the time domain, where $x(t)$ is the combined time domain signal, and $f(t)$ and $h(t)$ are two components, such that:-

$$x(t) = f(t) \times h(t) \quad \text{(2.14)}$$

Then the corresponding frequency domain signal $X(f)$, is given by

$$X(f) = F(f) * H(f) \quad \text{(2.15)}$$

where $*$ is the symbol for convolution. Note that in a similar fashion, a multiplication of factors in the frequency domain corresponds to a convolution in the time domain, giving rise to equations similar to (2.14) and (2.15) with exchanged operators.

The spectrum represented by $X(f)$, from equation (2.15), can be relatively easily visualised, as from convolution, the combined spectrum $X(f)$ is given by replacing all spectral lines of the spectrum $H(f)$ with a scaled version of the spectrum of $F(f)$. If the separate spectra from $H(f)$ and $F(f)$ are known in isolation, then the expected combined spectrum $X(f)$ can be obtained with relative ease. This is most simply done when either $F(f)$ or $H(f)$ are at a small number of discrete frequencies.

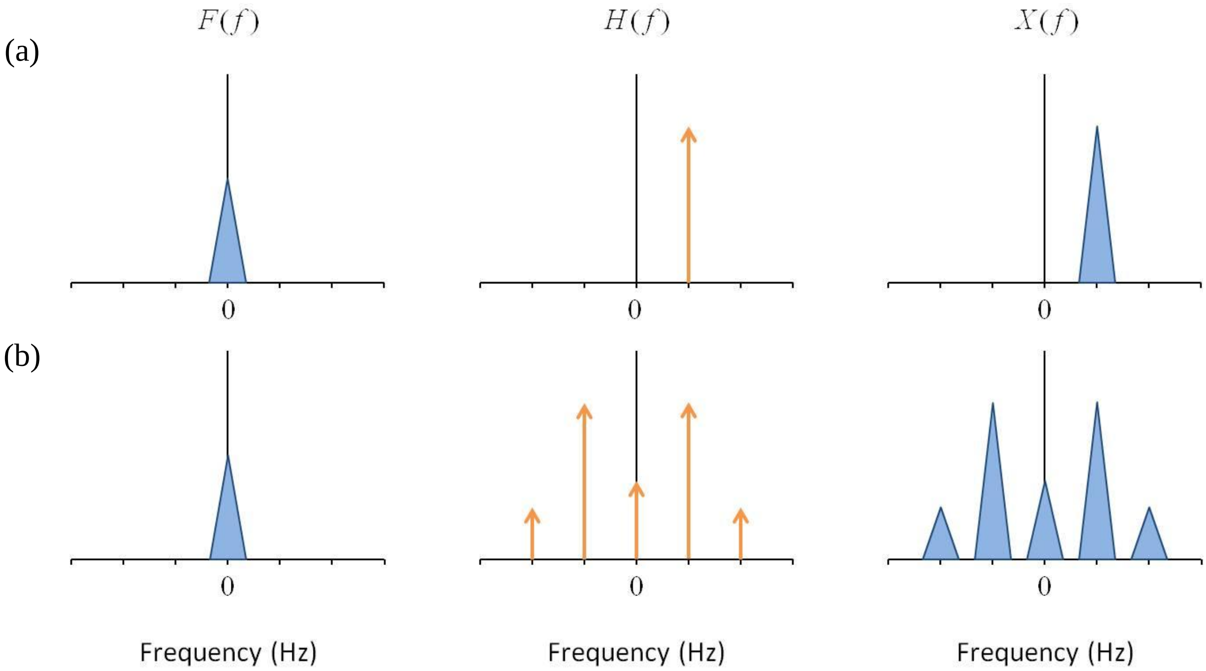


Figure 2.14 – Convolution of spectral components
(a) convolution with single discrete frequency
(b) convolution with multiple discrete frequencies

Figure 2.14 graphically shows the convolution of a spectral component $F(f)$ with another component $H(f)$ to give the combined spectrum $X(f)$, as per equation (2.5) on page 35. In (a), $H(f)$ consists of a single discrete (positive) frequency, so the combined $X(f)$ consists of a single instance of $F(f)$ scaled to the height of the discrete component from $H(f)$, replacing the single discrete value from $H(f)$. In (b), $H(f)$ consists of multiple discrete frequencies (positive and negative frequency), so each discrete component is replaced by a scaled version of $F(f)$ to give the combined $X(f)$.

2.4 Interpolation

The concepts in this section have been developed based on P.D. McFadden [34], and M. Norton and D. Karczub [33].

Interpolation is the method of estimating the value of a data point which lies between a given series of data points.

This type of data estimation is employed in many different and varied fields, often for analysing statistical data.

In digital signal analysis, interpolation is used to estimate the value of a signal at positions other than those at which the signal was initially sampled. For example it could be used to estimate the value of a signal at time 1.2 sec, when the signal was only sampled at 1 sec and then 2 sec.

Interpolation basically involves fitting an appropriate function to the known data points. A new data value lying between the known samples is then estimated by evaluating the fitted function at the new sampling point. The only real difference between the multitude of interpolation techniques is the type of function fitted to the data values. The choice of an appropriate interpolation function depends entirely on the characteristics of the data, such as whether the data is continuous or discontinuous, and whether the same function applies to the entire data series. There is no single interpolation method which is suitable for every situation.

Some examples of commonly used functions are logarithmic, exponential, polynomials of differing orders, Gaussian, sum of sine waves and Weibull functions.

For digital signal analysis, typically a single function is not applicable for an entire signal, so a series of functions of the same type are usually fitted to the signal data. Some common interpolation functions used are zero, first and second order polynomial functions, and cubic spline functions.

A zero order function is commonly called a zero order hold function, and the data between two points is simply the first data value held for the entire distance between the two points, basically making a series of step functions.

A first order interpolation function (linear interpolation) is simply a straight line fitted between two successive values.

Second and third order polynomial functions involve respectively parabolic and cubic functions being fitted to the data points. A different function of the same order is fitted to each group of data values, and typically the first and sometimes the second derivatives are made equal at the joints between the different functions – so that the transition from one function to the next is smooth and continuous.

For cubic spline interpolation, a series of cubic splines are fitted to the data. Each cubic spline function is fitted to four data points, and consists of 3 different cubic functions which are directly related to each other. The name comes from the ship building industry, where during ship construction it was necessary to bend wooden planks, or splines, around fixed pegs. The shape of a cubic spline is identical to the shape that a flexible beam would take on when bent around a series of fixed discrete points (i.e. a beam with point loads). Figure 2.15 shows the characteristic shape of a cubic spline function.

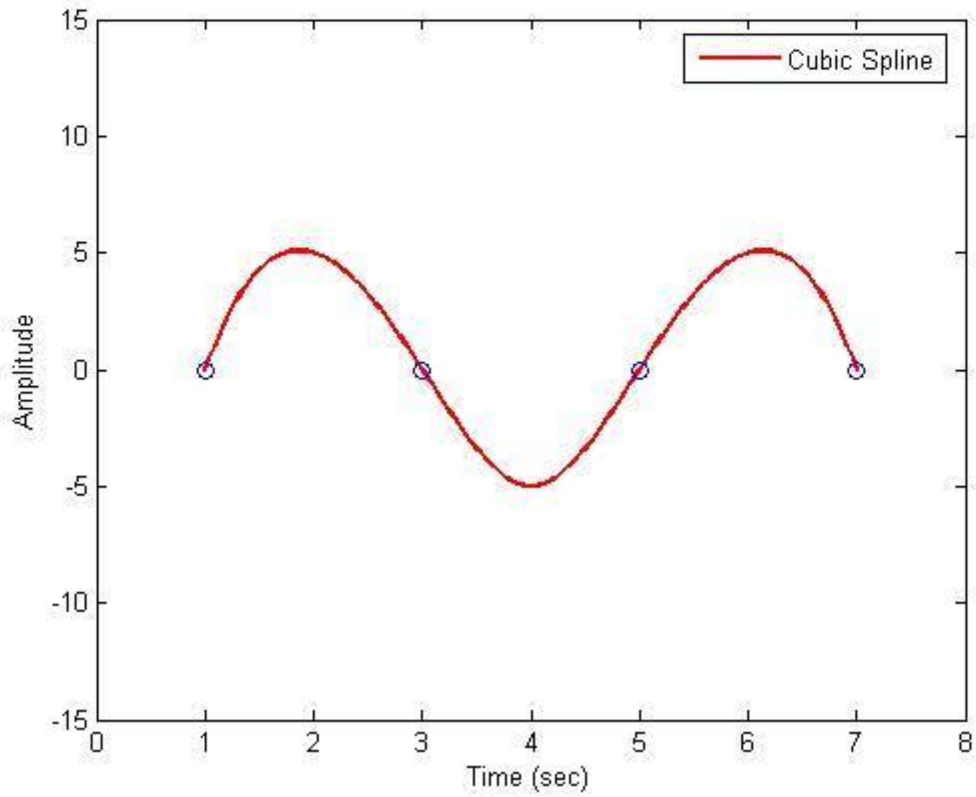


Figure 2.15 – Single cubic spline function

Different cubic spline functions are fitted successively down a signal when cubic spline interpolation is used. The first and second derivatives of the cubic splines are continuous at intermediate points, but the third derivative (the shear force for a beam) has step functions at the samples (point loads for a beam). Near the ends of the series of samples to be interpolated, values have to be specified for the slope and curvature (a beam would have zero curvature at the ends for example).

The choice of interpolation function to be used is often decided by the associated computation time. This is because more complicated functions take significantly longer to compute than simpler ones. Often the best choice of an interpolation function is the least complicated function which closely conforms to the signal characteristics, rather than a complex function which only marginally increases accuracy. However with increases in computing power this is increasingly becoming a smaller consideration in the choice of interpolation method.

For resample based order-tracking, Interpolation is typically used for two major stages. The first is to calculate the time values corresponding to equal increments of phase from the phase-time relationship. The second stage is to calculate the values of the vibration signal at these calculated time values.

The following series of figures shows graphically the interpolation steps for calculating time values corresponding to equal phase from a phase-time relationship, for the first stage of interpolation for resample based order-tracking. In this example first order interpolation functions have been used (straight lines between data points).

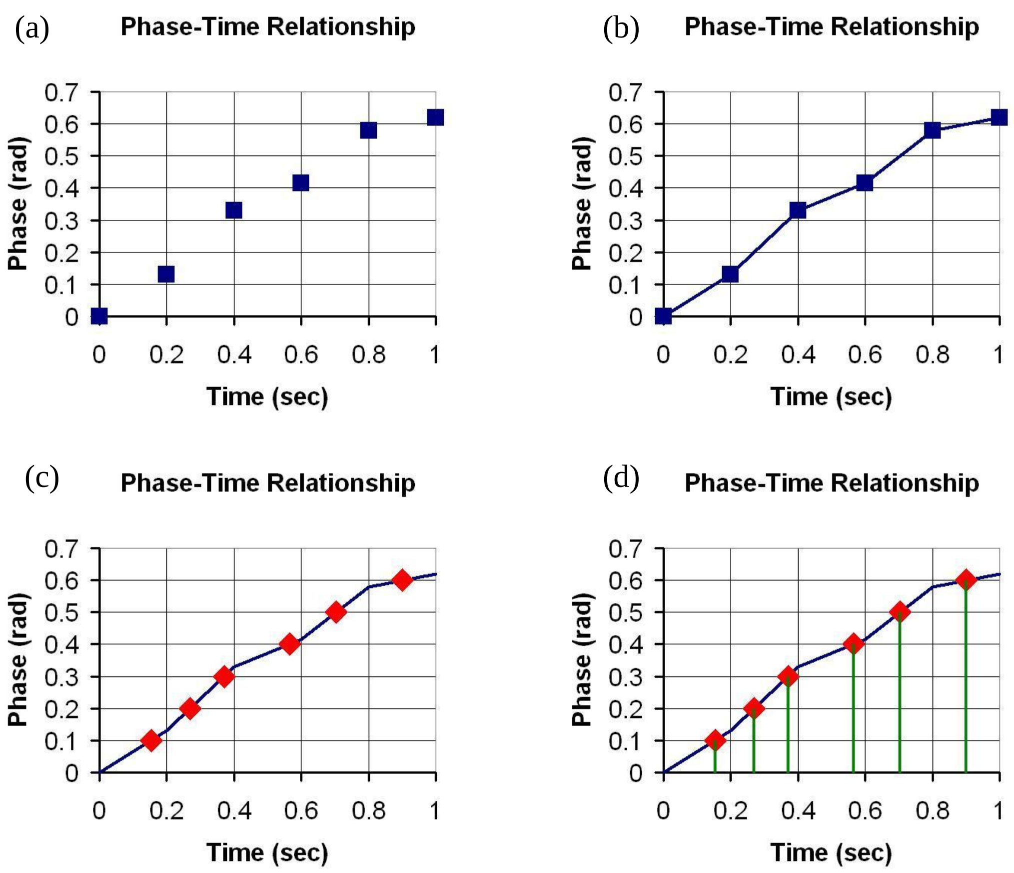


Figure 2.16 – Interpolation steps for a phase-time relationship. Linear interpolation (a) Initial Points plotted (b) First order functions fitted between data values (c) New locations with Equally spaced phase located (d) Time values calculated

Figure 2.16 (a) shows the initially known data points. These are equally spaced in time, and contain phase information at these points. Figure 2.16 (b) shows the first order functions, or straight lines, which have been fitted to the data points. These first order functions make up the interpolation function. Figure 2.16 (c) shows the new data points. These are equally spaced in phase, and have been located on the interpolation function. Figure 2.16 (d) shows the calculation of the time values of the new data values. These time values are then used for the second interpolation step to resample and order-track the corresponding vibration signal.

2.5 Modulation

For the field of resample based order-tracking, it is necessary to work with signals which have the characteristics of varying frequency. As such, a systematic method for describing a varying frequency signal is necessary to develop methods to work with this type of signal.

As stated in section 2.1 on page 34, Fourier analysis is used to describe a frequency spectrum by decomposing a signal into a sum of separate sinusoids. As a result, an ideal systematic description for varying frequency components is based on a sinusoid being varied.

A sinusoid with varying parameters is referred to as being modulated, and the extensive field of modulation theory has been developed to describe these types of signals. Historically, modulation theory has primarily been developed for the field of communications engineering, and extensive material is available on the time and frequency domain characteristics of frequency modulated signals in the context of communication.

The concepts of modulation in the following sections have been developed primarily based on two communications reference works, by R.E. Ziemer And W.H. Tranter [35], and M. Schwartz [36, 37]

Modulation theory as primarily developed for communications is focused on the modulation of a single sinusoidal carrier, as this is the optimum signal carrier to be used for communications. As a result, nearly all information available in the field is focused solely on single carrier modulation.

Varying components in a vibration signal are typically not based on the variation of a single sinusoidal component, but rather the variation of a base carrier signal consisting of a harmonic series of sinusoidal components. As such, the theory developed for communications is not immediately applicable. However, it is possible to make use of the extensively developed materials on modulation theory by identifying that a modulated vibration signal based on the variation of a harmonic series is the equivalent of the sum of the harmonic orders, where each harmonic order is expressed as a separate modulated signal. In this fashion modulation theory can be used to characterise vibration signals with varying parameters.

The simplest expression for a constant frequency sinusoid $y(t)$ is given by:-

$$y(t) = A_c \cos(2\pi f_c t + \phi_c) \quad \text{......................................} \ (2.16)$$

where A_c is the amplitude of the constant sinusoid, f_c is the frequency in Hz, and ϕ_c is the phase constant (so phase offset at time zero).

For modulation theory, where modulation is defined as the variation of a constant frequency sinusoid wave, with “carrier” frequency f_c, it can be seen that in equation (2.16) there are only two other parameters which can be varied.

This gives rise to two main modulation types.

Variation of the amplitude term gives rise to Amplitude Modulation (AM), the first main type of modulation. Variation of the angle term gives rise to angular modulation. For angular modulation, if a signal is varied based on phase it is called phase modulation, and if a signal is varied based on frequency (angular velocity, derivative of phase) it is called Frequency Modulation (FM). However both phase and frequency modulation are very similar, and practically for the application of order-tracking there is no difference. As such it is appropriate to treat them both in the same way. In this work, both methods of angular modulation will be discussed under the section title Frequency Modulation, as this type of angular modulation will be used through the rest of this thesis.

Finally, the term general modulation is used when a signal has both varying amplitude and angle parameters (combined amplitude and angle modulation).

The following material on modulation is broken up into six sections.

For the first section, amplitude modulation of a single sinusoidal carrier is discussed, as covered in most communications reference works, covering both time and spectral representations and properties.

For the second section, amplitude modulation of a single sinusoidal carrier will be extended to the case where the carrier is a harmonic series, as more generally applicable to vibration analysis and hence order-tracking.

For the third section, angle (frequency) modulation of a single sinusoidal carrier is discussed, as again commonly covered in many communications reference works, again both covering time and spectral representations and properties.

For the fourth section, frequency modulation of a single sinusoidal carrier is similarly extended to cover the case where the carrier is a harmonic series.

For the fifth section, general modulation of a single sinusoidal carrier will be discussed, where a carrier is both amplitude and frequency modulated.

Finally, for the sixth section, general modulation of a single sinusoidal carrier is similarly extended to the case where the carrier is a harmonic series.

2.5.1 Amplitude modulation for sinusoidal carrier

The first main modulation type is Amplitude modulation, where the amplitude term from a constant frequency sine wave is varied.

In the communications field the amplitude of the constant carrier sinusoid is set independently to the amplitude of the variation, as the carrier itself contains no information and is removed after transmission for communication signals, so the carrier amplitude can feasibly be set to any value. For physical systems, such as dealt with in this thesis work, the combined modulated signal is of interest, rather than only the variation component.

As a result, when dealing with physical systems, it is convenient to use a slightly different definition for amplitude modulation, more suited to dealing with the combined modulated signal, rather than definition commonly used in the communications field. Note only the definition which is suitable for physical systems, and used in this work, is described below.

So, for physical systems, it is convenient to define the amplitude of the carrier wave as unity with no dimensions, so the amplitude of the final modulated signal is defined solely by the sum of the amplitude variation and a DC offset (both with physical dimensions). The definition, that the carrier amplitude is dimensionless and equal to unity, will be used in this work for amplitude modulation.

An amplitude modulated constant frequency sinusoid $y(t)$, for carrier amplitude of unity, is given by:-

$$y(t) = [A_{DC} + A_m(t)]\cancel{A_c}\cos(2\pi f_c t) \quad \text{(2.17)}$$

where A_{DC} is the DC offset (which must be large enough so that the sum with $A_m(t)$ is always positive), $A_m(t)$ is the variation in amplitude around the constant value A_{DC}, and A_c is the amplitude of the carrier wave. Note that both $A_m(t)$ and A_{DC} have the same physical dimensions and units. It should also be noted that as $A_c = 1$, it can effectively be omitted from equation (2.17).

In communications theory, where the goal is to communicate a signal represented by variations, the $A_m(t)$ term is commonly referred to as a message signal. The constant frequency sinusoid $A_c\cos(2\pi f_c t)$ is commonly referred to as the carrier signal as this sinusoid carries the message. f_c is similarly called the carrier frequency. It should also be noted that the requirement, that the sum of A_{DC} and $A_m(t)$ is always positive, is not necessary for communications signals where the amplitude can be negative. This requirement is only necessary for physical mechanical systems, for example with a gear train, where zero or negative amplitude is not physically meaningful. Zero or negative amplitude would require the force to be either zero or negative, which is typically not possible for mechanical situations, such as in a gear train this would result in backlash rather than the physical system changing direction (corresponding to negative amplitude).

For amplitude modulation defined by equation (2.17), the message signal $A_m(t)$ plus the DC offset A_{DC} directly gives the envelope of the resulting amplitude modulated signal.

The time domain representation for an amplitude modulation is most easily shown graphically. The following are two different cases of amplitude modulation for different message signal inputs, showing the effects of amplitude modulation.

(a)

(b)

(c)

(d)

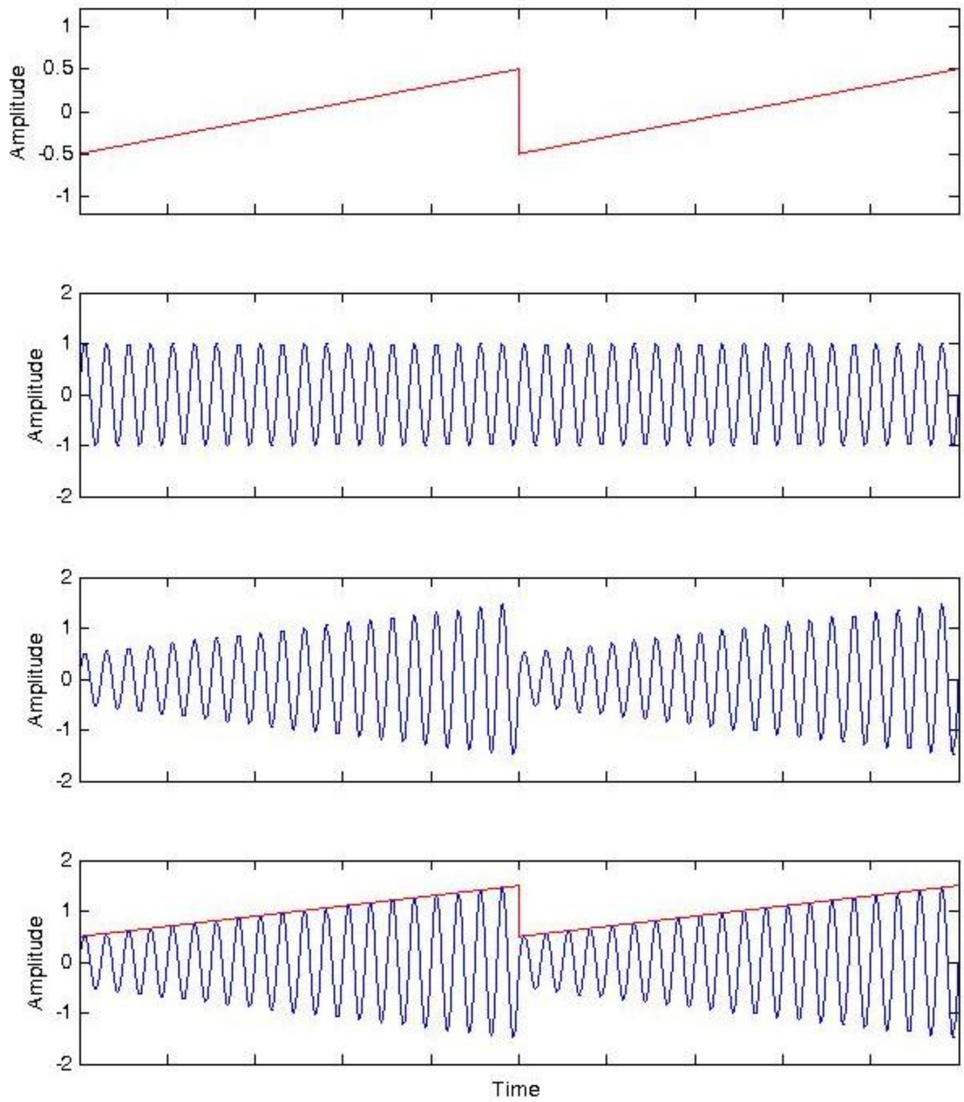


Figure 2.17 – Amplitude modulation for saw message signal
(a) message signal (b) constant frequency sinusoidal carrier
(c) amplitude modulated signal (d) amplitude modulated signal overlaid with message signal plus DC offset, indicating that the sum gives envelope of the modulated signal

Figure 2.17 shows the results of amplitude modulation, where the message signal is a saw tooth function. Figure 2.17 (a) shows the message signal, which in this case is a saw tooth function between -0.5 and 0.5 units amplitude. Figure 2.17 (b) shows the constant frequency sinusoidal carrier, with value between -1 and 1 (dimensionless). Figure 2.17 (c) shows the resulting amplitude modulated signal, when the carrier in (b) is multiplied by the message signal in (a) summed with DC offset of 1 unit. The variation in amplitude in the resulting signal is clearly evident. Figure 2.17 (d) again shows the amplitude modulated signal, overlayed with the message signal with DC offset $A_{DC} + A_m(t)$. As can be seen, the message signal plus DC offset forms the envelope of the resulting amplitude modulated sinusoid.

As a second example, the example of a sinusoidal message signal is shown.

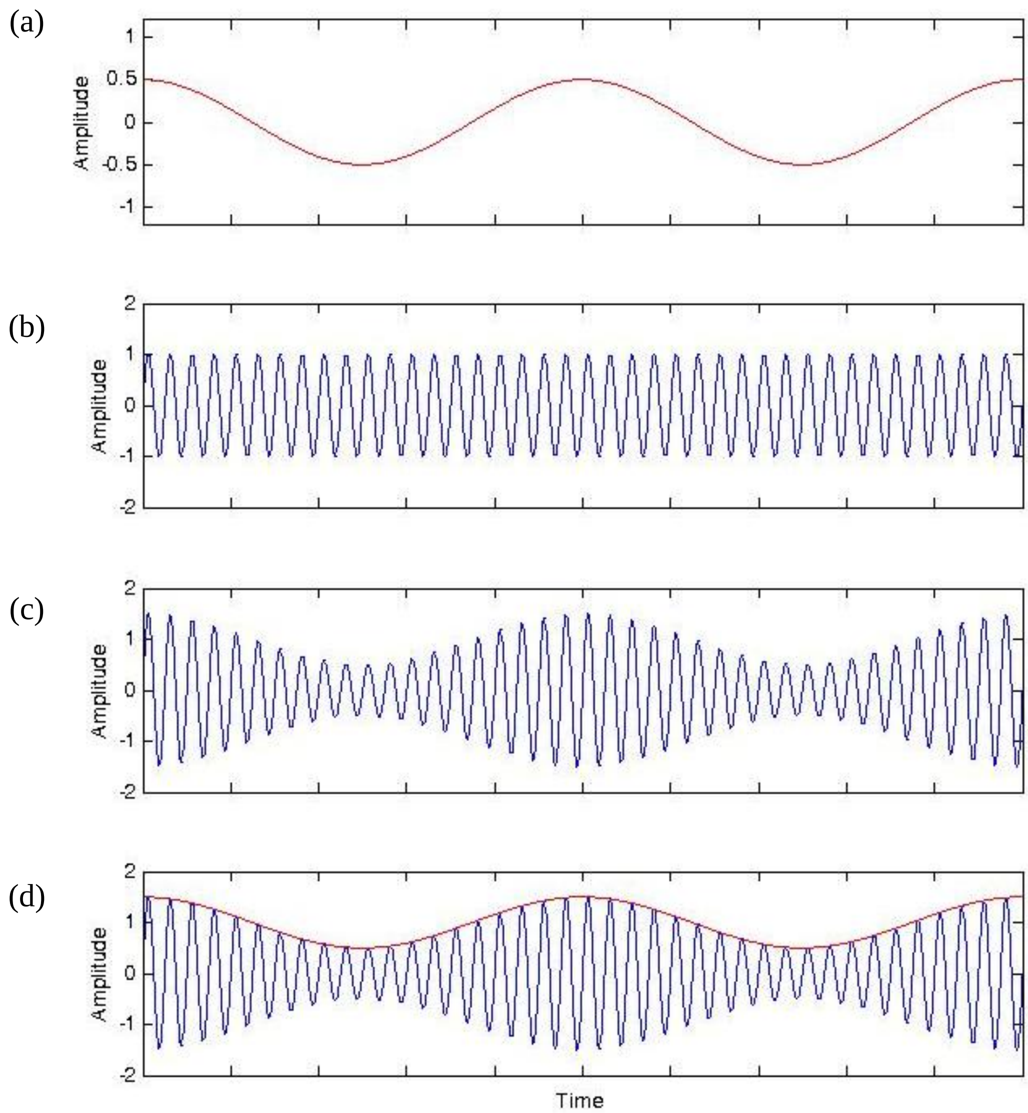


Figure 2.18 – Amplitude modulation for sinusoidal message signal
(a) message signal (b) constant frequency sinusoidal carrier
(c) amplitude modulated signal (d) amplitude modulated signal overlayed with sum of message signal plus DC offset

Figure 2.18 shows the amplitude modulated signal resulting when a sinusoidal message signal is used. Again the effects of the amplitude modulation are clearly visible in Figure 2.18 (c), and from Figure 2.18 (d) it can again be seen that the envelope of the amplitude modulated signal is directly given by the sum of the message signal with DC offset of 1.

To show the spectral properties of an amplitude modulated sinusoid, the simplest case to begin with is for a sinusoidal modulating signal, as shown in Figure 2.18. For this case, $A_m(t)$ can be expressed as:-

$$A_m(t) = A_{\mathrm{mod}} \cos(2\pi f_m t) \quad \text{......................................} \quad (2.18)$$

where A_{mod} is the amplitude of the sinusoid in the modulating signal, and f_m is the frequency of the same sinusoid. Equation (2.18) can be substituted into equation (2.17) on page 69, giving the amplitude modulated signal $y(t)$ of:-

$$y(t) = [A_{DC} + A_{\mathrm{mod}} \cos(2\pi f_m t)]\cos(2\pi f_c t) \quad \text{..........................} \quad (2.19)$$

By making use of the standard trigonometric identity:-

$$\cos(\theta)\cos(\varphi) = \frac{\cos(\theta - \varphi) + \cos(\theta + \varphi)}{2} \quad \text{..............................} \quad (2.20)$$

equation (2.19) can be expanded out to give:-

$$y(t) = A_{DC} \cos(2\pi f_c t) + \frac{A_{\mathrm{mod}}}{2} \cos(2\pi(f_c - f_m)t) + \frac{A_{\mathrm{mod}}}{2} \cos(2\pi(f_c + f_m)t) \quad . \quad (2.21)$$

As can be seen from equation (2.21), the amplitude modulated signal, for a sinusoidal message signal, is composed of the addition of three separate sinusoids. One sinusoid is located at original frequency of the carrier f_c, and two additional sinusoids are located around f_c at frequencies $(f_c + f_m)$ and $(f_c - f_m)$. In modulation theory, these two additional sinusoids are called sidebands, with the higher frequency sinusoid being the upper sideband, and the lower frequency sinusoid being the lower sideband.

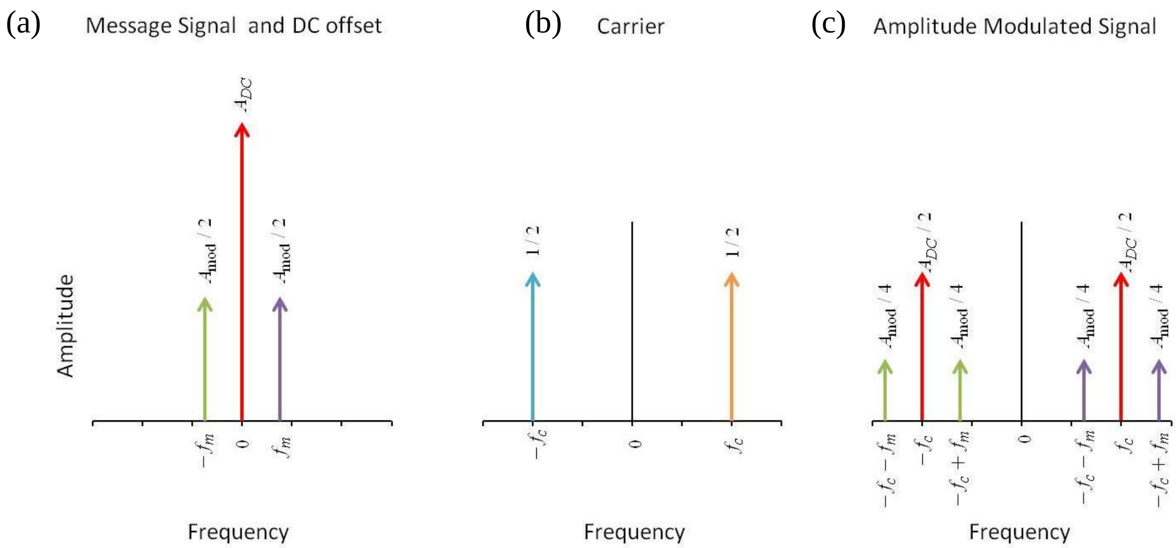


Figure 2.19 – Spectrum of amplitude modulated signal for sinusoidal message signal (a) Message signal and DC offset (b) Carrier (c) Amplitude modulated signal

Figure 2.19 shows the spectra for an amplitude modulated signal, for the case with a sinusoidal message signal. Figure 2.19 (a) shows the spectrum of the message signal summed with the DC offset, and (b) shows the spectrum of the carrier signal, where both are for a single sinusoid. Note that both spectra are plotted showing positive and negative frequency components. Figure 2.19 (c) shows the spectrum of the resulting amplitude modulated signal. As expected based on equation (2.21), the spectrum is made up from the three sinusoids. Note that both sidebands on the positive frequency side result from the positive component of the message signal and DC offset spectrum (a), and the sidebands on the negative frequency side result from the negative component of the message signal plus DC offset. It should be noted that the resulting spectrum in Figure 2.19(c) is the convolution of the message signal and DC offset in (a) with the carrier in (b), as both these components are multiplied in the time domain and so correspond to a convolution in the frequency domain. Convolution was described in detail in section 2.3 on page 57. Following the process of convolution, the amplitude of both the sidebands and DC offset component in the resulting signal (c) is half that from the message signal summed with the DC offset (a), as the components in (a) have been scaled by the carrier value from (b) of half unity.

For a message signal of a more general form (non-sinusoidal), the frequency spectrum shows a similar pattern to that from the case with a sinusoidal message signal.

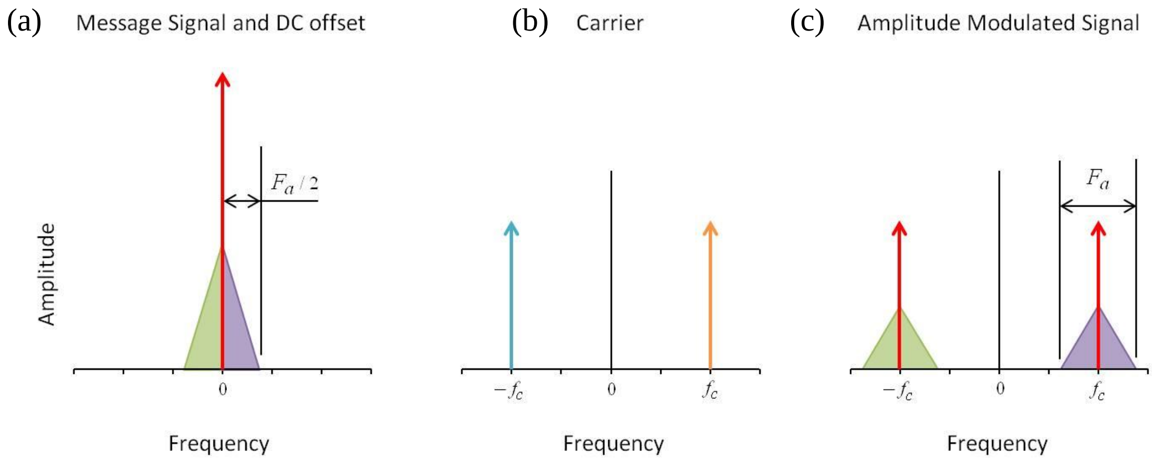


Figure 2.20 – Spectrum of an amplitude modulated signal for a non-sinusoidal message signal (a) Message signal (b) Carrier (c) Amplitude modulated signal

Figure 2.20 shows the spectra for amplitude modulation for a more general non-sinusoidal message signal. The message signal component in Figure 2.20 (a) is represented by a block of spectrum, where the bandwidth of the message signal is $F_a / 2$. Figure 2.20 (b) again shows the sinusoidal carrier component. In Figure 2.20 (c) it can be seen that the amplitude modulated signal is now made up of the DC offset component, and the spectrum of the message signal which has been shifted to be just above f_c, and then mirrored about f_c to form the information below the carrier frequency. The negative frequency components have been similarly formed around $-f_c$.

As can be seen in Figure 2.20, the bandwidth of the resulting spectrum of the amplitude modulated signal is given by F_a, where the bandwidth of the message signal is $F_a / 2$.

For the more general case shown here, in amplitude modulation theory the part of the message signal spectrum shifted above f_c is referred to as the upper sideband. Similarly, that part mirrored below f_c is referred to as the lower sideband. It should be noted that this is in contrast to the usage of the term sideband as used in Frequency modulation, as will be discussed in section 2.5.3 on page 80, and care should be taken to avoid confusing the different usage of the term sideband. In this thesis work, the term sideband will generally be referring to the definition used for frequency modulation, rather than that used for the topic of amplitude modulation.

2.5.2 Amplitude modulation with a harmonic series carrier

In the previous section 2.5.1, the theory of amplitude modulation was given for a sinusoidal carrier signal, where a single frequency sinusoid had its amplitude varied based on a message signal.

For a more complicated carrier signal, where the basic signal being varied consists of a harmonic series, such as a pulse train, amplitude modulation can be defined based on amplitude modulation of the multi-component carrier signal.

For this case, the combined amplitude modulated signal $y(t)$ can be expressed as the sum of a series of amplitude modulated sinusoidal carrier signals as given by equation (2.17) on page 69, for each harmonic order n of the carrier signal, where the carrier frequency for each harmonic order is given by nf_c, so that:

$$y(t) = \sum_{n=1}^{\infty} y_n(t) = \sum_{n=1}^{\infty} [A_{DC} + A_m(t)] C_n \cos(2\pi n f_c t) = [A_{DC} + A_m(t)] \sum_{n=1}^{\infty} C_n \cos(2\pi n f_c t) \quad (2.22)$$

where f_c is the frequency of the fundamental order of the carrier harmonic series (first harmonic), and nf_c is the frequency of the nth harmonic of the carrier harmonic series. C_n is the (dimensionless) amplitude scaling factor for carrier harmonic n, where the harmonics of the carrier may not have equal amplitudes. For example, if the carrier is a series of pulses, the C_n would correspond to the harmonics of that pulse shape (e.g. rectangular or half cosine). It is suggested that C_1 is scaled to unity. The amplitude modulating function $A_m(t)$ and DC offset A_c are assumed to be the same for the whole signal and thus for all harmonics.

Similarly to that discussed in section 2.5.1 on page 68, the time domain signal for the amplitude modulated signal $y(t)$ from equation (2.22) will have an amplitude variation based directly on the message signal (plus DC offset). Due to the similarities with the single carrier case, a specific illustration is not given here for a non-sinusoidal carrier being amplitude modulated.

What is different is the resulting spectrum of the amplitude modulated signal.

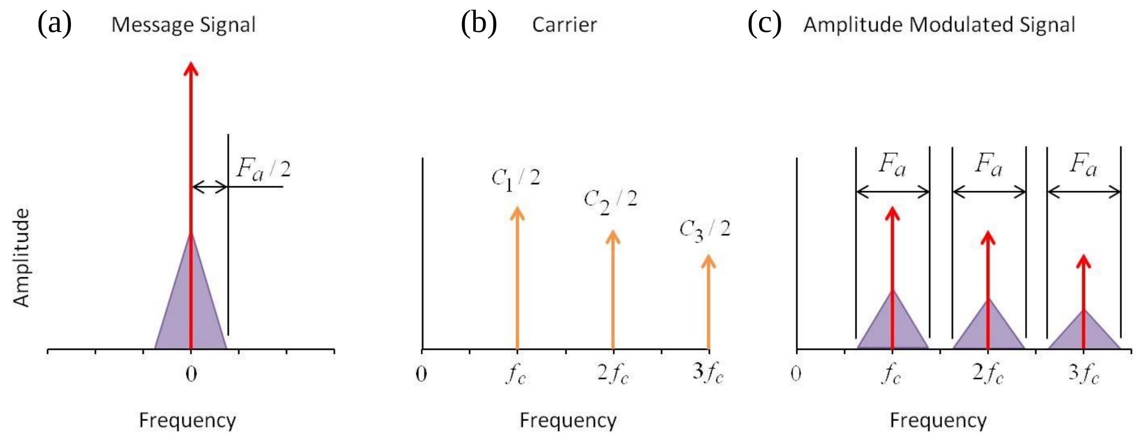


Figure 2.21 – Spectrum of an amplitude modulated signal for a harmonic series carrier with non sinusoidal message signal
(a) Message signal (b) Carrier (c) Amplitude modulated signal

Figure 2.21 shows the spectra for an amplitude modulated signal where the carrier is a harmonic series, with non sinusoidal modulation. Note that for the second and third spectra in (b) and (c), only one side of the spectrum has been shown, and negative frequencies have not been displayed. Figure 2.21 (a) shows the spectrum of a non-sinusoidal message signal, similarly to the case shown in Figure 2.20. Figure 2.21 (b) shows the spectrum for the first three orders of the carrier frequency which is a harmonic series, and would typically continue for all orders up to the Nyquist frequency. Figure 2.21 (c) shows the first three orders of the resulting amplitude modulated spectrum, and would continue for all the carrier harmonics of the carrier signal spectrum. As can be seen, each order of the carrier signal has now been replaced by the DC offset component which is surrounded by an upper side band, and a mirrored lower sideband.

Each order, consisting of a DC offset component plus a sideband pair, has an amplitude which has been scaled by C_n, and extends an equal amount around each harmonic order of the fundamental carrier frequency, and so the bandwidth of every harmonic order is identical, being F_a, where $F_a / 2$ is the bandwidth of the message signal.

2.5.3 Frequency modulation for sinusoidal carrier

The second main modulation type is produced by varying the phase angle of a constant frequency carrier, and is known as angle modulation.

As with amplitude modulation, angle modulation has primarily been studied in the field of communications for modulation of a sinusoidal carrier signal. As such, angle modulation of a sinusoidal carrier will first be discussed.

If a variation is expressed in terms of phase, the modulation of the signal is referred to as phase modulation. If a variation is instead expressed in frequency, which is effectively angular velocity, which is the derivative of phase, the modulation is referred to as frequency modulation.

It will however be shown that phase and frequency modulation are almost identical, and for practical purposes as applied to order-tracking (where the original message signal is unknown), can be treated the same.

A phase modulated signal $y(t)$ with a constant frequency sinusoid carrier is given by:-

$$A_c \cos(2\pi f_c t + \phi_m(t)) \quad \text{..} \quad (2.23)$$

where $\phi_m(t)$ is the variation in phase around the phase of the constant frequency sinusoid (carrier).

Similarly, a frequency modulated signal $y(t)$ with a constant frequency sinusoidal carrier is given by:-

$$y(t) = A_c \cos(2\pi f_c t + \int m(t)dt) \quad \text{.................................. (2.24)}$$

where $m(t)$ is the variation in frequency around the constant frequency sinusoid carrier.

As with amplitude modulation, $\phi_m(t)$ and $m(t)$ are both referred to as the message signal for their respective modulation types.

To illustrate the time domain representation of these signals, both modulation methods will firstly be shown using the same message signal (identical data as phase for one modulated signal, and frequency for the other signal, so $\phi_m(t) = m(t)$).

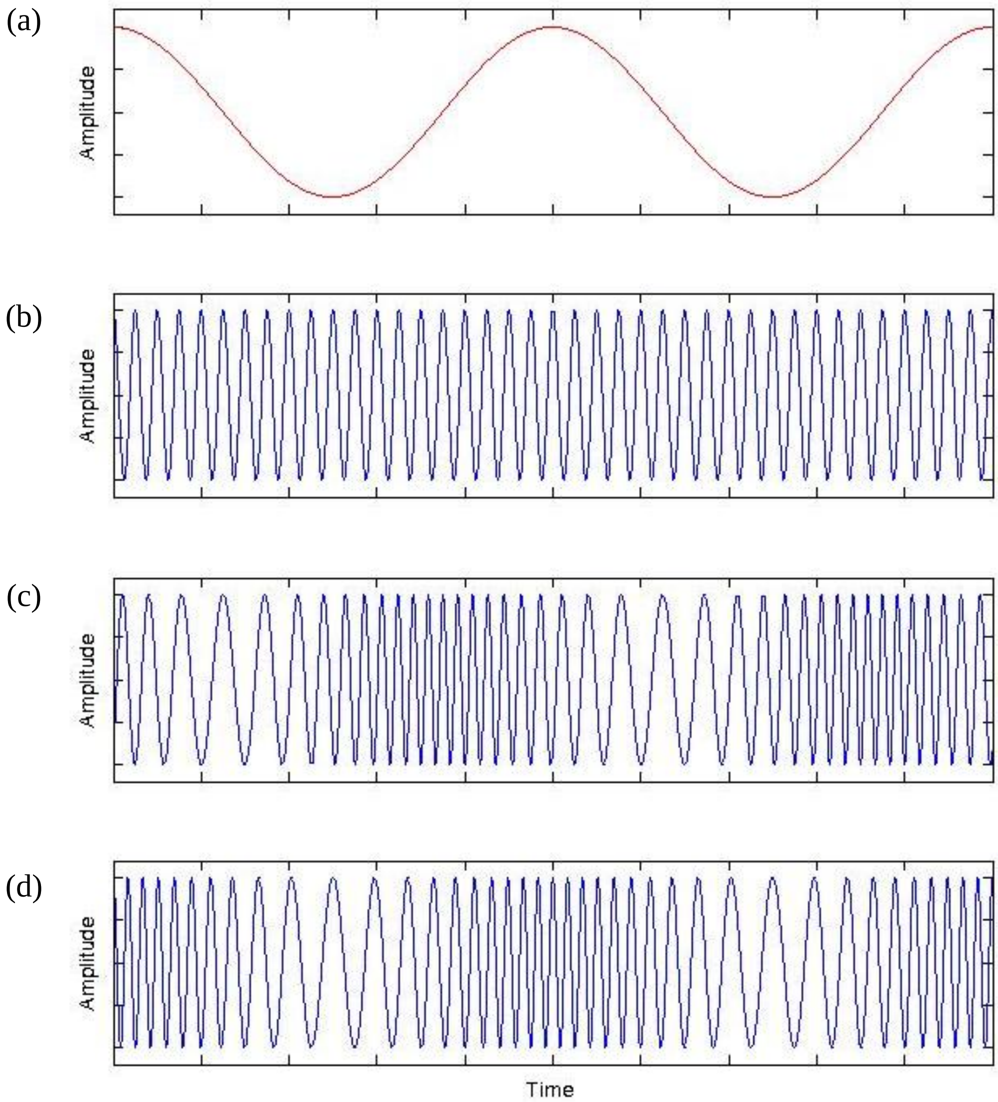


Figure 2.22 – Angle modulation for sinusoidal message signal
(a) Message signal (b) Constant frequency sinusoidal carrier
(c) Phase modulated signal (d) Frequency modulated signal

Figure 2.22 shows the results for angle modulation, when a sinusoidal message signal is used. Figure 2.22 (a) shows the sinusoidal message signal, and (b) shows the sinusoidal carrier frequency, similar to Figure 2.18 on page 72. Figure 2.22 (c) shows the results for the phase modulated signal, and it can be seen that the frequency of the carrier signal has now been varied, and the amplitude has remained unchanged (opposite to amplitude modulation). Figure 2.22 (d) shows the results for the frequency modulated signal, and it can be again seen that the frequency of the carrier signal is varying. Figure 2.22 (c) and (d) also show that the outputs for phase and frequency modulation are almost identical. For the phase modulation result in (c), the phase deviation is proportional to $\phi_m(t)$ (or $m(t)$), and in turn the frequency deviation is proportional to the derivative of the phase deviation. As a result, the phase modulated signal has the highest instantaneous frequency (closest spaced waveforms) when the message signal has the highest positive slope, and lowest instantaneous frequency (furthest spaced waveforms) when the message signal has the lowest slope. The frequency modulation result from (d) is directly proportional to the message signal $m(t)$ (or $\phi_m(t)$), so has the highest instantaneous frequency (closest spaced waveforms) when the message signal has a maximum, and lowest instantaneous frequency when the message signal has a minimum.

As can be seen in comparing Figure 2.22 (c) and (d), since the derivative of one sinusoid is another sinusoid, the main difference between the phase modulated signal and frequency modulated signal, for a single modulating frequency, is a phase shift.

For order-tracking applications, it is generally more useful to talk in terms of frequency deviation (resulting from changes in machine speed) rather than phase deviation, and so Frequency Modulation is the type of angle modulation which will be used to describe the systems in this thesis work. The following material in this section is based on frequency modulation, as defined by equation (2.24) on page 81.

From equation (2.24), the message signal $m(t)$ contains both the frequency deviation (variation) away from the carrier frequency value at each instant in time, as well as the frequency at which it changes, which known as the modulating frequency.

In order to separate the parameters for maximum frequency deviation from the carrier frequency, and the modulating frequency, it is common to factorise out the maximum frequency deviation value from the message signal. This means the message signal is constrained to have amplitude values in the range of $[-1,1]$ (normalised message signal), and the maximum frequency deviation could be brought out of the integral.

The time domain formula defining a frequency modulated signal can then be expressed as:-

$$y(t) = A_c \cos\left[2\pi f_c t + 2\pi D \int m_p(t)dt\right] \quad \text{............................. (2.25)}$$

where A_c is the constant amplitude of the carrier frequency, f_c is the constant frequency of the carrier (in Hz), D (Hz) is the peak deviation of the message signal from the carrier frequency (maximum variation), and $m_p(t)$ is the normalised message signal with amplitude range $[-1,1]$.

Not shown in equation (2.25) is the additional parameter W (Hz), which isdefined as the maximum modulating frequency, which is the maximum frequency of message signal $m_p(t)$ (baseband bandwidth of $m_p(t)$). Note that W is not simply the highest instantaneous frequency of the message signal, as the highest instantaneous frequency is typically lower than the baseband bandwidth. This is because for a non-sinusoidal variation, the message signal itself can be considered as another FM signal. As will be shown later in this section, the bandwidth of a FM signal is greater than the difference between maximum and minimum instantaneous frequency, and hence the baseband bandwidth is greater than the maximum instantaneous frequency.

When this formula is implemented in a computer in discrete form, the integral term is typically replaced by the cumulative sum of the message signal multiplied by the sampling period of the message signal. The discrete time domain definition of a frequency modulated signal would then take the form:

$$y(n) = A_c \cos\left(2\pi \frac{f_c}{f_s} n + \frac{2\pi f_d}{f_s} \times \text{cumsum}[m_p(n)\Delta t]\right) \text{ .... (2.26)}$$

where $1/f_s$ is the sampling period Δt of the message signal, with f_s being the sampling rate, and n is the discrete series corresponding to continuous time t $(= n\Delta t)$ in this instance.

To show the spectral properties of a frequency modulated signal, as with the amplitude modulation case in section 2.5.1 on page 68, it is easiest to begin with the simple case of a sinusoidal message signal, and then extend this case to general message signals.

For a sinusoidal case, the peak deviation value D, which is a constant for a sinusoidal message signal, is defined as f_d. Similarly, the modulation frequency W, which is again a constant, is defined as f_m.

For the sinusoidal modulation case the normalised message signal $m_p(t)$ is now:

$$m_p(t) = \cos(2\pi f_m t) \text{ .... (2.27)}$$

Substituting into equation (2.25) gives the frequency modulated signal with sinusoidal modulation $y(t)$ as:

$$y(t) = A_c \cos\left[2\pi f_c t + 2\pi f_d \int \cos(2\pi f_m t) dt\right] \text{ .... (2.28)}$$

where $y(t)$ is the FM signal, A_c is the constant magnitude of the carrier signal, f_c is the carrier signal frequency, f_d is the frequency deviation of the sinusoidally varying message signal (equivalent to D in Equation (2.25)), f_m is the modulation frequency or frequency of the sinusoidally varying message signal (equivalent to W).

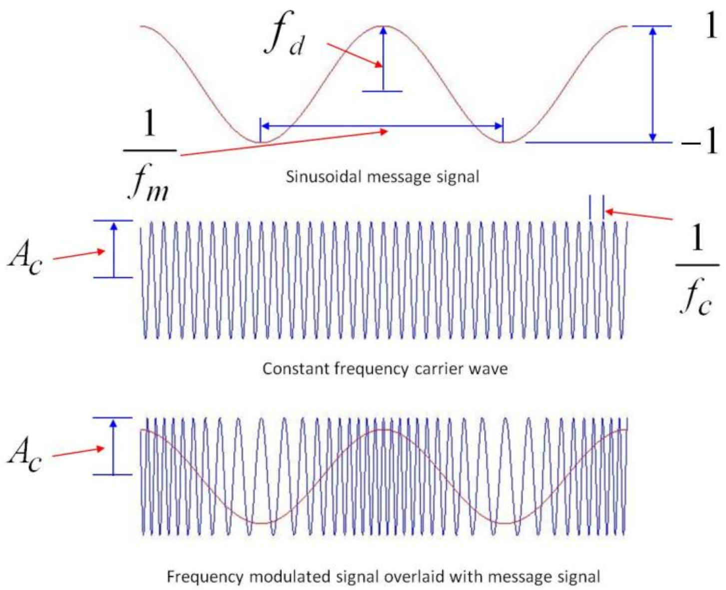


Figure 2.23 – Parameters of interest of a frequency modulated sinusoid with sinusoidal modulation

Figure 2.23 shows a frequency modulated sinusoid with sinusoidal modulation, with all the parameters of interest shown. For the PDOT methods developed in this work in the following chapters, the carrier frequency f_c will always be considered as the middle frequency value of the bandwidth, equidistant from the maximum and minimum values. f_d (and D) will then be the equal maximum positive and negative deviation away from the carrier frequency.

To determine the frequency spectrum of a frequency modulated signal with sinusoidal modulation, from equation (2.28), the signal can be expressed as a Fourier series, by using equation (2.2) on page 35, which gives $y(t)$ as:

$$
\begin{aligned}
y(t) = \quad & A_c J_0(\beta)\cos(\omega_c t) \\
& -A_c J_1(\beta)[\cos(\omega_c - \omega_m)t - \cos(\omega_c + \omega_m)t] \\
& +A_c J_2(\beta)[\cos(\omega_c - 2\omega_m)t + \cos(\omega_c + 2\omega_m)t] \\
& -A_c J_3(\beta)[\cos(\omega_c - 3\omega_m)t - \cos(\omega_c + 3\omega_m)t] \\
& + \ldots \\
& -A_c J_{k-1}(\beta)[\cos(\omega_c - (k-1)\omega_m)t - \cos(\omega_c + (k-1)\omega_m)t] \\
& +A_c J_k(\beta)[\cos(\omega_c - k\omega_m)t + \cos(\omega_c + k\omega_m)t]
\end{aligned}
$$

For $k = 0,1,2,...,\infty$ .... (2.29)

where ω_c and ω_m are the angular velocities corresponding to f_c and f_m, respectively.

β is called the modulation index, and is defined as:

$$\beta = \frac{f_d}{f_m} \quad \text{.. (2.30)}$$

$J_k(\beta)$ is called a Bessel function of the first kind, and is defined as:

$$J_k(\beta) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} e^{j(\beta \sin y - ky)} dy \quad \text{.................................. (2.31)}$$

After substituting equations (2.30) and (2.31) into equation (2.29), the frequency spectrum of the frequency modulated signal is then equal to the frequency spectrum of the Fourier series from equation (2.29).

The positive side of the frequency spectrum (with the negative side being a mirror of the positive, as it is a real signal) of the Fourier series $y(t)$ contains a carrier component, and an infinite number of sideband pairs.

The carrier frequency component is obtained when $k=0$, at frequency f_c and of magnitude $\frac{1}{2}A_c\left|J_0(\beta)\right|$. This is formed from the first line of equation (2.29).

The infinite number of sidebands are comprised of sideband pairs, with a pair of sidebands for each k, for the range $k=[1,\infty]$. Each sideband pair is from one line from equation (2.29), after the first line, which is the carrier frequency

The sidebands are spaced by f_m around the carrier frequency f_c, so that for the sidebands corresponding to $k=1$, the two sidebands would be located at the frequency values equal to f_c+1f_m and f_c-1f_m, respectively. In the general case, the two sidebands would be located at a frequency values of $f_c \pm k \times f_m$.

The magnitude of the sidebands is given by Bessel functions of the first order, such that the magnitude of the sideband pair corresponding to $k=1$ would be given by $\frac{1}{2}A_c\left|J_1(\beta)\right|$, and in the general case, the magnitude of the sidebands would be $\frac{1}{2}A_c\left|J_k(\beta)\right|$.

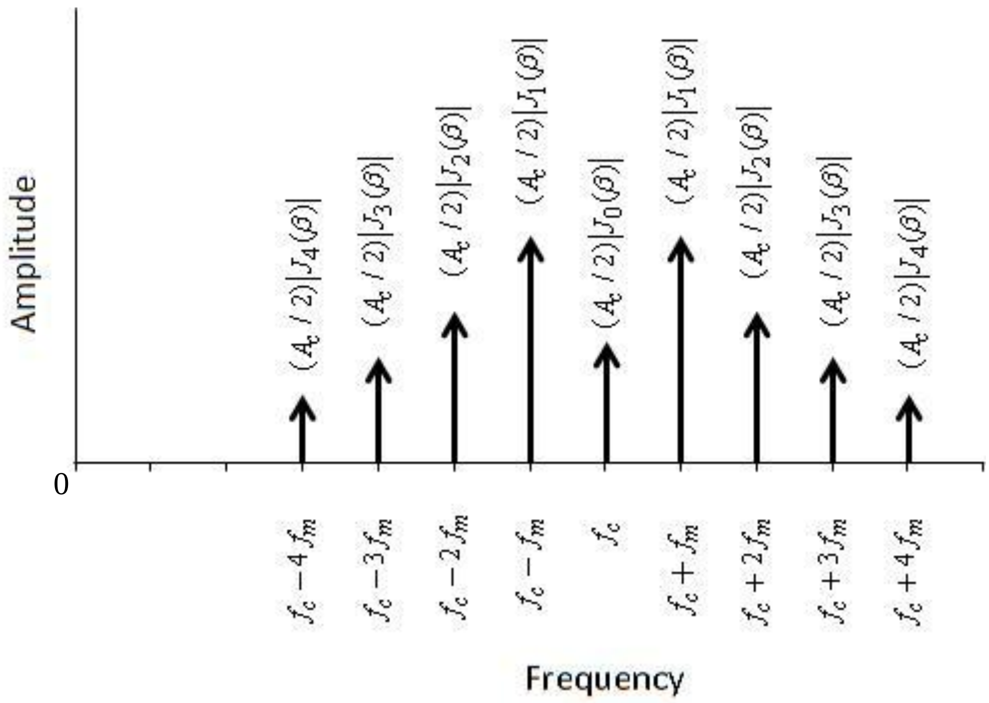


Figure 2.24 – Sideband magnitude and spacing for positive side of the Frequency spectrum of a Frequency modulated signal with sinusoidal modulation, showing first four sideband pairs

Figure 2.24 shows graphically the position and magnitude of the carrier frequency component at f_c, and the spacing and magnitude of the first four sideband pairs, corresponding to k = 0 to 4. This sideband pattern continues for all k values of the range $[1,\infty]$, to give an infinite number of sidebands.

It should be specifically noted that the definition of a sideband for frequency modulation differs to that for amplitude modulation, as presented in section 2.5.1 on page 68. For frequency modulation, each discrete frequency value spaced away from the carrier frequency is called a sideband, whereas for amplitude modulation all frequency values to one side of the carrier frequency are collectively referred to as one “sideband”.

As has been stated, the frequency spectrum of the frequency modulated signal with sinusoidal modulation contains an infinite number of sidebands, the magnitude of which is given by Bessel functions of the first type. Bessel functions of the first type are not linear, so different values of β result in dramatically different frequency spectra. There are even certain values of β for which $J_0(\beta)=0$, so a carrier component is not present, resulting in a markedly different spectrum.

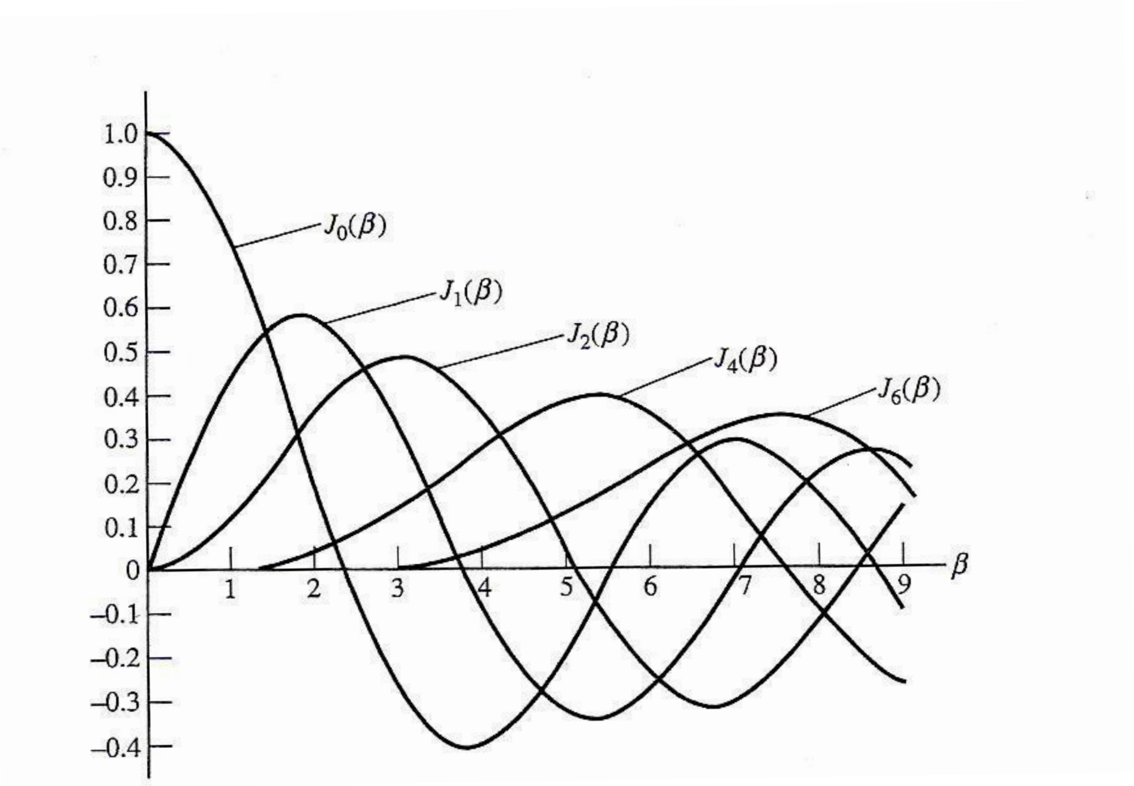


Figure 2.25 – First 7 Bessel functions of the first kind for different values of β. [35]

Figure 2.25 shows the first 7 Bessel functions for different values of β. The magnitudes of the sidebands for a given β value are proportional to the values of the Bessel functions at that β value, as per equation (2.29) on page 87, and above.

However, while Bessel functions result in a complicated spectrum, $J_n(\beta)$, for a given value of β, always tend to zero for increasing values of n. As a result, there are only a finite number of Bessel functions which are not approximately equal to zero, and hence there is a finite spread of frequency values, or bandwidth, which the non-zero sidebands span in the frequency domain, for any given β value. Sidebands which are not approximately equal to zero are commonly called significant sidebands.

This trend is best illustrated by firstly expressing a Bessel function of the first kind as a series expansion, which can be examined in the limit as n approaches ∞. Ziemer and Tranter [35] give this series expansion as:

$$J_n(\beta) \cong \begin{cases} \dfrac{\beta^n}{2^n n!}\left[1 - \dfrac{\beta^2}{2^2\,(n+1)} + \dfrac{\beta^4}{2 \cdot 2^4\,(n+1)\,(n+2)} - \cdots\right], \\ \sqrt{\dfrac{2}{\pi\beta}}\cos\left(\beta - \dfrac{n\pi}{2} - \dfrac{\pi}{2}\right), \quad \beta \gg 1 \end{cases} \quad \text{................ (2.32)}$$

The derivation of this series expansion is covered in mathematics reference works, such as Rukmangadachari and Keshava Reddy [38].

From this series (2.32), for large n it can be said that:

$$J_n(\beta) \propto \frac{\beta^n}{2^n n!} \quad \text{.. (2.33)}$$

Thus for a given β, the limit as n approaches ∞ is:

$$\lim_{n\to\infty} J_n(\beta) = 0 \quad \text{.. (2.34)}$$

This limit establishes that all frequency modulated signals will have a finite non-zero bandwidth. This can be seen in the following frequency spectra corresponding to different β values.

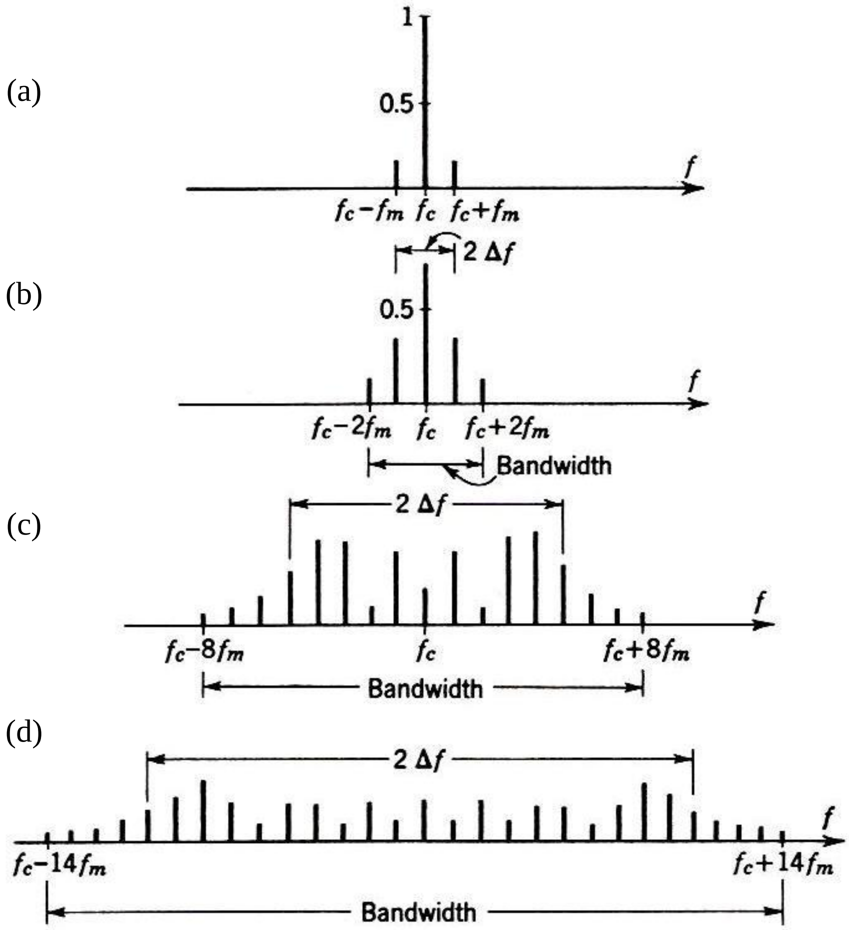


Figure 2.26 – Frequency spectra for sinusoidal Frequency modulated signals with different β values. [37]

(a) $\beta = 0.2$ (b) $\beta = 1$ (c) $\beta = 5$ (d) $\beta = 10$

It can be seen that in every frequency modulated signal in Figure 2.26 there is only a finite number of non-zero (significant) sidebands, and hence each signal has a finite bandwidth.

From Figure 2.26 it can be seen that calculating the bandwidth of a frequency modulated signal is not a trivial exercise. Above each case in Figure 2.26 is indicated the range $2\triangle f$, equal to $2f_d$, which is effectively the difference between the maximum and minimum instantaneous frequency of the signal. As can be seen, for all except the first case (a), visible sidebands extend beyond the range of $2\triangle f$, and so the bandwidth of a frequency modulated signal is not simply the difference between maximum and minimum instantaneous frequency.

As the bandwidth for a frequency modulated signal is theoretically infinite, and so the definition of bandwidth is based on what constitutes a significant sideband, specifying the bandwidth is not as straight forward as with amplitude modulation.

For amplitude modulation, only one definition of bandwidth is needed, as bandwidth is finite for amplitude modulation and so exactly deterministic. For frequency modulation, different definitions for a significant sideband result in different methods of determining bandwidth.

The simplest method to calculate the bandwidth of a frequency modulated signal is to use Carson's rule [39], which is an approximation of the bandwidth of a frequency modulated signal with a sinusoidal carrier.

Carson's rule is typically used for a frequency modulated sinusoid with non sinusoidal modulation, and so is not limited to the case of a sinusoidal message signal which is first being considered here.

For frequency modulation with a sinusoidal carrier, the bandwidth of the FM signal will be specified as F. Carson's rule thus defines bandwidth as:-

$$F = 2(D + W) \quad \text{...} \; (2.35)$$

where, as earlier defined, D is the peak frequency deviation of the message signal, and W is the highest frequency of the message signal (baseband bandwidth).

By including the introduced restriction that will be used within this thesis work, that f_c is the midpoint between the maximum and minimum frequency values of the message signal, Carson's rule from equation (2.35) can alternatively be expressed as being a range around the carrier frequency of:

$$F = \pm(D + W) \text{ around } f_c \quad \text{.....................................} \quad (2.36)$$

For the case of sinusoidal modulation, parameters f_d and f_m can be substituted for D and W in equations (2.35) or (2.36).

Bandwidth calculated by Carson's rule is said to contain 98% of the power of an FM signal, which means that the sum of the squares of the significant sidebands is 98% of the sum of the squares of all sidebands for a FM signal. This corresponds to the power outside the bandwidth being approximately 17dB lower than the power within the bandwidth.

However, examination of Figure 2.26 on page 92 shows that Carson's rule is effectively a relatively coarse definition for significant sidebands. For the examples shown, the bandwidth using Carson's rule is the equivalent of the range shown as $\triangle f$ plus the additional width of two sidebands spacing $2f_m$ (one sideband pair, so one sideband each way beyond $\triangle f$). As can been seen from Figure 2.26, only figures (a) and (b) have all visible sidebands contained within the bandwidth as defined by Carson's rule. For the cases (c) and (d), there are respectively two and three clearly visible sideband pairs extending beyond the bandwidth defined by using Carson's rule.

To calculate bandwidth to a greater degree of accuracy than with Carson's rule, a specific definition of a significant sideband needs to be developed.

For the case of a sinusoidal message signal, with a suitable definition of a significant sideband, the bandwidth can be calculated. For the sinusoidal case with sidebands of magnitude $\frac{1}{2}A_c\left|J_k(\beta)\right|$, it is possible to find the series of integer k values corresponding to all significant sidebands. The highest significant sideband order $k_{\max}$ is then the largest integer value of k corresponding to a significant sideband, based on the definition of a significant sideband used.

The bandwidth of the FM signal F for a sinusoidal message signal, is then found from the sideband spacing and the highest significant sideband order $k_{\max}$, such that:

$$F = 2k_{\max} f_m = \pm k_{\max} f_m \text{ around } f_c \quad \text{.............................} (2.37)$$

There are numerous definitions for a significant sideband, which are applicable in different situations. Many of the common definitions used in the communications field require direct access to parameters from the message signal, which is commonly not accessible for order-tracking applications, and so are not suitable to the work presented in this thesis. The two most common definition types, as presented in reference works, are discussed here.

The first common sideband definition type uses a percentage threshold, where a sideband is significant if its magnitude is greater than the percentage threshold multiplied by a reference value.

Various parameters can then be used as a reference value for this definition type, however there are three parameters commonly used.

The first reference value commonly used is the value of $\left|J_0(0)\right|$, which is unity, corresponding to the magnitude of the un-modulated carrier A_c.

The second is the magnitude of the carrier component for the applicable β value, which is $|J_0(\beta)|$.

The third is the largest magnitude of the Bessel functions for that particular β value, so the maximum value from the series $|J_o(\beta)|, |J_1(\beta)|, |J_2(\beta)|, ..., |J_k(\beta)|$ for $k = 0,1,2,...,\infty$. It should be noted that this reference value will be used to calculate bandwidths in this thesis work, and is used in section 3.4.1 on page 147.

These three reference values are each applicable to different situations, however the second definition is often avoided due to the fact that the carrier frequency is zero for some β values.

As an example, the following is the method of determining the value over which a sideband is significant for a percentage threshold value, using the three common reference value definitions.

For a given value $\beta = 4.5$, the value for $|J_0(\beta)| = |-0.3205| = 0.3205$, and the maximum $|J_k(\beta)|$ value is for $|J_3(\beta)| = |0.4247| = 0.4247$. If a percentage limit is to be set at 40dB, so that a sideband is significant if its magnitude is greater than 1% of the reference value,

- For the first reference value, the value of the un-modulated carrier, any sideband with magnitude greater than $0.01 \times |J_o(0)| = 0.01 \times 1 = 0.01$ would be significant.
- For the second reference value, the value of the carrier component, any sideband with magnitude over $0.01 \times |J_0(\beta)| = 0.01 \times 0.3205 = 0.0032$ would be significant.

- For the third reference value, the value of the maximum $|J_k(\beta)|$ component, any side band with magnitude over $0.01 \times \max|J_k(\beta)| = 0.01 \times 0.4247 = 0.0042$ would be significant.

The second common definition type for determining which sidebands are significant makes use of the fact that the power of a signal is preserved during frequency modulation. This means that the power of an un-modulated signal is equal to the sum of the powers of every sideband of a modulated signal. It is then possible to define significant sidebands as the series of sidebands which contain a set percentage of the original signal, such as 90 or 99% of the original signal power. This is the same principle as that from which Carson's rule, and equation (2.35) on page 93, were developed.

As an example, if a power percentage of 99% were to be used, significant sidebands would be determined by firstly working out the power contained in the carrier frequency component. The power from the first pair of sidebands corresponding to $k=1$ could then be added to the carrier frequency component, and then the power from the second pair of sidebands could be added to this total, and so on for increasing values of k. This addition would continue until the sum of the power just exceeded the percentage value multiplied by the power from the un-modulated constant frequency carrier wave (power of carrier is $|J_0(0)|^2 = 1^2 = 1$). Once the power sum is over the required power percentage value, this sideband pair corresponds to $k_{\max}+1$, with the preceding sideband pair being $k_{\max}$. $k_{\max}+1$ and any further sidebands are considered negligible.

Once all the significant sidebands k have been determined, based on a suitable definition of a significant sideband, it is an easy matter to work out the $k_{\max}$ value corresponding to the highest significant sideband pair. The bandwidth can then be calculated using equation (2.37) on page 95.

The frequency spectrum for a realistic frequency modulated signal which has a non-sinusoidal message signal, as per equation (2.25) on page 84, is more complex than for a frequency modulated signal with a sinusoidal message signal. However, the frequency spectra for both cases will be of approximately the same form.

For a frequency modulated signal with a non-sinusoidal message signal, the frequency spectrum will contain a carrier frequency component, and a series of individual sidebands. The magnitudes of the higher order sidebands will tend to zero, so the frequency spectrum information will again have a finite bandwidth.

The carrier frequency component will still be in the centre of the bandwidth of the signal, simply because in this thesis work it is defined as the frequency value half way between the maximum and minimum frequencies present in the frequency modulated signal. Note that this restriction is specific to this thesis work, and is not a definition for frequency modulation in general, as presented in reference works.

However the number of sidebands, and the sideband spacing, can differ significantly from a frequency modulated signal with sinusoidal modulation. Typically an FM signal with non sinusoidal modulation will have a spectrum containing a larger number of sidebands, with closer sideband spacing. The sideband spacing is also no longer necessarily uniform. It should also be noted that while theoretically the spectrum still contains discrete sidebands, for a discretely sampled spectrum from an FFT which has a finite frequency resolution, if the sideband spacing is lower than the frequency resolution, the frequency spectrum can smear into a single broad peak, or into multiple conjoined sections, rather than showing discrete sidebands.

Specifying the specific frequency spectral components for a case of non-sinusoidal modulation is extremely complex, and will not be attempted here. This is also beyond the scope of most reference works on the subject of frequency modulation, and is not readily available.

As the bandwidth for Carson's rule, from equation (2.35) on page 93, is defined in terms of the peak frequency deviation D and the maximum modulating frequency W, Carson's rule can be directly employed to calculate the bandwidth for a frequency modulated signal with a non-sinusoidal message signal. As with the sinusoidal message signal case, this definition of bandwidth will still be relatively coarse.

The method shown above of calculating bandwidth using Bessel functions cannot be directly employed for the case of a non-sinusoidal message signal. However, this method can be used to give an approximation of the bandwidth that is typically conservative.

To calculate this bandwidth approximation, the sinusoidal frequency deviation parameter f_d can be replaced by D, and the modulating frequency of the sinusoidal message signal can be replaced by W. For the approximation, the modulation index from equation (2.30) on page 87 is now expressed as:

$$\beta = \frac{D}{W} \quad \text{...} \quad (2.38)$$

In the same manner as described for a frequency modulated signal with sinusoidal modulation, the series of Bessel functions of the first type for $k = 0,1,2,...,\infty$

can be solved for this new β value. The number of significant sidebands can then be determined for an appropriate definition, and then the largest k value $k_{\max}$ corresponding to a significant sideband determined in the same manner as described for a frequency modulated signal with sinusoidal modulation.

It is important to note that sidebands will not necessarily be present in the exact same positions as for a frequency modulated signal with sinusoidal modulation, so these calculations will not actually correspond to the real number of significant sidebands present. However the value for $k_{\max}$ calculated in this manner can be used to approximate the bandwidth with:

$$F = 2k_{\max}W = \pm k_{\max}W \text{ around } f_c \quad \text{(2.39)}$$

Bandwidth calculations done in this manner are approximate, but are typically sufficiently accurate. As stated, this method of bandwidth calculation is generally conservative, but a safety factor can be used if slightly understating the bandwidth would be a problem.

2.5.4 Frequency modulation with a harmonic series carrier

In the previous section 2.5.3, the theory of frequency modulation was given for a sinusoidal carrier signal, where a single frequency sinusoid had its frequency varied based on a message signal.

For a more complicated carrier signal, where the basic signal being varied consists of a harmonic series, such as a pulse train, frequency modulation (or more correctly, modulation of pulse spacing) can be defined based on the frequency modulation of a sinusoidal carrier, in a similar fashion to amplitude modulation as discussed in section 2.5.2 on page 77. It should be kept in mind that the common factor relating all higher harmonics of the carrier is that the fundamental modulation is of the time spacing between events (e.g. the individual pulses of a tacho signal), and thus the phase variation corresponding to a given time variation is simply scaled by the harmonic order, and directly proportional to it.

So for this case, the combined frequency modulated signal $y(t)$ can be expressed as the sum a series of frequency modulated sinusoidal carrier signals as given by equation (2.25) on page 84, for each harmonic order n of the carrier signal, where the carrier frequency for each harmonic order is given by nf_c, so that:

$$y(t)=\sum_{n=1}^{\infty} y_n(t)=A_c\sum_{n=1}^{\infty} C_n \cos\left[2\pi n f_c t+2\pi n D\int m_p(t)dt\right] \quad \text{................ (2.40)}$$

where f_c is the frequency of the fundamental order of the carrier harmonic series (first harmonic), and nf_c is the carrier frequency of the nth harmonic. C_n is the amplitude scaling factor for harmonic n, where the harmonics of the carrier may not have the same amplitude scaling factor as the fundamental order. It should be noted that the carrier amplitude A_c has physical dimensions and units, and the amplitude scaling factor C_n is dimensionless. Note that as with amplitude modulation of a harmonic series carrier, in section 2.5.2 on page 77, it is suggested that C_1 is scaled to unity, so A_c gives the amplitude of the fundamental order $n = 1$.

From equation (2.40) it can firstly be seen that the frequency variation is not constant for all orders, which is dissimilar to the amplitude modulation case where the amplitude variations were independent of order. As a result, the spectrum of a higher harmonic is not simply a copy of the fundamental frequency component shifted by the carrier frequency, as is the case for amplitude modulation. The addition of the harmonic order term n to the frequency variation makes the frequency deviation term D proportional to harmonic order n.

However, it can also be noted from equation (2.40) that the normalised message signal remains constant for all harmonic orders, which means that the peak modulating frequency W is not proportional to harmonic order. This has the effect that the bandwidth of higher harmonics n is not just simply proportional to n.

To develop the spectral representation of each harmonic order, which combine to give the total spectrum, as with the case of a sinusoidal carrier from 2.5.3 it is easiest to begin with the simplest case of the message signal being sinusoidal. As with the case of a sinusoidal carrier, the general variation parameters D and W can be replaced with their constant counterparts f_d and f_m for a sinusoidal message signal.

The normalised sinusoidal message signal is again equal to equation (2.27) on page 85, which can be substituted into equation (2.40) to give the total signal $y(t)$ as:

$$y(t)=\sum_{n=1}^{\infty} y_n(t)=A_c\sum_{n=1}^{\infty} C_n\cos\left[2\pi n f_c t+2\pi n f_d\int\cos(2\pi f_m t)dt\right] \quad \text{............ (2.41)}$$

where $y_n(t)$ is one harmonic order of the total frequency modulated signal $y(t)$.

To determine the frequency spectrum of one order $y_n(t)$ of frequency modulated signal with sinusoidal modulation, from part of equation (2.41), the signal can be expressed as a Fourier series, by using equation (2.2) on page 35. This is in an identical fashion as with equations (2.29) , (2.30) and (2.31) on page 87, and results in:

$$\begin{aligned}
y_n(t)= \quad & A_c C_n J_0(\beta_n)\cos(n\omega_c t) \\
& -A_c C_n J_1(\beta_n)[\cos(n\omega_c-\omega_m)t-\cos(n\omega_c+\omega_m)t] \\
& +A_c C_n J_2(\beta_n)[\cos(n\omega_c-\omega_m)t+\cos(n\omega_c+\omega_m)t] \\
& -A_c C_n J_3(\beta_n)[\cos(n\omega_c-3\omega_m)t-\cos(n\omega_c+3\omega_m)t] \\
& + \ \dots \\
& -A_c C_n J_{k-1}(\beta_n)[\cos(n\omega_c-(k-1)\omega_m)t-\cos(n\omega_c+(k-1)\omega_m)t] \\
& +A_c C_n J_k(\beta_n)[\cos(n\omega_c-k\omega_m)t+\cos(n\omega_c+k\omega_m)t]
\end{aligned}$$

For $k=0,1,2,...,\infty$ .... (2.42)

where ω_c and ω_m are the angular velocities corresponding to f_c and f_m, respectively.

$J_k(\beta_n)$ is again the Bessel function of the first kind, now defined as:

$$J_k(\beta_n) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} e^{j(\beta_n \sin y - ky)} dy \quad \text{........................ (2.43)}$$

The main difference, when compared to equation (2.30) on page 87, is the modulation index β_n, which for the harmonic order n is now:

$$\beta_n = \frac{nf_d}{f_m} \quad \text{.. (2.44)}$$

This change to the modulation index has the resulting effect that the spectrum for each harmonic order $y_n(t)$ is not simply a proportional change based on n.

The spectrum for $y_n(t)$ resulting from equation (2.42), will have an identical form to that for a sinusoidal carrier signal, in that the positive spectrum consists of a carrier frequency component, with an infinite number of sidebands.

The spectrum in this case is again approximately finite, based on defining what constitutes a significant sideband. Although the sideband spacing is still equal to that from the sinusoidal carrier case, of f_m, the change in modulation index for different harmonics results in different sideband magnitudes. This results in a set of significant sidebands k^n which differs for each harmonic order n. In turn, the maximum sideband number $k^n_{\max}$ for each harmonic order n is now also different.

The bandwidth of harmonic order n, which is F_n, is thus given as:

$$F_n = 2k^n_{\max} f_m = \pm k^n_{\max} f_m \text{ around } nf_c \quad \text{............................ (2.45)}$$

As with the case for a sinusoidal carrier signal in the previous section 2.5.3, the bandwidth as calculated in equation (2.45) for a sinusoidal message signal can be extended to the case for a non-sinusoidal message signal, by replacing f_d with D, and f_m with W, to give an approximation of modulation index (similar to equation (2.38) on page 99) as:

$$\beta_n = \frac{nD}{W} \quad \text{......} \quad (2.46)$$

which can be used to calculate the bandwidth F_n for a non-sinusoidally modulated case using the modified form of (2.45):

$$F_n = 2k_{\max}^{n}W = \pm k_{\max}^{n}W \text{ around } nf_c \quad \text{......} \quad (2.47)$$

Although equation (2.45) (and (2.47)) alone just indicate that the bandwidth for different harmonic orders n will be different, by identifying that the modulation index from equation (2.44) (and (2.46)) is proportional to harmonic order n, a trend can be established for the bandwidth F_n. As can be seen from Figure 2.26 on page 92, increasing modulation index values result in increasing bandwidth of a frequency modulated signal. As such, as a result of the modulation index β_n being proportional to n, it can be said that the bandwidth F_n of $y_n(t)$ will increase as the harmonic order n increases.

This is more clearly evident if the alternative bandwidth measure of Carson's rule is used to calculate the bandwidth of $y_n(t)$. Carson's rule gives the bandwidth F_n for harmonic order n as:

$$F_n = 2(nD + W) = \pm(nD + W) \text{ around } nf_c \quad \text{......} \quad (2.48)$$

which clearly indicates that the bandwidth F_n increases for higher harmonic order n.

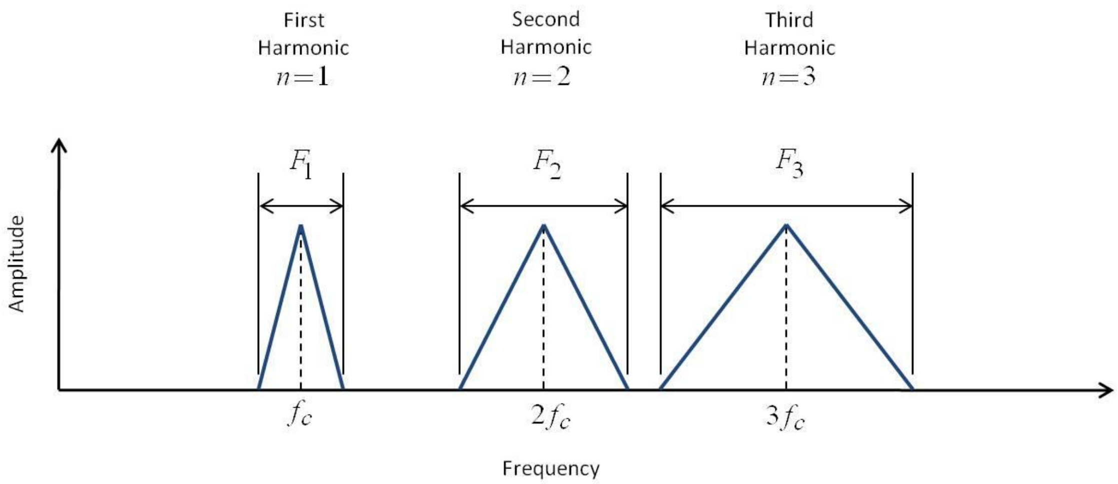


Figure 2.27 – Increasing bandwidth of F_n for increasing harmonic order n

Figure 2.27 graphically shows the bandwidths for the first three harmonics of a frequency modulated non-sinusoidal carrier, indicating the increasing bandwidth F_n for increasing harmonic orders n.

For the case where the peak frequency deviation D is much greater than the modulating frequency W, such that $D \gg W$, an approximation can be made that:

$$\beta_n = \frac{nD}{W} \approx \frac{nD}{nW} = \frac{D}{W} \quad \text{...} \ (2.49)$$

This approximation makes use of the fact that if W is very small, then nW is also very small, and so the multiplication by n does not significantly change the value of β_n. Use of this approximation of a constant β_n value for all orders results in a constant $k_{\max}^n$ value for all harmonics. This results in the bandwidth F_n of a harmonic being given by:

$$F_n = 2k_{\max}^n(nW) = 2nk_{\max}^n W = \ \pm nk_{\max}^n W \ \text{ around } \ nf_c \quad \text{................} \ (2.50)$$

This indicates that with this approximated modulation index, the bandwidth F_n for harmonic order n is approximately proportional to the bandwidth of the fundamental order F_1, such that:

$$F_n \approx nF_1 \quad \text{..} (2.51)$$

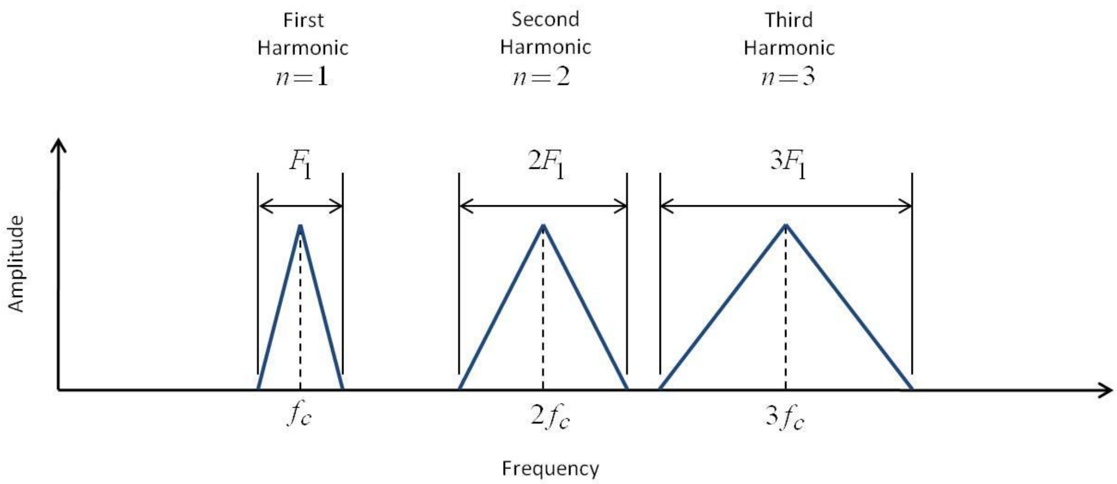


Figure 2.28 – Approximation that bandwidth is proportional to harmonic order

Figure 2.28 graphically shows the spectrum of a frequency modulated signal, where three harmonic orders are shown, indicating the proportionality of higher harmonics to the fundamental, when approximations are made based on the assumption that $D \gg W$. Note that bandwidths are now proportional to the bandwidth of the first harmonic, rather than being different for each harmonic order as in Figure 2.27.

It should be noted that while some formulae are presented in the body of this thesis work, in section 3.3 on page 137, and similar formulae have been presented by Bonnardot et al. [10] and Combet and Gelman [21], which are all based upon the assumption made for the modulation index in equation (2.49) for when $D \gg W$, resulting in a bandwidth proportional to harmonic order n, it is this authors recommendation that use of this assumption is preferably not used, and the bandwidth of each harmonic F_n calculated using the methods described earlier in this section to give the bandwidth equation (2.47) on page 105.

2.5.5 General modulation for sinusoidal carrier

General modulation is the name used when a signal is both amplitude and frequency (angle) modulated. This results in a more complicated signal than with either modulation type in isolation. As with amplitude and frequency modulation, initially the case for a sinusoidal carrier frequency will be discussed here.

A signal with general modulation and a sinusoidal carrier $y(t)$ can be expressed as:

$$y(t) = A_{DC} + A_m(t)\,\cos\left[2\pi f_c t + 2\pi D \int m_p(t)dt\right] \quad \text{.....................} \quad (2.52)$$

where $A_m(t)$ is the message signal of the amplitude modulation component, and $m_p(t)$ is the normalised message signal of the frequency modulation component. Note that as with amplitude modulation alone, as presented in sections 2.5.1 and 2.5.2, starting on page 68, the carrier amplitude A_c is now defined as being dimensionless and equal to unity, and so has been omitted from equation (2.52). It should also be noted that, as with the definition for amplitude modulation, the sum of $A_{DC} + A_m(t)$ must always be positive.

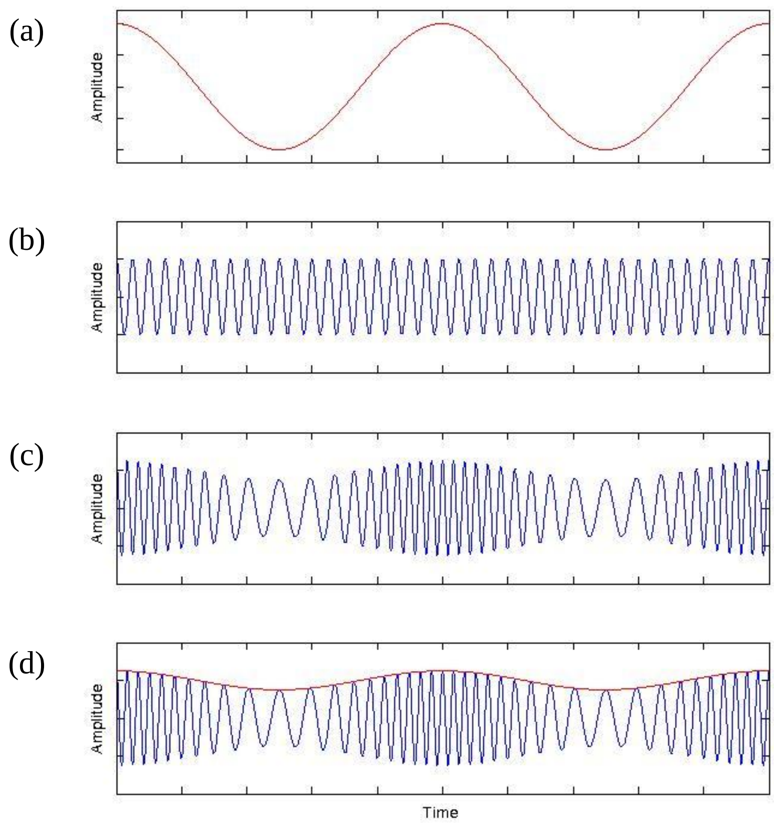


Figure 2.29 – General modulation for sinusoidal message signals
(a) message signal. Used for AM and FM (b) constant frequency sinusoidal carrier
(c) general modulated signal (d) general modulated signal overlayed with message signal indicating message signal gives envelope

Figure 2.29 shows a general modulation case, where a sinusoidal message signal has been used for both the amplitude modulation component and the frequency modulation component. Both message signals have the same frequency, and only differ in scaling factor used. Figure 2.29 (c) shows the general modulated signal, in which both amplitude modulation is visible in the changes to amplitude of the carrier, and frequency modulation with the change in carrier frequency. Figure 2.29 shows the message signal summed with the DC offset, showing that as for the amplitude modulation case, the envelope of the general modulated signal is equal to the sum of the message signal and DC offset used for the amplitude modulation component.

The spectral representation for a general modulated signal is complex, and a mathematical representation will not be presented here. However, it is possible to make use of the convolution theorem, as discussed in section 2.3 on page 57, to make generalisations on the bandwidth of the spectrum of a general modulated signal.

Looking at equation (2.52) on page 109, it can be seen that the general modulated signal $y(t)$ is basically the product of two terms, the amplitude modulation component and the frequency modulation component.

Based on the convolution theorem, the spectrum of $y(t)$ will be the convolution of the spectrum of the amplitude modulation component with the spectrum of the frequency modulation component.

If the convolution is thought of as the amplitude modulation component replacing every spectral line from the frequency modulation component, then every line in the frequency modulation component will be smeared out by the bandwidth of the amplitude modulation component.

The resulting total bandwidth F^t for the general modulated signal with a sinusoidal carrier will then be:

$$F^t = F + F_a \quad \text{...} \quad (2.53)$$

where F is the bandwidth of the frequency modulation component, which is smeared out in each direction by $F_a / 2$, where F_a is the bandwidth of the amplitude modulation component. The bandwidth F can be found using methods discussed for an

FM signal with a sinusoidal carrier as discussed in section 2.5.3 on page 80. Similarly, the bandwidth F_a can be found using methods discussed for an AM signal with sinusoidal carrier, as discussed in section 2.5.1 on page 68.

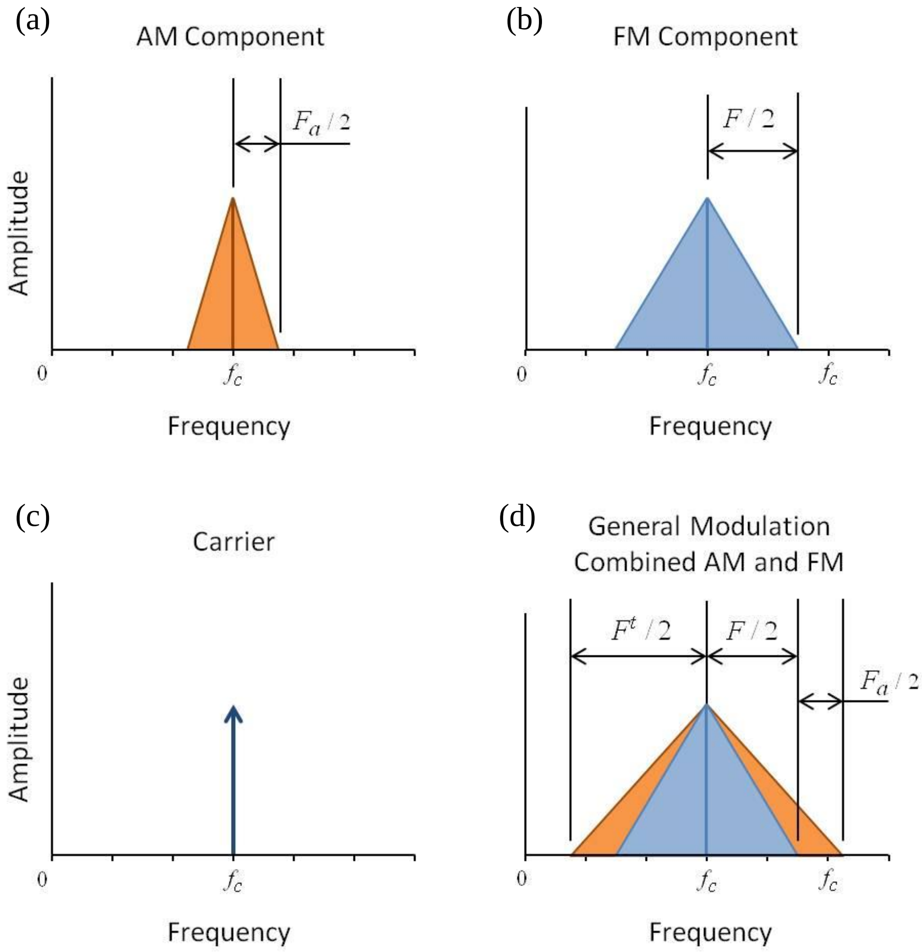


Figure 2.30 – Spectral representation showing bandwidth for general modulation of sinusoidal carrier

(a) AM component bandwidth (b) FM component bandwidth

(c) Sinusoidal carrier (d) Combined total bandwidth for general modulation

Figure 2.30 graphically shows the bandwidth relationship from equation (2.53). Figure 2.30 (c) shows the single frequency of the sinusoidal carrier. Figure 2.30 (a) shows the bandwidth of the amplitude modulation component separately, which is the bandwidth of an AM signal with sinusoidal modulation and carrier. Figure 2.30 (b) similarly shows the bandwidth of the frequency modulation component, which is the bandwidth of a FM signal with sinusoidal modulation and carrier. Figure 2.30 (d) shows the combined bandwidth F^t showing that it is the combined bandwidths from the AM and FM components.

2.5.6 General modulation for a harmonic series carrier

In the previous section 2.5.5, general modulation was presented for a sinusoidal carrier signal, where a single frequency sinusoid had both its frequency and amplitude varied based on message signals.

For a more complicated carrier signal, where the basic signal being varied consists of a harmonic series, such as a pulse train, general modulation can be defined based on the general modulation of a sinusoidal carrier case, in a similar fashion to amplitude modulation as discussed in section 2.5.2 on page 77, and frequency modulation as discussed in section 2.5.4 on page 101.

For this case, the general modulated signal $y(t)$ can be expressed as the sum a series of general modulated sinusoidal carrier signals as given by equation (2.52) on page 109, for each harmonic order n of the carrier signal, where the carrier frequency for each harmonic order is given by nf_c, so that:

$$y(t) = \sum_{n=1}^{\infty} y_n(t) = \sum_{n=1}^{\infty} \left(A_{DC} + A_m(t) \right) C_n \cos\left[2\pi n f_c t + 2\pi n D \int m_p(t) dt \right] \quad \text{........ (2.54)}$$

where $y_n(t)$ is one harmonic order of the total general modulated signal $y(t)$.

As can be seen in equation (2.54), the amplitude modulation component remains constant for different orders n (except for the amplitude scaling factor C_n for each order), but the frequency modulation component changes with n. This follows the pattern seen for amplitude modulation and frequency modulation in isolation.

The time domain signal $y(t)$ will be of a similar form to that seen in Figure 2.29 on page 110, with the exception that the base carrier will no longer be a sinusoid. Each carrier frequency component will be both amplitude modulated and frequency modulated, in an identical fashion to that seen in Figure 2.29.

As with the sinusoidal carrier case in section 2.5.5 on page 109, no attempt will be made here to represent the spectrum mathematically. As with the previous case, the general bandwidth characteristics of the spectrum can be determined.

The spectrum of the total signal $y(t)$ will be the sum of the spectra from each harmonic order, which is a general modulated signal with a sinusoidal carrier.

Each order is again formed from the convolution of the amplitude modulation component and frequency modulation component from one order.

As was seen in section 2.5.2 on page 77, the bandwidth of the amplitude modulation remains constant for different orders n. The bandwidth of the frequency modulation component will increase for increasing orders n, as was shown in section 2.5.4 on page 101.

Based on the convolution theorem, the total bandwidth F_n^t for the nth harmonic order is given by:

$$F_n^t = F_n + F_a \quad \text{...} \; (2.55)$$

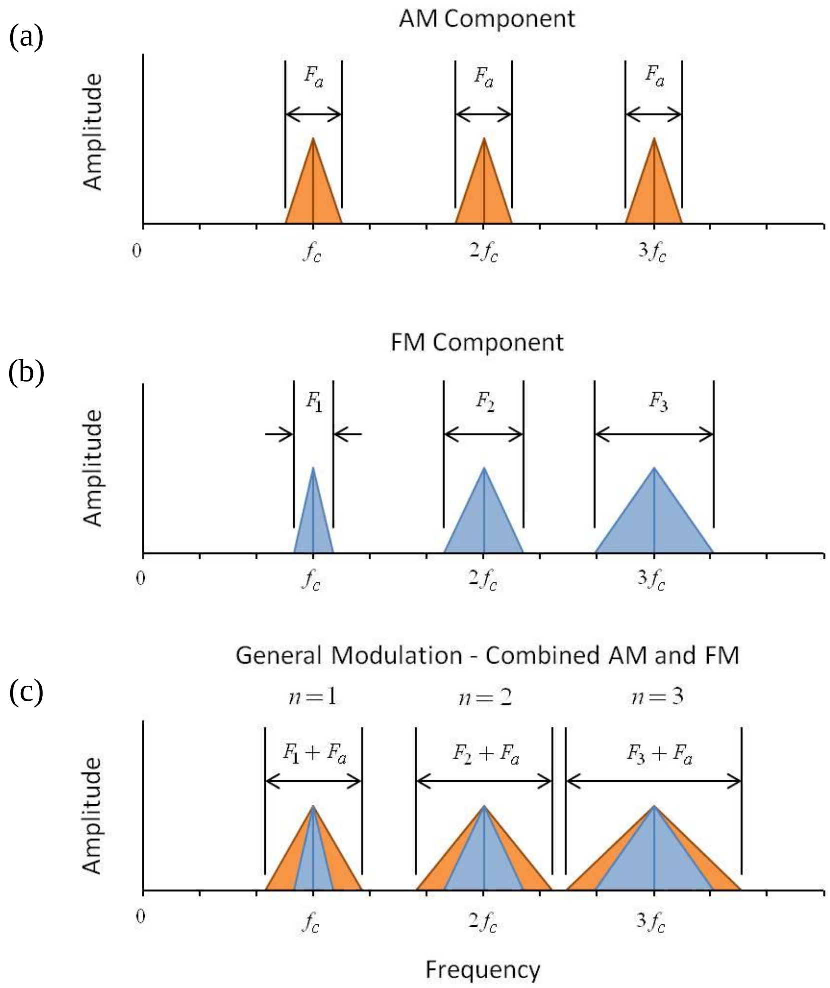


Figure 2.31 – Spectral representation showing bandwidth for general modulation of a non-sinusoidal carrier, showing first 3 harmonics
(a) AM component bandwidth (b) FM component bandwidth
(c) Combined total bandwidth for general modulation

Figure 2.31 graphically shows the bandwidth relationship from equation (2.55), for the first three harmonics $n=1$ to 3 (with scaling factors C_n all equal). Figure 2.31 (a) shows the bandwidth of the amplitude modulation component, and it can be seen that the bandwidth of each order is constant, matching the form seen for an AM signal with non-sinusoidal modulation. Figure 2.31 (b) shows the bandwidth of the frequency modulation component, which can be seen to be increasing for increasing order n. Figure 2.31 (c) shows the total bandwidth of each order F_n^t is the sum of the constant

AM bandwidth and the increasing FM bandwidth, resulting in a total bandwidth which increases with order n.

2.6 Demodulation

In the context of modulation, where a carrier signal is modulated by a message signal to give a modulated signal $y(t)$, demodulation is simply a method to recover the message signal from the modulated signal $y(t)$.

For a signal which is amplitude modulated, as discussed in sections 2.5.1 and 2.5.2, starting on page 68, recovery of the message signal is termed amplitude demodulation.

Similarly, for a frequency modulated signal, as discussed in sections 2.5.3 and 2.5.4, starting on page 80, recovery of the message signal is termed frequency demodulation. It should be noted that the demodulation method presented here, for frequency modulated signals, actually recovers the phase variation of the modulated signal, which is the integral of the frequency modulation message signal, and is so termed phase demodulation.

For a signal with general modulation, as discussed in sections 2.5.5 and 2.5.6, starting on page 109, this signal can be both amplitude and phase demodulated to recover the respective modulation signals.

The demodulation methods presented here are frequency domain based methods of demodulation. This material is based on Randall [31]. It should be noted that these are not the only methods to demodulate a signal, and for example, significantly different methods are more commonly used in the communications field. Demodulation methods for communications typically require real time demodulation, where frequency domain based demodulation is a post processing method unsuitable for real time demodulation.

As was discussed in the frequency analysis section 2.1 on page 34, a signal can be represented as the sum of a series of sinusoids. Each of these sinusoids can be in turn represented as the sum of two contra-rotating vectors. This was shown graphically in Figure 2.1 on page 37. For amplitude demodulation, it is desired to gain a measure of the amplitude of these contra-rotating vectors, and similarly for phase demodulation, the phase of the contra-rotating vectors.

For a real signal $y(t)$, the amplitude and phase of the contra-rotating vectors is not immediately accessible, as both have summed to a combined real sinusoidal result. However, use can be made of a vector equivalency presented in section 2.1.

This vector equivalency states that the sum of two contra-rotating vectors (positive and negative frequency) is equivalent to the real component of the positive frequency vector (which is complex) with double amplitude. This equivalency is illustrated in Figure 2.5 on page 43. The (complex) positive frequency vector in the time domain allows direct access to its amplitude and phase components.

As was discussed in 2.1, in relation to frequency editing, the positive vectors for a signal can be obtained by taking the FFT of the original signal $y(t)$, discarding all negative frequency information (corresponding to the negatively rotating one of the two contra-rotating vectors), which results in a single sided spectrum, and then transforming the single sided frequency information back to the time domain with an IFFT. This process allows access to the amplitude and phase information of a signal $y(t)$, and forms the basis of frequency based demodulation.

2.6.1 Phase demodulation

To conduct phase demodulation, from the frequency spectrum of the modulated signal $y(t)$, a one-sided (positive frequency only) demodulation band is selected, centred on the frequency information to be demodulated.

For a frequency modulated signal with a sinusoidal carrier, as discussed in section 2.5.3 on page 80, the demodulation band is centred on the carrier frequency f_c. To recover the message signal, the demodulation bandwidth must fully encompass the bandwidth of the (positive frequency component of the) frequency modulated signal, so no message signal information is missed.

For a frequency modulated signal with a harmonic series carrier, as discussed in section 2.5.4 on page 101, the demodulation band must fully encompass one, and only one, harmonic order of the spectrum of the FM signal. As each harmonic order contains a repetition of the message signal, as indicated in equation (2.40) on page 101, it is necessary to encapsulate only information from one harmonic order in the demodulation band. It should be noted that as the bandwidth of a harmonic order increases with harmonic order, eventually some harmonic orders will overlap in the spectrum of an FM signal. Any overlapped harmonic orders are unsuitable for demodulation with this method, as a demodulation band cannot be selected which only encompasses one harmonic order, and the overlap will give aliasing of the demodulated result.

As with the case of a sinusoidal carrier, the demodulation band is centred on the carrier frequency of the harmonic order to be demodulated. Figure 2.32(a) shows the selection of a demodulation band centred on a harmonic at carrier frequency f_c.

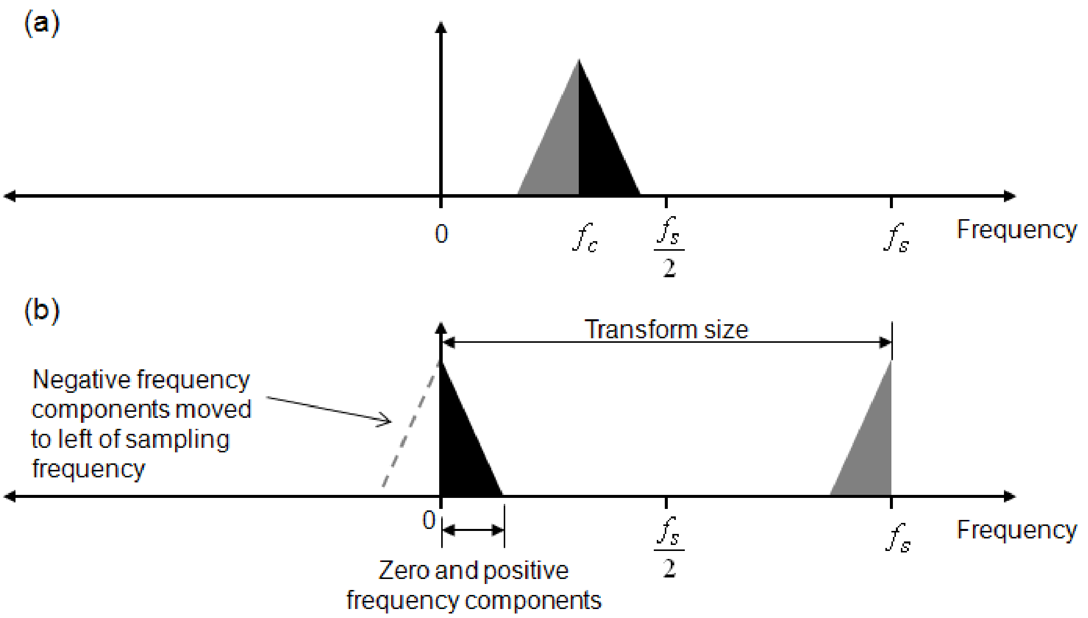


Figure 2.32 – Spectrum frequency shift in the frequency domain, where carrier frequency f_c is shifted to zero (modified, based on [31])
(a) original selected demodulation band (b) Shifted demodulation band

The demodulation band is then shifted around zero frequency, with the carrier frequency component centred at zero frequency, as seen in Figure 2.32(b). Because of the periodicity of FFT spectra, the negative frequency components must be shifted to just below the sampling frequency, for a frequency spectrum of the second general form, as illustrated in Figure 2.32(b). The frequency components must also be complex, not just the amplitudes as illustrated.

The shifted demodulation band is then inverse Fourier transformed, and the phase component of the result extracted. This results in what is termed wrapped phase, which results from the phase being determined as an arctangent, which can only be determined in the range $\pm\pi$. As a result the wrapped phase contains phase jumps when the phase would move outside this range. In order to correctly obtain the continuous phase of the signal, the wrapped phase must be unwrapped, where phase jumps of 2π are removed. As a final step, the phase needs to be divided by harmonic order n, if a harmonic order other than the first was demodulated. The resulting signal is the phase demodulation of the original modulated signal $y(t)$, and gives the message signal that modulated $y(t)$.

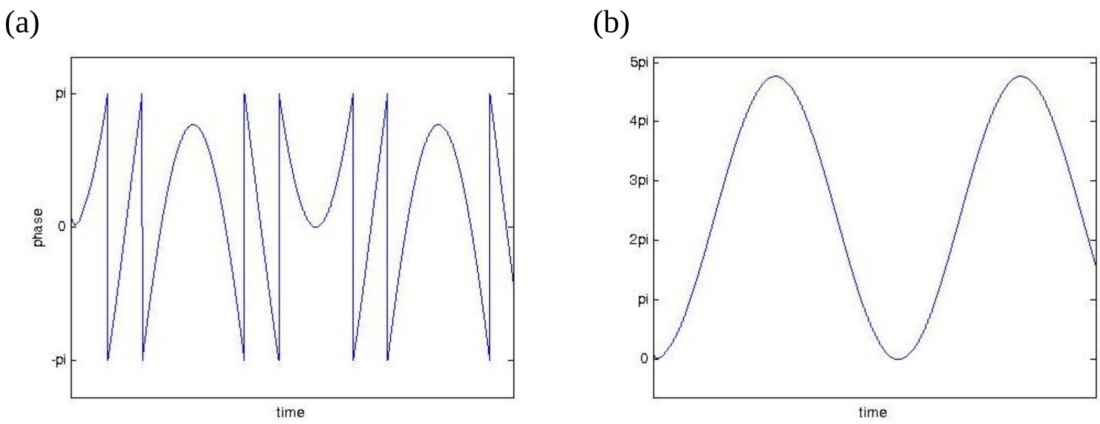


Figure 2.33 – Wrapped and unwrapped phase
(a) wrapped phase (b) corresponding unwrapped phase

Figure 2.33 (a) shows an example of wrapped phase, and (b) shows the corresponding unwrapped phase which no longer has discontinuities. The phase jumps in the wrapped phase in (a) are clearly visible at locations where the unwrapped signal in (b) would leave the $\pm\pi$ range.

The purpose of the frequency shift of the demodulation band in the spectrum, as shown in Figure 2.32 (b), is so the recovered phase information represents the message signal (variation around carrier). Without the frequency shift, the recovered phase would be the sum of the phase of the constant carrier frequency (a line with constant slope) and the message signal (variations around that straight line). The frequency shift sets the carrier frequency to zero, resulting in only variations around the carrier from the message signal remaining.

The frequency shift of the demodulation band also aids in the correct unwrapping of the signal. An un-shifted demodulation band would result in a steeper slope of the phase, due to the addition of the linear phase from the carrier signal, which could not be as reliably unwrapped due to likelihood of greater differences between adjacent phase values (a jump of $> \pi$ being taken as a jump to be removed)..

It should be noted that the shifted frequency band shown in Figure 2.32(b), on page 121, could have the sampling rate changed to any sampling rate which is greater than twice the maximum frequency of the shifted demodulation band. As was shown in section 2.2 on page 45, in regards to visual aliasing, the sampling rate can be varied to any amount by zero padding (or removing zeros), without any change to the signal providing aliasing does not occur due to the sampling rate being less than twice the highest frequency.

Increasing the sampling rate in this fashion is effectively a method of interpolation, as discussed in section 2.4 on page 59, as raising the sampling rate in the frequency domain gives resulting time domain data with closer spaced samples, which contain no errors (ideal interpolation). Demodulation effectively performs 'almost ideal' interpolation in transforming from the frequency domain back to the time domain, the only error coming from the very small residual arising out of the definition of "significant sidebands". This windowing in the frequency domain with a rectangular ideal filter is equivalent to sinc(x) interpolation in the time domain, but the sinc(x) function is infinitely long and must be truncated, giving indeterminate errors. This almost ideal interpolation hence gives an inherently better result than other interpolation methods, such as cubic spline interpolation, as discussed in section 2.4. The only limitation is that this type of interpolation can only be used to obtain a phase-time map with constant spacings in time (though to any desired degree of resolution), so for the determination of the sample times at constant phase intervals (varying in time), and for interpolation of the actual signals at those times, a second interpolation method must be used, but the resulting errors can be made arbitrarily small.

There is one critical issue which can prevent the correct phase being extracted by this described phase demodulation method. If the amplitude component (in the time domain), accompanying the phase component extracted with this method, goes to zero magnitude at any position, the phase becomes indeterminate at these positions. As a result, the phase extracted by the demodulation process at these positions contains discontinuities, and thus the extracted phase does not match the true phase of the signal.

So, as a result, in order to obtain the correct phase from the phase demodulation process, it is necessary for the amplitude component to always be greater than zero, for the entire demodulation band selected. It should be noted that only amplitudes corresponding to the demodulation band selected cannot become zero, and zero amplitudes from the rest of the spectrum cause no issues.

This is an additional reason that the definition of amplitude modulation, given in section 2.5.1 on page 68, and general modulation, given in section 2.5.5 on page 109, which is suitable for use with physical systems, specifies that the sum $A_{DC} + A_m(t)$, which gives the amplitude of an AM or general modulated signal, must always be positive. Without this constraint, the amplitude term could be zero, which would introduce phase discontinuities in any demodulated phase, and so the demodulated phase would be incorrect for a physical system.

2.6.2 Amplitude demodulation

Amplitude demodulation can be conducted in an almost identical fashion to phase demodulation, and both methods can be conducted concurrently for a general modulated signal.

For amplitude demodulation, the same frequency shifted demodulation band can be used which encompasses one harmonic order of the modulated signal, as shown in Figure 2.32 on page 121.

Once this signal is transformed back to complex time domain values, rather than taking the phase values as with phase demodulation, the amplitude values are taken. The result is the amplitude demodulated signal from the original signal $y(t)$, which is the message signal which modulated $y(t)$ (plus DC offset, which must be removed). Note that the signal in this case is not divided by harmonic order n, as with the phase demodulation case, as the message signal for amplitude modulation is equal across all harmonic orders, as indicated in equation (2.22) on page 77.

It should be noted that the frequency shift of the demodulation band is not necessary for amplitude demodulation, as the same result will be given regardless of the carrier frequency value. A slightly easier frequency shift method, suitable only for amplitude modulation, is presented in Randall [31]. However, in the context of order-tracking as presented in this work, amplitude modulation will typically only be present in combination with frequency modulation (general modulation), so a common demodulation method is preferable.

It should also be noted that, unlike with phase demodulation as discussed in 2.6.1, the correct amplitude is extracted even if the amplitude within a demodulation band becomes zero at a given position. This fact is invaluable in investigating possible phase discontinuities in a demodulated phase signal, as the corresponding demodulated amplitude is still correct if instances of zero amplitude have occurred. As a result, the corresponding demodulated amplitude signal can be evaluated to determine if the amplitude became zero at any point in a given demodulation band, causing a phase discontinuity in the associated demodulated phase.

Chapter 3 Basic Phase Demodulation based Order Tracking

Chapter Overview

The following chapter presents the basic PDOT method, which is extended into a generalised methodology in the following chapter.

The permissible speed variations for which the basic PDOT method can be employed are then established, and methods and guidelines are presented for evaluating if a given signal is suitable.

Finally, some basic simulated results are presented, showing the results possible with PDOT in the absence of contaminating factors.

3.1 Single stage of Phase Demodulation Based Order-Tracking

The PDOT method, in its simplest form, consists of a single stage of order-tracking to resample a varying frequency vibration signal from the time domain to the angular domain (with constant angular sampling). This order-tracking is done with the use of a synchronously recorded reference signal, which gives the map of rotation angle versus time for a varying speed component of interest (reference component). The resulting resampled signal is sampled at equal steps in the rotation angle of the varying speed reference component.

Typically, a reference component would correspond to a rotating shaft in a machine, where the rotational speed of other components is proportional to that of the reference shaft. Examples include any shaft in a gear train, where the rotation angle of other gears, bearings and shafts can be expressed relative to the reference shaft.

In nearly every case, a reference component will consist of a harmonic series. This harmonic series for a speed varying case can be considered as a frequency modulated signal with a harmonic series carrier, as was discussed in section 2.5.4 on page 101. The phase-time map used for the PDOT method is found from a member of the harmonic series of the reference component. The resulting resampled signal is then expressed in terms of the fundamental frequency or first harmonic (order) of the reference component.

A suitable reference signal to be used for order-tracking must be low pass filtered in a similar way to the acceleration signal. This is critical, as the PDOT method utilises the frequency information of the reference signal directly, and low pass filtering is required to avoid aliasing in the frequency domain.

Although the PDOT method can be used with different reference signal types, the simplest type to use is a synchronously recorded once per revolution pulse type tachometer, attached to a rotating shaft, which gives one pulse per revolution of the rotating shaft with fixed amplitude and aspect ratio. This reference signal type will be used initially to derive the operating limitations of one pass of the PDOT method. This method is effectively the same as presented in [1], except for changes to the methods for calculating the bandwidth of an FM signal, and subsequent changes to the speed variation limits, for which the modified method can be used.

3.2 Procedure for a single pass of order-tracking

The basic single stage PDOT method consists of two major steps. The first step is to construct a phase-time relationship from the reference signal by using phase demodulation. The second step is to use interpolation to resample the vibration signal at equal increments of phase.

3.2.1 Phase Demodulation of Reference Signal

The phase-time relationship is produced by extracting the phase modulation signal of one harmonic of the reference component using phase demodulation. In the context of frequency modulation, as presented in sections 2.5.3 and 2.5.4, starting on page 80, the phase-time relationship is the sum of the message signal and the phase of the carrier frequency component.

Phase demodulation has been discussed in detail in many works. A detailed description of the traditional method of phase demodulation is given in sections 2.6 and 2.6.1 starting on page 118. The phase demodulation method used with the PDOT method is almost identical to the traditional process given in the earlier sections

In summary, from the frequency spectrum of the reference signal, a one-sided (positive frequency only) demodulation band is selected, centred on the harmonic to be demodulated. The demodulation band must encompass an entire harmonic order of the reference signal (including modulation sidebands) for correct order-tracking. However, as the demodulation process is blind, ideally the demodulation band should contain no other components, or noise, which would corrupt the result. In particular it should not be corrupted by aliasing components from higher orders of the same signal, since as with other types of aliasing, once it occurs it cannot be removed by further processing.

Another requirement for correct order-tracking is that the selected demodulation band should contain only positive (time domain) amplitudes. As discussed in the phase demodulation section 2.6.1, starting on page 120, if the amplitude drops to zero within a demodulation band, the extracted phase contains discontinuities, and does not correctly match the true speed variation of the machine, which would result in erroneous order tracking.

The selected demodulation band is then shifted around zero frequency, with the carrier frequency component centred at zero frequency, as seen in Figure 2.32 on page 121. Note that the original sampling frequency f_s is retained for the PDOT application, as opposed to many other applications of phase demodulation, where it can be reduced to as little as twice the width of the band being demodulated. Retaining the original sampling frequency causes the phase demodulation to act as another interpolation stage, as will be discussed in the next section.

The shifted demodulation band is then inverse Fourier transformed, and the phase component of the result extracted and unwrapped.

In contrast to traditional phase demodulation, the linear phase progression of the original carrier component, which was shifted to zero frequency, is then added back to the phase result. In effect this gives the phase of the actual harmonic, as opposed to the phase variation around the carrier.

Finally, the resulting phase is divided by the harmonic order of the harmonic utilised, giving phase in terms of the fundamental order of the component of interest. This gives the instantaneous phase of the first harmonic of the reference signal vs. time, and is a first estimate of the general Phase-Time relationship.

It should be noted that while a similar result could in principle be produced without shifting the demodulation band, this shift is necessary to be able to reliably "unwrap" the phase to a continuous function of time, as discussed in section 2.6.1 on page 120. The actual phase-time map used for angular resampling must include the linear component corresponding to the carrier frequency, but the carrier frequency is known exactly and can be added back in with no error.

3.2.2 Interpolation

The second major step in the PDOT method is to use interpolation to conduct resampling. An overview of interpolation methods is given in section 2.4 on page 59. For the PDOT method the interpolation is done in two stages

In the first stage, interpolation is used with the Phase-Time relationship to determine time values corresponding to equal increments of phase. For the PDOT method, the phase sampling rate used is typically chosen so the final record length in samples is higher than the original record length. It must be such that the sampling rate is higher than the original in every part of the record, and must give an integer number of samples per 2π of phase (one rotation). It can also be desirable to begin the resampled record at a specific reference signal position, such as the (positive going) zero crossing or a threshold crossing point on a pulse, corresponding to a fixed rotation angle, to facilitate comparison between multiple signals from the same machinery. This specific position usually lies between two original samples, hence it is recommended to use linear interpolation to locate this value rather than take the closest sample.

It is assumed that the original signals, both the vibration and reference signals, have initially been suitably oversampled to avoid aliasing. As the low pass filtering is done in the time domain, once the signals are resampled into the order domain the constant frequency cut-off of the low pass filter becomes a variable limit in the order domain. Since the maximum frequency ratio of the reference component in the record is approximately 2:1, as will be shown in Section 3.3, it would generally be safe to oversample by this factor before order tracking, thus making sure the number of samples per period at the lowest speed in the range is still greater than before resampling.

In the second stage, interpolation is used to resample the vibration signal at the time values calculated in the first stage. This results in an order-tracked signal, with constant “frequency” components (i.e. orders) in the rotation angle domain.

Cubic spline interpolation was selected for the PDOT method, based on a study by P.D. McFadden [34] who found that the cubic spline method gave better accuracy and the minimum spectral leakage compared to other methods. This was also recommended by Bonnardot et al. [10].

Cubic spline interpolation can cause end effects at the start and end of a signal record, due to arbitrary specification of the slope at the two ends. These end effects can be removed by discarding a suitable number of samples at each end of the resampled signal record, but in any case the end effects from the FFT processing of finite records would normally dominate in determining how much to discard.

It should also be noted that, as mentioned in the previous section 3.2.1, the earlier step of phase demodulation effectively acts as another stage of interpolation for the Phase/Time map. As was discussed in the phase demodulation section 2.6.1 on page 120, phase demodulation effectively performs ‘almost ideal’ interpolation in transforming from the frequency domain back to the time domain, which gives a very high resolution initial Phase-Time relationship and significantly increases the effectiveness of using subsequent cubic spline interpolation, when compared to the use of cubic spline interpolation alone as with time based order-tracking methods [11, 12].

The benefits gained from phase demodulation could be further improved by oversampling to an even higher sampling rate, rather than retaining the original sampling rate f_s during the phase demodulation process as presented in section 3.2.1. This oversampling is limited only by the subsequent increase in computational cost.

3.3 Selection of Demodulation Band, and limits of permissible harmonic order bandwidth

As stated in Section 3.2.1 on page 131, for the phase demodulation process it is necessary to select a band in the frequency spectrum which contains one, and only one, order of the reference component. The choice of which order to use is dependent on the spectral properties of the reference component.

As mentioned previously, the simplest reference type to use is a synchronously recorded once per revolution pulse type tachometer, and this signal type is first considered. A once per revolution pulse tachometer is effectively a frequency modulated signal, with a harmonic series carrier. This type of frequency modulated signal was presented in detail in section 2.5.4 on page 101, and the spectral characteristics of this frequency modulation type are presented in detail there.

This spectrum consists of a number of harmonic orders, each of which can itself be considered as a frequency modulated signal with a sinusoidal carrier, such as discussed in section 2.5.3 on page 80. Each of these harmonic orders consists of a finite bandwidth centred at an integer multiple of the centre frequency of the fundamental harmonic. The bandwidth of each harmonic order increases with increasing harmonic order, which was shown illustratively in Figure 2.27 on page 106. This trend causes the bandwidths of higher harmonic orders to eventually overlap, which are then no longer suitable as a candidate for phase demodulation of a harmonic order.

For simplicities sake, it will first be assumed that the bandwidth of higher harmonics is proportional to the bandwidth of the fundamental harmonic order, such as presented by equation (2.51) on page 107. While this assumption is only valid for very low maximum modulating frequency, it is useful in illustrating the approximate limits to permissible bandwidth for a harmonic order which can be used for demodulation. The assumption also allows the development of simpler relationships giving the maximum permissible bandwidth of a harmonic. This bandwidth assumption leads to the spectral trend shown in Figure 2.28 on page 107.

Based on equation (2.51), it is found that the largest permissible non-overlapping bandwidth value for F, which is defined as F^{lim}, is possible when the fundamental harmonic is used for demodulation. If higher harmonics are used, the magnitude of the permissible bandwidth is reduced.

Figure 3.1 illustrates the maximum value of F^{lim} which avoids overlap of the phase modulation sidebands between the first two harmonics, for the case with no amplitude modulation. It is seen that this ensures at the same time that none of the sidebands around higher harmonics encroach on those around the first harmonic, although they do overlap all others.

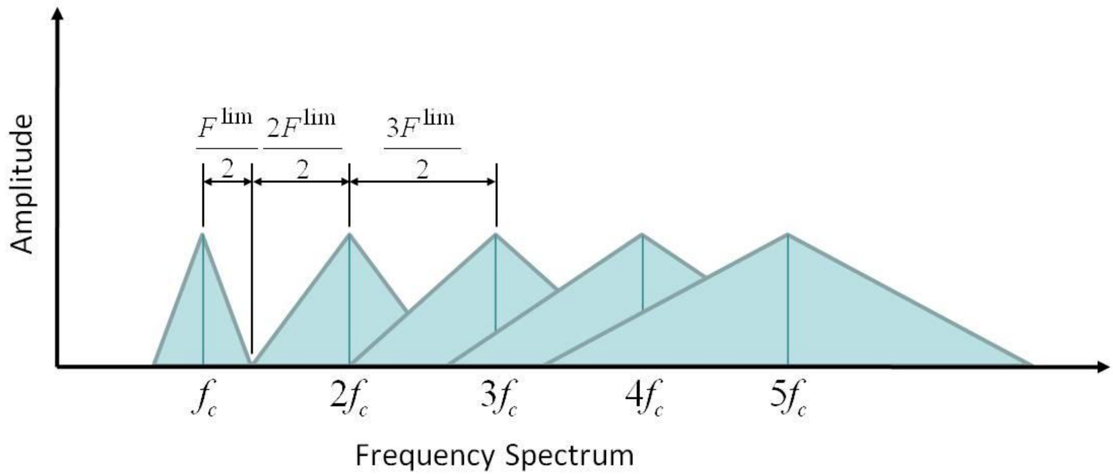


Figure 3.1 – Maximum bandwidth case for first harmonic showing extent of sidebands from higher harmonics, assuming proportional harmonic orders, for the case with no amplitude modulation

Figure 3.2 shows just the first two harmonics, for the same limiting case as in the previous figure.

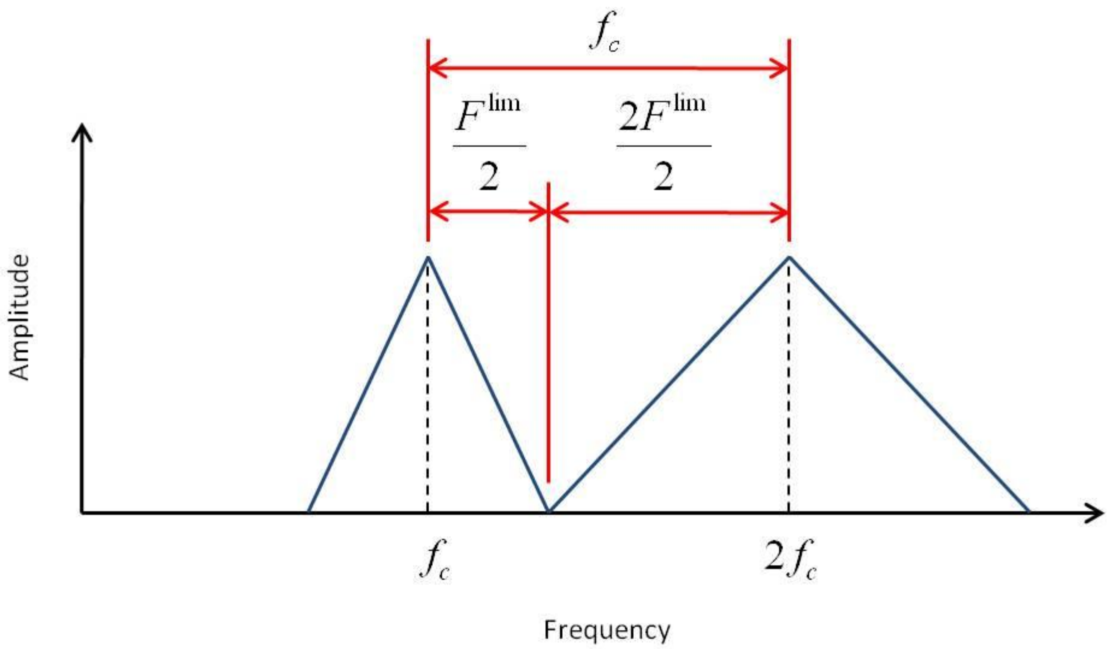


Figure 3.2 – Maximum bandwidth case for first harmonic, assuming proportional harmonic orders, for the case with no amplitude modulation

By identifying that the spacing between harmonic order centres is always f_c, the maximum value of $F^{\lim}$ for the limiting case shown in Figure 3.2 can be found algebraically, by:

$$\frac{F^{\lim}}{2}+\frac{2F^{\lim}}{2}=f_c \quad \text{...} \quad (3.1)$$

which can be solved for $F^{\lim}/2$ to give:

$$\frac{F^{\lim}}{2}=\frac{f_c}{3}=33\% \text{ (of } f_c\text{)} \quad \text{.......................................} \quad (3.2)$$

The full bandwidth $F^{\lim}$ can alternatively be expressed as a span around the carrier frequency f_c, so that:

$$F^{\lim}=\pm 33\% \text{ (of } f_c\text{) around } f_c \quad \text{....................................} \quad (3.3)$$

This thus defines the maximum bandwidth $F^{\lim}$ that can be accommodated by the PDOT method (a maximum ratio of 2:1 as stated earlier), using the assumption that harmonic bandwidths are proportional to the fundamental, and can only be achieved by demodulating the first harmonic of a series.

It is possible to use the same process to calculate the maximum permissible bandwidth term $F^{\lim}$ for the general harmonic order number n if a higher order than the first will be used for order-tracking. By identifying that the first overlap of harmonic n will be from the next harmonic $n+1$, and that the addition of the half-bandwidth of both harmonics equals f_c, gives the limiting case for harmonic n as:

$$\frac{nF^{\lim}}{2}+\frac{(n+1)F^{\lim}}{2}=f_c \quad \text{...} \quad (3.4)$$

which can be solved for the half-bandwidth $F^{\lim}/2$ to give the maximum permissible bandwidth for harmonic order number n,

$$\frac{F^{\lim}}{2}=\frac{f_c}{1+2n}=\frac{100}{1+2n}\% \text{ (of } f_c\text{)} \quad \text{...................................} \quad (3.5)$$

for the assumption that higher harmonics are proportional to harmonic order.

Similarly to equation (3.3), The full bandwidth $F^{\lim}$ can alternatively be expressed as a span around the carrier frequency f_c, so that:

$$F^{\lim}=\pm\frac{100}{1+2n}\% \text{ (of } f_c\text{) around } f_c \quad \text{.................................} \quad (3.6)$$

It should be noted that Equation (3.5) is similar in form to Equation (3) from Bonnardot et al. [10], but in that paper, and in Combet and Gelman [21], further simplifications are made based on the assumption that the order n is high, so that the overlap occurs approximately half way between order n and n+1.

The limiting bandwidth equation given in equation (3.5) allows for relatively simple calculation, as only the bandwidth of the first harmonic needs to be calculated. However, this limiting bandwidth is only valid for cases with extremely low maximum modulating frequency. Practically, even small maximum modulating frequencies will result in a bandwidth pattern which is not exactly proportional to harmonic order. As stated in section 2.5.4 on page 101, ideally the bandwidths of each harmonic should be calculated separately, rather than relying on the assumption of proportional bandwidths of higher harmonics.

As shown in section 2.5.4, the bandwidth of the nth harmonic can be calculated by either using Bessel function calculations, to give a bandwidth from equation (2.47) on page 105, or using Carson's rule, to give a bandwidth from equation (2.48) on page 105.

In a similar fashion to equation (3.4) and (3.5), the limiting bandwidth $F_n^{\lim}$ can be calculated for the general harmonic order number n, which is applicable to give the limiting bandwidth for the fundamental harmonic as well as higher ones. By identifying that the first overlap of harmonic n will be from the next harmonic $n+1$, and that the addition of the half-bandwidth of both harmonics equals f_c, gives the limiting case for harmonic n as:

$$\frac{F_n^{\lim}}{2} + \frac{F_{(n+1)}^{\lim}}{2} = f_c \quad \text{...} \ (3.7)$$

which can be solved for the half-bandwidth $F_n^{\lim} / 2$, to give the maximum permissible bandwidth for harmonic order number n,

$$\frac{F_n^{\lim}}{2} = f_c - \frac{F_{(n+1)}^{\lim}}{2} \quad \text{...} \; (3.8)$$

where $F_{(n+1)}^{\lim}$ is the bandwidth of the harmonic order $n+1$ for the limiting case.

While equation (3.8) is more difficult to calculate than equation (3.5), its application is not limited to very small maximum modulating frequencies, and so the usage of this limit calculation is preferable.

While these bandwidth calculations assume a reference signal does not contain any amplitude modulation, it is possible to extend them to the case where a reference signal is amplitude modulated. A reference signal which contains both amplitude and frequency modulation is said to have general modulation. General modulation is discussed in sections 2.5.5 and 2.5.6, starting on page 109. A typical reference signal which has a series of harmonic orders, can be considered as a generally modulated signal with harmonic series carrier, as presented in section 2.5.6. The bandwidth characteristics for this type of signal were shown illustratively in Figure 2.31 on page 116.

The total bandwidth of the harmonic order n, for a general modulated harmonic carrier signal, was presented in equation (2.55) on page 115 as the sum of the frequency modulation bandwidth of the harmonic order and the “effective” amplitude modulation bandwidth (which is constant for all orders). A specific definition of “effective” bandwidth is given in section 3.7 on page 167.

In a similar fashion to equations (3.5) and (3.8), the limiting bandwidth of the frequency modulation component $F_n^{\lim}$ can be calculated for the general harmonic order number n, which is applicable to give the limiting bandwidth for the fundamental harmonic as well as higher ones. By identifying that the first overlap of harmonic n will be from the next harmonic $n+1$, and that the addition of the half-bandwidth of both harmonics equals f_c, gives the limiting case for harmonic n as:

$$\left(\frac{F_n^{\lim}}{2}+\frac{F_a}{2}\right)+\left(\frac{F_{(n+1)}^{\lim}}{2}+\frac{F_a}{2}\right)=f_c \quad \text{.................................. (3.9)}$$

which can be solved for the half-bandwidth $F_n^{\lim}/2$ to give the maximum permissible bandwidth for harmonic order number n,

$$\frac{F_n^{\lim}}{2}=f_c-\frac{F_{(n+1)}^{\lim}}{2}-F_a \quad \text{...................................... (3.10)}$$

3.4 Defining bandwidth in terms of Frequency variation and modulation

With a limit to the permissible bandwidth of a reference signal harmonic, as given in the previous section 3.3, the limit to the permissible values of frequency variation and modulation can be determined by relating them to the bandwidth of a reference signal harmonic.

In sections 2.5.3 and 2.5.4, starting on page 80, it was shown that the bandwidth of a frequency modulated signal was a function of both frequency deviation and modulation frequency.

This was shown for both a frequency modulated signal with a sinusoidal variation, where bandwidth was shown to be a function of both frequency variation (deviation) f_d and modulating frequency f_m. For this case, the bandwidth of harmonic order n was shown to be related to these two parameters by equation (2.45) on page 104, based on Bessel function calculations, or alternatively by equation (2.48) on page 105, based on Carson's rule.

It was also shown that for non sinusoidal frequency modulation, that bandwidth is conservatively a function of the peak frequency variation (deviation) D and maximum modulating frequency W. For this case, the bandwidth of harmonic order n was shown to be conservatively related to these two parameters by equation (2.47) on page 105, based on Bessel function calculations, or alternatively again by equation (2.48), based on Carson's rule.

While these methods accurately relate bandwidth to the absolute values for frequency deviation and modulation, they present difficulties in developing general acceptability criteria for an order tracking method. As bandwidth was shown to be a function of frequency deviation and modulating frequency, and the acceptable limits to bandwidth are a function of carrier frequency, in effect the acceptability criteria for whether order-tracking can be conducted with a particular signal is dependent on three variables. However, this can be simplified with some definition changes, suitable for use when considering order-tracking. These changes are applicable to both bandwidth calculation methods, either the Bessel function based approach, or the Carson’s rule based approach.

3.4.1 Defining bandwidth in terms of Frequency variation and modulation using Bessel functions

The method of defining bandwidth in terms of Bessel functions was given in detail in section 2.5.3, starting on page 80. As a summary, for a frequency modulated sinusoid with sinusoidal frequency modulation, the frequency spectrum will consist of a carrier frequency component at f_c surrounded by a series of sideband pairs with a spacing of f_m (modulation frequency). There is theoretically an infinite number of sidebands, however as the order of the sidebands approaches infinity their magnitude approaches zero. By setting a limit on the dynamic range of the sidebands, a frequency modulated sinusoid has a finite bandwidth and hence a finite number of sideband pairs. The magnitude of the carrier and each sideband pair is proportional to the absolute value of a Bessel function of the first kind, $\left|J_k(\beta)\right|$, where k is the sideband index number, and β is the modulation index which was given in equation (2.30) on page 87 as $\beta = f_d / f_m$. f_d is the frequency deviation (variation) of the FM signal, and f_m is the modulating frequency.

With a definition of significant sidebands, for sidebands k with magnitude proportional to $\left|J_k(\beta)\right|$, it is possible to find the series of integer k values corresponding to all significant sidebands. The highest significant sideband order $k_{\max}$ is then the largest integer value of k corresponding to a significant sideband.

The bandwidth of the FM sinusoid with sinusoidal modulation is then found from the sideband spacing and the highest significant sideband order $k_{\max}$, which was given as equation (2.37) on page 95.

The bandwidth of a FM sinusoid with non-sinusoidal modulation can be approximated, by replacing f_d with D, the peak frequency deviation, and f_m with W, the maximum modulating frequency, giving the modulation index $\beta = D/W$ as per equation (2.38) on page 99. Using the modified modulation index, the bandwidth approximation can be calculated in the same fashion as for a FM sinusoid with sinusoidal modulation, which is given in equation (2.39) on page 100.

For a signal which is a frequency modulated harmonic series carrier, each harmonic order of the signal can be considered as a separate frequency modulated sinusoid. This was discussed in detail in section 2.5.4, starting on page 101. The modulation index for a harmonic order is proportional to harmonic order, so that for harmonic order n, the modulation index is $\beta = nf_d / f_m$ for the case of sinusoidal modulation, as given in equation (2.44) on page 104, or $\beta = nD/W$ for the case of non-sinusoidal modulation, as given in equation (2.46) on page 105. The bandwidth of a harmonic order n is then calculated in an identical fashion as for a FM sinusoidal carrier using the associated modulation index, with equation (2.45) on page 104 giving the bandwidth of a harmonic order for sinusoidal modulation, and equation (2.47) on page 105 giving the bandwidth for non sinusoidal modulation.

In order to simplify acceptability criteria for phase demodulation based order-tracking, it is possible to express the frequency deviation and modulating frequency as a percentage of the carrier frequency f_c. For a frequency modulated sinusoid with sinusoidal modulation f_d and f_m can be replaced with both terms represented as a percentage of the carrier, given by $f_{d\%c}$ and $f_{m\%c}$, respectively. Similarly, for a frequency modulated sinusoid with non-sinusoidal modulation, D and W can be replaced with both terms expressed as a percentage of the carrier, given by $D_{\%c}$ and $W_{\%c}$, respectively.

So for a FM sinusoid with sinusoidal modulation, the modulation index becomes:

$$\beta = \frac{f_{d\%c} \times \cancel{f_c}}{f_{m\%c} \times \cancel{f_c}} = \frac{f_{d\%c}}{f_{m\%c}} \quad \text{......................................} \ (3.11)$$

which can be seen to be identical to the modulation index as given in equation (2.30) on page 87.

The bandwidth of the FM sinusoid with sinusoidal modulation can now also be expressed as a percentage of carrier frequency, $F_{\%c}$, so equation (2.37) on page 95 becomes:

$$F_{\%c} = 2k_{\max} f_{m\%c} \ (\% \text{ of } f_c) = \ \pm k_{\max} f_{m\%c} \ (\% \text{ of } f_c) \text{ around } f_c \quad \text{.........} \ (3.12)$$

Similarly, for a FM sinusoid with non-sinusoidal modulation, the modulation index becomes:

$$\beta = \frac{D_{\%c} \times \cancel{f_c}}{W_{\%c} \times \cancel{f_c}} = \frac{D_{\%c}}{W_{\%c}} \quad \text{..} \ (3.13)$$

which is again identical to the modulation index as given in equation (2.38) on page 99, when the parameters are not expressed as a percentage.

In a similar fashion to the case of sinusoidal modulation, the bandwidth of the FM sinusoid with non-sinusoidal modulation can be expressed as a percentage of carrier frequency, $F_{\%c}$, so equation (2.39) on page 100 becomes:

$$F_{\%c} = 2k_{\max}W_{\%c} \text{ (\% of } f_c) = \pm k_{\max}W_{\%c} \text{ (\% of } f_c) \text{ around } f_c \quad \text{..........} \ (3.14)$$

And in an identical fashion, the equations for a frequency modulated harmonic series carrier can also be updated with parameters expressed as a percentage of the carrier frequency, so:

For the case of FM of a harmonic series carrier, with sinusoidal modulation, the modulation index given in equation (2.44) on page 104 becomes:

$$\beta = \frac{nf_{d\%c} \times \cancel{f_c}}{f_{m\%c} \times \cancel{f_c}} = \frac{nf_{d\%c}}{f_{m\%c}} \quad \text{......................................} \ (3.15)$$

which is identical to the original modulation index. The resulting bandwidth equation (2.45), on page 104, for harmonic n can then be expressed as a percentage of the carrier frequency, $F_{n\%c}$, so:

$$F_{n\%c} = 2k_{\max}^{n} f_{m\%c} \text{ (\% of } f_c) = \pm k_{\max}^{n} f_{m\%c} \text{ (\% of } f_c) \text{ around } nf_c \quad \text{.......} \ (3.16)$$

For the case of FM of a harmonic series carrier, with non-sinusoidal modulation, the modulation index given in equation (2.46) on page 105 becomes:

$$\beta = \frac{nD_{\%c} \times \cancel{f_c}}{W_{\%c} \times \cancel{f_c}} = \frac{nD_{\%c}}{W_{\%c}} \quad \text{......................................} \ (3.17)$$

which is identical to the original modulation index. The resulting bandwidth equation (2.47), on page 105, for harmonic n can then be expressed as a percentage of the carrier frequency, $F_{n\%c}$,so:

$$F_{n\%c} = 2k_{\max}^{n}W_{\%c} \text{ (\% of } f_c) = \pm k_{\max}^{n}W_{\%c} \text{ (\% of } f_c) \text{ around } nf_c \quad \text{........} \ (3.18)$$

These equations of bandwidth based on Bessel functions can then be substituted into the limitation equations as presented in section 3.3 on page 137, to give an indication as to whether a combination of frequency deviation and modulating frequency will result in a non-overlapping bandwidth which can be successfully used with phase demodulation, for the PDOT method.

As can be seen, these bandwidth equations rely on determining the maximum significant sideband of a signal harmonic, which is either $k_{\max}$ or $k_{\max}^{n}$ in the above equations. Multiple criteria are used in the literature to define significant sidebands. These different methods are presented in detail in section 2.5.3 starting on page 95.

For the phase demodulation based order-tracking methods, the selected sideband definition which is used is a percentage threshold, based on the largest modulus of the Bessel function value. This sideband definition is the optimum type to use with the PDOT method. This corresponds to the first common sideband definition type, with third reference value, as presented in section 2.5.3.

Practically, suitability of a signal for order-tracking is determined from the spectrum of a reference component. As mentioned previously, a reference component typically consists of a frequency modulated harmonic series carrier. As this spectrum consists of multiple harmonic orders, it is impractical to calculate the power of just one harmonic order. This precludes the use of a significant sideband definition based on a percentage of signal power, the definition typically used for communication signals, which was the second common sideband definition type generally used for FM signals listed in section 2.5.3.

Out of the three common reference values used with a percentage threshold based significant sideband definition, the first two definitions are not desirable for use with the PDOT method. The first common reference value of $|J_0(0)|$, which in practice corresponds to the magnitude of the unmodulated carrier A_c is typically unknown for order-tracking applications, and so cannot be used. For most order-tracking applications generally only the final frequency modulated reference component is known, and so prior knowledge of the carrier and message signal parameters is not available, as they would be with communications applications. The second common reference value of $|J_0(\beta)|$, which in practice is the amplitude of the carrier component in the frequency modulated signal, is undesirable as this value is equal to zero for some values of β, which makes any corresponding percentage threshold also zero in this instance. As a result, the third common reference value type is used for the PDOT method, as mentioned above.

In practice, this sideband definition means that for this work a sideband is considered significant if the ratio of its magnitude to that of the sideband with the highest magnitude is within the specified dynamic range.

In this work, two different Dynamic Range (DR) values, 40dB and 20dB, are used as examples, but other values could be used with the PDOT method depending on the specific application.

Generally, the spectrum of a reference signal which contains a separable order, suitable for use with the PDOT method, will visually show clear separation of the order(s) to be used for order-tracking. This separation will approximately correspond to the DR being taken for significant sideband definition. It should be noted that in the spectrum, the overlap of sidebands from adjacent harmonics will be additive, which can change the visible DR from that used to define a significant sideband (which assumes each harmonic order is in isolation). In the worst case, the effects of overlap will result in a visible DR which is 6 dB smaller than that used for an equivalent significant sideband definition, but this is extremely unlikely. It would only occur if these (discrete) sidebands coincided exactly, and were in phase. A more likely decrease for a general modulation spectrum is 3 dB. Based on this, if a spectrum visually shows separation with a given DR, then it will be slightly conservative.

The separation of harmonic orders, based on DR, is more clearly shown illustratively.

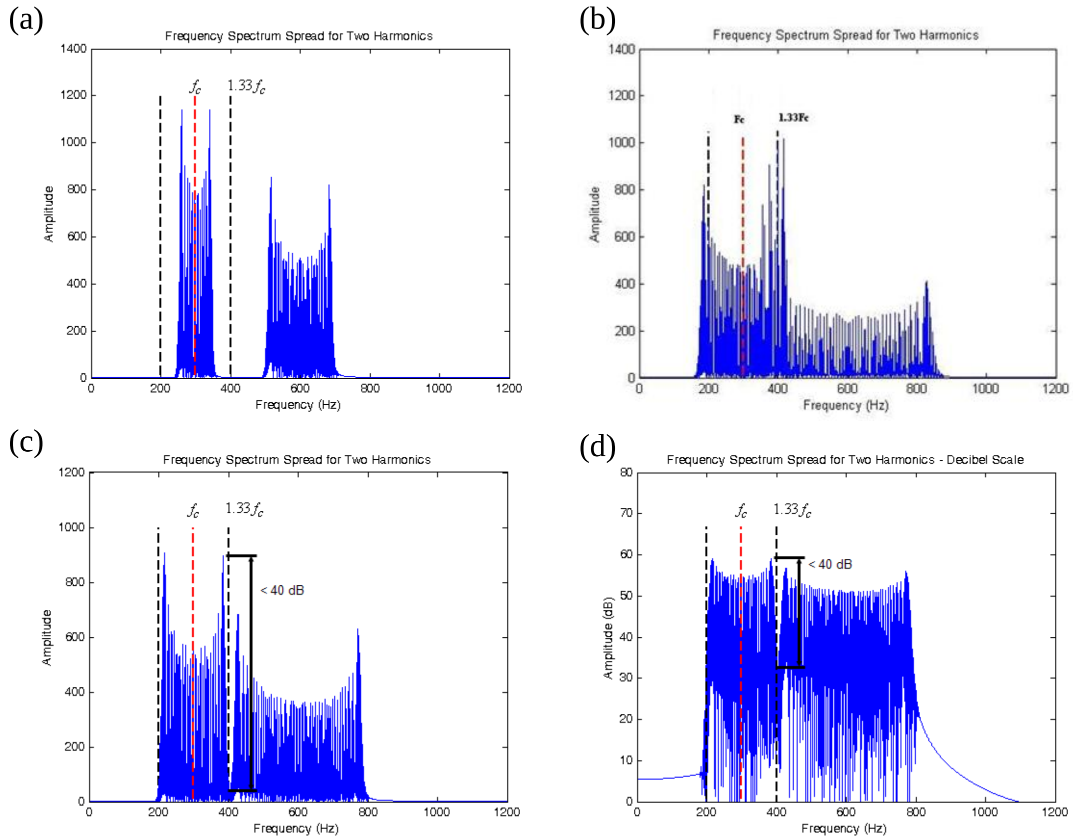


Figure 3.3 – Spread of sidebands for different frequency variations
(a) Spread of two non-overlapping harmonics (b) Spread of two overlapping harmonics
(c, d) Spread of two harmonics with overlap between dynamic ranges 20 and 40 dB
(c) Linear scale (d) dB scale

Figure 3.3(a) shows a case where the two harmonics of the reference signal shown are clearly separated using a DR of 40dB or 20dB, and could be successfully demodulated. The distribution of the two sets of sidebands has a similar shape, but the width of the sideband spread for the second harmonic is greater than that of the first harmonic. Figure 3.3(b) shows a case where the two harmonics of the reference signal shown have clearly overlapped. In this case, sidebands from the second harmonic are present within the DR (both 20 and 40dB) of the first harmonic, which is taken from the maximum value of the first harmonic order.

In Figure 3.3(c) and (d), where (c) has a linear scale and (d) shows the same case with a dB amplitude scale, the first and second orders are just beginning to overlap. In this instance the 20dB criterion on DR would be satisfied, but not the 40dB criterion.

With this definition of a significant sideband, it is possible to define the bandwidth of an FM reference component using the Bessel function approach, which in turn allows the determination of acceptable frequency deviation and modulation parameters which can be used for order tracking.

3.4.2 Defining bandwidth in terms of Frequency variation and modulation using Carson's Rule

An alternative approach to calculating the bandwidth of a frequency modulated signal is to use Carson's rule [39], which is an approximation of the bandwidth of a frequency modulated sinusoidal carrier. The usage of Carson's rule to determine bandwidth of a FM signal was discussed in detail in section 2.5.3, which starts on page 80.

Bandwidth calculated by Carson's rule is said to contain 98% of the power of an FM signal, which corresponds to the power outside the bandwidth being approximately 17dB lower than the power within the bandwidth. While this DR is not directly comparable to those used with Bessel function based calculations, as given in the previous section, due to differences in definitions, Carson's rule generally gives a less accurate approximation of bandwidth than given with Bessel function calculations.

The advantage of using Carson's rule is that it allows for significantly simpler bandwidth calculations, than using Bessel functions.

The bandwidth of an FM sinusoid with non-sinusoidal modulation is simply given by equation (2.35) on page 93, or equivalently by equation (2.36). The bandwidth of a harmonic order of a FM harmonic series with non-sinusoidal modulation is similarly given by equation (2.48) on page 105.

Similarly to the Bessel function approach given in the previous section, it is possible to express the frequency deviation and modulating frequency as a percentage of the carrier frequency to simplify bandwidth acceptability calculations.

With this change of parameters, for a FM sinusoid with non-sinusoidal modulation, the bandwidth from Carson's rule using equations (2.35) and (2.36) can alternatively be expressed as a percentage of the carrier frequency, $F_{\%c}$, so that:

$$F_{\%c} = 2(D_{\%c} + W_{\%c}) \text{ (\% of } f_c) = \pm(D_{\%c} + W_{\%c}) \text{ (\% of } f_c) \text{ around } f_c \quad \text{... (3.19)}$$

With the same change in parameters, for a harmonic order of a FM harmonic series with non-sinusoidal modulation, the bandwidth from Carson's rule using equation (2.48), , can alternatively be expressed as a percentage of the carrier frequency, $F_{n\%c}$, so that:

$$F_{n\%c} = 2(nD_{\%c} + W_{\%c}) \text{ (\% of } f_c) = \pm(nD_{\%c} + W_{\%c}) \text{ (\% of } f_c) \text{ around } nf_c \quad (3.20)$$

3.5 Acceptable values of frequency variation and modulation frequency, for no amplitude modulation

With the bandwidth calculations from Sections 3.4.1 and 3.4.2, starting on page 147, it is possible to determine the values of $f_{d\%c}$ or $D_{\%c}$ (frequency variation, or deviation) and $f_{m\%c}$ or $W_{\%c}$ (modulating frequency) which can be used with the PDOT method, by equating these with the bandwidth limitation calculations in section 3.3 on page 137. Note that these calculations shown here assume the reference signal is not amplitude modulated. The inclusion of amplitude modulation is covered in the later section 3.7.

Firstly, the bandwidth limitation equation for a harmonic order, given by equation (3.7) on page 142, can be updated to reflect the limiting bandwidth being expressed as a percentage of carrier frequency (i.e. $f_c = 100\%$), where the limiting bandwidth is $F_{n\%c}^{\lim}$ for the nth harmonic, to give:

$$\frac{F_{n\%c}^{\lim}}{2} + \frac{F_{(n+1)\%c}^{\lim}}{2} = 100\% \quad \text{.......................................} \ (3.21)$$

With this limitation, it can be seen that for a harmonic order n to be acceptable for use with order tracking, then:

$$\frac{F_{n\%c}}{2} + \frac{F_{(n+1)\%c}}{2} \leq 100\% \quad \text{.......................................} \ (3.22)$$

For bandwidth calculations based on Bessel functions, the bandwidth values for harmonic orders n and $(n+1)$, given by equation (3.18) on page 150, can be substituted into equation (3.22), resulting in:

$$\frac{\cancel{2}k_{\max}^{n} f_{m\%c}}{\cancel{2}} + \frac{\cancel{2}k_{\max}^{(n+1)} f_{m\%c}}{\cancel{2}} \leq 100\% \text{ (of } f_c\text{)} \quad \text{......................... (3.23)}$$

This equation can be solved, using different DR definitions to give different values of $k_{\max}^{n}$, to determine if a combination of modulation parameters (frequency deviation and modulating frequency) is acceptable for use with the PDOT method. Note that $k_{\max}^{n}$ is a function of $f_{d\%c}$, so while not explicitly listed, equation (3.23) in effect includes the frequency deviation parameter. As stated previously, $f_{d\%c}$ and $f_{m\%c}$ can be exchanged for the parameters $D_{\%c}$ and $W_{\%c}$, respectively to give an approximation for acceptable parameters for a non-sinusoidal modulation case.

For bandwidth calculations based on Carson's rule, the bandwidth values for harmonic orders n and $(n+1)$, given by equation (3.20) on page 157, can be substituted into equation (3.22), resulting in:

$$\frac{\cancel{2}(nD_{\%c} + W_{\%c})}{\cancel{2}} + \frac{\cancel{2}([n+1]D_{\%c} + W_{\%c})}{\cancel{2}} \leq 100\% \text{ (of } f_c\text{)} \quad \text{.............. (3.24)}$$

which can be factorised to give:

$$(2n+1)D_{\%c} + 2W_{\%c} \leq 100\% \text{ (of } f_c\text{)} \quad \text{................................ (3.25)}$$

For the case of sinusoidal modulation, to be comparable to equation (3.23), $D_{\%c}$ and $W_{\%c}$ can be replaced by $f_{d\%c}$ and $f_{m\%c}$, respectively, in equation (3.25).

It should be emphasised that actual acceptability of a combination of modulation parameters to be used with the PDOT method should preferably be based on inspection of the spectrum of the reference signal to check for suitable separation of orders based on a DR. These bandwidth calculations to determine suitability of an order should only be used as a guide.

3.6 Acceptability chart for first harmonic

With the acceptable values for frequency deviation and modulation, given by equations in the previous section, it is possible to construct an acceptability chart indicating maximum allowable speed deviation for different maximum modulating frequencies.

The acceptability charts shown here are for the first harmonic order of a reference component ($n = 1$), which due to the spreading of higher harmonics gives the maximum values for which the PDOT method can be used. As stated earlier, higher harmonics will have increasingly restricted modulation parameters. Also note that these calculations assume the reference signal is not amplitude modulated, the inclusion of which would lower the acceptable modulation parameters.

Similar acceptability charts for the harmonics $n = 2$ to $n = 10$ are shown in Appendix A on page 423.

To construct each acceptability chart, an array of parameters for $f_{d\%c}$ and $f_{m\%c}$ were constructed using a resolution of 0.1%. Values for $f_{d\%c}$ spanned from 0.1 to 40%. Values for $f_{m\%c}$ spanned from 0.1 to 50%. Every combination of these parameters in the array was tested against the appropriate acceptability equation.

For the first acceptability chart, Bessel function based calculations were performed, using a DR of 40dB, with equation (3.23) on page 159 for harmonic order $n = 1$.

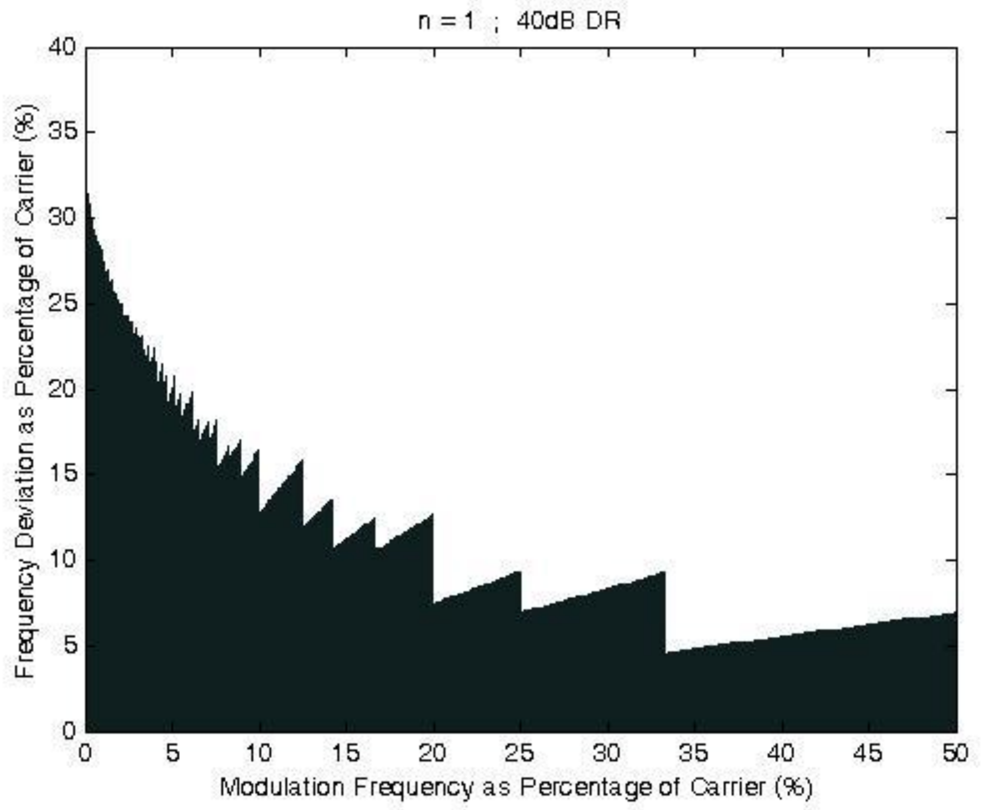


Figure 3.4 – Combinations of frequency deviation and modulating frequency permissible for the PDOT method, for the first harmonic order, using Bessel function calculations and a 40dB DR

Figure 3.4 shows the acceptable values of frequency deviation and modulating frequency which are permissible for the PDOT method for the first harmonic, based on Bessel function calculations with a 40dB DR. All acceptable combinations of frequency deviation and modulating frequency are shown in black. It should be noted that the discontinuous nature of the chart is not an error, but a result of the integer jumps in value $k_{\max}$, for both the first and second harmonic orders, as the largest non-negligible sideband in each order changes.

For the second acceptability chart, Bessel function based calculations were performed, using a DR of 20dB, with equation (3.23) on page 159. This chart is shown in Figure 3.5.

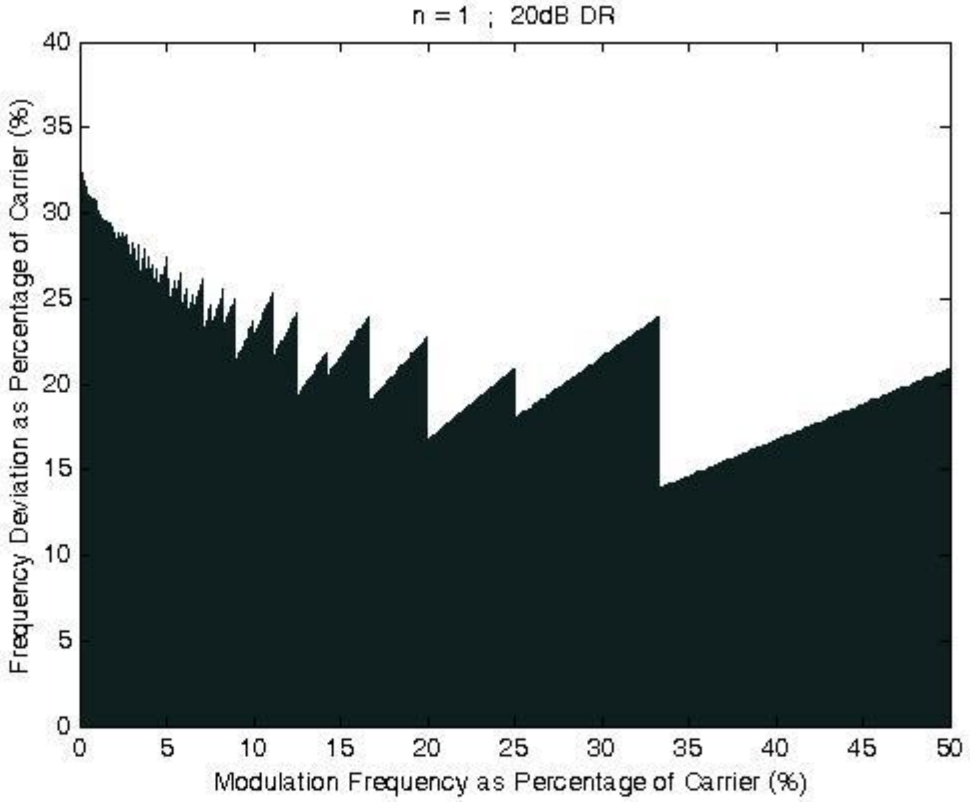


Figure 3.5 – Combinations of frequency deviation and modulating frequency permissible for the PDOT method, for the first harmonic order, using Bessel function calculations and a 20dB DR

For the third acceptability chart, Carson’s rule based calculations were performed, with equation (3.25) on page 159.

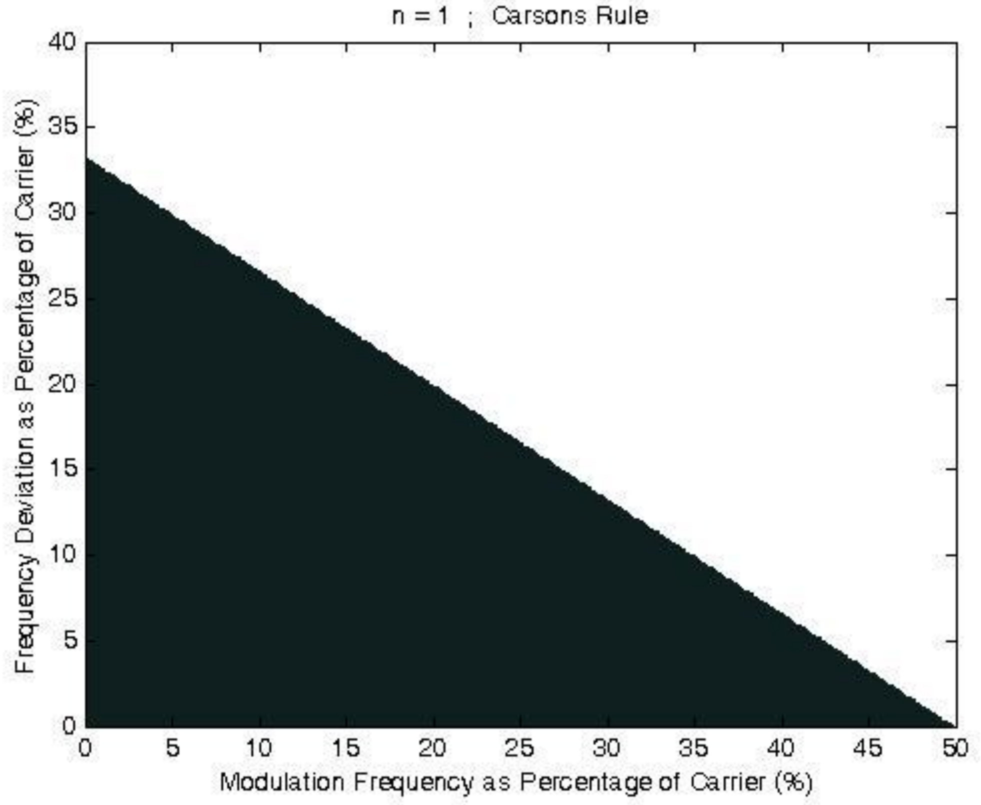


Figure 3.6 – Combinations of frequency deviation and modulating frequency permissible for the PDOT method, for the first harmonic order, using Carson’s rule

Figure 3.6 shows the permissible values of frequency deviation and modulation given using Carson’s rule. It should be noted that this acceptability limit is simply a linear curve in this instance, due to the simplicity of equation (3.25).

In order to compare the different acceptability combinations given in the different acceptability charts, the envelope of each chart is plotted in Figure 3.7.

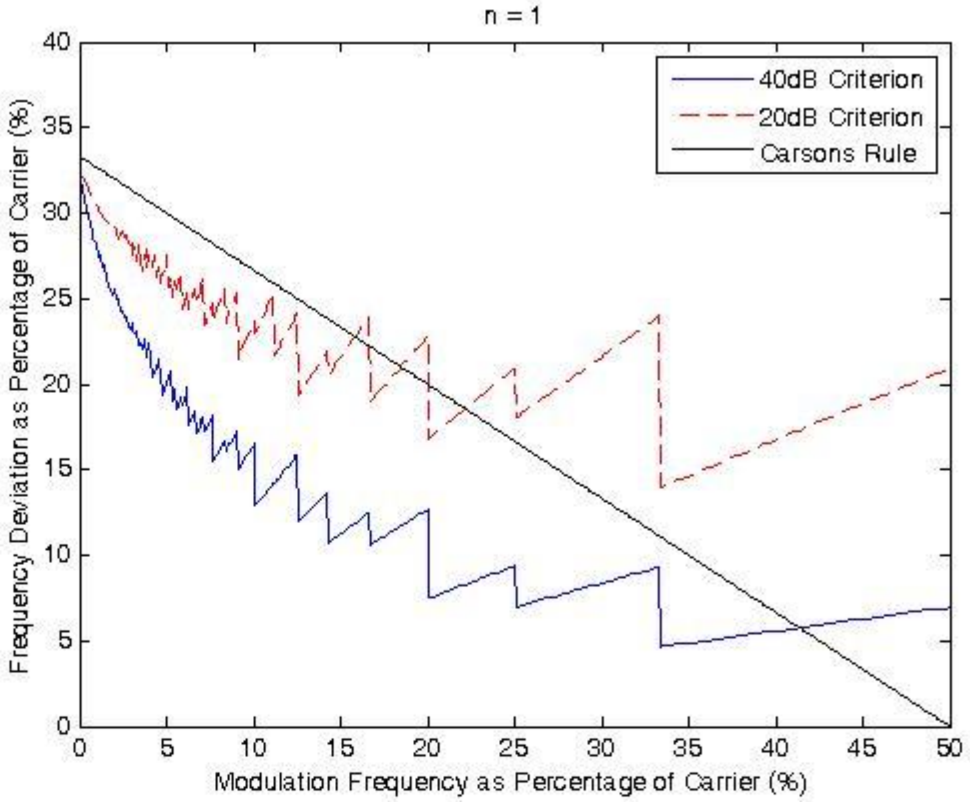


Figure 3.7 – Envelope of combinations of frequency deviation and frequency modulation permissible for the PDOT method, using the Bessel function criterion for 40dB DR and 20dB DR and Carson's rule

As can be seen in Figure 3.7, the different methods of determining permissible bandwidth result in different acceptable values. As these different methods effectively indicate acceptability of different levels of harmonic order overlap, using each definition will result in differing quality to the resulting order-tracking of a signal. This will be shown in the following section 3.8.

It should be noted that acceptability charts similar to Figure 3.4 to Figure 3.7 were previously presented in [1, 6, 9], however these previous acceptability charts were all developed based on the assumption that the bandwidth of a harmonic order is proportional to harmonic order, so using bandwidth limitation equation (3.5) on page 141. The acceptability charts presented in this work do not make use of the assumption of proportionality of harmonic orders, so are based on equation (3.7) on page 142. As mentioned in section 3.3 on page 137, it is preferable to not make use of the assumption of proportionality of the bandwidth of harmonic orders, and as such it is recommended that the acceptability charts presented in this thesis work be used in preference to those published in my prior works [1, 6, 9].

3.7 Acceptable values of frequency variation and modulation frequency, with amplitude modulation

As mentioned previously, the acceptability equations in Section 3.5 on page 158 are based on the assumption of no amplitude modulation being present in the reference component, as is common with tachometer and encoder type signals. The following acceptability charts in section 3.6 on page 161are based on the equations in section 3.5, and so also assume no amplitude modulation is present.

For reference signals which contain amplitude modulation, such as vibration response signals, the acceptability equations in section 3.5 cannot be used directly to determine if the first harmonic is suitable for order-tracking.

However, acceptability equations can be given including amplitude modulation, in a similar fashion to that shown in section 3.5.

In section 3.3 on page 137, bandwidth limitation equations were also given which include the case of a reference signal also being amplitude modulated. This is specifically covered by the bandwidth limitation equation (3.9) on page 144.

So, with the bandwidth calculations from Sections 3.4.1 and 3.4.2 starting on page 147, it is possible to determine the values of $f_{d\%c}$ or $D_{\%c}$ (frequency variation, or deviation) and $f_{m\%c}$ or $W_{\%c}$ (modulating frequency) which can be used with the PDOT method, by equating these with the bandwidth limitation equation (3.9).

Firstly, the bandwidth limitation equation for a harmonic order, given by equation (3.9), can be updated to reflect the bandwidth terms being expressed as a percentage of carrier frequency (i.e. $f_c = 100\%$), where the limiting FM bandwidth is $F_{n\%c}^{\lim}$ for the nth harmonic, and AM bandwidth is $F_{a\%c}$, to give:

$$\left(\frac{F_{n\%c}^{\lim}}{2} + \frac{F_{a\%c}}{2}\right) + \left(\frac{F_{(n+1)\%c}^{\lim}}{2} + \frac{F_{a\%c}}{2}\right) = 100\% \quad \text{......................} \ (3.26)$$

With this limitation, it can be seen that for a harmonic order n to be acceptable for use with order tracking, then:

$$\frac{F_{n\%c}}{2} + \frac{F_{(n+1)\%c}}{2} + \frac{\cancel{2}F_{a\%c}}{\cancel{2}} \leq 100\% \quad \text{...............................} \ (3.27)$$

It should be noted that Equations (3.9), and subsequently (3.26) and (3.27), assume that the total bandwidth of the first harmonic is given by the sum of the bandwidth of the AM component and the FM component, which is over conservative if the ‘true’ AM bandwidth is used.

The first harmonic of the reference component will consist of the convolution of the spectra of the AM and FM components, as discussed in section 2.5.6 starting on page 114. While the bandwidth of the AM component is finite, the bandwidth of the FM component is theoretically infinite and so only defined for a given DR. Hence, the convolved result is also theoretically infinite (convolution with an infinitely long spectrum gives an infinitely long result), and so it is only fully meaningful for the convolved result to also be defined in terms of the same DR used to define FM.

So, the total bandwidth of the first harmonic of the reference component is then the bandwidth of the convolved result (infinite) to the specified value of DR. As a result, the total bandwidth will always be somewhat less than the sum of the FM and ‘true’ AM component bandwidths for a given DR.

However, it is possible to define the “effective” bandwidth F_a of the amplitude modulation component as that added amount (to the nominal bandwidth of the FM component alone) which gives the same dynamic range (e.g. 20 dB or 40 dB) for the convolved spectrum, as that which defined the nominal FM bandwidth. The total bandwidth of the nth harmonic will then be the sum of the bandwidth of the “effective” AM bandwidth F_a and the frequency modulation bandwidth F_n, and this F_a can be used with Equation (3.9) on page 144. This “effective” AM bandwidth can be expressed as a percentage of the carrier frequency $F_{a\%c}$.

In practice, determining the “effective” AM bandwidth will be complicated unless the AM and FM spectra are known approximately and can be convolved. Visual inspection would in general be simpler.

For prediction with unknown signals, conservative values can always be obtained for the AM bandwidth term F_a, by using the “true” AM bandwidth, with:

$$\frac{F_a}{2} = f_a \quad \text{......} \quad (3.28)$$

where f_a is the maximum modulating frequency of the amplitude modulation component. Thus, a bandwidth limit $F_n^{\lim}$ calculated when using the “true” F_a, from equation (3.28), will give lower permissible values than if the “effective” AM bandwidth were used.

This approximation of the AM bandwidth term, in equation (3.28), can alternatively be expressed as a percentage of the carrier frequency $F_{a\%c}$, to give:

$$\frac{F_{a\%c}}{2} = f_{a\%c} \quad \text{......} \quad (3.29)$$

where $f_{a\%c}$ is f_a expressed as a percentage of the carrier frequency f_c.

So, for bandwidth calculations based on Bessel functions, the bandwidth values for harmonic orders n and $(n+1)$, given by equation (3.18) on page 150, can be substituted into equation (3.27) on page 168. The approximate "true" AM bandwidth in percentage form, from equation (3.29), can also be substituted, resulting in:

$$\frac{\cancel{2}k_{\max}^{n} f_{m\%c}}{\cancel{2}} + \frac{\cancel{2}k_{\max}^{(n+1)} f_{m\%c}}{\cancel{2}} + 2 f_{a\%c} \leq 100\% \text{ (of } f_c\text{)} \quad \text{................ (3.30)}$$

This equation can be solved, using different DR definitions to give different values of $k_{\max}^{n}$, to determine if a combination of modulation parameters (frequency deviation, modulating frequency of the FM component, and maximum modulating frequency of the AM component) is acceptable for use with the PDOT method. As stated previously, parameters $D_{\%c}$ and $W_{\%c}$ can be exchanged for $f_{d\%c}$ and $f_{m\%c}$, respectively, to give an approximation for acceptable parameters for a non-sinusoidal modulation case. Note that $f_{a\%c}$ is the maximum modulating frequency of the AM component, so is equally applicable to sinusoidal and non-sinusoidal modulation cases.

For bandwidth calculations based on Carson's rule, the bandwidth values for harmonic orders n and $(n+1)$, given by equation (3.20) on page 157, can be substituted into equation (3.27) on page 168, resulting in:

$$\frac{\cancel{2}(nD_{\%c} + W_{\%c})}{\cancel{2}} + \frac{\cancel{2}([n+1]D_{\%c} + W_{\%c})}{\cancel{2}} + 2 f_{a\%c} \leq 100\% \text{ (of } f_c\text{)} \quad \text{....... (3.31)}$$

which can be factorised to give:

$$(2n+1)D_{\%c} + 2(W_{\%c} + f_{a\%c}) \leq 100\% \text{ (of } f_c\text{)} \quad \text{........................ (3.32)}$$

For the case of sinusoidal modulation, to be comparable to equation (3.30), $D_{\%c}$ and $W_{\%c}$ can be replaced by $f_{d\%c}$ and $f_{m\%c}$, respectively, in equation (3.32).

It should again be emphasised that actual acceptability of a combination of modulation parameters to be used with the PDOT method should be based on inspection of the spectrum of the reference signal to check for suitable separation of orders based on a given DR. The above bandwidth calculations to determine suitability of an order should only be used as a guide. This is particularly true when the approximate "true" AM bandwidth is used in the calculations.

3.8 Simulated testing with a frequency modulated sine wave

In order to evaluate the basic PDOT method, it was first tested with simulated signals. This was conducted in order to evaluate its success in the absence of external contaminating factors.

To begin with a simple case, an identical reference signal and signal of interest were constructed, in the form of a frequency modulated sinusoid with sinusoidal modulation, of the same form as defined by Equation (2.28) on page 85. Results for a single harmonic were studied in detail in [6, 9], and are not shown here.

Another test signal was then generated with ten harmonics (frequency modulated harmonic series carrier), by Equation (2.40) on page 101, with no other frequency components or noise. All harmonics were cosines to match a pulse type tacho signal. Note that the signal was frequency modulated only, and no amplitude modulation was included.

The two reference signals are shown in Figure 3.8. The reference signal with ten harmonics (Figure 3.8(b)) was used to order-track itself and also a data signal corresponding to the first harmonic only (Figure 3.8(a)).

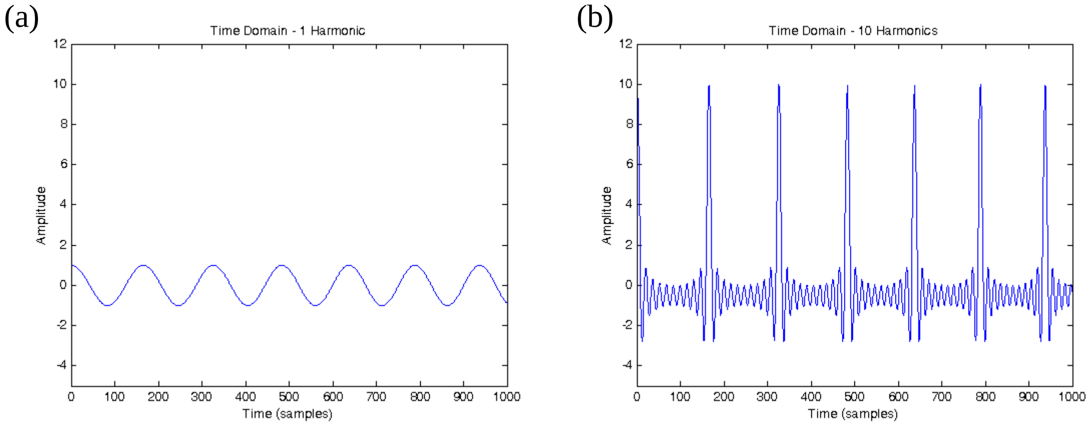


Figure 3.8 – Time domain plots of the reference signals
(a) single harmonic (b) ten harmonics

Note that the effects of the frequency modulations are not visible in these plots as the displayed time lengths are too short relative to the frequency modulation parameters.

In order to evaluate the effects of using the three bandwidth definitions for order-tracking, three different simulated tests were conducted. For the same value of $f_{m\%c}$, the maximum allowable value of $f_{d\%c}$ was calculated using each bandwidth definition from Figure 3.7 on page 165.

For the following tests the following common signal parameters were used:-

$$f_s = 50 \text{ kHz}, \; f_c = 300 \text{ Hz}, \; A_c = 1, \text{ Time } t = 5 \text{ s}, \; f_{m\%c} = 2\%$$

The maximum allowable $f_{d\%c}$ for the Bessel function criterion with 40dB DR, was found to be 25.6%, using equation (3.23) on page 159, so $f_{d\%c} = 25\%$ was used for the simulation.

The maximum allowable $f_{d\%c}$ for the Bessel function criterion with 20dB DR, was found to be 29.3%, using equation (3.23), so $f_{d\%c} = 29\%$ was used for the simulation.

For the Carson's rule criterion, the maximum allowable $f_{d\%c}$ was 32%, so this was used for the simulation.

For each of these cases, the ten-harmonic reference signal was used to conduct order-tracking on both the single harmonic data signal and the reference signal itself, so both signals can be compared.

Ideally, the order-tracked result spectrum should consist of one single discrete frequency at order 1 for the data signal, and 10 discrete frequency values at orders 1-10 for the reference signal.

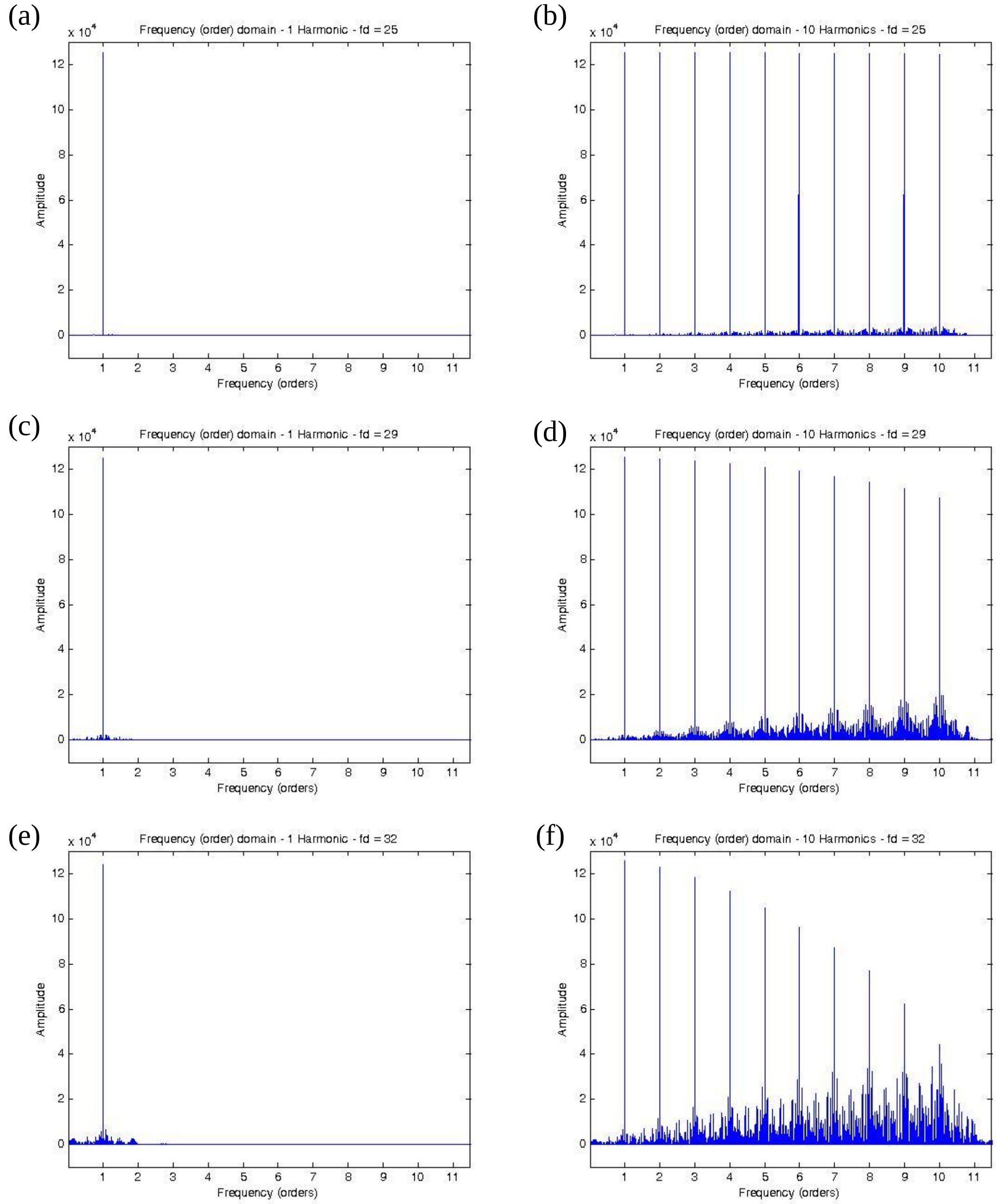


Figure 3.9 – Order-tracked spectra for simulated signal for
(a, b) $f_{d\%c} = 25\%$ (40 dB DR) (c, d) $f_{d\%c} = 29\%$ (20 dB DR)
(e, f) $f_{d\%c} = 32\%$ (Carson's rule) (a, c, e) Data signal with 1 harmonic
(b, d, f) Reference signal with 10 harmonics

Figure 3.9 shows the order spectra for the order tracked results for one and ten harmonics, the three limiting criteria 40dB DR, 20dB DR and Carson's rule. From these spectra it can be concluded that the Bessel function criterion with 40dB DR is the only one that gives completely acceptable results with no significant aliasing, though the other two give separation of the first few orders, which may be acceptable for some applications, in particular if no further improvement by iteration is desired.

It should be noted that the harmonics of the signal all had equal magnitude, and so a C_n value of 1, which although not typical of harmonic patterns seen in real signals, helps to illustrate the increasing smearing of higher harmonics with incorrect order tracking.

It should be emphasised that these simulated results showing the effects of overlap are for ideal FM signals with no other contaminating components. The effects of noise and amplitude modulation are illustrated in some later examples.

Finally, it should be noted that similar simulation results were previously presented in [1], however different maximum values were calculated in that case, based on the assumption that the bandwidth of a harmonic order is proportional to harmonic order (as per equation (3.5) on page 141), and so the previous simulated results in [1] were shown for different $f_{d\%c}$ values.

Chapter 4 Generalised methodology for phase demodulation based order tracking

Chapter Overview

The following chapter expands the basic PDOT method, presented in Chapter 3, into the generalised methodology for PDOT.

In this chapter a number of options are presented, which can be employed in multiple combinations in a modular fashion, to implement PDOT tailored to different applications.

These options include:

- A variety of reference signal types, including using no external reference
- Using one or multiple stages for order-tracking
- Using a segmented approach, for very large speed variations
- Using a reversible approach, allowing a signal to be modified in the order domain, and then transformed back to the time domain

4.1 Applied phase demodulation based Order-Tracking

The basic PDOT method as described in Chapter 3, starting on page 127, forms the core of a generalised methodology for order-tracking signals for a variety of applications. Four main features allow the PDOT to be used in a generalised fashion.

The first feature is that the method can be used with a variety of reference signal types, as discussed below, which have different limitations, and result in different order-tracked results which may be suitable for different applications.

The second feature is that while the basic PDOT method can only be employed with one harmonic order at a time, it is possible to implement the method multiple times in series on a range of harmonic orders. This is advantageous, as low harmonic orders allow a larger speed variation to be order-tracked, but higher harmonics generally result in more accurate order-tracking due to having a higher resolution.

The third feature is that the PDOT method can be employed in a segmented fashion for signals with very large speed variations which cannot be order-tracked with basic PDOT. A signal can be separated into multiple segments, each of which containing a permissible speed range for use with the basic PDOT method. Each segment can be order-tracked separately using the PDOT method, and then all segments recombined to give an order-tracked result for the entire signal.

The fourth feature is that the PDOT method is reversible, so is suitable for signal processing techniques which require operations both in the order domain, and the time domain. An example presented in this work is the filtering of order-based components from a signal, before analysis in the time domain.

The main advantage of the PDOT method, compared with most others, is that the variation in phase (and frequency) with time is completely defined by its spectral representation. This holds provided that aliasing, caused by overlapping of spectral sidebands from higher orders (or in fact over zero frequency for a single order), is avoided. With other methods some sort of curve is being fitted to a series of points without necessarily corresponding to the true speed variation. The above discussion has defined the conditions for which the aliasing can be avoided, and shown that a 40 dB DR criterion is satisfactory for most practical purposes. This will be taken up in the examples of application of the method in Chapter 5 and Chapter 6.

4.2 Reference Signal Type

The only two essential requirements for a reference signal are that it be low pass filtered, and that the included reference component is exactly phase-locked to the component of interest. Note that this means that pulses for example must have constant aspect ratio.

Any reference signal which satisfies these requirements is suitable for use with the PDOT method. However, the exact nature of the reference signal has an impact on the order-tracked result, and so different reference signals are advantageous for different applications.

The properties of three common reference signal types are discussed here, although these are not the only possibilities.

4.2.1 Once-per-Revolution Pulse Tacho

One of the most common reference signal types is a once-per-revolution tacho signal. This signal gives one single pulse per revolution of a rotational component. This might come from detection of a raised section or hole by a proximity transducer (key phasor), or a photoelectric sensor detecting passage of a reflecting or black patch on a shaft. Such pulses often have both constant amplitude and aspect ratio, making them ideal for the PDOT method, however this is not always true, and a specific tacho device should be evaluated for suitability. Furthermore, the spectrum of a tacho signal generally has minimal noise and no other frequency components which could contaminate the harmonics of the reference component, which allows a tacho signal to be employed with the full modulation parameters as shown in Section 3.5 on page 158.

As a tacho signal only contains effectively one phase point per rotation, the reference component only represents speed variations up to half the tacho fundamental order. When used for order-tracking, the low frequency variations of the shaft will be compensated for, but any high-frequency variations will remain. This can be advantageous for analysis which utilises higher-frequency variations, such as phase demodulation to detect gear cracks. It can however be disadvantageous for situations where all frequency variations need to be compensated for.

4.2.2 Encoder

An encoder signal is very similar to a tacho signal, but contains multiple pulses per revolution of the component of interest. As with a tacho, the pulses typically have constant amplitude and aspect ratio, and the spectrum is free of noise and contaminating components, giving comparable advantages. An encoder requires attachment to a shaft, and so can only be used in situations where shafts are accessible. If the encoder attachment is via a flexible coupling this will limit the upper frequency order to which PDOT can be applied.

As an encoder contains multiple pulses per revolution, it does contain information of higher frequency variations, up to half the order of the encoder pulses per revolution. An encoder can be advantageous for situations where all frequency variations need to be compensated for, but disadvantageous where it may be desired to retain information, for example about deterministic frequency modulation of gearmesh components, which would otherwise be removed by a high order encoder.

4.2.3 Using the Signal Itself

It is also possible to use the data signal itself as a reference signal, where the reference component harmonic is extracted from the data signal. This has the advantage that a separate sensor is not needed. For vibration signals, an accelerometer can generally be fitted on the machinery housing, and so can be used for cases where a shaft suitable for a tacho or encoder mounting is physically inaccessible. The data signal also provides a continuous measure of the reference component phase, so a full range of frequency variations can be compensated for.

A data signal has the disadvantage of generally containing a significant amount of noise, and other frequency components, so that it can be difficult to find a separable reference component harmonic, which can further restrict the allowable frequency modulation parameters.

A data signal will generally also include amplitude modulation of the reference component (general modulation), increasing the bandwidth of a harmonic and so further restricting allowable modulation parameters.

A major potential problem is that for rapid speed changes, the response signal does not directly follow the input forcing function, because of the impulse response time of the system, and so the frequency of the response is not fully aligned with the current instantaneous speed. This problem has been thoroughly investigated by Borghesani et al [40], but does not affect many applications, such as wind turbines, where the inertia is so large that the response is effectively immediate.

As a result of these factors, the suitability of using the response data signal directly is highly case specific, and each individual application needs to be evaluated separately.

4.3 Multi-Stage Approach

As stated in Section 4.1, it is possible to implement the PDOT method multiple times in series to refine the order-tracking process, and so the number of iterations to be used can be varied for different applications.

In order to implement a multistage approach, order-tracking at each stage is conducted on both the signal of interest as well as the corresponding reference signal. The order-tracked reference signal is then used as the basis of the next stage of order-tracking. For the first stage, a low harmonic order is used allowing for a large permissible speed variation range, which removes most speed fluctuations from the reference component, leaving only residual speed variations on higher harmonics, making initially overlapping higher harmonics separable and thus usable by the PDOT method. Successively higher and higher harmonics can be used, in each case further refining the order-tracked result. Figure 4.1 graphically shows the procedure for order-tracking, extended to N stages.

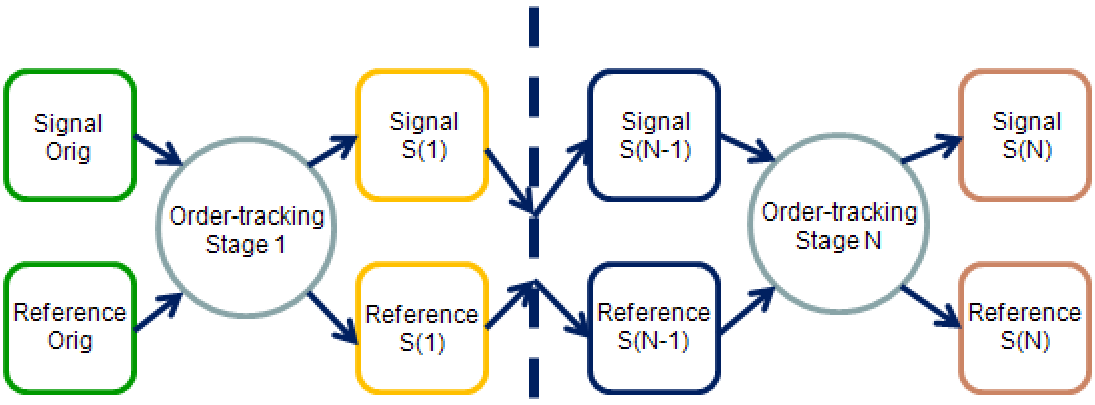


Figure 4.1 – Multi-stage approach to Order-Tracking

Figure 4.2 shows the specific procedure for using the data signal as a reference signal to perform multi-stage order-tracking, extended to N stages. It should be noted that a new reference component harmonic is extracted at each stage, rather than the initially extracted harmonic being used for all stages.

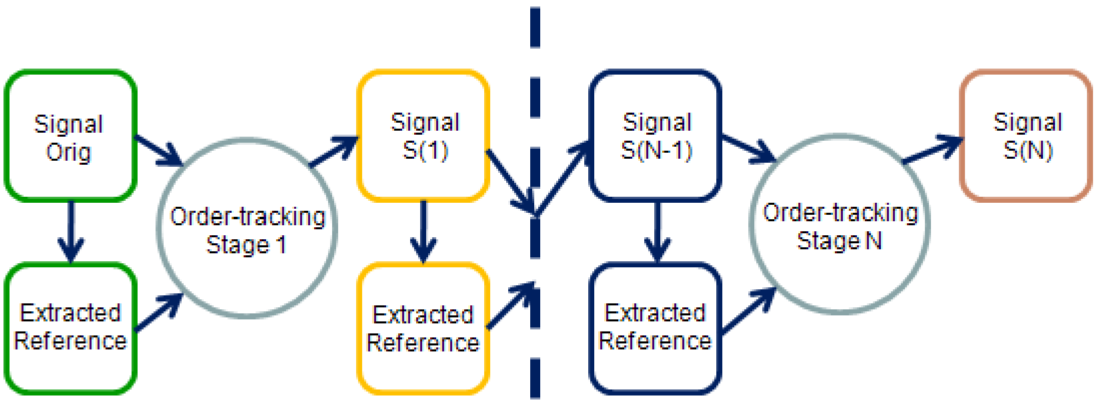


Figure 4.2 – Multi-Stage approach to Order-Tracking using an Extracted Reference Signal

Any number of stages could be utilised to progress to higher and higher frequency harmonics, but the errors introduced by the PDOT method would be compounded with each stage. Where the speed variations are small, and thus amplitude modulation of the reference component negligible, these errors have been found to be many orders of magnitude smaller than the residual modulation typically remaining, and have allowed successive improvement over several iterations.

The required number of stages for a given application will generally depend on the subsequent MCM analysis to be employed, which will determine to what harmonic order accurate results are required. For a last stage, a harmonic order greater than the required maximum order needed for further analysis should be used. The number of stages will then be one more than the number required to make the harmonic to be used for a last stage separable, as each stage will remove further residual frequency variations and allow access to higher harmonic orders.

A complicating factor for the multi-stage approach is the presence of amplitude modulation in the reference component, such as typically found with response signals over a larger speed range. Amplitude modulation is generally a result of forcing components, such as gearmesh harmonics, passing through fixed structural resonance frequencies. The amplitude modulation is initially completely separate from phase modulation, as shown in Equation (2.52) in the section 2.5.5 on General modulation for sinusoidal carrier, starting on page 109, and so has no effect on the phase demodulation process used for the first stage of order-tracking, apart from being part of the bandwidth of the reference component harmonic. However, after the first stage of order-tracking the amplitude modulation becomes frequency modulated in proportion to the speed variation that was compensated for. When another stage of the PDOT method is implemented, the selected frequency band around the reference component harmonic used for phase demodulation will result in the extracted phase being the combination of the desired residual speed variation to be compensated for, and the frequency modulation introduced in the amplitude modulation component. As the demodulation process is blind, these different phase variation sources cannot be separated, and so the combined phase variation will be compensated for, resulting in incorrect order-tracking when using multiple stages. The amount of amplitude modulation directly effects the degradation of the order-tracking result when using multiple stages, with larger amplitude modulations resulting in poorer results. As amplitude modulation tends to increase with greater speed variation, this can have a further limiting effect on the allowable speed variation parameters, and in cases with large amplitude modulation a multi-stage approach may not be feasible. Another important point is that if the purpose of the order-tracking is to perform TSA, then any amplitude modulation will result in a non-representative synchronous average, with a constant amplitude, which will leave a deterministic component included in the residual signal when subtracted from the total signal.

4.4 Segmented approach to order-tracking

The basic PDOT method presented in Chapter 3 is limited to an acceptable range of speed variations which can be successfully order-tracked. These acceptable ranges were discussed in detail in sections 3.5 and 3.6, starting on page 158.

It is possible to order-track signals with larger speed variations, greater than those permissible for the basic PDOT method, by making use of record segmenting, which is a previously unpublished method.

In essence, this involves segmenting the signal into multiple parts, each of which covers a speed range less than the maximum allowable with the basic PDOT method. The segments are separately order-tracked, and then recombined, resulting in an order-tracked result of the entire original signal.

To undertake the order-tracking of the data segments, the reference signal to be used is identically segmented, so there is a separate reference signal segment corresponding to each data segment. The reference signal segments are then used to order-track the corresponding data segments.

This segmenting process is applicable for use with general varying speed signals captured during machinery operation, as with the basic PDOT method, but is also applicable to the analysis of run-up and run-down situations.

The following steps and considerations are necessary to implement the segmentation based order-tracking approach, and each will be discussed in the following sections.

Firstly a plot of the speed vs. time of the machine for the entire signal is necessary. From this speed plot, the locations and size of the segments can be determined, based on the speed variation parameters of the signal. Overlapping of segments is necessary to suppress end effects, which in turn necessitates the use of windowing. The segments then need to be suitably low pass filtered to prevent aliasing. Finally, to facilitate correct rejoining of the segments after order-tracking, it is necessary to enforce that all segments are resampled on a common phase axis, in addition to an identical sampling rate, so the sampling points from adjacent segments correctly line up.

Finally, an optimisation is discussed of the multi-stage approach to order-tracking, when used in conjunction with segmentation.

4.4.1 Obtain suitable speed plot

In order to conduct the segmenting, the first requirement is to obtain a plot of speed vs. time (speed plot) for the fundamental order of the signal component which is to be used for order-tracking. If the speed is obtained from a higher order, this should be divided by the harmonic order, to give a speed vs. time plot for the fundamental order.

Typically, a fundamental shaft harmonic would be preferred for use as a reference component, as opposed to higher harmonics, as this would allow a greater speed range to be covered per segment. However, any suitable harmonic of a reference component could be used, as with the basic PDOT method. It should be emphasised that the speed plot used to conduct segmenting should be normalised to the fundamental order, even if a higher order component is to be used as the reference signal.

A speed plot can be produced via any suitable method, and will not be discussed in detail in this work, as prior literature contains extensive work on this subject, and it will inevitably be application specific. The speed plot can be significantly coarser than the reference signal needed for order-tracking, as the speed plot is only used for segmenting and not for order-tracking itself, and so does not have to be produced from the same source. A coarser speed plot will simply require a greater factor of safety in determining the segmenting locations. The speed plot should have units of Hertz, to facilitate calculations described in the following sections.

As the PDOT method produces the best order-tracked result using an external reference signal, such as a tachometer or encoder, it would be desirable to construct the speed plot based on the pulse timings from one of these reference signal types if available.

On the other hand, if tacho-less order-tracking is to be conducted using only the response signal itself, a speed plot could alternatively be produced by using peak detection from a spectrogram plot of the response signal, such as described by Urbanek et al. [25].

4.4.2 Calculate speed variation parameters, to determine segment size

The desired segmenting locations are determined from the plot of speed vs. time. The maximum segment size is dependent on the maximum permissible speed variation parameters which can be accommodated using the basic PDOT method. So, in order to determine the size of a segment, it is first necessary to determine the speed variation parameters for a given segment.

The maximum permissible speed variation parameters for the basic PDOT method were presented in section 3.5 on page 158. The permissibility of a segment was shown to be related to three speed parameters, the maximum percentage frequency deviation $D_{\%c}$, maximum percentage modulating frequency $W_{\%c}$, and carrier frequency f_c.

As the frequency deviation and modulating frequency limits are non-dimensionalised based on the carrier frequency f_c, this parameter must first be determined for a segment being considered for permissibility. Values for the other two parameters can then be determined, and then permissibility of a segment subsequently determined.

4.4.2.1 Determining carrier frequency and maximum frequency deviation values

The carrier frequency f_c can simply be taken as the midpoint between the maximum and minimum speeds of the segment being considered. So if the maximum speed in a segment is defined as $f_{\max}$, and the minimum as $f_{\min}$, then the carrier frequency f_c is simply:

$$f_c = \frac{f_{\max} + f_{\min}}{2} \quad \text{... (4.1)}$$

It should be specifically emphasised that the carrier frequency of a specific segment is based on the maximum and minimum frequencies in that segment, and so every segment will have a different carrier frequency.

With the carrier frequency determined, the maximum frequency deviation D is then given by:

$$D = f_{\max} - f_c = f_c - f_{\min} = \frac{f_{\max} - f_{\min}}{2} \quad \text{........................... (4.2)}$$

The maximum percentage frequency deviation, as a percentage of the carrier, is hence given by:

$$D_{\%c} = \frac{100D}{f_c} = \frac{100(f_{\max} - f_c)}{f_c} = \frac{100(f_c - f_{\min})}{f_c} = \frac{100(f_{\max} - f_{\min})}{2f_c} \quad \text{.......... (4.3)}$$

4.4.2.2 Determining maximum modulating frequency value

Determining the maximum percentage modulating frequency $W_{\%c}$ for a segment is more complicated that finding the other two parameters. In communications theory, for general periodic signals, the maximum modulating frequency W is defined as the baseband bandwidth of the message signal (where the speed plot is the message signal in this application). As such, the maximum modulating frequency can normally be found by taking the FFT of the message signal, and locating the highest frequency in the spectrum of the message signal.

However, the typical basic speed profile in a segment which will be encountered in segmentation based PDOT order-tracking is not an oscillation. Typically the speed plot of a segment will approximately consist of either a ramp up, ramp down, or a positive or negative half-sine wave, as shown in Figure 4.3.

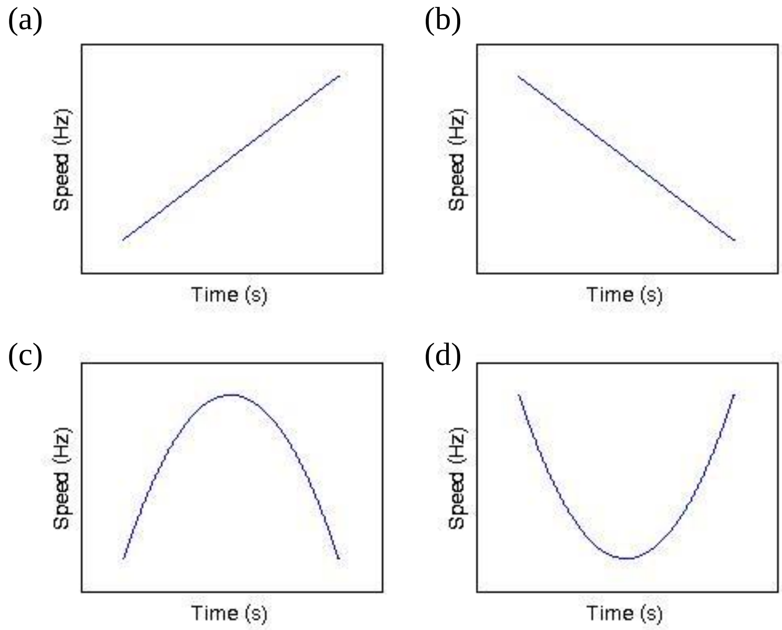


Figure 4.3 – Characteristic shape of typical speed plot segments
(a) Ramp up (b) Ramp down
(c) Positive half-sine-wave (d) Negative half-sine wave

As these basic speed profiles for a segment are all non-periodic, so not oscillating, the FFT of the speed plot for these basic speed profiles will be dominated by wrap-around errors where the FFT algorithm has enforced periodicity. As a result, the FFT is not suitable to determine contributing frequencies from the non-oscillating component of a signal, in order to evaluate the maximum modulating frequency W (and hence $W_{\%c}$). As an example, this is illustrated in Figure 4.4 for the basic speed profile cases of a ramp up, and a positive half-sine wave.

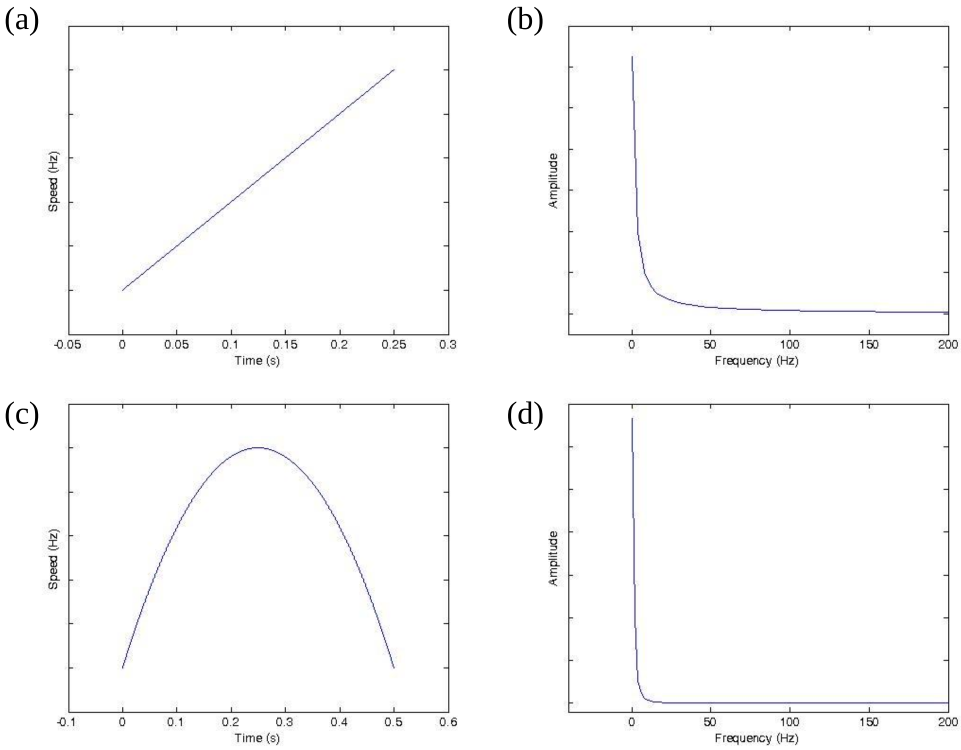
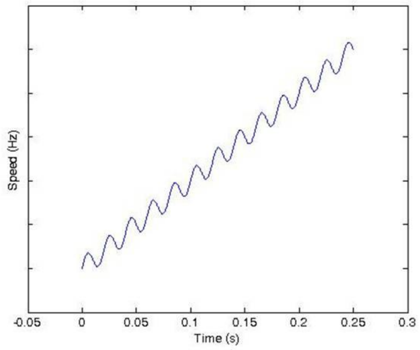


Figure 4.4 – Typical speed plot and corresponding spectrum
(a, b) Ramp up (c, d) Positive half sine wave
(a, c) Speed plot (b, d) Corresponding spectrum (low frequency)

Figure 4.4 (a) shows the speed plot for the basic speed profile of a ramp up. Figure 4.4 (b) shows a zoom of the low frequencies of the spectrum of the ramp up segment. As can be seen in (b), the spectrum only displays the effects of the wrap-around error which manifests as a steep decay in the low end of the spectrum. As the spectrum contains no meaningful frequency information for the non-oscillating ramp up segment, it is not possible to determine the maximum modulating frequency from the spectrum in this instance. Similarly, the spectrum for the positive half-sine wave segment in (d) also only displays the effects from the wrap-around error, and so contains no meaningful frequency information. It should be noted that the decay is more rapid in the spectrum for a half-sine wave, and so the effects of wrap-around errors are present in a smaller frequency range. The spectra from the ramp down and negative half-sine wave segments have the same characteristics as the spectra from the ramp up and positive half-sine wave, respectively, and so have not been illustrated here.

If a speed plot segment consists of the combination of one of the typical speed plot shapes, and oscillating variations around the basic speed plot shape, then it is often possible to determine the maximum modulating frequency from the spectrum of the message signal (speed plot segment) in the traditional fashion. This is because oscillating components must have a higher frequency than the non-oscillating component from the basic speed plot shape. As it is only necessary to find the maximum modulating frequency from the spectrum, this is given by extent of the spectral information of the higher frequency oscillating component, and it doesn't matter that the spectral information from the non-oscillating basic speed plot shape component is not correct, provided that it does not mask the oscillating component.

(a)



(b)

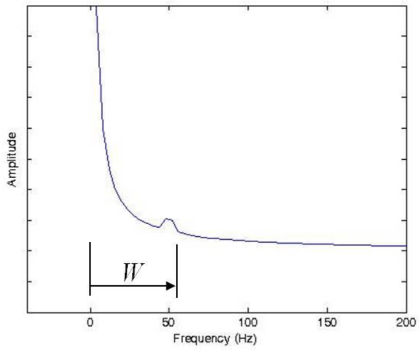


Figure 4.5 – Ramp up and oscillation; signal and corresponding spectrum
(a) Ramp up and oscillating variation (b) Spectrum

As an example, Figure 4.5 (a) shows a speed plot segment which consists of the combination of a ramp up basic shape and a 50 Hz oscillating component. Figure 4.5 (b) shows the corresponding spectrum, where the spectral information from the variation component is visible at 50 Hz, in addition to the decay from the underlying ramp component. As seen in (b), the value for the maximum modulating frequency W can be taken from the spectrum of the message signal in the traditional manner (baseband bandwidth, after neglecting the wrap-around error component), in this example.

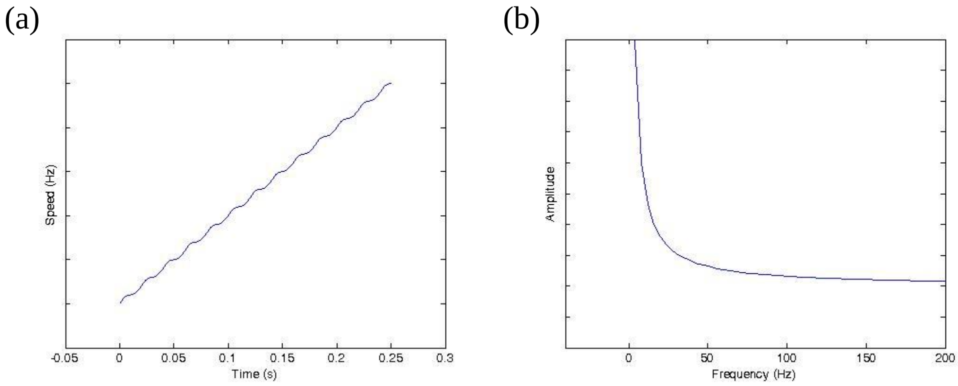


Figure 4.6 – Ramp up and small oscillation; signal and corresponding spectrum
(a) Ramp up and small oscillating variation (b) Spectrum

Figure 4.6 shows an example where the additional oscillating component is masked in the spectrum (b), and in this case the maximum modulating frequency cannot be determined from the spectrum. Figure 4.6 (a) again consists of the combination of a ramp up basic shape and a 50 Hz oscillating component, however the 50 Hz oscillating component is five times smaller than that shown in Figure 4.5 (a). Figure 4.6 (b) shows the corresponding spectrum, and it can be seen that the oscillating component at 50 Hz is completely masked by the wrap-around error from the underlying ramp-up component.

For cases which consist of the combination of one of the typical speed plot shapes, and oscillating variations around the basic speed plot shape which is unmasked, so a value of W can be found from the spectrum, the maximum modulating frequency is then given simply as:

$$W_{\%c} = \frac{100W}{f_c} \qquad (4.4)$$

As stated, for cases which consist solely of one of the basic speed plot shapes, such as shown in Figure 4.4 on page 197, or which contain additional oscillations which are masked, such as shown in Figure 4.6, it is not possible to determine the maximum modulating frequency from the spectrum in the traditional fashion. In these instances, it is necessary to estimate the maximum modulating frequency based on the speed plot directly in the temporal domain.

4.4.2.2.1 Finding equivalent sinusoid for maximum modulating frequency

Estimating the maximum modulating frequency directly from the speed plot can be accomplished by finding the equivalent constant frequency sinusoid, with frequency f_{meq}, corresponding to the speed plot segment. The maximum modulating frequency W is then found as $W = f_{meq}$.

The equivalent sinusoid has amplitude f_{deq}, and so takes the form:

$$y(t) = f_{deq} \sin(2\pi f_{meq} t) \quad \text{...} \ (4.5)$$

The equivalent sinusoidal values are found by fitting a section of a constant frequency sinusoid to the speed plot segment, so that the maximum and minimum frequencies from the speed plot segment span one quarter of a period of the equivalent constant frequency sinusoid. This is illustrated in Figure 4.7 for the different typical speed plot segments.

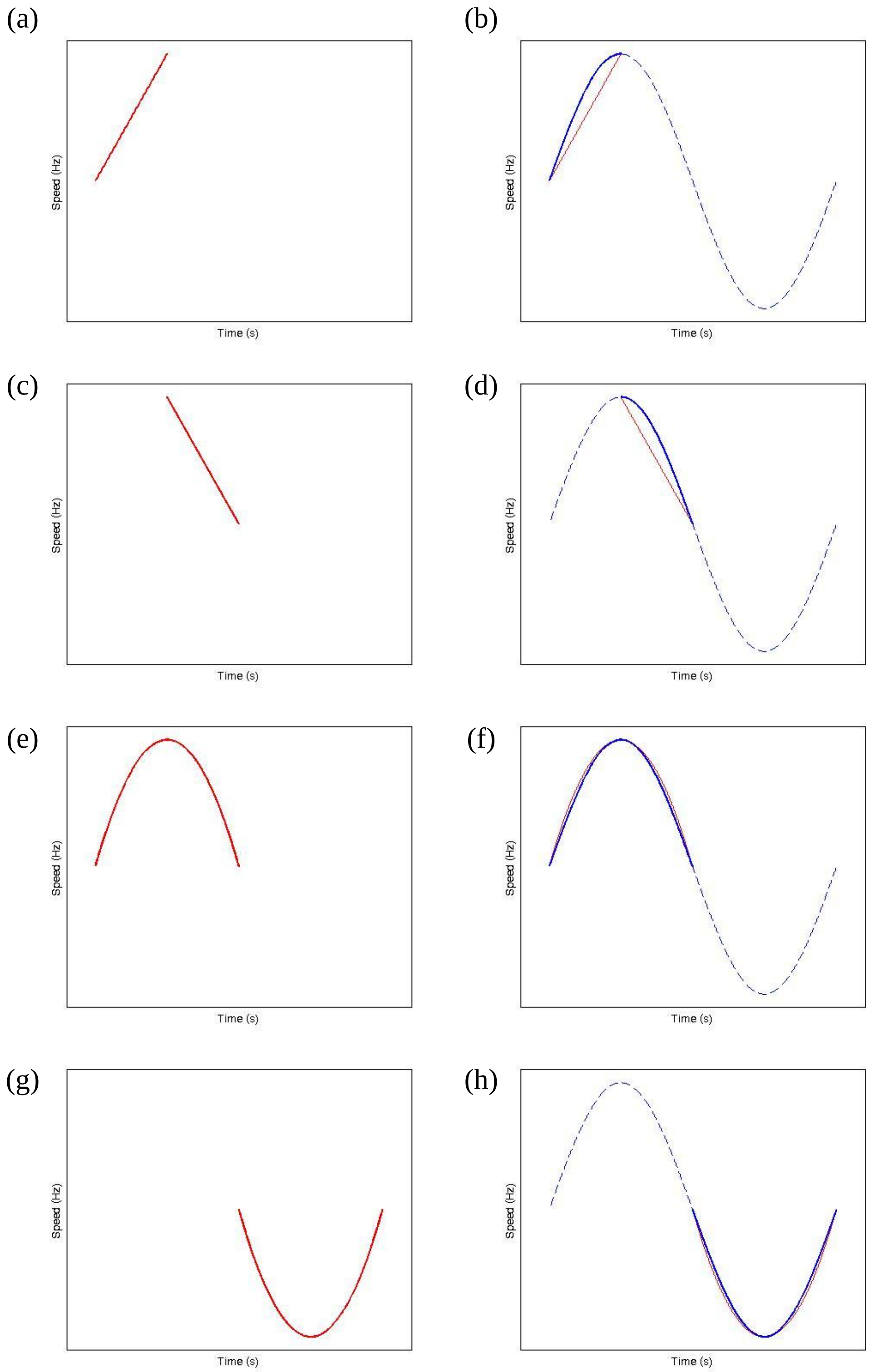


Figure 4.7 – Speed plot segments fitted to equivalent constant frequency sinusoid section

(a, c, e, g) Speed plot segments (b, d, f, h) Equivalent sinusoid section

(a, b) Ramp up (c, d) Ramp down

(e, f) positive half sinusoid (g, h) negative half sinusoid

Figure 4.7 shows the typical speed plot segments on the left, and the equivalent sinusoidal section fitted to the typical speed plot segment on the right. The typical speed plot segment is shown as a solid red line, and the equivalent sinusoidal section is shown as a solid blue line. As can be seen, in each case the maximum and minimum frequency from the speed plot spans one quarter of the equivalent sinusoid. This has the result that the ramp up and ramp down segments are equivalent to one quarter of a period of a constant frequency sinusoid, and the positive and negative half sinusoid speed plot segments are equivalent to half of a period of a constant frequency sinusoid.

For all cases, the equivalent sinusoid is defined as having an amplitude of f_{deq}, and frequency f_{meq} resulting in a period of $1/f_{meq}$, and quarter-period of $1/4f_{meq}$. These parameters are illustrated in Figure 4.8.

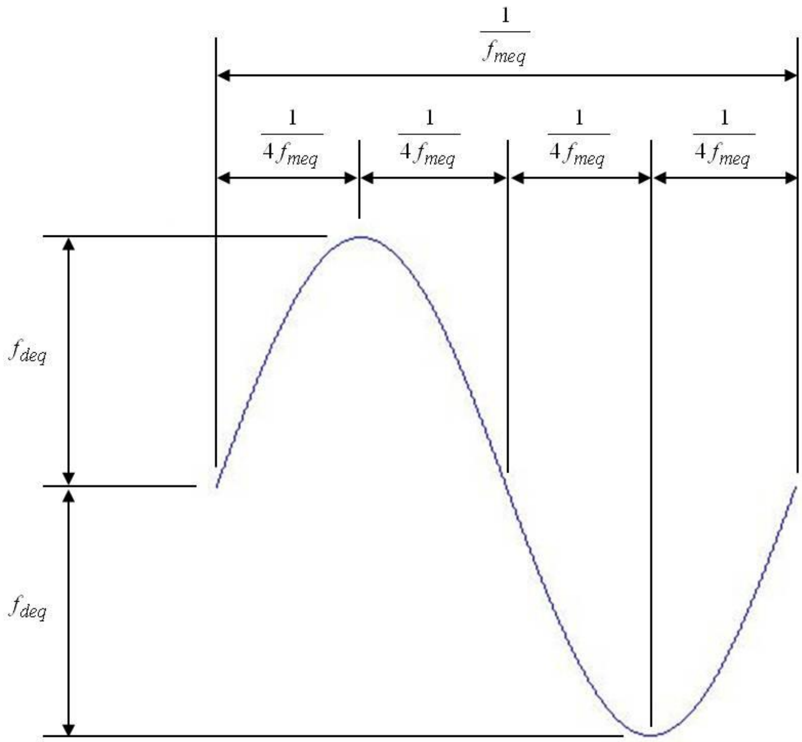


Figure 4.8 – Parameters of interest from equivalent sinusoid

For the speed plot segments, each segment has a maximum frequency $f_{\max}$ and minimum frequency $f_{\min}$, as previously defined, resulting in a frequency span of $f_{\max} - f_{\min}$. Using equation (4.2) on page 194, this frequency span is equal to $f_{\max} - f_{\min} = 2D$. The time span corresponding to this frequency span between the maximum and minimum values is defined as t_m. These parameters are shown illustratively in Figure 4.9.

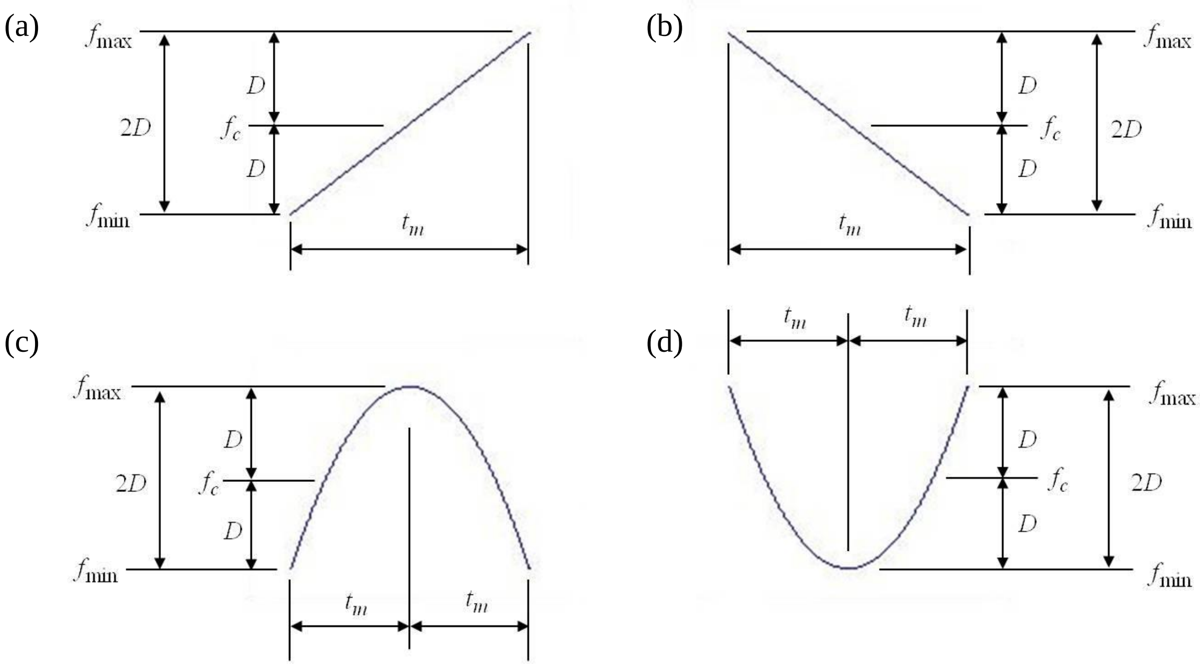


Figure 4.9 – Parameters of interest on typical speed plot segments
(a) Ramp up (b) Ramp down
(c) Positive half-sine-wave (d) Negative half-sine wave

By equating the parameters for the speed plot segments and the corresponding equivalent sinusoidal parameters together, it is possible to gain a value for the equivalent modulating frequency f_{meq}, and subsequently the approximate maximum modulating frequency W. This can be accomplished by using one of two proposed approaches.

The first approach is to match the average slope of the speed plot segment and equivalent sinusoid together, and the second is to match the maximum slopes together.

4.4.2.2.2 Matching average slope to find equivalent modulating frequency

For the first approach, the average slope of the speed plot between maximum and minimum frequency values is equated to the average slope between the centre and maximum frequency of the equivalent quarter-period sinusoidal segment. This results in the average slope from the basic speed plot being on the left, and the average slope of the quarter-period sinusoidal segment being on the right of the following equation:

$$\frac{f_{\max} - f_{\min}}{t_m} = \frac{f_{deq}}{\frac{1}{4 f_{meq}}} \qquad (4.6)$$

From this equation, it is possible to equate the top and bottom of each side together, which effectively results in the ‘rise’ of both segments being equal, and the ‘run’ of both segments being equal, resulting in:

$$f_{\max} - f_{\min} = 2D = f_{deq} \qquad (4.7)$$

and

$$t_m = \frac{1}{4 f_{meq}} \qquad (4.8)$$

Equation (4.8) can be solved for f_{meq}, and consequently W, as:

$$W = f_{meq} = \frac{1}{4 t_m} \qquad (4.9)$$

This method of equating the average slopes of the speed plot and equivalent sinusoid is most suitable for use where the speed profile almost exactly matches one of the typical speed profiles, and no significant deviation in slope is present, as only in this case will the determined equivalent constant modulating frequency, based on the 'average', correspond to the equivalent maximum modulating frequency.

With a value found for the maximum modulating frequency W, the maximum percentage modulating frequency $W_{\%c}$ can then simply be found using Equation (4.4) on page 199.

4.4.2.2.3 Matching maximum slope to find equivalent modulating frequency

For the second alternative approach, the maximum slope of both the speed signal and the equivalent sinusoid are equated together, which can then be solved to find f_{meq}.

This approach is better suited for cases where the speed profile differs away from the corresponding typical speed profile while still maintaining the same characteristic shape (and so not oscillating), so that the local slope of the speed plot segment is greater (in parts) than the corresponding local slope on typical speed plot segment.

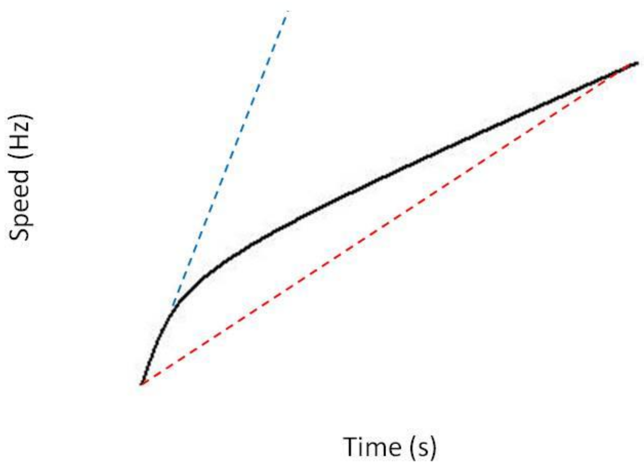


Figure 4.10 – Example of speed plot which differs from one of the basic speed profiles
Black: Speed plot Red: 'Average' value Blue: slope of steepest part of speed plot

Figure 4.10 shows an example speed plot for this situation, where the speed plot (in black) differs from the basic run-up profile. In this case the 'average' between maximum and minimum frequency values (in red), which would have been matched with the previous approach, is a poor representation of the maximum frequency of the speed plot, which dominates at the steeper initial segment of the speed plot (slope of this section shown in blue).

The approach of matching the maximum slope is also more suitable for cases where the speed plot contains an additional oscillating component, where the oscillating component is masked, such as the example shown in Figure 4.6 on page 199. This approach allows the determination of f_{meq} based on the maximum slope of the speed plot segment, which should correspond to the equivalent maximum modulating frequency for the cases which deviate away from the basic speed profiles, rather than the ‘average’ modulating frequency which would be gained with the previous approach presented in section 4.4.2.2.2.

To accomplish this method of matching the maximum slope, firstly the (absolute) maximum slope $S_{\max}$ of the speed plot segment is found, by finding the maximum value of the derivative of the speed plot (derivative of instantaneous speed), using any appropriate method.

A very basic method to calculate the maximum of the derivative of the speed plot would be to use of a first order difference equation, by first taking the (absolute value of the) difference $|\triangle f|$ between each sample in the speed plot segment, for all samples in the segment, and then finding the maximum of these (absolute) difference values $\triangle f^{\max}$. The local (absolute maximum) slope $S_{\max}$ of the speed plot at this location is then the maximum difference value divided by the sampling period T_s, where the sampling period is the inverse of sampling frequency, so $T_s = 1/f_s$. The absolute maximum slope $S_{\max}$ is hence given by:

$$S_{\max} = \frac{\triangle f^{\max}}{T_s} = \frac{\triangle f^{\max}}{\frac{1}{f_s}} = \triangle f^{\max} f_s \quad \text{.............................. (4.10)}$$

As previously stated, usage of a first order difference equation to find the maximum slope $S_{\max}$ as per Equation (4.10), is just one alternative to calculate the maximum of the derivate of instantaneous speed, and any appropriate method could be used to find $S_{\max}$.

Then, starting with the equation of the equivalent constant frequency sinusoid given in equation (4.5) on page 201, the slope S^{eq} of the equivalent sinusoid is found by taking the derivative, so:

$$S^{eq} = \frac{dy}{dt} = 2\pi f_{deq} f_{meq} \cos(2\pi f_{meq} t) \quad \text{.............................. (4.11)}$$

The maximum slope $S^{eq}_{\max}$ is present when the cosine term is at the maximum of 1 at $t = 0$, so:

$$S^{eq}_{\max} = 2\pi f_{deq} f_{meq} \quad \text{.. (4.12)}$$

The maximum slope of the speed plot $S_{\max}$ and the equivalent sinusoid $S^{eq}_{\max}$ can then be equated together, and the result solved for f_{meq} (and consequently W), giving:

$$W = f_{meq} = \frac{S_{\max}}{2\pi f_{deq}} \quad \text{.. (4.13)}$$

By using equation (4.7) on page 206, the frequency deviation parameter f_{deq}, which is the amplitude of the equivalent constant sinusoid in equation (4.5), can be replaced by $2D$ or $f_{\max} - f_{\min}$, which corresponds to the total frequency span of the speed plot being matched to a quarter-period of the equivalent sinusoid. As a result, Equation (4.13) becomes:

$$W = f_{meq} = \frac{S_{\max}}{4\pi D} = \frac{S_{\max}}{2\pi(f_{\max} - f_{\min})} \quad \text{.............................. (4.14)}$$

Again, with a value found for the maximum modulating frequency W, the maximum percentage modulating frequency $W_{\%c}$ can then simply be found using Equation (4.4) on page 199.

It should be specifically noted that utilising the approach of matching the maximum slope will in effect give a more conservative value for W and $W_{\%c}$, than the values gained by using the first approach of matching the average slope as presented in the previous section. As such, it is appropriate to consider results gained with the second approach as having a greater factor of safety than those gained with the first approach, which will be discussed in the following sections.

4.4.3 Determine segment sizes and locations

With values for the speed variation parameters $D_{\%c}$, $W_{\%c}$ and f_c, it is possible to check if the speed variations in a given segment are permissible for use with the basic PDOT method, using the methods described in Chapter 3, starting on page 127.

Acceptability charts when using the fundamental order of a reference component, for a given combination of speed variation parameters, were presented in section 3.6 on page 161. Additional acceptability charts for higher orders are presented in Appendix A on page 423. The acceptability chart corresponding to the harmonic order of the reference component, to be used for order-tracking, should be used to calculate permissible parameters. Three different acceptability, or permissibility, limits are presented, based on different criteria. The limits using these criteria are shown in Figure 3.7 on page 165 for the fundamental, or first, order of a reference component, and the limits are shown for higher orders in the additional acceptability charts in Appendix A on page 423. In summary, combinations of parameters lying below the 40 dB Sideband definition criterion line in Figure 3.7 will result in the best order-tracking result. The 20 dB Sideband definition criterion line and the Carson's rule criterion line give inferior order-tracking results, while allowing greater speed ranges per segment and hence a smaller number of required segments.

On the subject of which criterion is suitable for use with the segmented method, sometimes it is preferable to make use of the limit given by the 40dB criterion with better accuracy, however since many applications where order tracking of a whole record is desired, with large speed variations, would not in general have very high requirements on accuracy, one of the weaker criteria could be used. As an example, it would not be meaningful to carry out synchronous averaging for a widely varying signal, because of the very great changes in amplitude along the record.

It should be noted that as the limits to speed variation are based on a speed ratio, then segment sizes will not remain constant. For the same percentage allowable speed ratio, smaller absolute speed ranges will apply at lower speeds (for a constant sweep rate), and segment sizes will be correspondingly smaller.

It should also be noted that, in principle, it would require an infinitely small segment size to be able to go all the way down to zero frequency, or the stopped state for a machine. As such it is necessary to set a lower frequency cut-off f_{cut} (or $f_{cut\%}$ as a percentage of the maximum speed of the signal), below which order-tracking will not be conducted, for signals which approach zero frequency. However, since the vibration levels at very low speed are often negligible, the loss of information at very low frequency is usually of little consequence.

So, once an appropriate acceptability criterion has been selected to evaluate the permissibility of a combination of speed parameters, and a low frequency cut-off level has been chosen (if required), it is possible to determine the segment sizes and locations. This is conducted in an iterative testing process, and different sized segments from the overall speed profile can be tested, until permissible segments are found which span the entire speed profile (excluding sections below the low frequency cut-off).

One complication which can arise is if the regions of the signal below the low frequency cut-off f_{cut} are present in the middle of a signal, rather than at either end as with a run-up or run-down speed profile. In this instance it is necessary to treat the total signal as two (or more) independent signals, which are separately order-tracked using a segmented approach as described here, and unfortunately the independent signals cannot be recombined back to the entire original signal.

It should be noted that, as with the basic PDOT method, final permissibility of a given combination of speed variation parameters (and so a given segment) should be based on examination of the spectrum (of the actual signal, not the corresponding speed plot) for visual separation based on an appropriate dynamic range, as discussed in Chapter 3.

Finding segments to span an entire signal is quite laborious, as each change in segment size will inevitably result in a different frequency span, which will result in a different carrier frequency, which in turn will change the modulation parameters which are a percentage of the changed carrier frequency. This complicates the testing of multiple segment sizes for permissibility.

In principle, in order to minimise the errors introduced by the segmenting process, it is desirable that a signal should be segmented into the smallest number of segments as possible. In order to ensure the minimum number of segments are used, each segment should span as wide a speed range as possible. Finding optimum segment sizes which are not only permissible, but also close to the maximum permissible modulation values, is even more laborious than solely finding permissible segment sizes.

However, the errors introduced by the number of segment joins are typically orders of magnitude less than the errors introduced by aliasing due to overlap with higher reference signal harmonic orders which occurs when larger speed variation parameters are present in a single segment (as observed in results collected to date). As such, it is typically counter-productive to optimise segment size to have the largest possible speed range and size, and equal or more accurate results can be gained with using a smaller speed range per segment, and hence a greater number of smaller sized segments. While this should not be taken to extremes, it is possible to simplify calculations to determine segment sizes.

It is possible to simplify the process of determining permissible segment sizes and then locations, by having smaller speed variations within each segment, and hence a greater number of total segments. This concept is best considered as using a greater factor of safety in calculating permissible speed variation parameters for a segment.

4.4.4 Simplified segment size calculations, using a larger factor of safety

In order to simplify the segment size calculations, the goal is to use a greater factor of safety in calculating permissible speed variation parameters for a segment. The goal is to introduce a combined factor of safety when evaluating the speed variation parameters.

As the selection of different methods in different stages of the calculations results, in effect, in different factors of safety for each stage, then the overall factor of safety will be the combination of the factors of safety from the different steps. As a result, it should be kept in mind that if an overly conservative method is selected for one step, then other steps will not require as large a factor of safety to give a combined factor of safety. If overly conservative methods are used at every stage, then the combined factor of safety can be excessive, and hence counter-productive in improving calculations.

The first step in simplifying the segment size calculations, is to determine a common conservative value for $W_{\%c}$, which is appropriate to be used in calculations for all segments, rather than finding the maximum modulating frequency separately for each segment. In effect, this is finding a value of $W_{\%c}$ with a greater factor of safety. It is then possible to find a corresponding maximum permissible value of $D_{\%c}$ using the acceptability criteria. Conservative segments can then be more simply found, all with a common percentage frequency deviation value.

To accomplish this, firstly the maximum (non-percentage) modulating frequency value for W (= f_{meq}) can be determined in an identical fashion as described above for each segment, except using the entire valid speed range. This can be accomplished using any of the described methods as appropriate, which were either finding the baseband bandwidth from the spectrum of the message signal (if possible), as described in section 4.4.2.2 on page 195, or by finding the equivalent modulating frequency by matching average slope, as described in section 4.4.2.2.2 on page 206, or by matching the maximum slope, as described in section 4.4.2.2.3 on page 208. Of these methods, finding the baseband bandwidth from the spectrum should give the correct result (factor of safety = 1), using the average slope approach should give the correct or an optimistic result (factor of safety ≤ 1), and using the maximum slope approach should give correct or conservative results (factor of safety ≥ 1).

With the maximum absolute value for W found, a value can then be calculated for the maximum percentage modulating frequency to be used. The most conservative value (highest factor of safety) will be gained when the carrier frequency is the lowest possible. So with taking carrier frequency f_c to be equal to the lowest (valid) frequency in the entire signal $f_{\min}$, so that:

$$f_c = f_{\min} \quad \text{.. (4.15)}$$

then the most conservative modulating frequency value $W_{\%c}$ possible for the signal is found by substituting equation (4.15) into equation (4.4), on page 199, giving:

$$W_{\%c} = \frac{100W}{f_{\min}} = \frac{100\, f_{meq}}{f_{\min}} \quad \text{...................................... (4.16)}$$

A less conservative value would be gained by using the carrier frequency as the midpoint between the maximum and minimum frequencies from the entire (valid) speed range, so matching the definition used in earlier calculations. $W_{\%c}$ would then be given by equation (4.4). This value for $W_{\%c}$ has a lower factor of safety, and would typically be conservative for the upper half of the speed range, and optimistic for the lower half.

With a common value for $W_{\%c}$, the maximum permissible value for $D_{\%c}$ can be found, using the methods described in section 3.5 on page 158. This is conducted using one of three acceptability criteria, the usage of each would correspond to a different factor of safety. In ranking the three criteria, the 40dB Bessel function based criterion gives the most conservative results (largest factor of safety), followed by the 20dB criterion, and then the Carson's rule criterion which gives the least conservative results (lowest factor of safety). Note that a common value for $D_{\%c}$ could be chosen to be less than the maximum permissible value indicated by the acceptability criteria chosen, which would in effect result in an additional factor of safety >1.

With the selection of a common value for $D_{\%c}$, this can be used to determine the size of every segment to span the total frequency range.

With the value for $D_{\%c}$, it can be identified that the maximum and minimum frequencies of a segment, respectively $f_{\max}$ and $f_{\min}$ are both related to the carrier (centre) frequency, so that:

$$f_{\max} = \frac{(100 + D_{\%c})}{100} f_c \quad \text{.......................................} \quad (4.17)$$

and

$$f_{\min} = \frac{(100 - D_{\%c})}{100} f_c \quad \text{...} \quad (4.18)$$

Equation (4.18) can be rearranged for f_c,

$$f_c = \frac{100 f_{\min}}{(100 - D_{\%c})} \quad \text{...} \quad (4.19)$$

which can be substituted into equation (4.17), giving:

$$f_{\max} = \frac{(100 + D_{\%c})}{\cancel{100}} \times \frac{\cancel{100} f_{\min}}{(100 - D_{\%c})} = \frac{(100 + D_{\%c}) f_{\min}}{(100 - D_{\%c})} \quad \text{................} \quad (4.20)$$

Equation (4.20) can alternatively be rearranged to solve for $f_{\min}$, giving:

$$f_{\min} = \frac{(100 - D_{\%c}) f_{\max}}{(100 + D_{\%c})} \quad \text{.......................................} \quad (4.21)$$

Segments which fully span a signal can then be found, by making use of equations (4.20) and (4.21).

Starting with either the highest frequency or lowest frequency in the entire signal, which will be respectively $f_{\max}$ or $f_{\min}$, the alternate maximum or minimum frequency limit for the first segment can be found using either equation (4.20) or (4.21) respectively. The (continuous) section of signal bounded by these two values $f_{\max}$ and $f_{\min}$ forms the first segment. Working out from the boundaries of this segment, the edge(s) of the first segment become either a maximum or minimum for the next segment along from the first. The span for the neighbouring segments can again be determined using either equation (4.20) or (4.21) as appropriate. This process should be continued until segments are found which fully span the entire signal.

For cases which require a low frequency cut-off, the process should be continued until a $f_{\min}$ value is found which is less than the selected low frequency cut-off value f_{cut}. For this final segment which spans the low frequency cut-off value, two different approaches can be employed, depending on how the cut-off frequency should be treated.

For the first approach, the low frequency cut-off f_{cut} can be treated as a hard cut-off, and the final segment is truncated at the low frequency cut-off, rather than spanning to $f_{\min}$. As the final segment is truncated, care must be taken in ensuring that the segment still contains a complete overlapped section of signal to mate with the neighbouring segment. If the truncation has resulted in an incomplete overlapped section with the adjacent segment (segment size less than the minimum size of a suitable overlap), the segment should be discarded, and the previous segment treated as the final segment.

For the second approach, the low frequency cut-off f_{cut} can be treated as a guide, and a 'complete' segment is retained for the final segment, spanning between $f_{\max}$ and $f_{\min}$, rather than truncating the segment.

Ultimately, the choice as to whether the low frequency cut-off f_{cut} (or $f_{cut\%}$) is used as a hard limit, or as a guide, is application specific, and depends on why the lower-frequency cut-off was chosen. It is suggested, by this author, that unless a hard cut-off is specifically required, that the cut-off be used as a guide and a 'complete' final segment be retained.

Note that care must be taken when calculating segment ranges, if the speed changes direction during a segment, as it would for the half-sinusoid basic speed plot types. As the speed changes direction, the inflection point may become either the maximum $f_{\max}$ or minimum $f_{\min}$ value for the segment in which the speed direction changes, rather than the limit remaining at the segment boundary, which changes the segment calculation.

This described method for determining segment locations is significantly less complicated than the iterative testing process described in section 4.4.3 on page 212, and is more appropriate for use in an automated fashion.

4.4.5 Overlapping segments and windowing

While non-overlapped segments could be treated, this would inevitably introduce discontinuities in reconstructing the total order-tracked signal from each segment. This is primarily because of wrap-around errors from the FFT analysis process, where each segment is treated as joined into a loop, repeated periodically, but is also affected by the end-effects associated with the cubic-spline interpolation used with the PDOT method.

In order to minimise the errors resulting from end effects, signal overlapping is used.

However, to maintain a uniform weighting of the recombined signal after treatment, the weighting of each signal must decrease to zero towards the end of each segment (where the wraparound errors are greatest) at the same time as the weighting on the same part of the signal in the other segment increases correspondingly. This is illustrated in Figure 4.11. An appropriate weighting or window function is required to taper the overlapped sections.

As it is necessary for the combined values of the windows to always add to unity in the overlapped sections, so as to ensure no distortion of the signal, complementary half-Hanning windows are the suggested windowing function to use over the overlapped sections as illustrated in Figure 4.11. This is because the sum of half-Hanning windows in opposite directions is always unity when they are the same length and overlap at the 50% point. The half-Hanning windows will still add to unity, even after resampling to the order domain provided both are resampled in an identical fashion, which is a requirement in any case since the signal section is the same for the overlapping sections of each segment, and the resampling is always to the same number of samples per rotation, even if the original time sampling rate is different. Thus, each overlapped section should be windowed with a half-Hanning window, with a unity weighting over the non-overlapping sections of each segment.

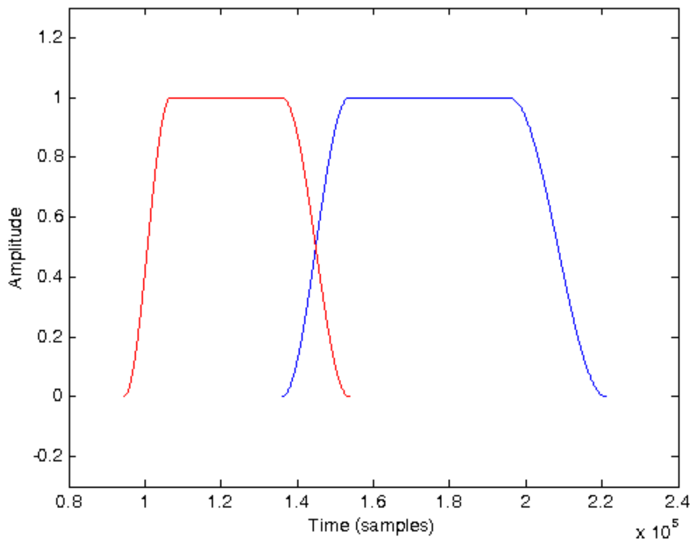


Figure 4.11 – Windowing functions for two adjacent segments

Figure 4.11 shows an example of such windowing functions for two adjacent overlapping segments.

It is desirable to keep the overlapping of segments to the minimum needed to suppress errors due to the wraparound errors from FFT processing, and cubic spline end effects, since this will minimise the total number of segments required to adequately span the original signal.

The optimum amount of overlapping needed is application specific, however as an example the case presented in section 5.5 on page 284 gave good results with overlapped sections of approximately 10% of the segment length at each end, so roughly 20% of the segment length consisted of overlapped sections. It should be specifically emphasised that the overlapped section of two adjacent segments must be identical size in time. Thus, it is recommended to define the overlapped region as a percentage of the length of the smaller of the two adjacent segments. This will typically result in a segment containing a differently sized overlap region at each end, rather than equally sized overlap regions.

It should be noted that the overlapped sections should be included in the maximum permissible speed span calculation for the segment, which was not included in the methods presented in the previous sections.

For the simplified method of calculating the segment sizes presented in section 4.4.4 on page 216, where the worst case $W_{\%c}$ and corresponding constant $D_{\%c}$ were found, and segment sizes were based on this, there are two different alternative modifications which could each include the effects of overlap when determining segment size.

The first modification is that a factor of safety can be subtracted from $D_{\%c}$ to cater for the overlap percentage. The calculations presented to find the segment boundaries will alternatively locate the midpoint of the overlapped sections (breakpoint), and overlapping will span around the calculated position of the breakpoint.

The second modification is that once the boundary of a segment is located, the border of the next segment should be moved back into the previous segment by the overlap percentage. The next segment's span will begin from within the previous section, rather than at a common boundary as originally presented in the previous section 4.4.2 on page 193.

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 and cause incorrect order-tracking when the amplitude gets too small, which would be induced by windowing the reference signal.

Thus, for the segmentation method presented here, it is suggested that the reference signal segments to be used for the order-tracking of the corresponding data signal segments are not windowed.

If it is desired to obtain an order-tracked version of the reference signal itself for further analysis, regardless of whether it is an external tacho or response signal, it is suggested to make a copy of the reference signal and then treat this as an extra data signal to be order-tracked.

It should be noted that there is no necessity to window the start of the first segment, or the end of the last segment in a signal, as these signal sections will not be recombined with any others. It is preferable that these sections not be windowed during the order-tracking process, though it should be remembered that there will be errors near the limits of these non-windowed segments, as there would be at the signal ends with the basic PDOT method. Note these end effects are automatically suppressed with the windowed segments, which have almost zero amplitude at the ends. These end sections containing errors, at the limits of the non-windowed segments, should be discarded for optimum results (as with the basic PDOT method), but this should only be done once the entire order-tracking process is complete, and the final signal has been recombined. Ideally the discarding of parts of the signal at the ends, if implemented, should be allowed for in the original recordings. For example, in the case of a run-up, a section of signal at constant speed could be recorded at the start and end of the signal, which can be discarded after processing without loss of information.

If it is desirable for the final order-tracked signal to be windowed before subsequent analysis, as is common with analysis of signals with large speed variations, this windowing should also be conducted on the recombined signal after order-tracking is complete.

4.4.6 Low pass filtering

Order-tracking any signal introduces the complication of a time-varying sampling rate. While a data signal is typically low pass filtered to prevent aliasing when sampled in the time domain, such that no frequency components extend beyond the Nyquist frequency (half the sampling frequency), this will not necessarily hold true after resampling to the angular domain.

Typically, the order-tracking process introduces a time-varying sampling rate directly proportional to the local speed of the machine.

The basic PDOT method presented in Chapter 3 relies on initial oversampling by a factor of two before selecting a phase sampling rate which gives a resulting resampled signal equal in length to the original signal oversampled by a factor of two. With a maximum speed ratio of 2:1, the oversampling effectively gives head-room into which the variable sampling rate can spill, without ever giving aliasing of the resampled signal. As the segmenting method presented here is intended for overall speed variations greater than 2:1, just an initial oversampling factor will not be sufficient to prevent aliasing over the entire signal. The segment with the highest speed range will be okay with oversampling by a factor of two, however in lower speed segments the signal will have to be low-pass filtered before resampling to remove potential aliasing components above the highest order of interest.

It should be noted that a signal could alternatively be re-sampled at a significantly greater angular sampling rate, so the lowest speed in the entire signal would not be aliased after resampling. However, this would result in a dramatically larger order-tracked signal, which would be much more difficult to analyse, and so simply using a higher angular sampling rate alone would often not be practical for most signals, with current computing power.

So for the segmented approach to order-tracking, as with the basic PDOT method, every segment should be initially oversampled by a factor of 2.

The order based sampling rate can then be chosen, which will give the same record length for the resampled highest speed segment, as the record length of the original highest speed segment oversampled by a factor of two. This is the same method as described in Chapter 3 for the basic PDOT method.

However, as the full speed plot of the signal is available for the segmented approach, the minimum permissible phase sampling rate to avoid aliasing could also be alternatively calculated. This is done by firstly finding the minimum frequency $f_{\min}$ present in the segment containing the maximum frequency for the entire (complete) signal, i.e. the 'highest speed' segment. This minimum frequency multiplied by Nyquist frequency of the order based sampling rate must be equal or greater than the Nyquist frequency for the temporal sampling rate to prevent aliasing. So, for an order based sampling rate (samples per period of the reference frequency) $f_{s\phi}$, and temporal sampling rate f_s, the above inequality is:

$$f_{\min} \times \frac{f_{s\phi}}{2} \geq \frac{f_s}{2} \quad \text{..} \; (4.22)$$

This can be solved for the order based sampling rate, to give:

$$f_{s\phi} \geq \frac{f_s}{f_{\min}} \quad \text{..} \; (4.23)$$

Any integer order based sampling rate which satisfies inequality (4.23) could then be selected to use, however it is suggested than the minimum order based sampling rate be used, unless a higher rate is specifically desired, as greater sampling rates will result in longer final order-tracked signals. A greater sampling rate would, for example, help minimise visual aliasing, as discussed in 2.2, which could be desired.

Each segment after the highest should then be low-pass filtered before order-tracking. Ideally the low pass filter would be based on orders of shaft speed, but as stated, this is not possible to implement before order-tracking (as this would require variable frequency filtration). Instead a constant frequency limit is selected to act as a low pass filter. The filter limit should be based on orders of the lowest speed in a segment, as this is the critical component which will first become aliased.

The best way to do the low-pass filtration is in the frequency domain, by simply setting frequency components above the selected low pass filter frequency value to zero. This gives ideal filtration with no phase distortion (FFT processing is effectively non-causal, since the second half of each time record also represents negative time).

Ideally, the low pass filter level should be based on orders for all segments. The maximum order at the lowest speed in a segment should be less than half the order based sample rate used for the order-tracking process. So, for an order based sampling rate (samples per period of the reference frequency) $f_{s\phi}$, the low pass filter threshold (in orders) $f_{lpf\phi}$ must be:

$$f_{lpf\phi} < \frac{f_{s\phi}}{2} \qquad (4.24)$$

To calculate the time domain low pass filter value to be used in the time domain for each segment, the lowest speed value in a specific segment $f_{\min}$ should be identified. The low pass filter threshold f_{lpf} (in Hertz) for the segment is then given by:

$$f_{lpf} = f_{\min} f_{lpf\phi} \qquad (4.25)$$

As an example, Figure 4.12 shows the low pass filter thresholds for a ramp-up signal, where the highest speed segment on the right has the highest low pass filter threshold, and lower speed segments to the left have increasingly lower filter thresholds.

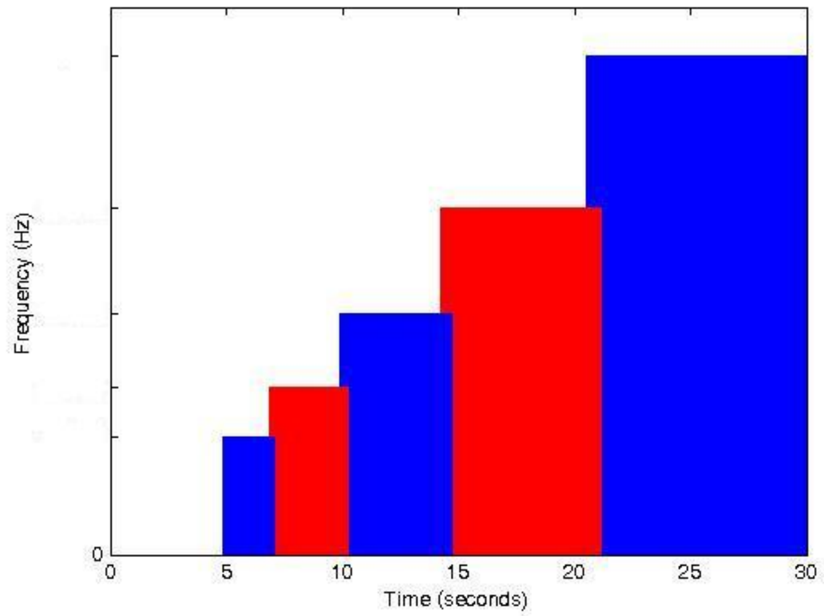


Figure 4.12 – Low-pass filter envelope for a ramp up signal

It should be noted, that as seen in Figure 4.12, the different low pass filter thresholds effectively overlap where the segments overlap. This has the potential to cause anomalous results in the overlapped section, for frequencies which are present for only one segment, but this will be smoothed by the windowing process, and will not affect results at and below the lower of the two filter thresholds.

It should also be noted that the resulting low pass filtered time signals will have a constant frequency limit. Once order-tracking has been conducted, this low pass filter limit will be variable in the order domain. For ramp signals, the low pass filter limit becomes a saw tooth pattern in the order domain of the recombined signal. This will result in only partial frequency information being available for some orders, so it may be desirable to low pass filter the final order-tracked signal in the order-domain with a constant order filter, to remove the partial order components.

To determine the highest complete order present in the final signal, which will give the minimum constant order filter threshold level suitable to remove all partial orders from the final order-tracked signal, it is first necessary to find the highest complete order value for each segment separately.

The highest whole (complete) order value $f_{w\phi}$ (in orders) can be found for each segment, with the time-domain low pass filter threshold value f_{lpf} used to filter the specific segment, and the maximum speed in the specific segment $f_{\max}$, by:

$$f_{w\phi} = \frac{f_{lpf}}{f_{\max}} \quad \text{...} \ (4.26)$$

Note that for the highest speed segment, which is not low pass filtered with the described filtering scheme above, the analogue low pass filter threshold from the original sampling of the data will give the value for f_{lpf}, for this specific segment. While theoretically, this would be the Nyquist frequency value $f_s / 2$, in practice analog low pass filters begin to affect a signal below this value. A rough conservative rule of thumb with analog low pass filters is to assume a filter threshold of $f_{lpf} = 0.4 f_s$, with frequency information above this threshold partially filtered, and this value is most likely suitable to be used as the f_{lpf} value with calculations for the highest speed segment to identify the highest complete order. However, some care should be taken to check this value is suitable for a given data acquisition system.

With a value of $f_{w\phi}$ calculated for every segment, the lowest value of $f_{w\phi}$ from all the segments represents the minimum constant order low pass filter threshold (in orders) suitable to remove all partial order information from the final recombined order-tracked signal.

4.4.7 Rejoining order-tracked segments

In order to rejoin the segments it is necessary to resample the segments using a common phase axis. To accomplish this, it is necessary to find a common point in adjacent segments, and set these positions to have common phase.

This common point could be anywhere within the overlap region between two segments. However, both segments will contain end effects at their extremes which should be avoided to prevent an incorrect phase value being used. The optimum common point to use is thus the midpoint of the overlap region.

In order to line up segments, during the order-tracking process of the first (left-most, lowest in time) segment, the phase value corresponding to the time sample at which the midpoint of the overlap occurs can be recorded. During the order-tracking of the next neighbouring segment, once the phase-time relationship has been constructed, the phase at the time sample corresponding to the break point in the new segment can be forced to have the previously recorded phase value from the previous segment, by adding or subtracting a constant value to the entire phase-time relationship. This then aligns the two segments. This also makes the implementation of a common phase axis easier, as the phase sample positions can be determined from this common point.

Once all segments are separately order-tracked, on a common phase axis so the segments are offset as they were before order-tracking, they can simply be added together to form the final order-tracked signal. Due to the use of half Hanning windows on each segment overlap, the overlapped sections will correctly sum to the order-tracked original signal. At this time, a final windowing function can be applied, if appropriate, to the order-tracked result before subsequent analysis, such as spectrum analysis, is conducted.

At this time, it would also be appropriate to employ a constant order based low pass filter to remove partial order information from the final order-tracked signals, if desired, as was discussed at the end of section 4.4.6 on page 227.

4.4.8 Modification to multi-stage method

The segmentation based order-tracking approach can be used with the other described methods of the generalised PDOT method presented in this chapter. This includes the multi-stage approach to order tracking described in section 4.3 on page 185.

It is certainly possible to refine the results of each individual segment with multiple stages of order-tracking, before finally recombining the segments back into an order-tracked result of the original signal. However as an alternative, it is very possible that the segments can be recombined after just the initial stage of order-tracking, and then further stages of order-tracking can be conducted on the recombined total signal.

As the bulk of the speed variations of the signal will be compensated for after one stage of order-tracking, even with very large speed variations present with the segmented method presented here, higher non-overlapped harmonics of the signal should be accessible for use for later stages even in the recombined signal. As such, there is likely no advantage in leaving the signal segmented after the first (or even an intermediate) stage of order-tracking. Better results should be produced by recombining the segments after the first (or an intermediate) stage, and then further stages of order-tracking conducted on the recombined whole.

Please note that no examples of the use of multi-stage order-tracking combined with the segmented approach are presented in this thesis.

4.5 Reversible order-tracking

The PDOT method can be used in a reversible fashion, so that an order-tracked signal can be resampled back to a constant temporal sampling rate. Using the PDOT method in a reversible fashion is suitable for applications where a signal contains both fixed frequency content (independent of running speed), and order based frequency content, and it is desired to analyse both in a complementary fashion.

As a typical application, a signal can be transformed to the order-domain using the basic PDOT method, then order based frequency content can be separated from the signal using an appropriate method, and then the remaining fixed frequency content can be resampled back to the original time domain, reversing the original order-tracking. This allows both types of frequency content to be analysed in isolation, in their respective domain, without contamination from the other frequency content type.

As summarised in section 1.2.3 on page 19, the basic concept of reversible order-tracking, and reversing transforms more generally, was first presented by Lembregts et al. in 1996 [15-18] (note all four papers are practically identical). However, while results were presented showing an order-tracking process being reversed, no information was provided as to what order-tracking method was utilised, and only a cursory methodology for a reversible process was given. The first non-cursory presentation of a reversible order-tracking process was given in 2005 by Groover et al. [14], who presented a reversible process for use with a time domain based computed order-tracking method. In a similar fashion to the work by Groover et al., the PDOT method can also be used in a reversible fashion. Note that the use of PDOT in a reversible fashion has not previously been published.

Reversing an order-tracked signal back to the time domain with the PDOT method can be simply accomplished, by keeping a record of the phase-time relationship used during the forward order-tracking process. The phase-time relationship can be interchangeably used to transform between phase and time, and so the order-tracked signal, which has constant phase sampling, can then simply be re-interpolated at the varying phase samples which correspond to the original time samples. This will produce a signal with the same constant (in time) sampling rate as the original signal.

The only restriction is that the order-tracked signal should not be extrapolated, and so any original time samples at the ends of the signal, which fall outside the span of the order-tracked signal, cannot be recovered. However, typically these out of bounds samples will be extremely limited in number, and the loss of the small amount of information will have a minimal impact on the analysis, provided it is recognised that the final signal transformed back to the time domain will have a slightly delayed starting time (offset), and end slightly earlier, when compared to the original signal.

The reversible approach to using PDOT can be employed when combined with any of the other methods presented in this Chapter 4, however some extra considerations are necessary for the reversible approach when combined with either a multi-stage, or segmented approach to PDOT.

When using a reversible approach, in combination with a multi-stage approach to PDOT as described in section 4.3 on page 185, it is not necessary to reverse back through every stage used for the forward order-tracking process by making use of each phase-time relationship. Rather, the phase values corresponding to the original time sampling can be progressively updated to the new phase sampling scheme for each stage during the forward order-tracking process, which is an extra step conducted in parallel with the order-tracking process, by using interpolation with the phase-time relationship found for each stage. Once order-tracking is complete at the nth stage, the signal can simply be resampled at the phase values for the nth stage which correspond to the original time sampling, in order to transform the order-tracked signal from the nth stage back to the original time sampling in one step. Conducting the reversal of the order-tracking in one stage will help to minimise the cumulative errors introduced with each stage of (forward or reverse) order-tracking. Please note that no examples of combining reversible and multi-stage approaches to PDOT are presented in this thesis.

Using a reversible approach, in combination with a segmented approach as described in section 4.4 on page 189, is more complicated. As a complete phase-time relationship is not formed for the whole signal with segmented order-tracking, it is not possible to reverse the order-tracking process in one step. As with the forward transform of order-tracking, it is necessary to segment the signal. Each segment can then be reversed back to the time domain, and then the segments recombined to give a combined signal in the time domain.

The complete (modified) order-tracked signal can be broken up into segments, in the order domain, in the identical locations where the (unmodified) order-tracked segments were originally recombined. This allows each segment to be reverse order-tracked using information retained during the forward order-tracking process, in the same manner as described above for a non-segmented implementation of reversible PDOT.

As with the forward order-tracking process, it is necessary to window the segments so they can be correctly recombined. While it is possible to apply a half-Hanning window in the order-domain to the segments before reversing the order-tracking, in a similar fashion as used in the forward order-tracking process as described in section 4.4.5 on page 222, it is easier to first resample the segments back to the time domain (with no windowing), and then apply half-Hanning windows in the time domain to each segment, before recombination of the windowed segments. This allows the identical half-Hanning windows to be applied, as were used for the forward order-tracking process, which removes the necessity of re-calculating suitable windowing functions. Note that the windowing functions should be truncated to account for missing samples which could not be extrapolated, however the minor truncation of the data, and hence the window function, will have little effect as the missing values at the segment ends would have been tapered to zero if still present. Once windowing is applied in the time domain, the segments can then be recombined to produce the final time domain signal, with the order-tracking reversed. Please note that no examples of combining reversible and segmented approaches to PDOT are presented in this thesis.

Chapter 5 Experimental testing of PDOT methods, with forward order-tracking

Chapter Overview

In this chapter, the results of using the PDOT method in various order-tracking applications are shown.

These examples include:

- Bearing and gear diagnostics, in the presence of small and large speed variations
- Employing multiple stages of PDOT to improve the results of order tracking, for gear diagnostics
- Using a segmented approach to PDOT, to successfully order-track a run-up signal with a very large speed variation

Please note that the presented examples all have low modulating frequency (f_m or W values). This is a result of the mechanical applications selected to illustrate the PDOT method, and the PDOT method itself can be used with higher modulating frequencies than those present in the examples presented in this chapter.

5.1 Multiple iterations with small speed variation and no amplitude modulation

Where speed variation is small, so that the amplitude modulation effect is negligible, a multi-stage PDOT method can be successfully employed using an extracted reference component from the data signal with minimal complications.

This situation was previously studied [7], and some results are shown here.

The aim of the study was to investigate alternative order-tracking methods to improve subsequent time synchronous averaging (TSA) results, when a reference signal cannot be directly coupled to the shaft of interest.

TSA is described in detail by McFadden [41]. The TSA method produces an averaged signal for one period of rotation of the component of interest, containing deterministic components only. By conducting the TSA for one period of rotation of the gears in the signal, the deterministic gear response can be separated from the original signal. TSA is extremely sensitive to speed fluctuations, and so accurate order-tracking is a necessity for successful employment of TSA.

The experimental system was a Larzac gas turbine engine at DSTO Laboratories, Melbourne, and the measurement data was supplied by DSTO. Note that detailed descriptions for this engine cannot be presented due to confidentiality requirements. A vibration signal was captured from an accelerometer mounted on the casing of the high pressure (HP) stage compressor. Signals were captured at three different loading values, of 75%, 85% and 95% of the maximum rated load of the engine. A sixty pulse-per-rev tachometer was mounted on an accessory gearbox shaft, which provided a reference signal in proportion to the speed of the HP shaft of interest. However, the ratio between the accessory and HP shaft was only known to 4 significant figures. This was only sufficiently accurate to allow partial removal of the first few harmonics of the HP shaft by synchronous averaging, leaving the higher order harmonics in place.

A number of methods were tested to improve the estimation of the gear ratio, including curve fitting the instantaneous frequency and progressive phase of the HP shaft speed and the tacho signal over various lengths of record, and using various harmonics of each. The criterion used to judge the accuracy of the estimate was termed the “Separation Index” (SI), defined as the ratio of the total mean square value of the separated deterministic components to the mean square value of the residual signal after subtracting them, given as:

$$SI = \frac{\dfrac{\sum_{i=0}^{n_y} y_i^2}{n_y}}{\dfrac{\sum_{i=0}^{n_x} x_i^2}{n_x}} \qquad (5.1)$$

where y is the frequency spectrum of the TSA signal, x is the frequency spectrum of the residual signal, and n_x and n_y is the number of samples in x and y, respectively.

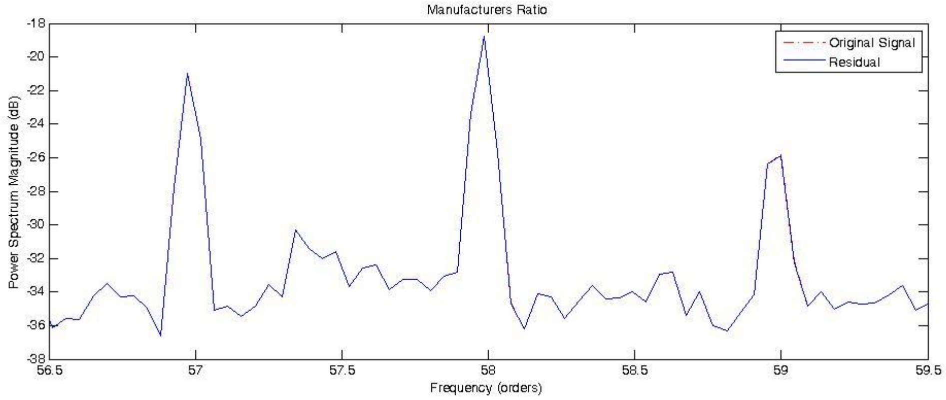
Assuming that the deterministic and residual components are independent, use of the correct ratio should maximise the numerator of this quotient, and simultaneously minimise the denominator. The actual value of the Index would of course be data dependent, and would vary for each measurement point and operating condition, but for a particular measurement should be maximised when the optimal separation of deterministic and random components is achieved.

One order-tracking approach investigated was to extract reference components of the HP shaft directly from the vibration signal, to conduct a multi-stage application of the PDOT method.

The multi-stage PDOT method was implemented using three stages, using the 1^{st}, 43^{rd} and 141^{st} HP shaft orders.

The success of the order-tracking process was evaluated by analysis of the subsequent TSA, where the residual signal should contain minimal HP shaft components, resulting in a high SI value. The success could also be judged qualitatively by visually comparing the original and residual spectra, to check for removal of gear components from the residual signal.

(a)



(b)

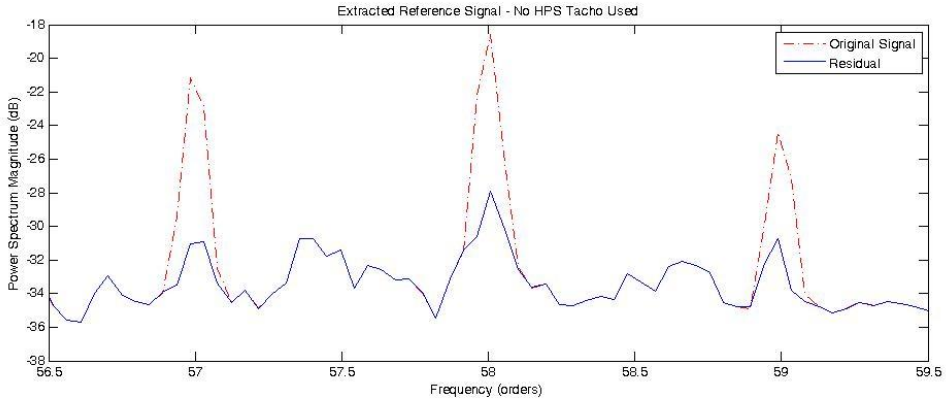


Figure 5.1 – PSD plot of original and residual signal, using
(a) the manufacturer’s gear ratio (b) Extracted reference signal

Figure 5.1 show comparison power spectral density (PSD) plots of the order-tracked signal, and the residual signal after the TSA was subtracted from the order-tracked signal, zoomed over three (high) orders for clarity. Figure 5.1(a) shows the results when the manufacturer’s supplied ratio was used with traditional order-tracking methods, and it can be seen that TSA removed none of the high frequency HP shaft orders. Figure 5.1(b) shows results when the multi-stage PDOT method was used to order-track the signal before TSA, and it can be seen that a significant portion of the HP shaft components were removed by TSA and are absent in the residual. The separation indices from all methods tested in the study showed that the multi-stage PDOT method, using a reference component extracted from the vibration signal, had the second highest quantitative measure of success out of nine alternative order-tracking procedures evaluated in [7]. The highest measure of success from the study was gained by initially employing time-domain order-tracking using the accessory shaft tachometer, before again using PDOT with an extracted reference component from the vibration signal. However, the improvement using the tacho as a first stage was only marginally better than the results shown for the multi-stage PDOT method, using only the response signal, and it was concluded that both methods were approximately equal in effectiveness.

Following the study [7], results with higher SI values were produced by iteratively estimating the gear ratio, to be employed with the tachometer signal, with the goal of maximising the SI. It is interesting to note that the optimum result implied a different gear ratio (to 8 significant figures) for three different load cases, emphasising the fact that the gear set is not rigid as commonly assumed.

These results show the successful implementation of a multi-stage PDOT method using an extracted reference component, in the absence of amplitude modulation effects.

5.2 Single stage PDOT with a large speed variation

This method is applicable where the full removal of speed variation is not important, such as with bearing faults, where characteristic frequencies typically vary by 1-2%. For this example, studied in [5], experimental signals were captured from a gearbox with seeded bearing faults, which I captured myself, with results for one seeded fault type shown here. These results, and some of the accompanying figures in this section, were originally produced by my co-author Bob Randall in [5], but using my code for the order-tracking, which was the core of this analysis.

A detailed overall description of the gearbox is presented by Hiroaki Endo [42] (Appendix B), which includes illustrations, and is not repeated here for brevity.

The gearbox was operating with two 32-tooth spur gears at a mean shaft speed of 6 Hz, with various amounts of speed variation including ±10% and ±25%. The period of the frequency sweep was 5s, making the modulation frequency W = 0.2 Hz. The signals analysed are with 25 Nm torque load, resulting in a radial bearing load of about 250 N. One bearing had seeded local faults, and the cases analysed here are for an inner race fault. A once-per-rev tacho signal was available for use as a reference signal, and the data signal could also be used as a reference signal.

For this analysis, only one stage of order-tracking was used as this was sufficient to successfully implement subsequent bearing diagnostics, and further multi-stage refinement was unnecessary.

The PDOT method was initially implemented using the tacho signal as the reference signal, and then separately using a reference component extracted from the acceleration signal. Results for a ±10% and ±25% speed variation are shown.

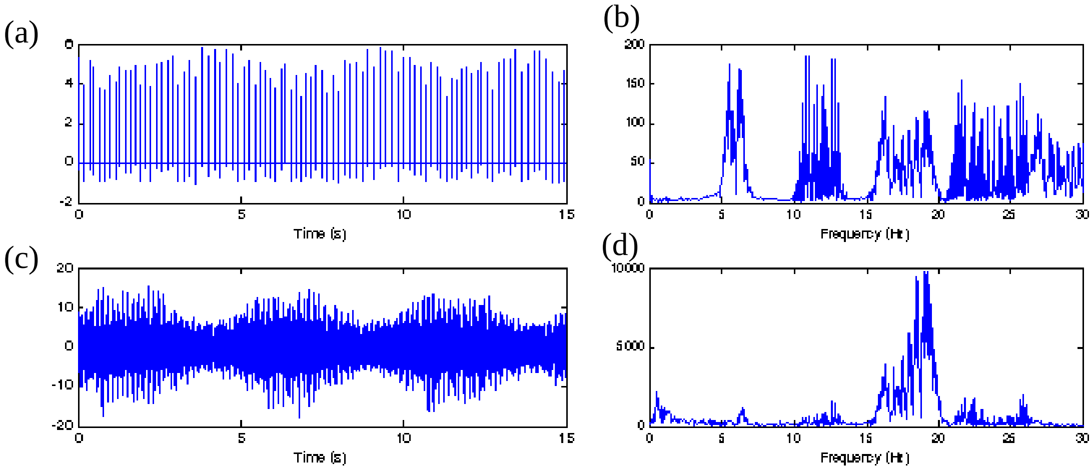


Figure 5.2 – Signals with ±10% speed variation (From [5])
(a, b) tacho (c, d) acceleration (a, c) time signals (b, d) spectra

Figure 5.2 shows time signals and corresponding spectra for the case with ±10% speed variation. The acceleration signal was measured on the casing above the faulty bearing. From Figure 5.2(a) and (c), amplitude modulation of the vibration signal at $f_a = 0.2$ Hz ($= W$), with period of 5 s, is clearly visible. From Figure 5.2 (b) it can be seen that the first three harmonics do not have overlapping sidebands and can be used for demodulation. Figure 5.3 shows a spectrogram of the tacho signal, which again indicates the first three harmonics can be used, and clearly shows the ±10% speed variation, with a period of 5 s corresponding to modulation frequency of $W = 0.2$ Hz.

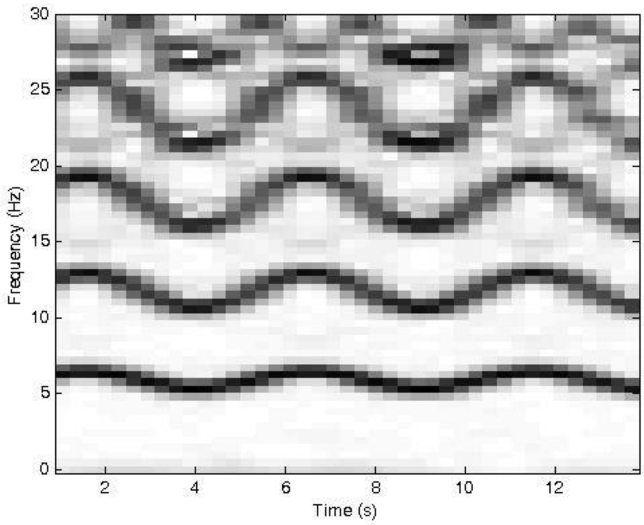


Figure 5.3 – Spectrogram of tacho signal for ±10% speed variation

In this instance the PDOT method was implemented using the third harmonic of the tacho signal, and then the third harmonic of the acceleration signal. In this case the acceleration signal has very weak first and second harmonics, unsuitable for use, which is likely a result of recent refurbishment including balancing and realignment, whereas typically these would be stronger and could also be used for order-tracking.

(a)

(b)

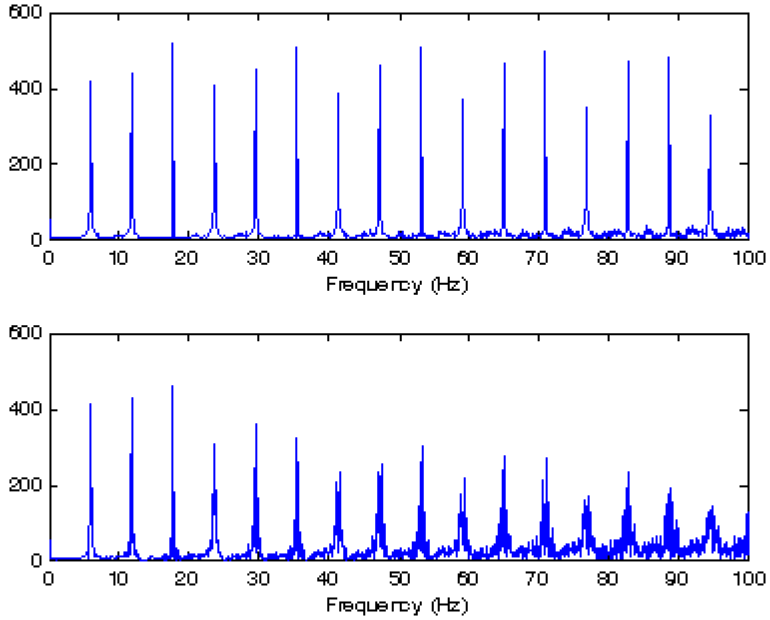


Figure 5.4 – Spectra of order tracked tacho signal using 3rd harmonic of (From [5]): (a) tacho signal (b) acceleration signal

Figure 5.4 shows the resulting spectrum of the tacho signal after order-tracking. It can be seen that each reference component harmonic in Figure 5.4(a) has resolved into a discrete peak indicating the success of the order-tracking process. The tacho signal is shown rather than the acceleration signal, as this gives a clearer picture of the removal of speed fluctuations. The frequency axis is scaled in terms of the mean speed of 6 Hz, but it is actually an order axis, which can be obtained by division by 6. In Figure 5.4(b) the acceleration signal was used and gave a slightly poorer result, but with the low harmonics still well resolved. A further iteration could have been carried out using say the 9th harmonic at about 54 Hz, but this was not necessary for subsequent bearing diagnostics.

In order to detect the seeded bearing fault, envelope analysis was conducted, as described in [43].

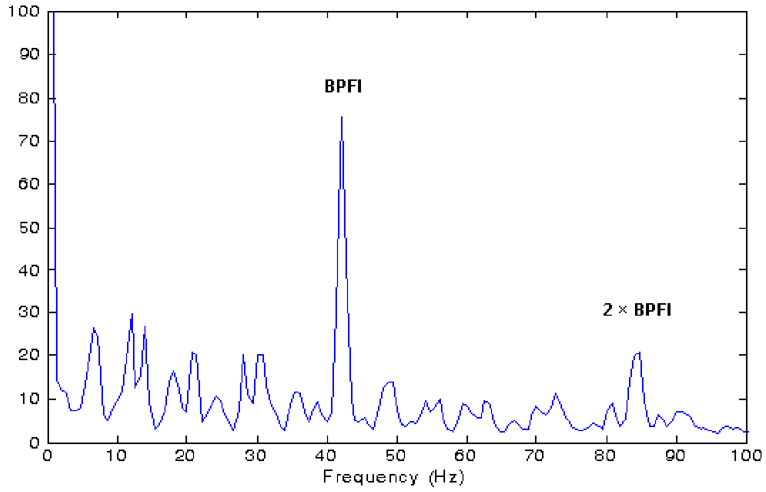


Figure 5.5 – Spectrum of the squared envelope of the order tracked signal showing harmonics of BPFI as well as harmonics and sidebands spaced at shaft speed 6 Hz (From [5])

Figure 5.5 shows the averaged spectrum of the squared envelope of the acceleration signal, demodulated in a high frequency band which maximised the kurtosis of the transmitted signal. Figure 5.5 shows results based on the reference component from the acceleration signal; however, virtually identical results were produced when order-tracking was conducted with the tacho signal. The frequency axis is once again based on mean rotational speed. The first two harmonics of BPFI (ballpass frequency, inner race) are seen, along with low harmonics of shaft speed (6 Hz), and sidebands spaced at this frequency around BPFI, indicating that an inner race fault has been detected. It should be noted that in the direct FFT spectrum of the envelope, there were modulation sidebands spaced at a mean value of 0.2 Hz because of the amplitude modulation seen in Figure 5.2(c) on page 248, but these are smeared because of the conversion to order scale, and the coarser resolution of the averaged spectrum.

Figure 5.6 shows the time signals and spectra for the ±25% speed variation case.

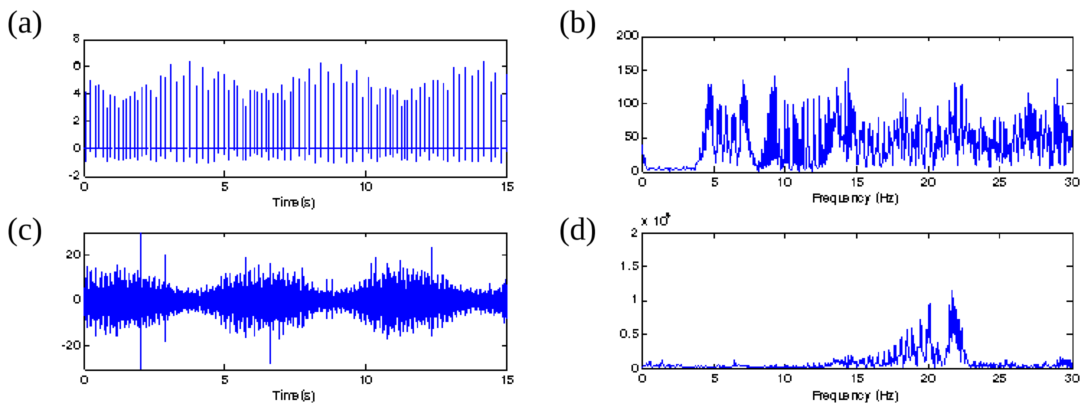


Figure 5.6 – Signals with ±25% speed variation (From [5])
(a, b) tacho (c, d) acceleration (a, c) time signals (b, d) spectra

From Figure 5.6(b) it can be seen that only the first harmonic in the tacho spectrum is separated from other harmonics, and was used for order-tracking.

Figure 5.6(d) contains no separable reference component harmonics, as the third harmonic has clearly smeared into neighbouring components, and again the first and second harmonics are very weak and cannot be used. In this instance, the PDOT method cannot be implemented using the acceleration signal, and this highlights the increased difficulty inherent in using an extracted reference signal over a recorded reference signal.

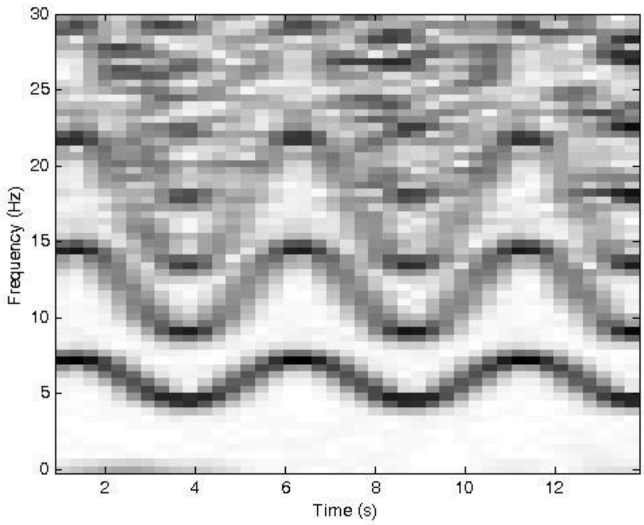


Figure 5.7 – Spectrogram of tacho signal for ±25% speed variation

Figure 5.7 shows the spectrogram for the tacho signal, which again indicates only the first harmonic can be used for order-tracking in this instance, as well as clearly showing the ±25% speed variation.

Figure 5.8 shows the order-tracked results, when using the tacho.

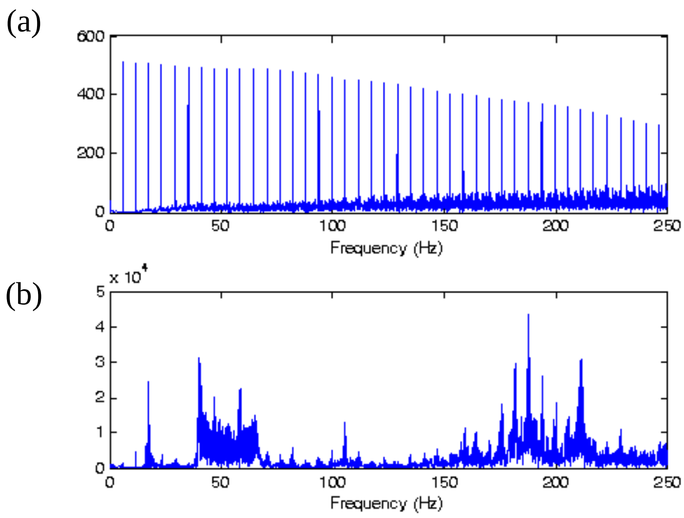


Figure 5.8 – Spectra after order tracking (From [5])
(a) Tacho signal (b) Acceleration signal

From Figure 5.8(a) it can be seen that each reference component harmonic has resolved into roughly a discrete peak indicating the success of the order-tracking process. The low harmonics of the acceleration signal in Figure 5.8(b) are now well separated, and the peak near 192 Hz corresponds to the gearmesh frequency, surrounded by modulation sidebands at shaft speed. These are still a little smeared, but by comparison with the equivalent harmonics of the tacho signal, it appears that most of the speed variation has already been removed, and the residual sidebands are due to amplitude modulation (a fixed frequency of 0.2 Hz in the time domain, but smeared in the angle domain). Even so, the correction given by this single stage of order tracking was sufficient to allow subsequent bearing diagnostics to be employed, and a similar result was achieved to that in shown in Figure 5.5, on page 251, indicating the successful detection of the bearing inner-race fault.

5.3 Multi-stage PDOT for a large speed variation

This method is applicable when more detailed analysis is required, such as for gear diagnostics. For this example, studied in [4], experimental signals were captured from the same gearbox as used in Section 5.2 on page 247, with seeded gear faults, with results for one fault type shown here.

The gearbox was operating with two 32-tooth spur gears at a mean shaft speed of 6 Hz, with various amounts of speed variation including ±10% and ±25%. The period of the frequency sweep was again 5s, making the modulation frequency W = 0.2 Hz. The signals analysed are with 25 Nm torque load. One gear had a seeded local tooth root crack. A once-per-rev tacho signal was available for use as a reference signal, and the data signal could also be used as a reference signal.

For this analysis, multiple stages of order-tracking were used where possible, as accurate high frequency information was desired for the subsequent gear diagnostics, and further refinement was necessary. Subsequent gear fault diagnostics were also implemented.

5.3.1 PDOT using tacho signal

The PDOT method was initially implemented using the tacho signal as the reference signal. Results for both the ±10% and ±25% speed variation are shown. With the tacho signal it was possible to conduct order-tracking in multiple stages to improve the order-tracking result.

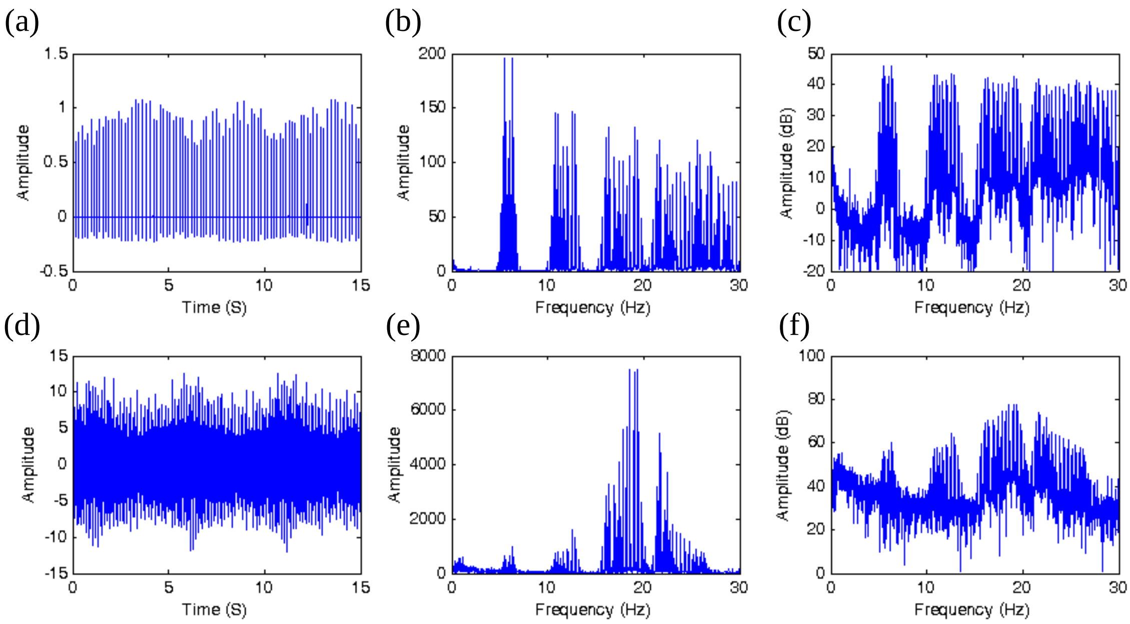


Figure 5.9 – Signals with ±10% speed variation
(a, b, c) tacho (d, e, f) acceleration (a, d) time signals (b, e) spectra (c, f) dB spectra

Figure 5.9 shows the initially captured signals for the ±10% case. Only 15 seconds of the 60 second records are shown in the time domain for visual clarity. From the spectra (b) and (e) it can be seen that the shaft harmonics are clearly modulated due to the speed variation. From (b) it can also be seen that the first harmonic of the reference signal is separable, giving initial indications it can be used for order tracking. It can also be noted that both the tacho signal (a) and acceleration signal (d) show clear amplitude modulation with a period of 5s visible, i.e. $f_a = 0.2Hz$. It is not clear why the tacho signal is amplitude modulated, and this typically would not be the case with a tacho.

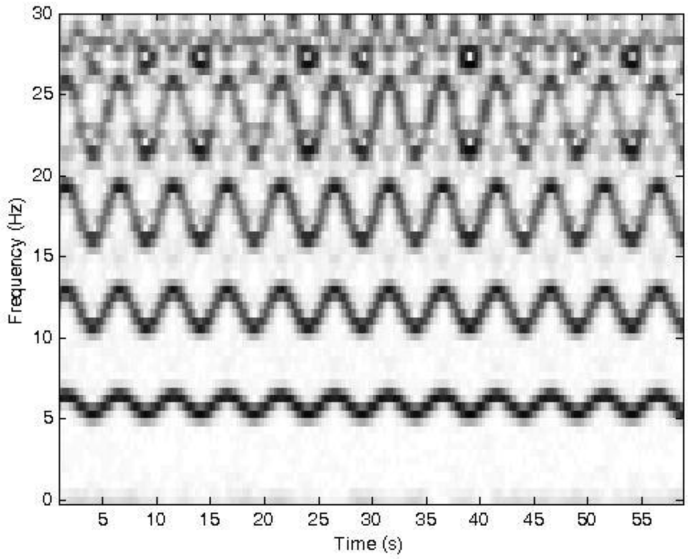


Figure 5.10 – Spectrogram of tacho signal for ±10% speed variation

Figure 5.10 shows a spectrogram of the tacho signal, which again indicates that the first harmonic of the tacho is separable, as well as showing the ±10% speed variation.

To illustrate the use of the bandwidth acceptability criteria, as set out in Sections 3.5, 3.6 and 3.7, starting on page 158, the modulation parameters for this signal are used to determine the suitability of the first harmonic for the ±10% signal.

The known parameters of the signal are $f_c = 6Hz$,. $D_{\%c} = 10\%$, $W_{\%c} = (0.2Hz / 6Hz) \times 100 = 3.33\%$, and $f_{a\%c} = (0.2Hz / 6Hz) \times 100 = 3.33\%$.

As a first check, the modulation parameters $D_{\%c} = 10\%$ and $W_{\%c} = 3.33\%$ can be used to check the suitability of using the first harmonic by using Figure 3.7 on page 165.

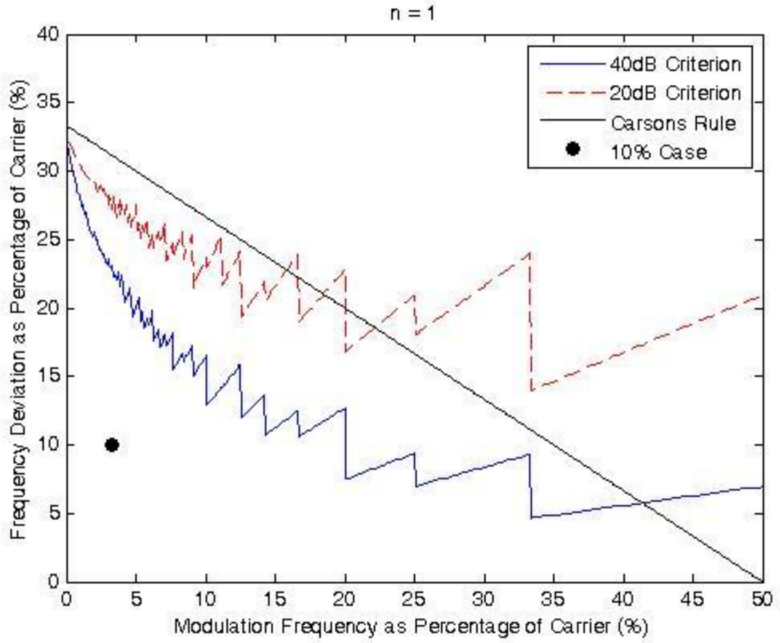


Figure 5.11 – Checking suitability of ±10% case on Figure 3.7

Figure 5.11 shows the location in Figure 3.7 as a solid circle for the specific modulation parameters being considered. As this point is below the acceptability envelope for all three bandwidth limits, this indicates that the first harmonic is suitable.

This first check does not consider the AM of the reference signal. As seen in Figure 5.9(a) the tacho signal is amplitude modulated, and so amplitude modulation will also be considered. This is done using equations from section 3.7 on page 167, for the case of the first harmonic $n=1$. This check is done using the three different presented acceptability criteria, the 40dB and 20dB DR based on Bessel function calculations, and the Carson's rule criterion, to determine if the first harmonic is suitable for use with order-tracking for the given modulation parameters.

For the criteria based on Bessel functions, using a 40dB and 20dB dynamic range, firstly the equivalent sinusoidal variation parameters to be used in calculations are gained by using the non-sinusoidal parameters, so that $f_{d\%c} = D_{\%c}$ and $f_{m\%c} = W_{\%c}$.

Firstly using a 40dB criterion for defining significant sidebands, the highest significant sideband order $k_{\max}^{n}$ is found for the first harmonic $n=1$ and second harmonic $(n+1)=2$. Using the method described in section 2.5.4 on page 101, these are found to be $k_{\max}^{1}=7$ and $k_{\max}^{2}=11$. Values can then be substituted into the left side equation (3.30) on page 170 to check if the given parameters are equal or less than the 100% limit given by the right side of the equation. This results in:

$$k_{\max}^{n} f_{m\%c} + k_{\max}^{(n+1)} f_{m\%c} + 2 f_{a\%c} = 7\times 3.33 + 11\times 3.33 + 2\times 3.33 = 66.6\% \quad .... (5.2)$$

where $66.6\% \leq 100\%$, indicating these modulation parameters are permissible using a 40dB criterion when amplitude modulation is included.

In an identical fashion using a 20dB criterion for defining significant sidebands, the highest significant sideband order $k_{\max}^{n}$ is found for the first harmonic $n=1$ and second harmonic $(n+1)=2$. Using the method described in section 2.5.4 on page 101, these are found to be $k_{\max}^{1}=5$ and $k_{\max}^{2}=9$. Values can then be substituted into the left side equation (3.30) to check if the given parameters are equal or less than the 100% limit given by the right side of the equation. This results in:

$$k_{\max}^{n} f_{m\%c} + k_{\max}^{(n+1)} f_{m\%c} + 2 f_{a\%c} = 5\times 3.33 + 9\times 3.33 + 2\times 3.33 = 53.28\% \quad .... (5.3)$$

where $53.28\% \leq 100\%$, indicating these speed parameters are permissible using a 20dB criterion when amplitude modulation is included.

To check if order-tracking is permissible based on the Carson's rule criterion, values are substituted into the left side equation (3.32) on page 170 to check if the given parameters are equal or less than the 100% limit given by the right side of the equation. This results in:

$$(2n+1)D_{\%c} + 2(W_{\%c} + f_{a\%c}) = (2\times1+1)\times10 + 2(3.33+3.33) = 43.32\%$$

where $43.32\% \leq 100\%$, indicating these speed parameters are permissible using the Carson's rule criterion when amplitude modulation is included.

So in this instance, all three criteria indicate that order-tracking is permissible for this case, when amplitude modulation is considered.

It should be noted that a more restrictive bandwidth criterion results in a larger bandwidth (so higher percentage), as larger numbers of sidebands are considered significant which increases the bandwidth F.

As stated in Section 3.5 on page 158, final acceptability should preferably be made based on inspection of the reference signal spectrum for separation of the first harmonic. From Figure 5.9(c), on page 256, which shows the tacho signal on a dB scale, it can be seen that the first harmonic centred at approximately 6Hz clearly shows greater than 40dB reduction of sidebands between orders, indicating the first harmonic is suitable for order-tracking based on the two Bessel function based acceptability criteria. Note that the Carson's rule criterion cannot be tested by inspection of the spectrum.

Based on the bandwidth calculations, and subsequent inspection of the spectrum, the first harmonic of the tacho was used for the first stage of the multi-stage order tracking.

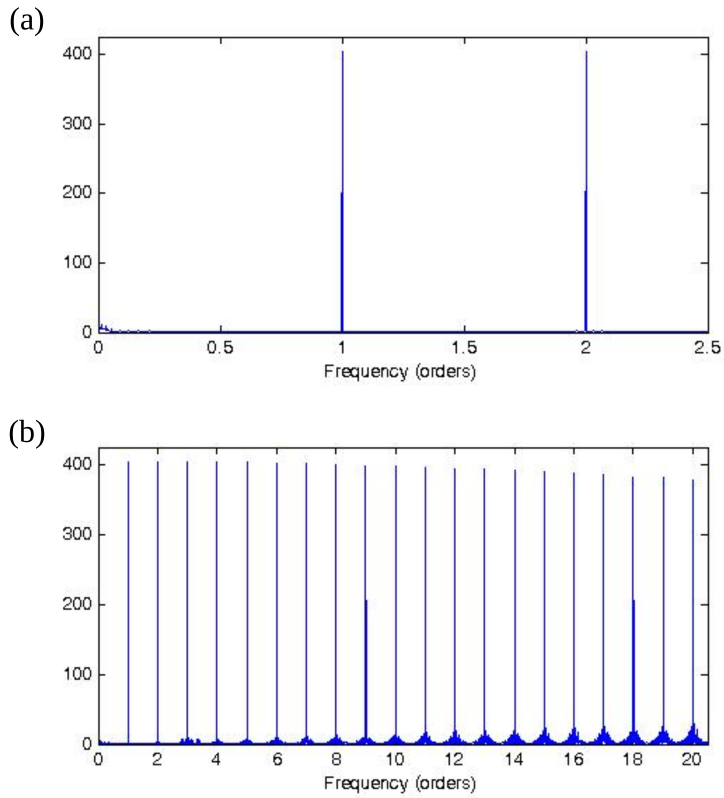


Figure 5.12 – Spectrum of order-tracked tacho after 1 stage for ±10% case
(a) showing first 2 harmonics (b) showing first 20 harmonics

Figure 5.12 shows the result of the order tracked tachometer signal. The ideal order-tracking would result in a single discrete component for each harmonic present. As can be seen, the first harmonic has been order-tracked to one discrete value, and there are progressively increasing residual modulation sidebands at the higher harmonics, as was expected from one stage of order-tracking.

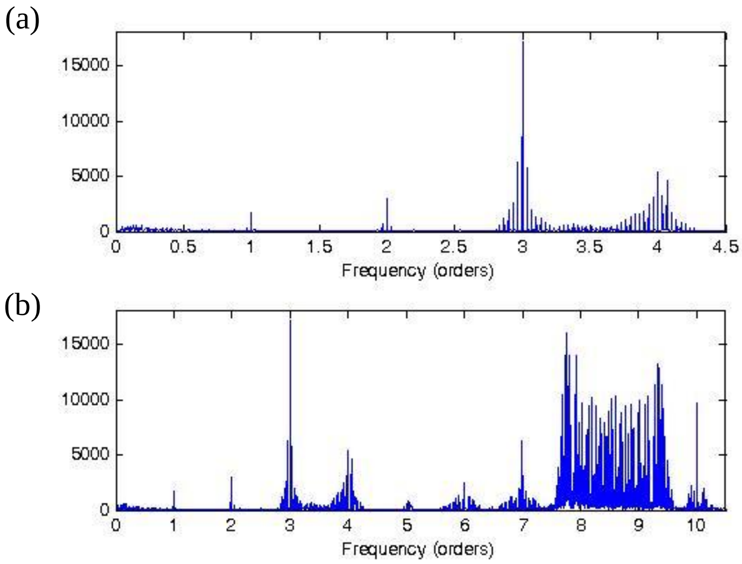


Figure 5.13 – Spectrum of order-tracked acceleration signal after 1 stage for ±10% case (a) first 4 orders (b) first 10 orders

Figure 5.13 shows the spectrum of the corresponding acceleration signal after one stage of order-tracking. Figure 5.13 firstly shows that the frequency modulation has largely been removed when compared to Figure 5.9(e) on page 256. The remaining modulation sidebands are primarily due to the amplitude modulation present in the signal. The large smeared component between orders 7.5 and 9.5 in Figure 5.13(b) is a constant frequency 50 Hz component which becomes smeared in the order domain. Ideally constant frequency components should be removed before order-tracking to prevent this, as they can mask true order components.

The signal was completely order-tracked using 4 stages of order-tracking. The multiple stages used the 1^{st}, 5^{th}, 21^{st} and 151^{st} harmonics in sequence.

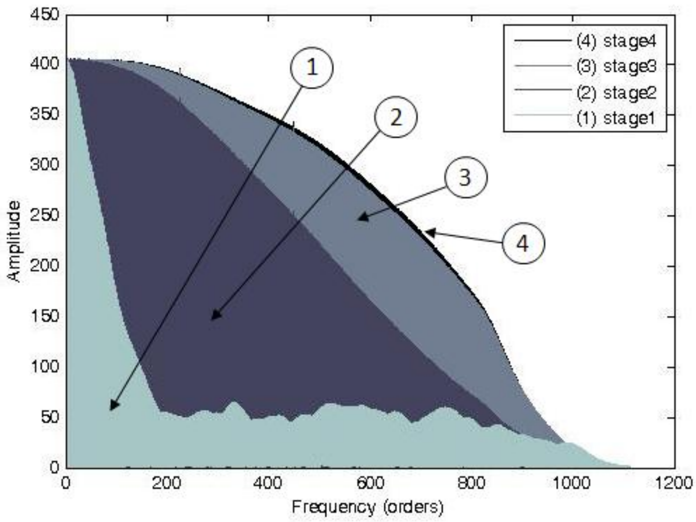


Figure 5.14 – Comparing Spectra of order-tracked tacho from all four stages for ±10% case

Figure 5.14 compares the tacho spectra from all four stages of order-tracking, effectively showing the envelope of each spectrum. Figure 5.14 clearly shows the progressive improvement of the higher harmonics in later stages of order-tracking. As the harmonics become less smeared and tend to discrete orders they increase in amplitude because the total energy of each harmonic becomes concentrated in one frequency line. This can be seen in the envelope of the order spectrum, with later stages providing smaller improvements in the order-tracking.

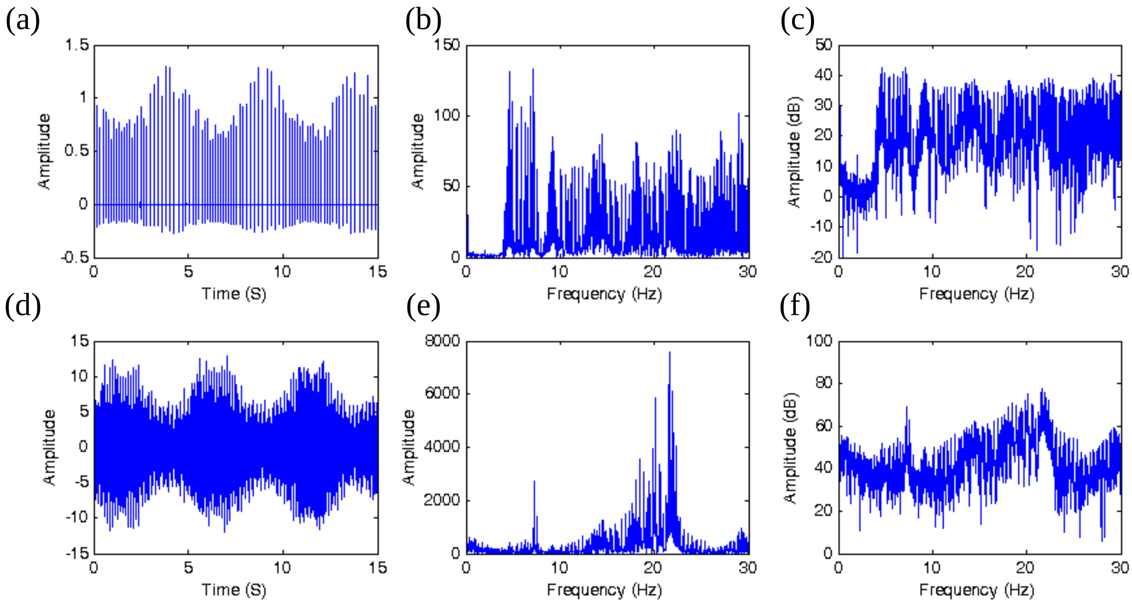


Figure 5.15 – Signals with ±25% speed variation
(a, b, c) tacho (d, e, f) acceleration, (a, d) time signals (b, e) spectra (f, f) dB spectra

Figure 5.15 shows the original signals for the ±25% case, and in (b) the larger speed variation is clearly visible. Figure 5.16 shows a spectrogram of the tacho signal, which also shows the ±25% speed variation.

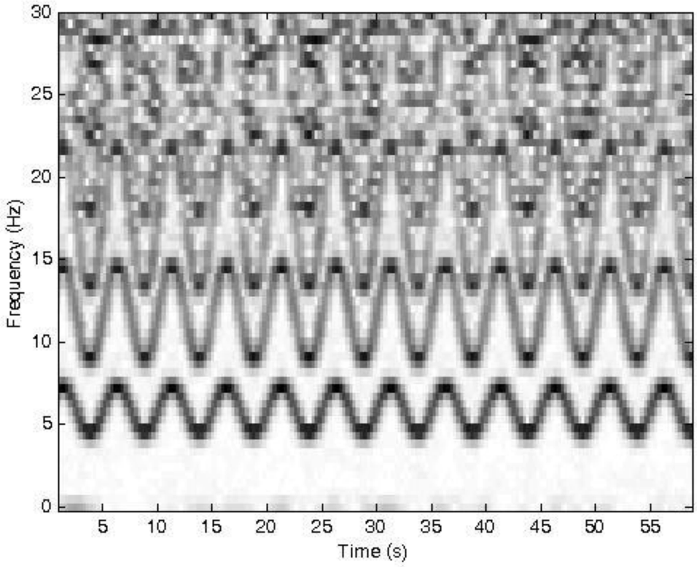


Figure 5.16 – Spectrogram of tacho signal for ±25% speed variation

Similarly to the ±10% case, the first harmonic was checked for suitability for use with order-tracking using the bandwidth acceptability criteria.

For this case, the known parameters are $f_c = 6Hz$, $D_{\%c} = 25\%$, $W_{\%c} = 3.33\%$, and $f_{a\%c} = 3.33\%$. Firstly, the modulation parameters $D_{\%c}$ and $W_{\%c}$ were checked with Figure 3.7 on page 165.

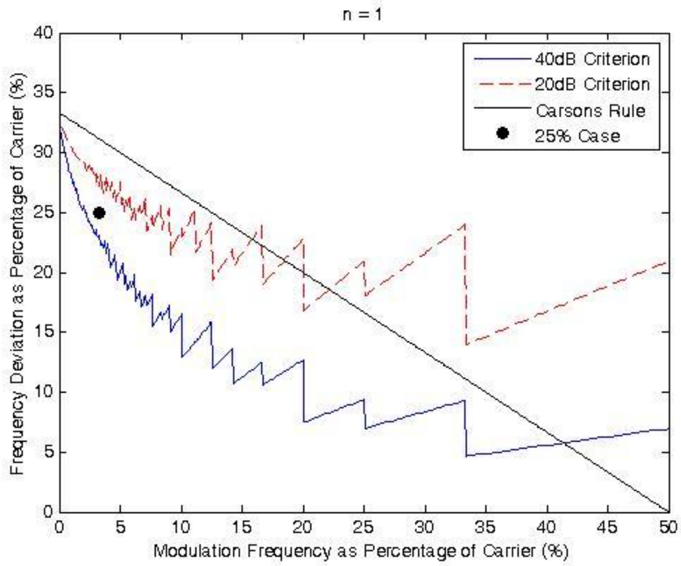


Figure 5.17 – Checking suitability of ±25% case on Figure 3.7

Figure 5.17 shows the location in Figure 3.7 as a solid circle for the specific modulation parameters being considered. In this instance, the Carson's rule and 20dB criteria indicate acceptability, while the 40dB criterion is not met. From these acceptability criteria, the combination of modulation parameters for the ±25% case appear to be near the limits of what can be successfully used with PDOT method, and order-tracked results would be expected to be poorer than those gained with the ±10% case above.

As for the 10% case, the 25% case can also be checked with the inclusion of the amplitude modulation component using equations (3.30) and (3.32) on page 170. While this is not specifically needed in this instance as the first check with Figure 3.7 has already indicated this is a marginal case, the check with the acceptability equations is shown here as an additional example of their usage.

As with the 10% example, this check is done using the three different presented acceptability criteria, the 40dB and 20dB DR based on Bessel function calculations, and the Carson's rule criterion, to determine if the first harmonic $n=1$ is suitable for use with order-tracking for the given modulation parameters.

For the criteria based on Bessel functions, using a 40dB and 20dB dynamic range, firstly the equivalent sinusoidal variation parameters to be used in calculations are gained by using the non-sinusoidal parameters, so that $f_{d\%c} = D_{\%c}$ and $f_{m\%c} = W_{\%c}$.

Firstly using a 40dB criterion for defining significant sidebands, the highest significant sideband order $k_{\max}^{n}$ is found for the first harmonic $n=1$ and second harmonic $(n+1)=2$. Using the method described in section 2.5.4 on page 101, these are found to be $k_{\max}^{1} = 13$ and $k_{\max}^{2} = 22$. Values can then be substituted into the left side equation (3.30) to check if the given parameters are equal or less than the 100% limit given by the right side of the equation. This results in:

$$k_{\max}^{n} f_{m\%c} + k_{\max}^{(n+1)} f_{m\%c} + 2 f_{a\%c} = 13 \times 3.33 + 22 \times 3.33 + 2 \times 3.33 = 123.21\% \quad . (5.4)$$

where $123.21\% > 100\%$. This indicates these modulation parameters are not permissible using a 40dB criterion when amplitude modulation is included, which matches the result when the check was made with Figure 3.7 on page 165, as shown in Figure 5.17.

In an identical fashion using a 20dB criterion for defining significant sidebands, the highest significant sideband order $k_{\max}^{n}$ is found for the first harmonic $n=1$ and second harmonic $(n+1)=2$. Using the method described in section 2.5.4 on page 101, these are found to be $k_{\max}^{1} = 11$ and $k_{\max}^{2} = 19$. Values can then be substituted into the left side equation (3.30) to check if the given parameters are equal or less than the 100% limit given by the right side of the equation. This results in:

$$k_{\max}^{n} f_{m\%c} + k_{\max}^{(n+1)} f_{m\%c} + 2 f_{a\%c} = 11 \times 3.33 + 19 \times 3.33 + 2 \times 3.33 = 106.56\% \quad .. (5.5)$$

where $106.56\% > 100\%$. This now indicates that these speed parameters are not permissible using a 20dB criterion. In this instance, the inclusion of amplitude modulation has changed the permissibility of the modulation parameters to being unacceptable, where the check on Figure 3.7 on page 165, which did not take into account amplitude, initially indicated permissibility. This illustrates the importance of including the effects of amplitude modulation for marginal cases, and not solely relying on a check with the acceptability charts, for situations where the reference signal is amplitude modulated.

To check if order-tracking is permissible based on the Carson's rule criterion, values are substituted into the left side equation (3.32) on page 170 to check if the given parameters are equal or less than the 100% limit given by the right side of the equation. This results in:

$$(2n+1)D_{\%c} + 2(W_{\%c} + f_{a\%c}) = (2 \times 1 + 1) \times 25 + 2(3.33 + 3.33) = 81.66\% \quad ..... (5.6)$$

where $81.66\% \leq 100\%$, indicating these speed parameters are permissible using the Carson's rule criterion when amplitude modulation is included.

The final acceptability was determined by inspection of the reference signal spectrum for separation of the first harmonic. From Figure 5.15(c) on page 264, which shows the tacho signal on a dB scale, it can be seen that the first harmonic centred at approximately 6Hz only shows an approximate reduction of sidebands between the first and second order of 15-20dB.

As a result, the ±25% case is expected to give comparably worse results than with the ±10% case.

The ±25% signal was again completely order-tracked using 4 stages of order-tracking. The multiple stages used the 1st, 5th, 21st and 151st harmonics in sequence.

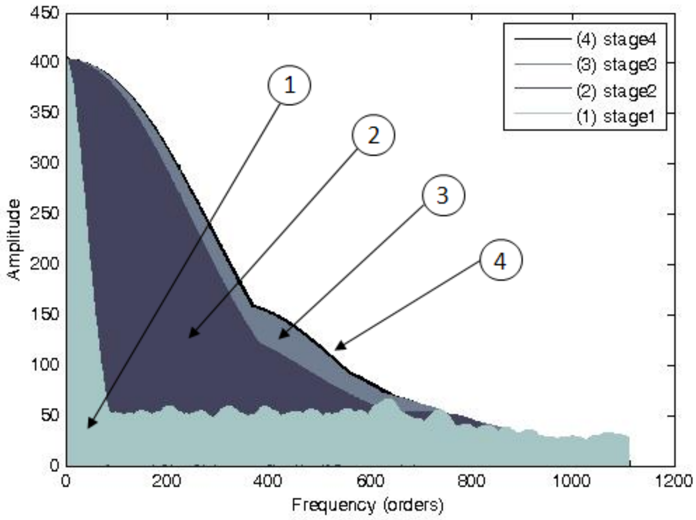


Figure 5.18 – Comparing Spectra of order-tracked tacho from all four stages for ±25% case

Figure 5.18 compares the tacho spectra from all four stages of order-tracking, effectively showing the envelope of each spectrum. Figure 5.18 clearly shows the progressive improvement of the higher harmonics in later stages of order-tracking in the ±25% case, though the lower amplitudes of higher harmonics indicate comparatively poorer order-tracking than when the 40dB criterion is met.

From the results of the multi-stage order-tracking for both the ±10% and ±25% speed variation signal, the multistage approach has successfully order-tracked the signals, with successive improvement seen at each stage, though with a poorer result when the speed variation is at the limits for which the PDOT method can be used.

In order to further evaluate the success of the multistage order-tracking process, the order-tracked results were subjected to a common gear analysis technique, in order to detect the seeded gear fault.

The first analysis process was to apply TSA to the order-tracked signals. By conducting the TSA for one period of rotation of the gears in the signal, the gearmesh of the 32 gear teeth should be visible as 32 equally spaced peaks if the prior order-tracking was successful, and the gear errors were completely uniform. Note that the triggering for TSA was linked to the rotation angle of the gears, and not synchronised in any way with the speed variation. Thus the results of TSA represent the average amplitude in each rotation cycle, averaged over twelve periods of the speed variation (with corresponding amplitude modulation).

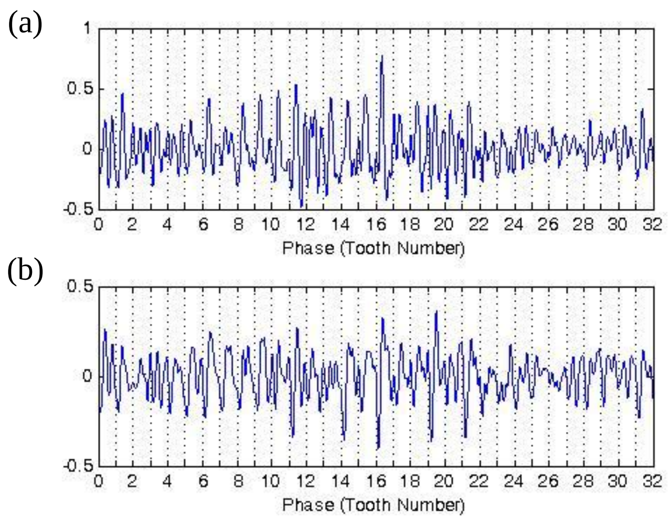


Figure 5.19 – TSA results (a) ±10% case (b) ±25% case

Figure 5.19 shows the results of the TSA for the two speeds. The figures shown were low pass filtered to just above the 4th gearmesh harmonic (128th order) in order to improve the visual clarity of the gearmesh. In the TSA one pulse per tooth is clearly visible, showing the meshing of each individual tooth. In this case two pulses per revolution are also seen in some zones. This happens because the second harmonic of gearmesh is stronger than the first in the overall averaged spectra. That the gearmesh can be correctly identified in the TSA results is a clear indication that the multi-stage order-tracking process is suitable for gear analysis when a tachometer is available. The slightly coarser result with the ±25% case results from the comparatively poorer order-tracking conditions.

To further evaluate the order-tracking process the technique of demodulating the gearmesh in the TSA was used, as described by McFadden [44]. The 2nd gearmesh harmonic was demodulated, and a bandwidth of ±15% of the 2nd harmonic of gearmesh was found to give optimum visual results.

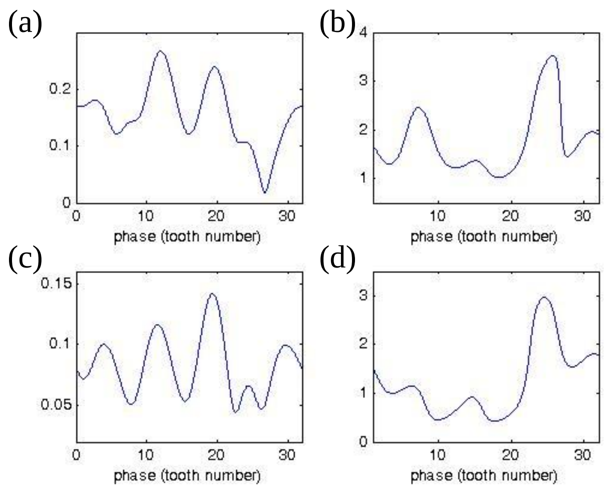


Figure 5.20 – Demodulated gearmesh amplitude and phase
(a, b) 10% case, (c, d) 25% case, (a, c) amplitude (b, d) phase

From Figure 5.20(b) and (d), it can be seen that there is a large phase change at tooth number 25 in both speed cases. This change in phase is typical for a gear tooth root crack, the seeded fault. That the fault can correctly be located in the demodulated gearmesh phase is further indication that the underlying multi-stage order-tracking process was successful. It should be noted that the extra variations seen in the demodulated gearmesh phase, such as at the approximate tooth number six seen in Figure 5.20(b) and (d), can be expected from a typical phase demodulated gearmesh from even a healthy gear. Similar variations are visible in Figures 3(c) and 4(c) of [44]. In the case presented here the seeded fault location was known, and could be distinguished from other variations on this basis. In normal MCM diagnosis, Figure 5.20(b) and (d) would ideally be compared to results from a healthy running state baseline signal to correctly identify a new fault.

5.3.2 PDOT using Extracted reference component from vibration signal

Using the same data as shown in Section 5.3.1, order-tracking was also conducted using an extracted reference component from the acceleration signal.

For the ±10% speed case, for a first stage the 3rd shaft harmonic from the acceleration signal was checked for suitability. From Figure 5.9(e) and (f) on page 256, it can be seen that the first two shaft harmonics are very low in amplitude, and show a low dB separation. As such the first two harmonics were not used for order-tracking, similarly to the bearing experimental results as shown in Section 5.2 on page 247.

This case was also checked using the bandwidth acceptability criteria. The parameters were identical to those used for the ±10% case in Section 5.3.1, with the exception that the third harmonic $n=3$ is being considered.

As a first check, the modulation parameters $D_{\%c}=10\%$ and $W_{\%c}=3.33\%$ can be used to check the suitability of using the third harmonic $n=3$ by using the acceptability chart for the third harmonic. The acceptability charts for harmonics 2-10 are shown in Appendix A on page 423 and the chart for the third harmonic $n=3$ is Figure A.13 on page 431.

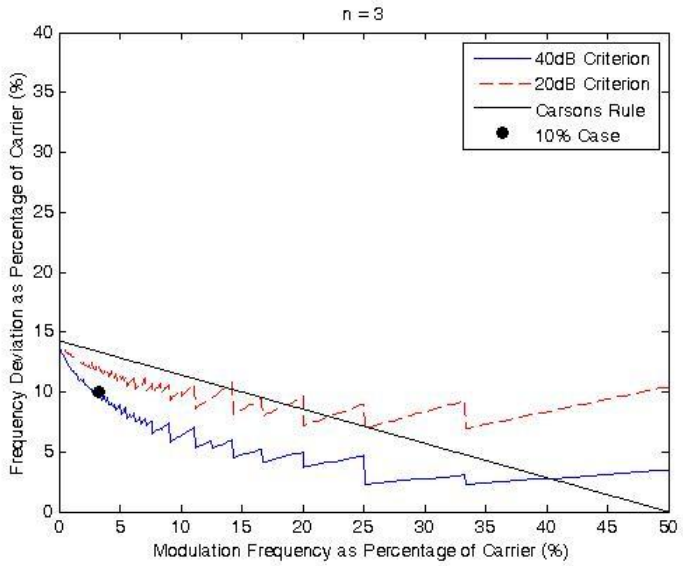


Figure 5.21 – Checking suitability of ±10% case for third harmonic on Figure A.13

Figure 5.21 shows the location in Figure A.13 as a solid circle for the specific modulation parameters being considered. As this point is right near the acceptability envelope for the 40dB criterion, this indicates that using the third harmonic will be another marginal case, and so poorer results are expected.

As for the cases where the tacho signal was used, the 10% case using the third harmonic can also be checked with the inclusion of the amplitude modulation component using equations (3.30) and (3.32) on page 170.

As with the cases from section 5.3.1, this check is done using the three different presented acceptability criteria, the 40dB and 20dB DR based on Bessel function calculations, and the Carson's rule criterion, to determine if the third harmonic $n = 3$ is suitable for use with order-tracking for the given modulation parameters.

For the criteria based on Bessel functions, using a 40dB and 20dB dynamic range, firstly the equivalent sinusoidal variation parameters to be used in calculations are gained by using the non-sinusoidal parameters, so that $f_{d\%c} = D_{\%c}$ and $f_{m\%c} = W_{\%c}$.

Firstly using a 40dB criterion for defining significant sidebands, the highest significant sideband order $k_{\max}^{n}$ is found for the third harmonic $n = 3$ and fourth harmonic $(n+1) = 4$. Using the method described in section 2.5.4 on page 101, these are found to be $k_{\max}^{3} = 15$ and $k_{\max}^{4} = 18$. Values can then be substituted into the left side equation (3.30) on page 170 to check if the given parameters are equal or less than the 100% limit given by the right side of the equation. This results in:

$$k_{\max}^{n} f_{m\%c} + k_{\max}^{(n+1)} f_{m\%c} + 2 f_{a\%c} = 15 \times 3.33 + 18 \times 3.33 + 2 \times 3.33 = 116.55\% \quad \text{.. (5.7)}$$

where $116.55\% > 100\%$. This indicates these modulation parameters are not permissible using a 40dB criterion when amplitude modulation is included, which is expected given the case was at the limit of permissibility for the 40dB criterion on the acceptability chart, as shown in Figure 5.21, before the amplitude modulation component was included in calculations.

In an identical fashion using a 20dB criterion for defining significant sidebands, the highest significant sideband order $k_{\max}^{n}$ is found for the third harmonic $n = 3$ and fourth harmonic $(n+1) = 4$. Using the method described in section 2.5.4 on page 101, these are found to be $k_{\max}^{3} = 12$ and $k_{\max}^{4} = 16$. Values can then be substituted into the left side equation (3.30) to check if the given parameters are equal or less than the 100% limit given by the right side of the equation. This results in:

$$k_{\max}^{n} f_{m\%c} + k_{\max}^{(n+1)} f_{m\%c} + 2 f_{a\%c} = 12 \times 3.33 + 16 \times 3.33 + 2 \times 3.33 = 99.9\% \quad \text{... (5.8)}$$

where $99.9\% \leq 100\%$, indicating that these speed parameters are just below the limit for permissibility with the inclusion of amplitude modulation, and so permissible, using a 20dB criterion.

To check if order-tracking is permissible based on the Carson's rule criterion, values are substituted into the left side equation (3.32) on page 170 to check if the given parameters are equal or less than the 100% limit given by the right side of the equation. This results in:

$$(2n+1)D_{\%c} + 2(W_{\%c} + f_{a\%c}) = (2\times 3+1)\times 10 + 2(3.33+3.33) = 76.66\% \quad ..... (5.9)$$

where $76.66\% \leq 100\%$, indicating these speed parameters are permissible using the Carson's rule criterion when amplitude modulation is included.

Final acceptability was determined by inspection of the spectrum in Figure 5.9(f) on page 256, which shows an approximate DR separation of the sidebands between the third and fourth order of approximately 35dB, indicating that order-tracking should still be reasonably valid.

The acceptability calculation for using a 20dB criterion, in equation (5.8), indicated that the visual separation in Figure 5.9(f) could be expected to be 20dB rather than 35dB. This discrepancy illustrates how the speed calculations tend to be conservative when the 'true' amplitude modulation is used in as an approximation in the calculations, rather than using the 'effective' amplitude modulation, as discussed in section 3.7 starting on page 167.

Thus, the third harmonic of the vibration signal was used for a first stage of order-tracking, when using the signal itself as the reference signal, and Figure 5.22 shows the results.

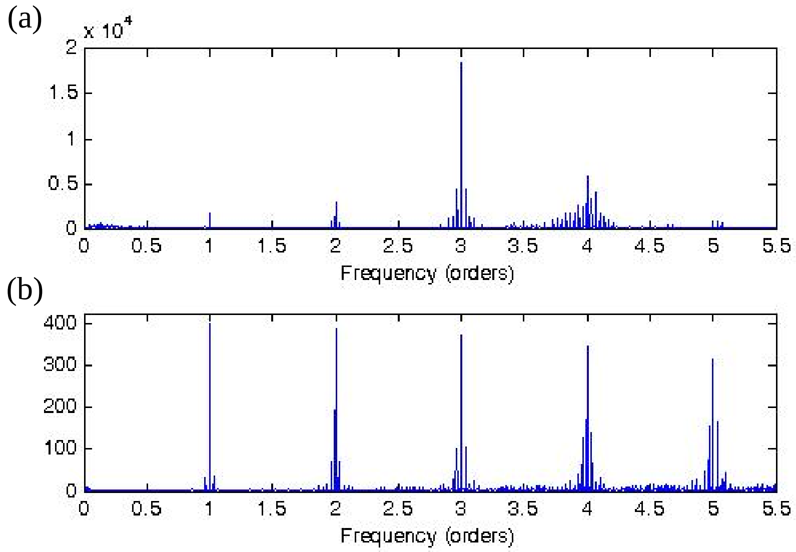


Figure 5.22 – Spectrum of order-tracked signals after 1 stage for ±10% case, 5 orders (a) acceleration (b) tacho

As can be seen the order-tracking has been relatively successful and order components can be seen, however both spectra contain significant modulation sidebands. As sidebands are also visible in the tachometer spectrum, and not just the acceleration signal, these sidebands are believed to indicate residual frequency modulation and not solely amplitude modulation.

For a second stage, several higher harmonics were tried for suitability, but it was found that all the visually separable higher harmonics gave a slightly worse order-tracked result than that from the first stage. This appears to be a consequence of the large amplitude modulation present in the acceleration signal, much larger than that present in the tachometer signal, and in this case means a multi-stage approach cannot be successfully employed despite being desirable for subsequent analysis.

As for the case in Section 5.3.1 on page 256, due to the weak first harmonic in the acceleration signal, the ±25% variation signal did not contain any separable harmonics suitable for order-tracking, as seen in Figure 5.15(e) and (f) on page 264, and so this case could not be tested when using the signal itself as the reference signal.

The first strong harmonic in this case is again the third, and has clearly smeared with the fourth harmonic. That the third harmonic is expected to be unsuitable for the speed parameters is quickly illustrated by using the acceptability chart for the third harmonic, Figure A.13 on page 431.

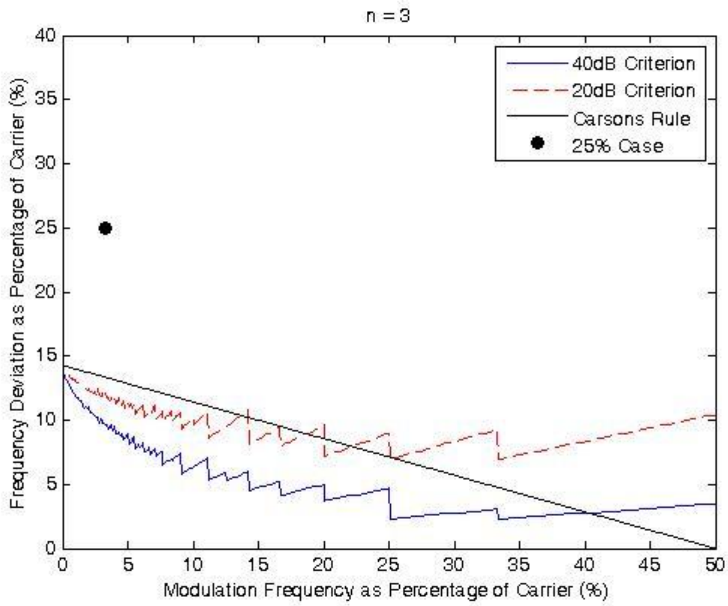


Figure 5.23 – Checking suitability of ±10% case for third harmonic on Figure A.13

Figure 5.23 shows the location in Figure A.13 as a solid circle for the specific modulation parameters being considered, and in this instance the third harmonic is clearly non-permissible with all three acceptability criteria for the 25% signal.

In order to evaluate the success of the first stage of order-tracking for the ±10% variation case, the signal was again further analysed with both TSA and gearmesh demodulation, as in Section 5.3.1 on page 256, but with just the one stage.

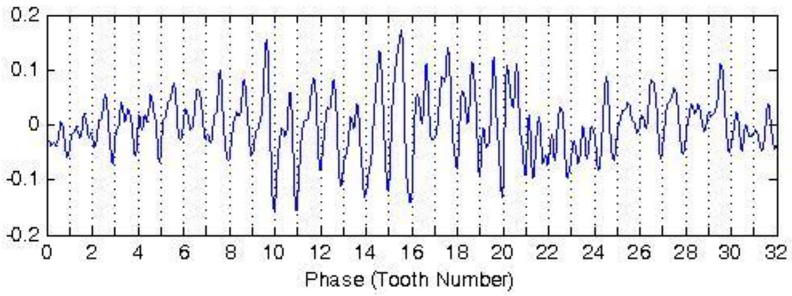


Figure 5.24 – TSA results for ±10% case

Figure 5.24 shows the results of the TSA. The figure shown was again low pass filtered to just above the 4th gearmesh harmonic. In the TSA one pulse per tooth is visible, and while not as clear as with the tachometer order-tracking cases, still shows the meshing of each individual tooth indicating that the order-tracking was successful enough for basic TSA analysis.

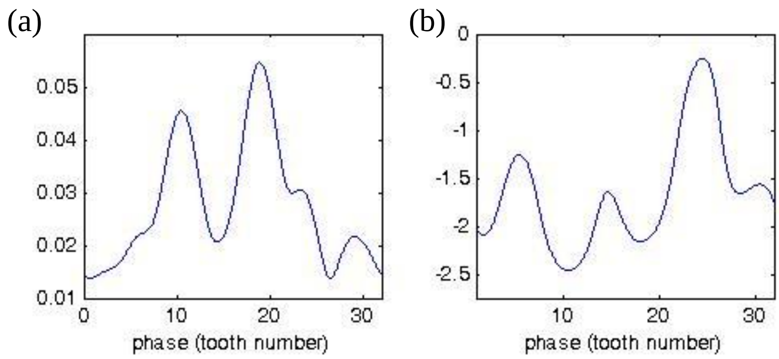


Figure 5.25 – Demodulated gearmesh amplitude and phase
(a) amplitude (b) phase

From Figure 5.25(b), it can be seen that there is again a large phase change at tooth number 25. This change in phase is typical for a gear tooth root crack, the seeded fault. That the fault can correctly be located in the demodulated gearmesh phase is further indication that the results of one stage of the order-tracking process, while not ideal, were successful enough to allow basic gearmesh demodulation analysis for gear fault detection.

5.4 Multi-stage PDOT for a moderate speed variation with minor amplitude modulation

The following example (in [1], originally from [45]) shows what can be achieved for a case with moderate speed variation, a variable speed wind turbine, using a reference signal extracted from the acceleration response. These results, and the accompanying figures presented in this section, were produced by the authors of [45] from privileged data, but using my code for the 2-stage order tracking, which was the heart of the procedure. Over a five minute period the speed varied by perhaps as much as 20%, but by inspection a section of signal was chosen which had very little amplitude modulation, where it was suspected the speed was more stable. This was confirmed using an STFT spectrogram. After analysis it was shown that the variation was 4-5% in that record.

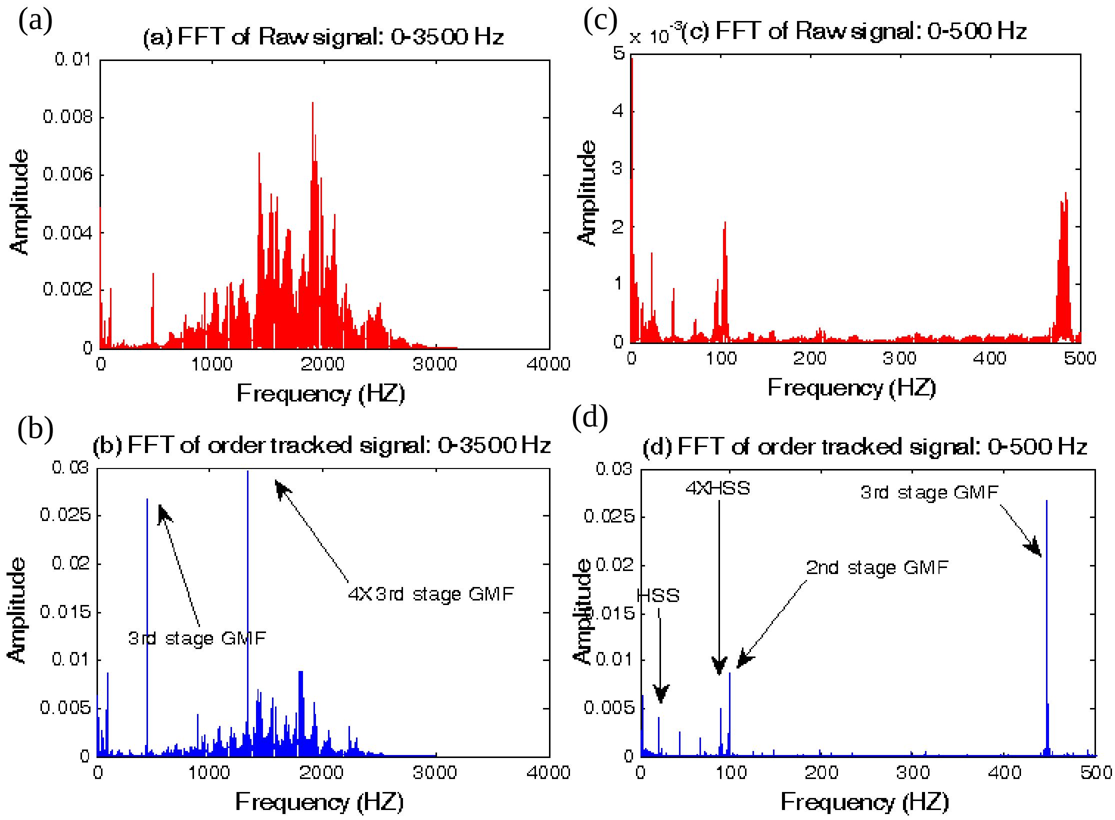


Figure 5.26 – FFT spectra of the raw and order tracked signals (from [45]):
(a) Raw signal (b) Order tracked signal (c) Raw (0-500 Hz)
(d) Order tracked (0-500 Hz)

Figure 5.26(a) shows the original spectrum over a range covering several harmonics of the highest gearmesh frequency at approx. 472 Hz, and Figure 5.26(c) over a range up to its first harmonic, but also showing the second stage gearmesh frequency. Figure 5.26(b) and (d) show the results of two stages of order-tracking (frequency scale no longer accurate because it is actually an order scale). Both sets of gearmesh harmonics and the harmonics of the high speed shaft (HSS) speed have become single components.

With this degree of removal of speed variation, it was possible to carry out a very fine analysis of the spectral structure over the full frequency range, allowing determination of the numbers of teeth on most gears. Figure 5.27 below shows the results of adjusting a finely tuned harmonic cursor to the set of harmonics thought to be the HSS speed. A potential gearmesh frequency occurs at the 20th harmonic of this shaft (approx. 23.61 Hz). In Figure 5.28 it is shown that this same frequency component coincided to 8 significant figures with the 113th harmonic of another series spaced at approx. 4.178 Hz, confirming that the numbers of teeth in that stage were 20 and 113. The suspected second stage gearmesh occurred at harmonic order 25 of that series, and matched equally well with the 71st harmonic of another series presumed to be at the sun gear speed. Thus, the numbers of teeth in both stages could be deduced blind.

This confirms that as long as amplitude modulation is limited, multi-stage PDOT can give very fine resolution spectra even for substantial speed variation.

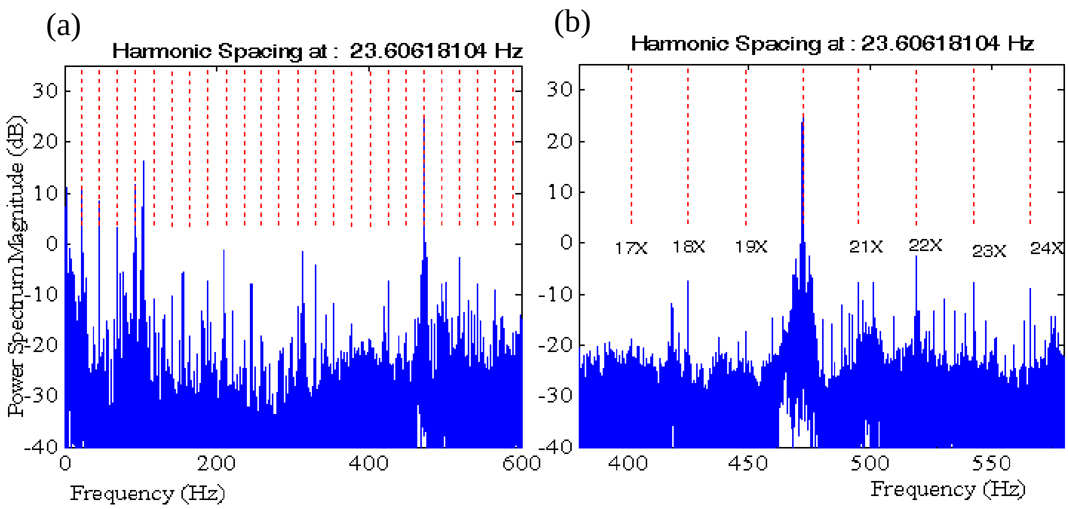


Figure 5.27 – Showing a series of harmonics spaced at 23.61 Hz, which include the final gearmesh frequency (472.1236208 Hz) as harmonic No. 20 (from [45])

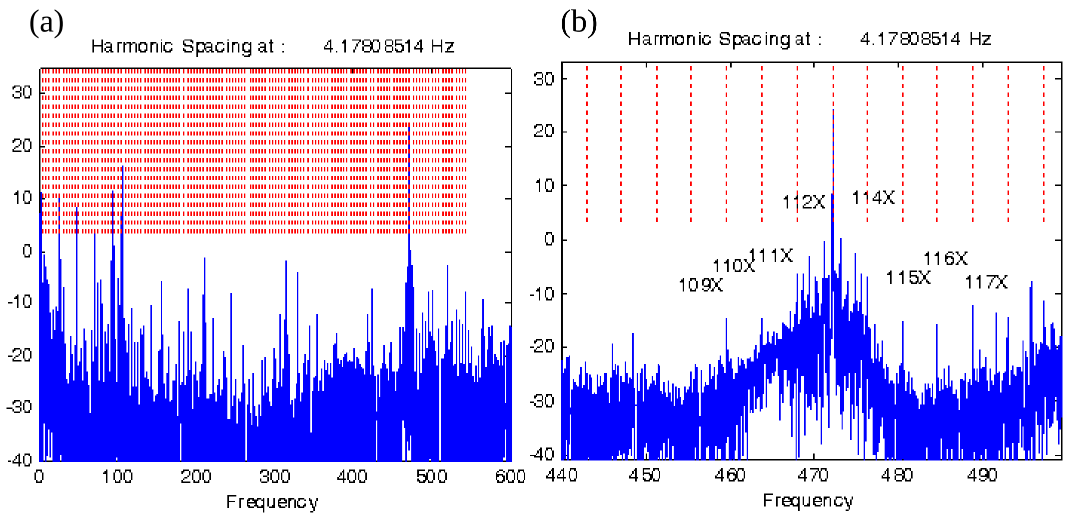


Figure 5.28 – Showing a series of harmonics spaced at 4.178 Hz, which include the second gearmesh frequency (472.1236208 Hz) as harmonic No. 113 (From [45]).

5.5 Segmented order-tracking for a Run-up

For cases where the speed range exceeds those permissible with the basic PDOT method, it is possible to employ the segmented approach to order-track a signal.

This is demonstrated with the following example, which is for a run-up situation. Note that this example has not previously been published.

For the example, a signal was generated to resemble the response on the casing of a machine during a run-up situation. This example was captured from the gearbox casing (with no internals) taken from the gearbox rig used for the examples shown in sections 5.2 and 5.3, starting on page 247. The casing was suspended with elastic and excited with a shaker. A force signal was recorded at the location of the shaker, and an additional acceleration signal was also recorded.

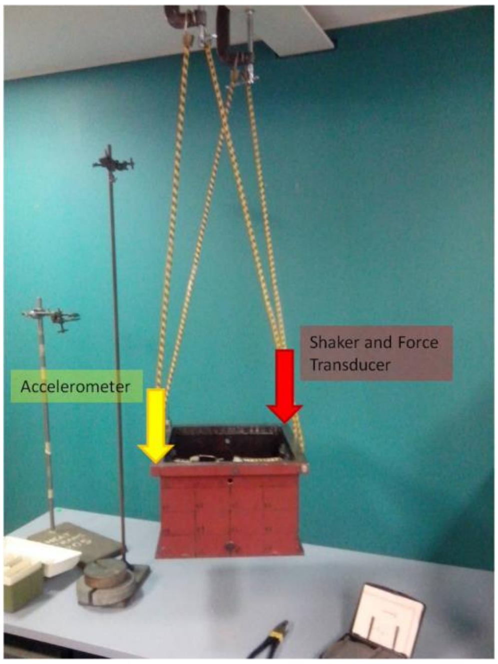


Figure 5.29 – Suspended gearbox casing, showing Shaker, Force transducer, and accelerometer locations

Figure 5.29 shows the gearbox casing suspended with elastic. The position of the shaker used to excite the casing, and the corresponding force transducer location, are indicated by the red arrow on one corner of the top rim of the casing, orientated vertically. The position of the accelerometer is also indicated by the yellow arrow, also on the top rim of the casing and orientated vertically, but on the opposite corner to the shaker position.

The excitation signal used to excite the shaker, was a 30 second pseudo-pulse input, this being formed by adding 10 cosine ‘harmonics’ together to form a pulse signal, of the characteristic form shown in Figure 5.30.

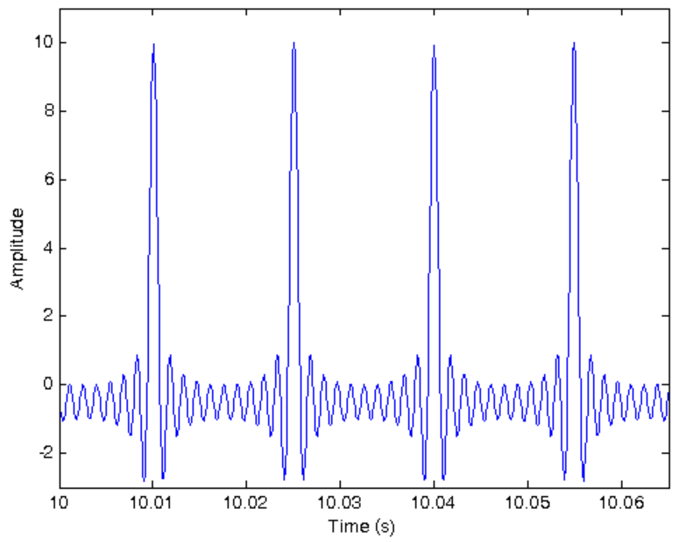


Figure 5.30 – Characteristic pulses of input signal

The excitation signal was modulated with a linear ramp ‘message signal’ as shown in Figure 5.31, corresponding to the casing being excited during a run-up situation from zero to 200 Hz over the 30 seconds. The input signal was constructed with a sampling rate of 10 kHz.

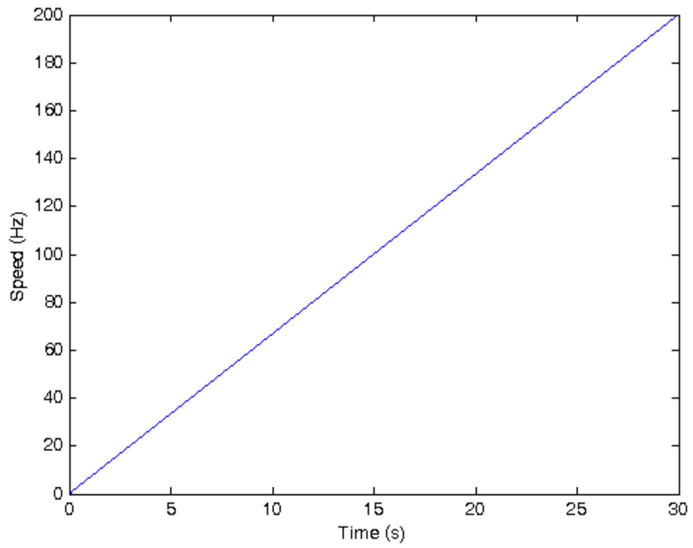


Figure 5.31 – ‘Message signal’, or (non-percentage) speed plot, of excitation signal

As stated previously, both a force signal and an acceleration signal were recorded. Both these signals were captured with a sampling rate of 16,384 Hz.

The force signal was of almost identical form to the excitation signal, and also roughly approximated the signal which would normally be gained from a pulse tachometer, so it was used as the reference signal for the order-tracking process.

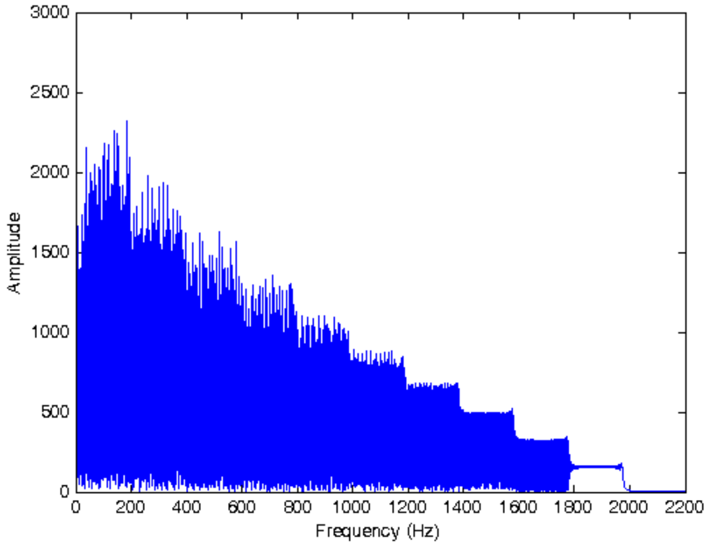


Figure 5.32 – Spectrum of reference (force) signal

Figure 5.32 shows the spectrum of the reference signal, from which it is clear that none of the ten harmonics are separable, as expected with a speed range above ±30%, confirming that the basic PDOT method cannot be directly employed, and a segmentation based approach as presented here is required.

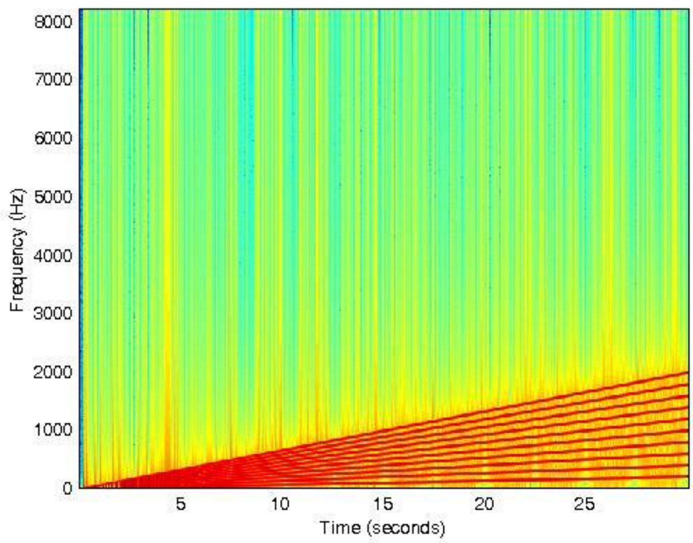


Figure 5.33 – Spectrogram of reference (force) signal

Figure 5.33 shows the spectrogram for the force signal, where the ramp-up speed variation along the signal is more clearly visualised. The ten radial lines emanating from the origin correspond to the ten harmonic orders present in the force signal.

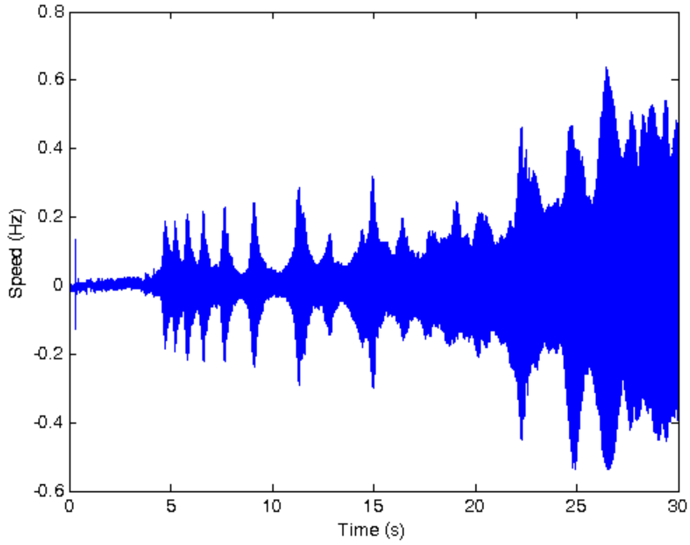


Figure 5.34 – Acceleration signal, time domain

Figure 5.34 shows the time domain plot for the recorded acceleration signal, and the effects of the ramp speed variation are clearly visible as a generalised ramp increase in amplitude of the plot. The effects of structural resonances being excited are also clearly shown by the intermittent peaks present in the time domain plot.

(a) (b)

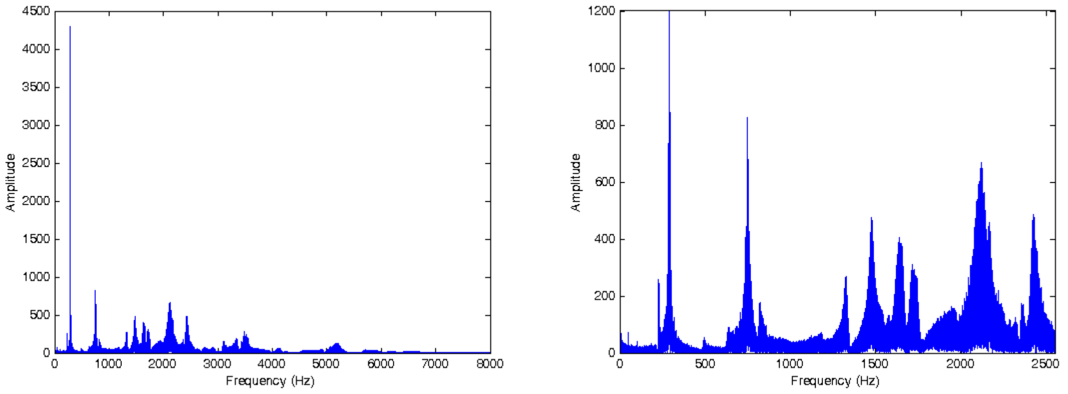


Figure 5.35 – Spectrum of acceleration signal

(a) Overall (positive) spectrum (b) Zoomed spectrum for low frequency

Figure 5.35 shows the corresponding spectrum for the acceleration signal, with (b) being a zoom of the lower frequencies of the full spectrum shown in (a). From Figure 5.35, it can be seen that, as with the reference signal, extensive smearing is present in the spectrum, to a degree that the basic PDOT method could not be employed.

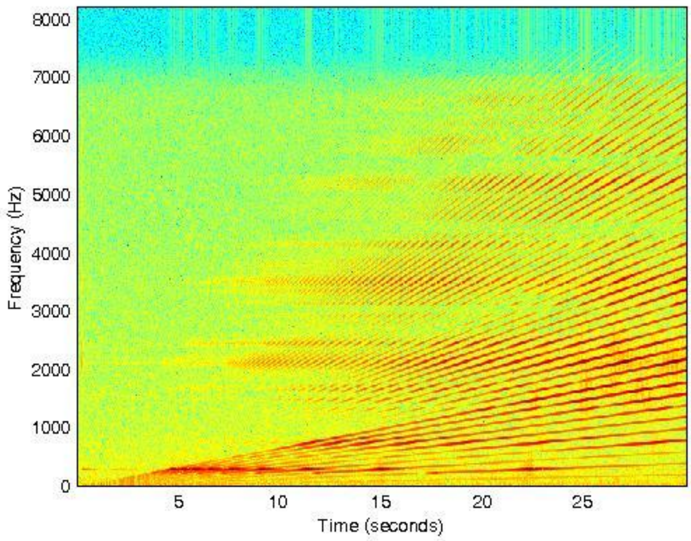


Figure 5.36 – Spectrogram of acceleration signal

Figure 5.36 shows a spectrogram plot of the acceleration signal. As with the force signal spectrogram from Figure 5.33 on page 288, harmonic orders are present as radial lines emanating from the origin. As a result of distortion, a greater number of harmonics have been excited in the response acceleration signal than the ten present in the input signal, resulting in an increased number of visible radial lines. The excitation of constant structural resonance frequencies is also visible in the spectrogram, as horizontal bands. It can also be noted that the effects of the analog low pass filter from the data acquisition system are visible in the upper frequencies, beginning at approximately 7 kHz, where the signal has been progressively filtered leading up to the Nyquist frequency.

Following the procedure set out for segmented order-tracking, as presented in section 4.4 on page 189, the first step for the segmented order-tracking process is to obtain a suitable speed plot.

For this example, the message signal used to modulate the excitation signal was selected to use as the speed plot. This signal roughly corresponds to the output signals, though not exactly, as this signal has not been modified by the transfer function resulting from the shaker and so is not exactly phase-locked to the force and acceleration signals. As such the message signal is a good approximation to the use of a coarse speed plot, which is permissible for use as the speed plot with the segmented order-tracking approach, as described in section 4.4.1 on page 191. In addition, the message signal has a different sampling rate to the recorded force and acceleration signals, which is again permissible with the segmented order-tracking approach.

To convert the message signal to the correct form suitable for use as the ‘speed plot’, the signal was first resampled from the original 10 kHz sampling rate, to the same 16,384 Hz sampling rate common to the force and acceleration signals. As the message signal is non-periodic, ‘ideal’ interpolation in the frequency domain cannot be employed as the resulting signal would be dominated by wrap-around errors. Instead cubic spline interpolation was used for the resampling of the message signal to 16,384 Hz.

In addition, the resampled message signal was changed to have a percentage scale, with the maximum speed corresponding to 100%, to give the ‘speed plot’ to be used for the order-tracking process. The resulting speed plot is shown in Figure 5.37.

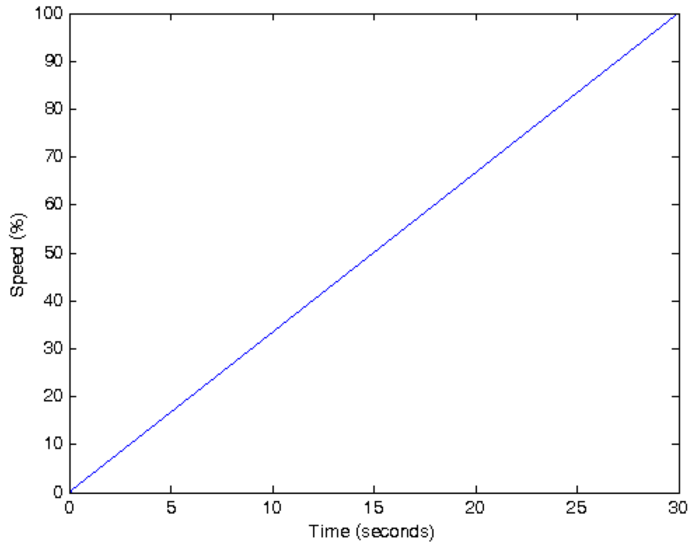


Figure 5.37 – Speed Plot

The second major step of the segmented based order-tracking method is to find suitable segment locations, based on the speed variation parameters.

For this example, the simplified process described in section 4.4.4 on page 216 is used.

The very first step was to select a suitable low-frequency cut-off $f_{cut\%}$, as the speed plot extends down to zero frequency. A nominal lower cut-off value of $f_{cut} = 20\%$ was selected for this example, which on the speed plot corresponds to, approximately, the first 6 seconds of the signal being discounted from the analysis.

Next, to calculate a common value for the maximum modulating frequency value W, two different approaches were presented in sections 4.4.2.2.2 and 4.4.2.2.3, starting on page 206. As the speed plot exactly matches the first typical speed plot shape, as shown in Figure 4.4 (a) on page 197, the most appropriate method to use in this example is to find the equivalent sinusoidal modulating frequency, by matching the average slope, as described in section 4.4.2.2.2 on page 206.

Firstly, the maximum and minimum frequencies from the speed plot are found to be 100% and 20% (after lower bound has been discounted), which when multiplied by the maximum frequency value of 200 Hz, gives the values $f_{\max} = 200\text{ Hz}$ at $t = 30\text{ s}$, and $f_{\min} = 40\text{ Hz}$ at $t = 6\text{ s}$. The time span between these two frequencies is found to be $t_m = 30 - 6 = 24\text{ s}$. The time span value t_m can then be used with equation (4.9) on page 206, to give:

$$W = f_{meq} = \frac{1}{4t_m} = \frac{1}{4 \times 24} = 0.01\text{ Hz} \quad \text{...............................} \ (5.10)$$

To convert W into the maximum percentage modulating frequency value $W_{\%c}$, the most conservative carrier frequency value will be used for this example. As given by equation (4.15) on page 217, this is:

$$f_c = f_{\min} = 40\text{ Hz} \quad \text{...} \ (5.11)$$

With this definition of the carrier frequency, the maximum percentage modulating frequency value $W_{\%c}$ can be found using equation (4.16) on page 217, as:

$$W_{\%c} = \frac{100W}{f_{\min}} = \frac{100 \times 0.01}{40} = 0.025\% \ \ (\text{of } f_c) \quad \text{........................} \ (5.12)$$

For this example, the first harmonic of the reference signal will be used for order-tracking, as this harmonic allows for the greatest speed variations per segment, and hence minimises the amount of segmenting necessary. While it would normally be possible to use Figure 3.7 on page 165 to find a suitable value for $D_{\%c}$ corresponding to the value of $W_{\%c}$, in this case the value of $W_{\%c}$ is smaller than the first $W_{\%c}$ value of 0.1% present in Figure 3.7, which was made with a 0.1% resolution. In the same manner as used to create Figure 3.7, a supplementary acceptability chart showing acceptable combinations of $D_{\%c}$ and $W_{\%c}$, using the first harmonic, with low $W_{\%c}$ values down to 0.01% was created. This supplementary chart is in Appendix A.2.1 on page 427, Figure A.5.

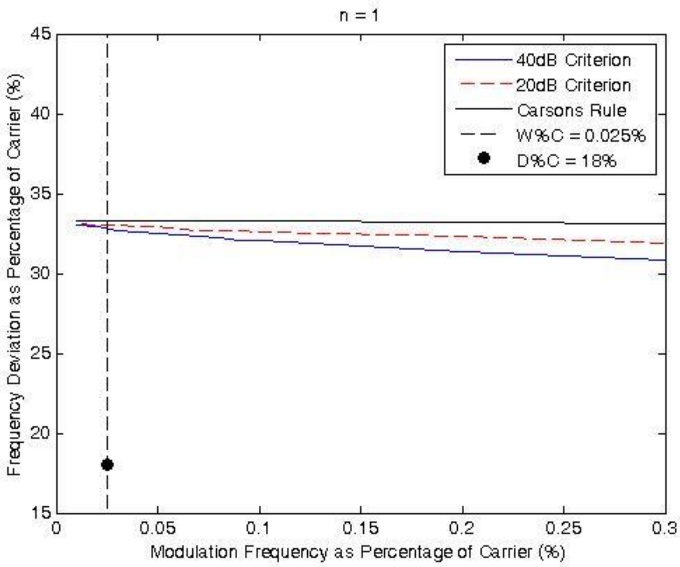


Figure 5.38 – Finding suitable $D_{\%c}$ value on supplementary Figure A.5

Figure 5.38 shows this example case, on the supplementary acceptability chart Figure A.5 for the first harmonic. The vertical dashed line indicates the value $W_{\%c} = 0.025\%$, and as can be seen in Figure 5.38, the limits to $D_{\%c}$ for each of the three criteria at $W_{\%c} = 0.025\%$ are all approximately 33%.

As the goal for the simplified method to calculating segment locations, as described in section 4.4.4 on page 216, is to select a very conservative value for $D_{\%c}$, the value of $D_{\%c} = 18\%$ was chosen to be used in this example. The value for $W_{\%c} = 0.025\%$ and $D_{\%c} = 18\%$ is indicated by the solid black circle in Figure 5.38, and it can be seen that this value lies well below the limits of all three criteria, and so represents a suitable conservative value of $D_{\%c}$ to be used for the segmenting process.

This selected conservative value for $D_{\%c}$ represents the use of a significant factor of safety, and so is suitable to be used to find the breakpoint between each segment (midpoint of the overlapped section connecting each segment), which is one of two alternative methods as discussed in section 4.4.4 on page 216. The final segments will contain a greater speed range than $D_{\%c} = 18\%$ once overlapping is introduced, but the final segments will still have a speed range less than the $D_{\%c} \leq 33\%$ limit indicated by Figure 5.38, due to the selection of a suitable conservative initial $D_{\%c}$ value.

To calculate the breakpoints between each segment, it is suitable to start at the highest speed in the speed plot, being the right most sample of the record. The initial left boundary ('breakpoint') of the right most segment was first found, by using equation (4.21) on page 219, to find the minimum speed of the segment, which corresponds to the left boundary position (breakpoint). Working from right to left (so high frequency to low), the segment breakpoints between each segment were then successively calculated, by using the minimum frequency $f_{\min}$ from the adjacent segment to the right, as the maximum frequency $f_{\max}$, to calculate the corresponding minimum frequency $f_{\min}$ for the current segment. $f_{\min}$ then corresponds to the next break point, or boundary, of the current segment.

This process was continued until a $f_{\min}$ value was found which was less than the $f_{cut\%} = 20\%$ lower frequency cut-off value previously selected. In this case the low frequency cut-off was treated as a guide, and the $f_{\min}$ value was retained for the final segment, so the segment contains a full frequency range of $\pm 18\%$, rather than truncating the segment at the $f_{cut\%} = 20\%$ lower frequency cut-off.

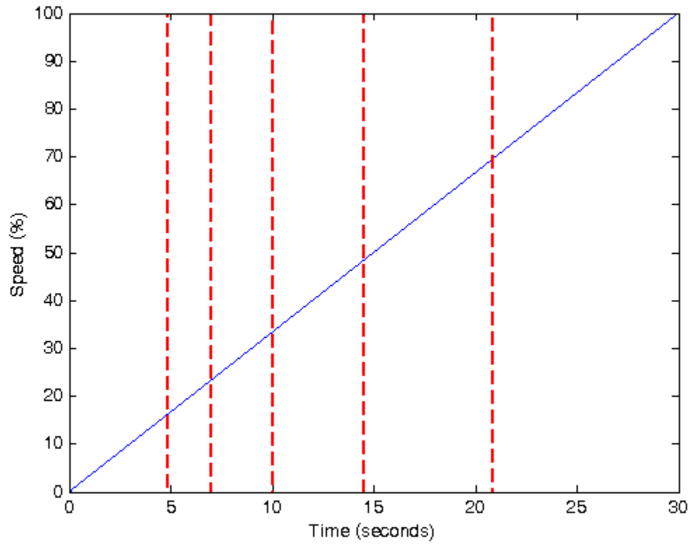


Figure 5.39 – locations of segment 'breakpoints'

Figure 5.39 shows the calculated positions for the segment breakpoints. As can be seen, the choice of a frequency span of $\pm D_{\%c} = \pm 18\%$ results in a total of five segments in this case, spanning between each breakpoint, starting from the right. Note that the sixth span between $t = 0$ and $t = 4.8$ s corresponds to signal below the cut-off frequency, and is discarded from further analysis. It can be seen that the left most segment extends below the initially selected low-frequency cut-off, with the leftmost breakpoint having a value of $f = 16.2053$ %.

For the size of the signal overlap, an overlap percentage of $\pm 5\%$ (of the smaller of the two adjacent spans between breakpoints) around the breakpoint was selected for each end, so 10% overlap at each end. This overlap means roughly 20% of the final segment length consisted of overlapped sections. As will be shown, usage of this overlap percentage gave correct results, and so other overlap percentages were not tested in this case.

(a)

(b)

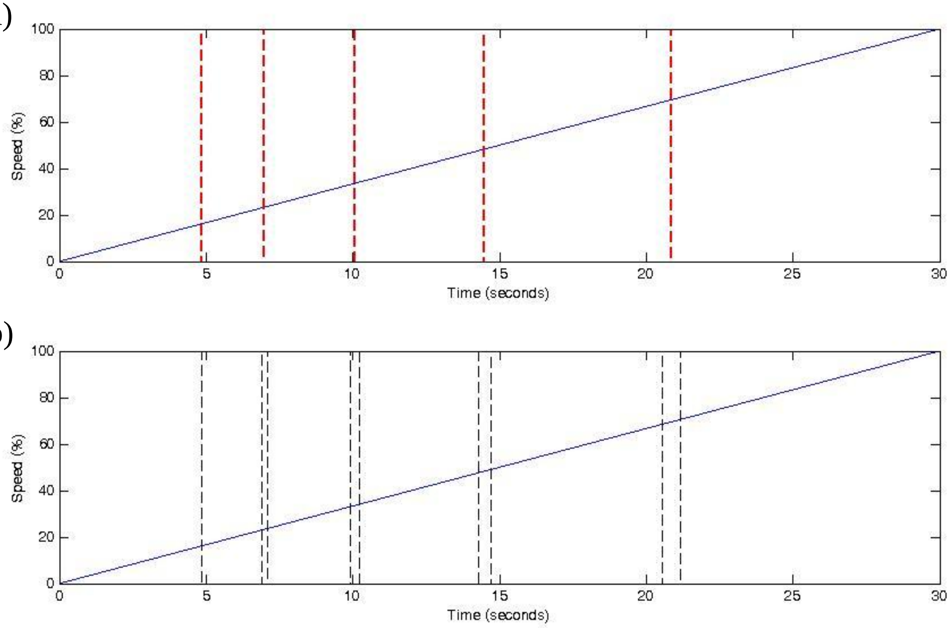


Figure 5.40 – Segment overlap positions

(a) breakpoints from Figure 5.39   (b) Overlap positions around breakpoints

Figure 5.40 (b) shows the locations of the overlap positions, which extend $\pm 5\%$ equally around the previously found break locations shown in (a). Note that as the overlap is a percentage, the overlapped regions become smaller with the corresponding decreasing segment size.

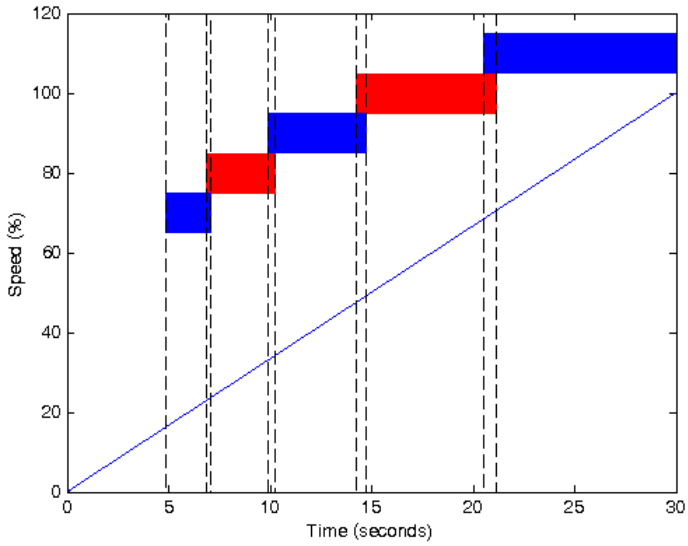


Figure 5.41 – Final segment locations between overlap positions

Figure 5.41 shows the spans of the final segment sizes, with the alternating blue and red bars, which clearly illustrates the overlap of adjacent segments.

It should be noted that for the leftmost segment, the previously calculated breakpoint was simply used as the leftmost segment boundary. As the leftmost segment is not overlapped with an adjacent segment at the left boundary, there was no need to further extend the span of this final segment, which would extend the lowest frequency value even further below the nominally selected cut-off frequency value $f_{cut\%}$.

At this stage, the original signals are broken up into the five separate segments, at the indicated segment boundary positions.

The next step is to window each segment with a half-Hanning window at the ends of each segment, which fully spans the overlapped regions, as described in section 4.4.5 on page 222.

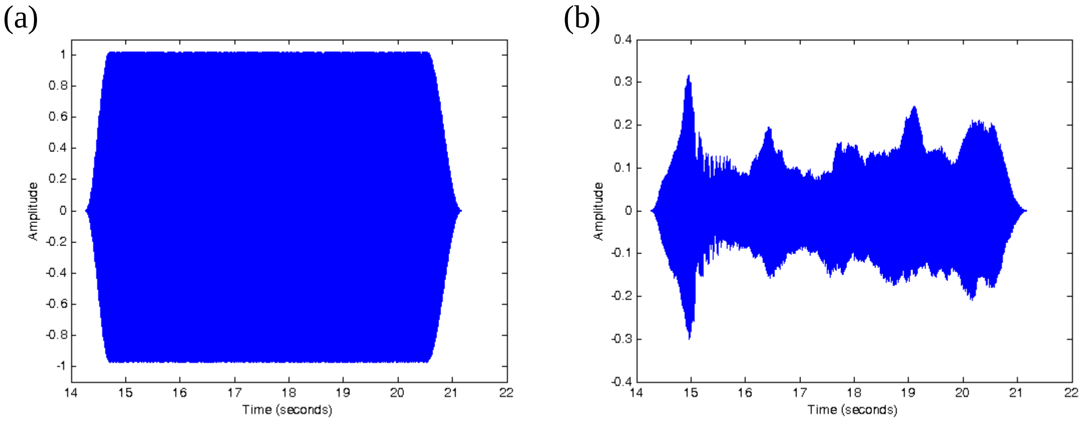


Figure 5.42 – Windowing of segments of data signals
(a) Duplicated force signal (b) acceleration signal

Figure 5.42 shows the windowed data signal segments, for the fourth segment counting from the left in Figure 5.41. Figure 5.42(b) shows the windowed acceleration signal segment, and the effects of the half-Hanning window can be seen at either end, tapering the signal down to zero amplitude. Figure 5.42(a) shows the windowed force signal (individual pulses not visible due to the resolution), where the effects of the half-Hanning window are more clearly visible in the envelope of the signal. It should be noted that the windowed force signal segments in (a) are from a duplicate copy of the force signal. As described in section 4.4.5 on page 222, the force signal segments to be used as reference signals (not pictured) are not windowed with half-Hanning windows, and remain unmodified. This is to prevent problems with the demodulation of the reference signal, which can occur when the amplitude of a signal decreases to zero, which would be induced at the segment boundaries by windowing the reference signal segments.

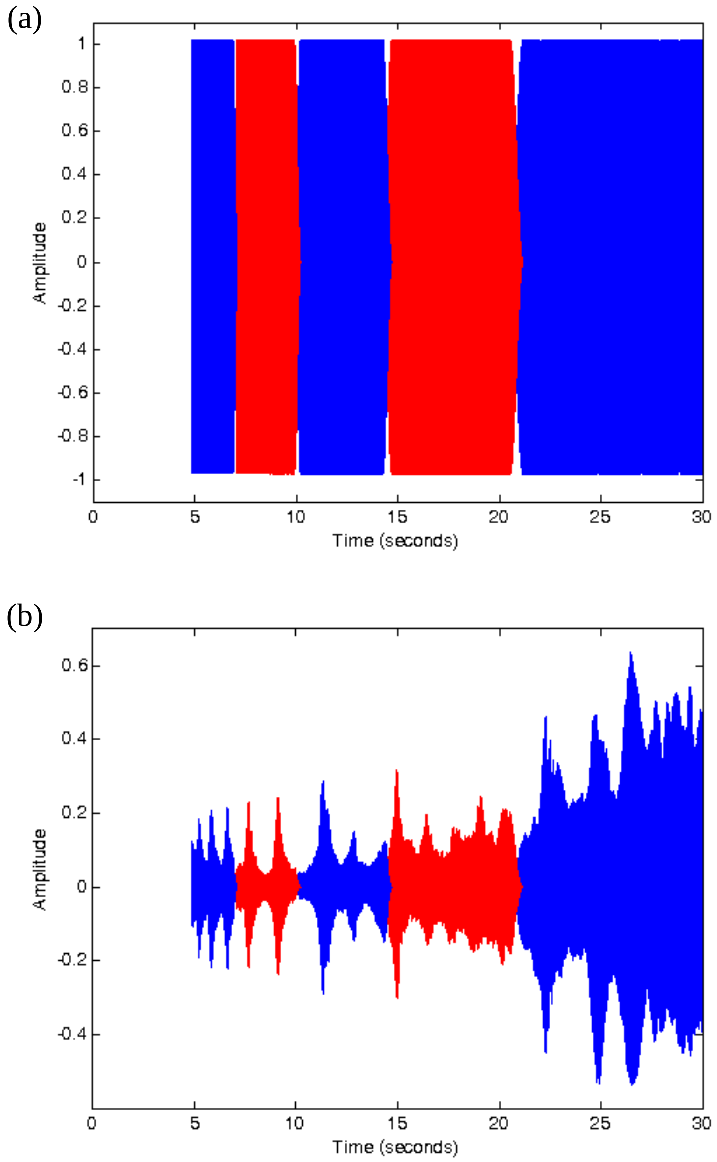


Figure 5.43 – Overlayed windowed segments

(a) Duplicated force signal (b) acceleration signal

Figure 5.43 (a) shows all five segments for the duplicate windowed force signal, overlayed on each other. It can be seen here the windowing has been applied between all the segment boundaries, but not implemented at the start of the first segment or at the end of the last segment, as suggested in section 4.4.5 on page 222.

The next step is to determine a suitable order based sampling rate, to be used for the order-tracking process. In this case, it was chosen to use the minimum order based sampling rate which would not cause aliasing of the highest speed segment, as described in section 4.4.6 on page 227. To calculate the minimum order based sampling rate, firstly the minimum frequency of the highest speed segment was identified, which corresponds to the lower segment limit of the fifth segment counting from the left (rightmost segment). The minimum frequency $f_{\min}$ was found to be 68.4317%, corresponding to $f_{\min} = 136.8635 \text{ Hz}$. This minimum frequency value, along with the data signal sampling rate $f_s = 16384 \text{ Hz}$, can be substituted into inequality (4.23), to give:

$$f_{s\phi} \geq \frac{f_s}{f_{\min}} = \frac{16384}{136.8635} = 119.71 \text{ samples/rev} \qquad (5.13)$$

Hence, the minimum integer order-based sample rate selected for use, which satisfies inequality (5.13) is $f_{s\phi} = 120$ samples/rev (orders).

With this determined order-based sample rate, the appropriate low pass filter thresholds for filtering all segments, except the highest, can be calculated. Firstly, the highest permissible low pass filter threshold value in orders $f_{lpf\phi}$ was selected, so only the minimum amount of signal information necessary is removed, by filtration, to prevent aliasing. By using inequality (4.24) on page 229, the largest permissible value for the low pass filter threshold in orders is $f_{lpf\phi} = 60$ orders (samples/rev).

Then by finding the minimum frequency value $f_{\min}$ for the four remaining segments, equation (4.25) on page 230 was used to calculate the low pass filter thresholds for the four remaining segments.

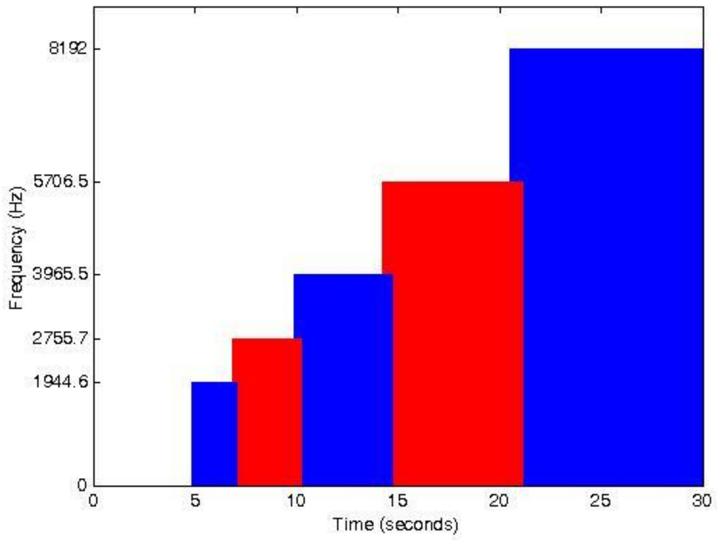


Figure 5.44 – Low pass filter thresholds for segments

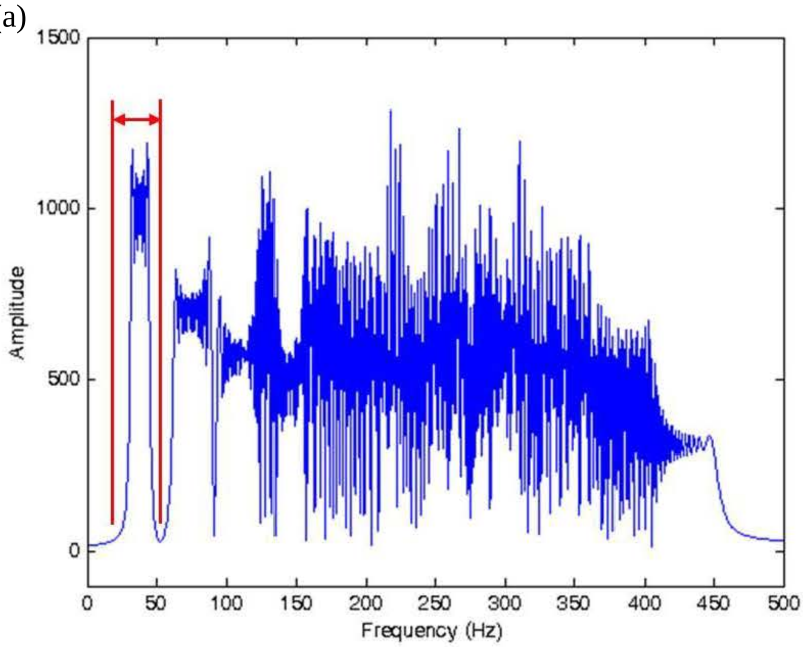
Figure 5.44 shows the calculated low pass filter thresholds for each segment, which is indicated by the maximum frequency value of each pass-band, which are shown as alternating red and blue areas. It should be noted that as the rightmost, highest frequency, segment is unfiltered, it retains the original low pass filter characteristic of the Nyquist sample rate, which is half the original sample rate $f_s / 2 = 8192$ Hz, from the original recording of the data signals.

With these segment limits, the individual segments for each of the data signals can be low pass filtered, using the FFT method described in section 4.4.6 on page 227. It should be noted that, while the reference signal segments were not windowed, it is still necessary to low pass filter these segments. In total, the low pass filtering is conducted on the acceleration signal segments, reference signal segments, and duplicate force signal segments.

As described in section 4.4.6, at the segment overlaps, there is an overlap of different low pass filter levels. This can be seen in Figure 5.44, where the low pass filter ‘blocks’ for each segment overlap. In the overlapped regions, there will only be partial frequency content for the higher frequencies only present in one of the two overlapping segments.

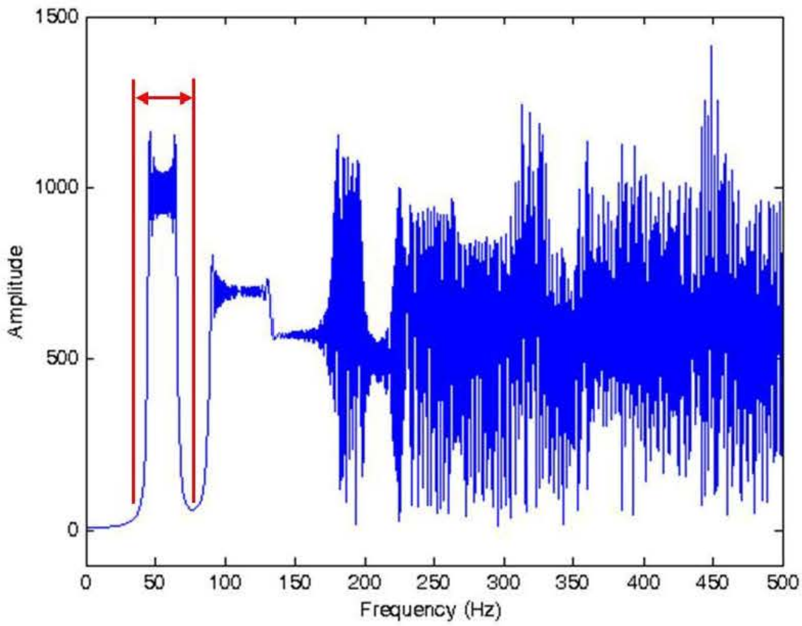
Finally, as a check that the final segments are suitable for order-tracking, the spectrum of the reference signal segments was each examined to check that the first harmonic order in each segment does not overlap with higher harmonics, as occurred in the original reference signal before segmenting as illustrated in Figure 5.32 on page 287.

(a)



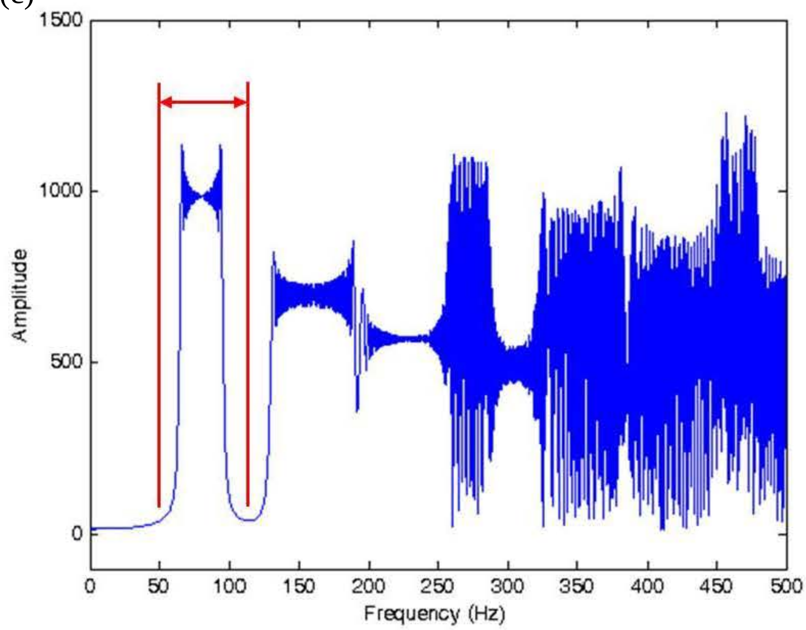
1500
1000
500
0
Amplitude
0
50
100
150
200
250
300
350
400
450
500
Frequency (Hz)

(b)

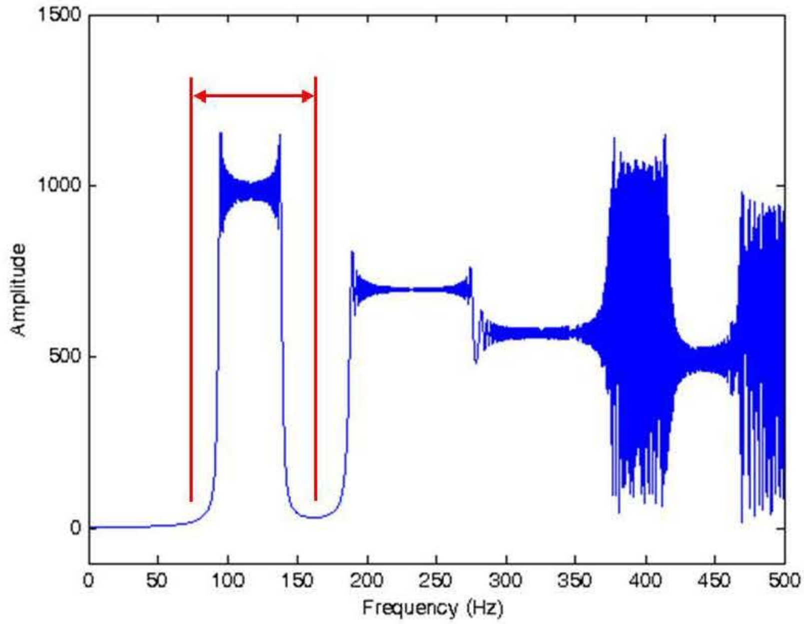


1500
1000
500
0
Amplitude
0
50
100
150
200
250
300
350
400
450
500
Frequency (Hz)

(c)


1500
1000
500
0
Amplitude
0 50 100 150 200 250 300 350 400 450 500
Frequency (Hz)

(d)


1500
1000
500
0
Amplitude
0 50 100 150 200 250 300 350 400 450 500
Frequency (Hz)

(e)

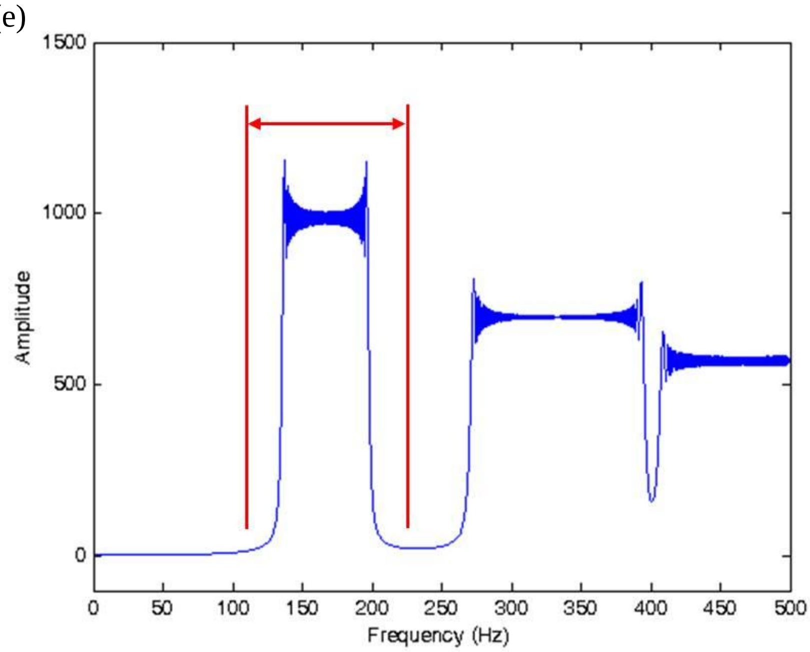


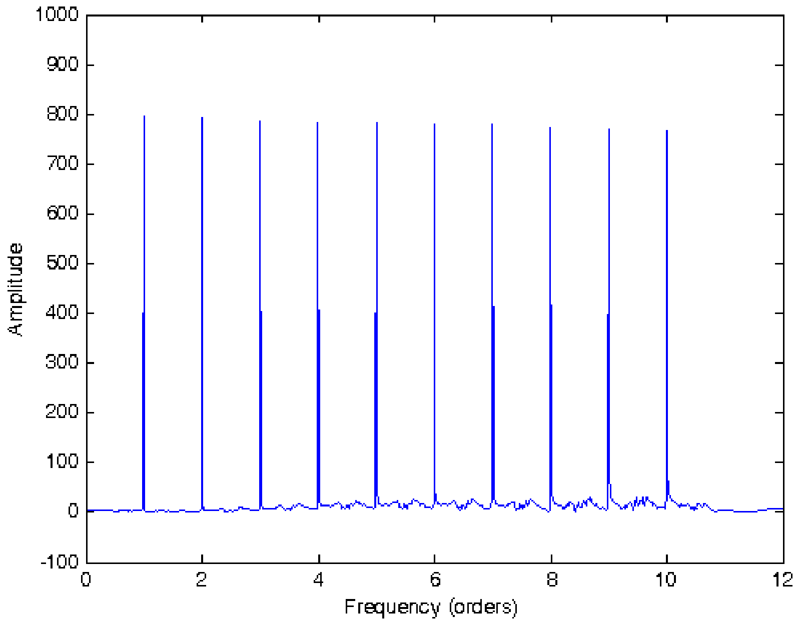
Figure 5.45 – Zoom of the spectrum of reference signal segments, showing demodulation band around first harmonic order
(a) segment 1 (b) segment 2 (c) segment 3 (d) segment 4 (e) segment 5

Figure 5.45 shows the spectrum of each reference signal segment, with the segments numbered from left to right according to their corresponding location on Figure 5.41 on page 298. As can be seen in each of the figures (a) through (e), in each case the first harmonic order is clearly separate from the adjacent second harmonic order, and so can be used for order-tracking. The shifted and increasingly sized speed range of the first harmonic order in each segment is also clearly evident, as the segment number increases, which corresponds to the different and increasing absolute speed ranges present in each segment as indicated in Figure 5.41. On each figure (a) through (e), the span of the demodulation bandwidth to be used for order-tracking, as described in section 3.2.1 on page 131, is indicated in red.

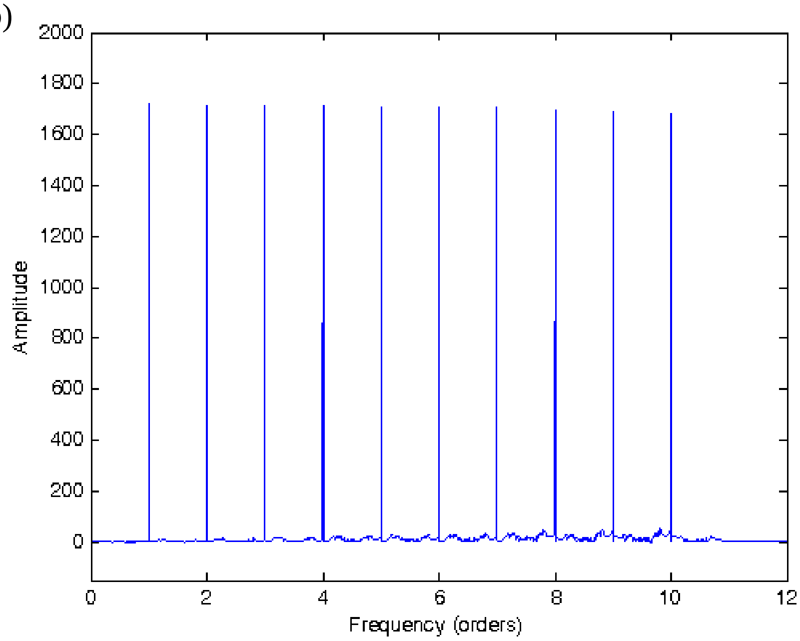
With all the signal segments correctly windowed (as required), low pass filtered, and each reference signal segment successfully checked for suitability, each reference signal segment can be used to order-track the corresponding acceleration and duplicate force signal segments. This is conducted using the basic PDOT method, as described in Chapter 3. However, a modification to the basic PDOT method is required, such that it is required to keep track of the common phase points between each segment, to facilitate subsequently rejoining each segment, as described in section 4.4.7 on page 233. This was accomplished by tracking the phase values corresponding to the previously found break-points, shown in Figure 5.39 on page 296, and then forcing these common phase locations to the same value in each segment during the order-tracking process, as described in section 4.4.7. Using a common phase value between segments, also has the added bonus of ensuring the sampling locations in different segments are kept in alignment, and are not offset from each other (while maintaining the same order based sampling rate).

To check the success of the order-tracking process, firstly the spectra for each reference signal segment were examined. With successful order-tracking, each spectrum will consist of discrete signal components at each harmonic order.

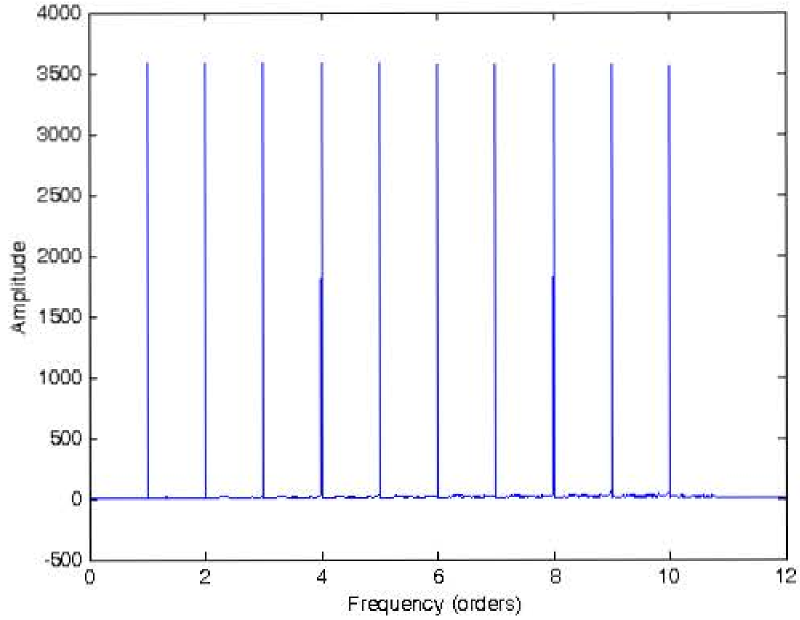
(a)



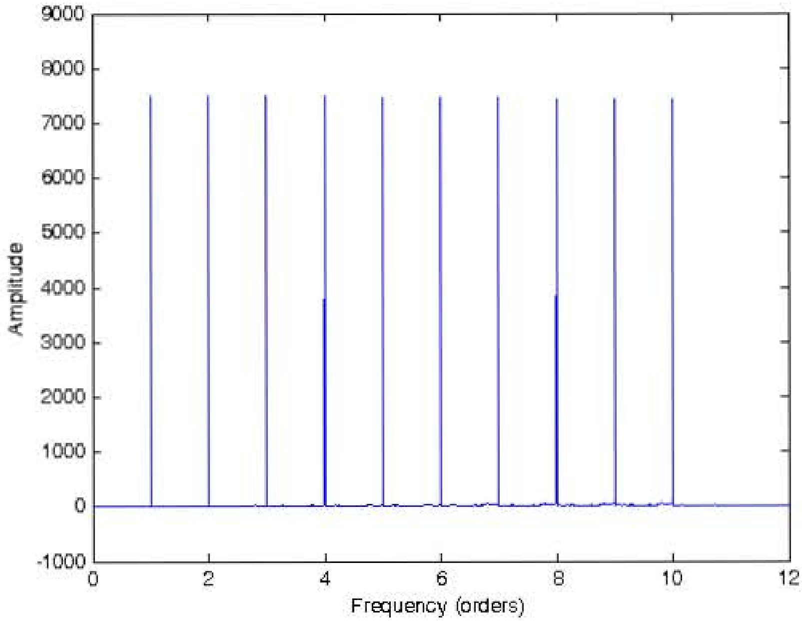
(b)



(c)


4000
3500
3000
2500
2000
1500
1000
500
0
-500
Amplitude
0
2
4
6
8
10
12
Frequency (orders)

(d)


9000
8000
7000
6000
5000
4000
3000
2000
1000
0
-1000
Amplitude
0
2
4
6
8
10
12
Frequency (orders)

(e)

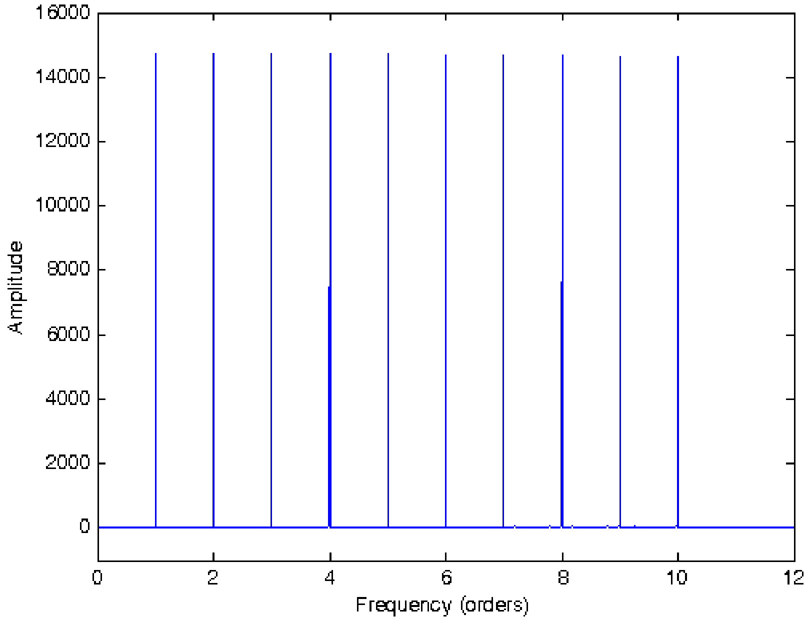


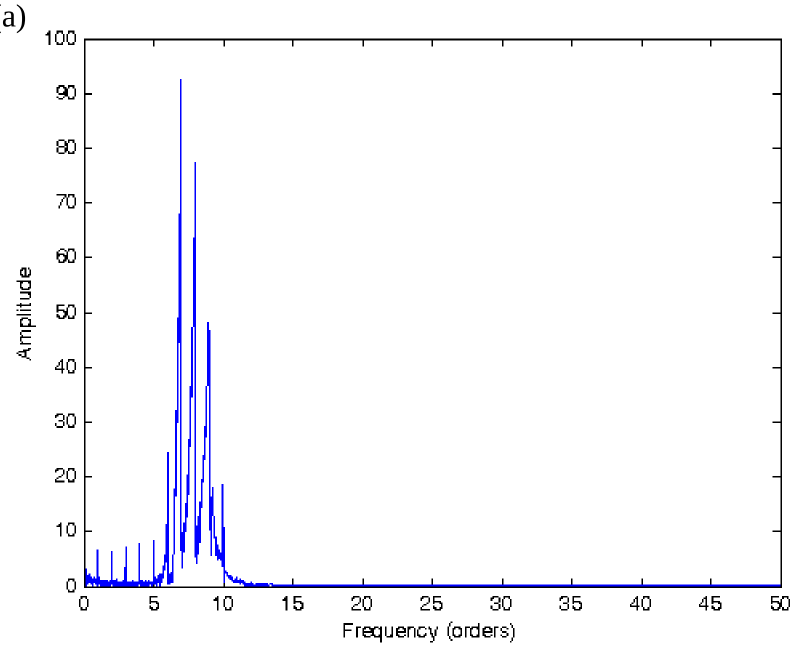
Figure 5.46 – Spectrum of order-tracked reference signal segments, showing all ten discrete orders are present

(a) segment 1 (b) segment 2 (c) segment 3 (d) segment 4 (e) segment 5

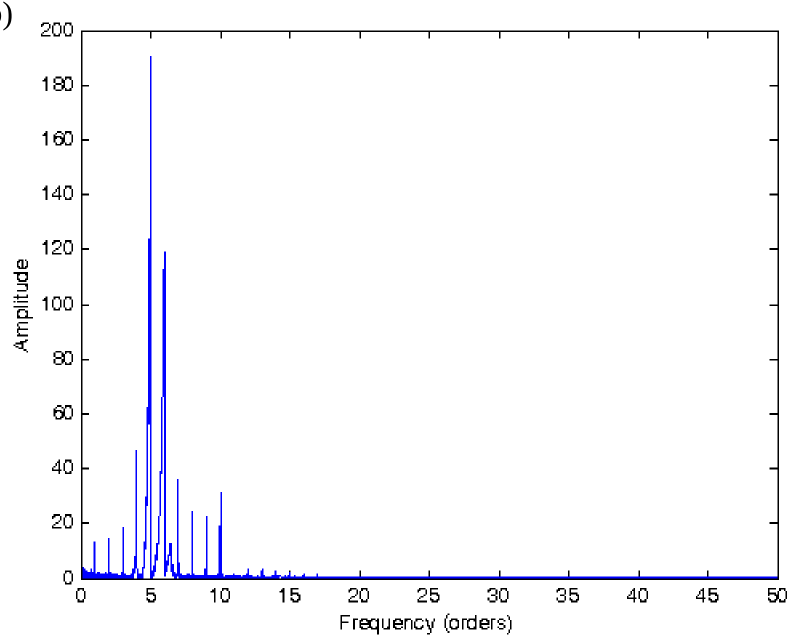
Figure 5.46 shows the spectrum of each reference signal segment, with the segments numbered in the same fashion as with Figure 5.45. The smeared harmonic orders visible in the original spectra from Figure 5.45, have now been correctly order-tracked to discrete peaks in Figure 5.46. In each spectrum in Figure 5.46, discrete peaks are now present at the first ten order values, corresponding to the input signal being constructed from ten harmonic orders, which indicates that each reference signal segment has been correctly order-tracked. It should be noted that the noise level present at non-integer orders, which is visually greatest in the first segment and decreases with harmonic order, in fact remains constant. As higher speed segments are not only longer (in time) records, but also have higher rotational speed, the number of rotations per segment varies with the square of the speed. This in itself increases the signal/noise ratio of the discrete spectrum components, but additionally, as can be seen on the vertical axes in Figure 5.46, the discrete components are increasing in amplitude with an increase in speed, which also increases them relative to the noise.

The next step is to check the spectra of the order-tracked acceleration signal segments, to see the extent to which the smearing of the original spectrum shown in Figure 5.35, on page 289, has been minimised or removed in the segments.

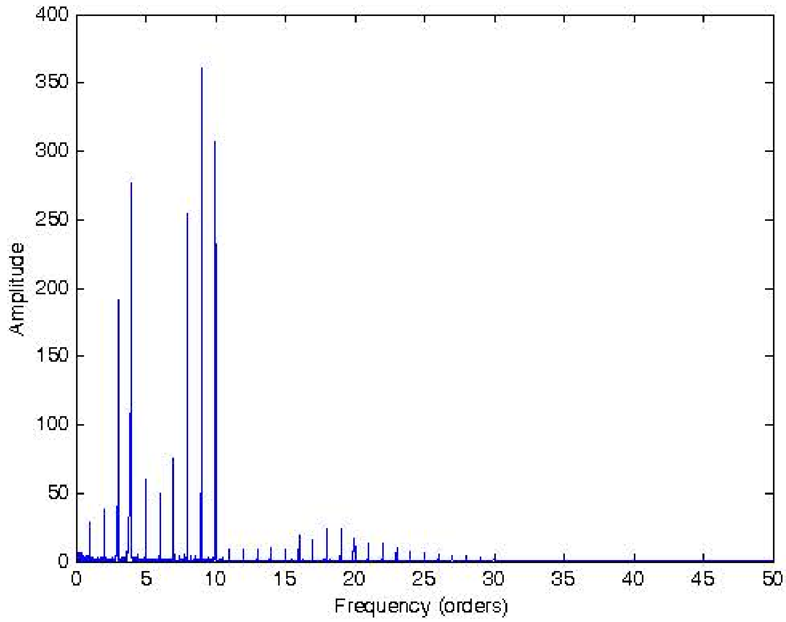
(a)



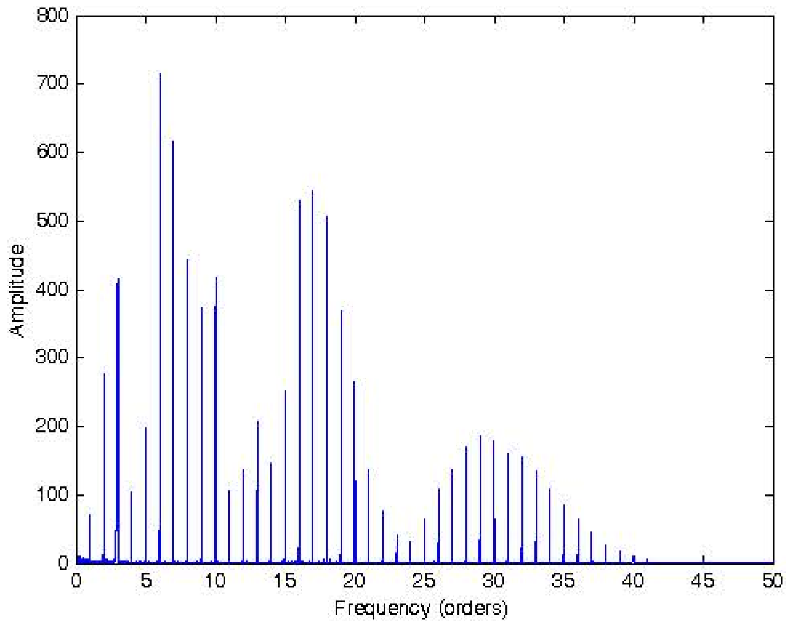
(b)



(c)


400
350
300
250
200
150
100
50
0
Amplitude
0 5 10 15 20 25 30 35 40 45 50
Frequency (orders)

(d)


800
700
600
500
400
300
200
100
0
Amplitude
0 5 10 15 20 25 30 35 40 45 50
Frequency (orders)

(e)

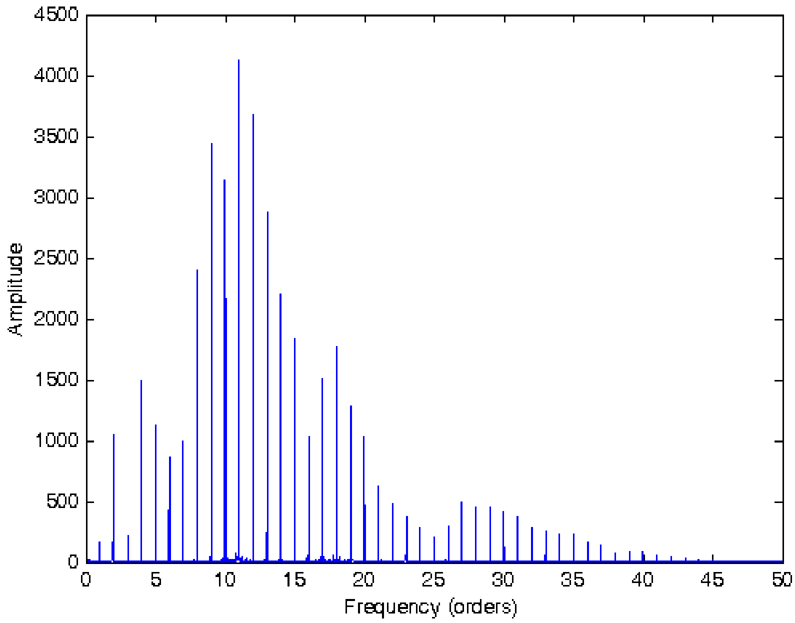


Figure 5.47 – Spectrum of order-tracked acceleration signal segments, showing discrete orders are present

(a) segment 1 (b) segment 2 (c) segment 3 (d) segment 4 (e) segment 5

Figure 5.47 shows the spectrum of each order-tracked acceleration signal segment, again using the same numbering scheme as with Figure 5.45. Firstly, in every acceleration segment (a) through (e), discrete components are now visible at order values, and nearly all smearing has been removed when compared with the original spectrum in Figure 5.35 on page 289, indicating that order-tracking has been successful. The order-tracked spectra all contain significant differences in the number of orders excited, as well as the magnitude of different orders. This is a result of the differing (linear) speed range present in each segment resulting in different (fixed frequency) structural resonances being excited by different excitation orders, which changes which harmonic orders are excited and the magnitude of the response at these orders between different segments. There is more smearing of the response harmonics than of the corresponding reference signal harmonics, which is attributable to the fact that the response signals are subject to amplitude as well as frequency modulation.

With the correct order-tracking of the individual segments, the segments can then be aligned correctly and summed together to form the order-tracked result of the entire original signal (excluding the section below the low frequency cut-off). As stated above, during the order-tracking process the phase values of the break points between segments were recorded to facilitate the rejoining of the individual segments. This allowed for the correct offset of the phase-time relationship for each segment, as used during the order-tracking process, described in Chapter 3.

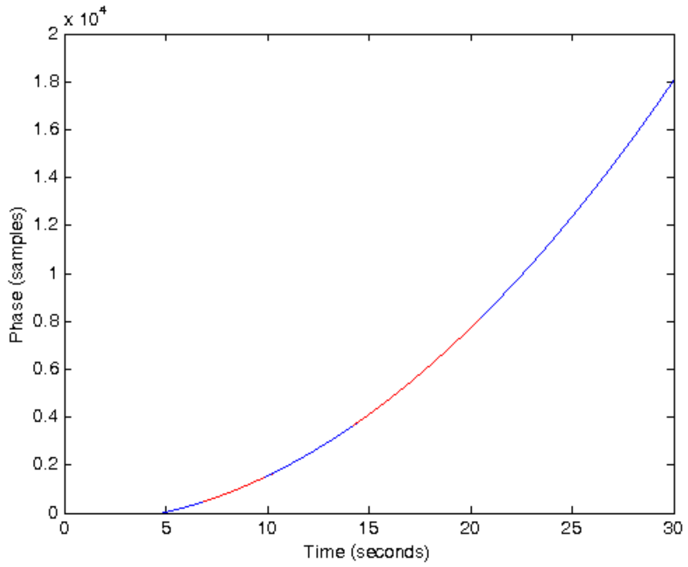


Figure 5.48 – Overlayed time-phase relationships from each segment

Figure 5.48 shows the phase-time relationships for each of the five segments overlaid on each other with the correct phase offset as determined by the common breakpoints. In this specific case, where the speed/time relationship is linear, the phase/time relationship is its integral, a parabola. The phase-time relationships for the individual segments are shown by alternating red and blue lines. As can be seen, with the correct phase offset recorded, for each breakpoint, the separate phase-time relationships all correctly line up, to visually represent the complete phase-time relationship, which could not be extracted from the original complete signal using the basic PDOT method due to there not being a separable harmonic order present in the complete reference signal. It should be specifically noted that the separate phase-time relationships have not been combined in Figure 5.48, but only visually overlayed. The separate phase-time relationships overlap in the overlap regions, which were originally shown on Figure 5.40 (b), on page 297, and Figure 5.41, on page 298. Each of the phase-time relationships contain end-effect errors at both ends, which necessitates the use of windowing the segments in order to minimise the effect of incorrect order-tracking at the end of each segment from the incorrect phase-time relationship, however this is not visible in Figure 5.48 due to the lack of resolution in the figure.

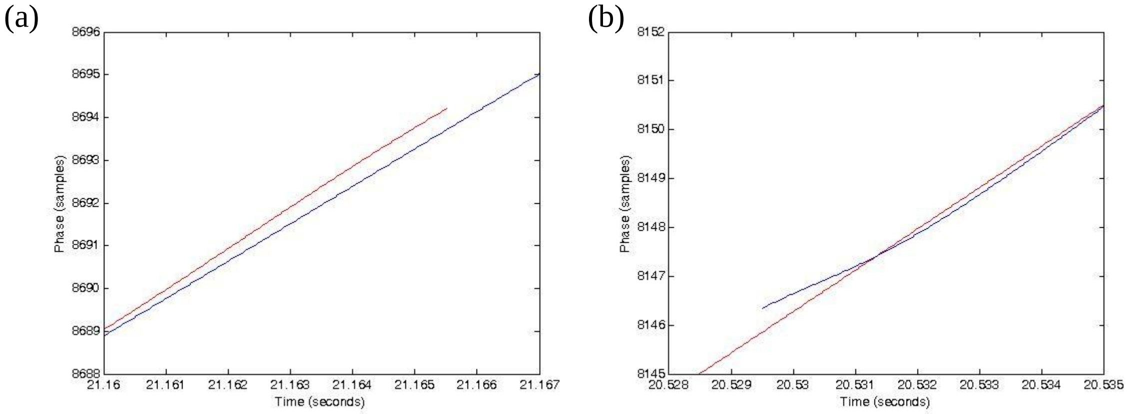


Figure 5.49 – Zoom of Figure 5.48, showing end effect errors on the phase-time relationship for different segments
(a) end of segment 4 (red) (b) start of segment 5 (blue)

Figure 5.49 shows different zoom plots of Figure 5.48, to illustrate the typical end effects present on each of the phase-time relationships from the different segments. In (a), the right hand side, or end, of the phase-time relationship from the fourth segment is shown in red, and has clearly deviated away from the correct phase-time relationship present from the overlapped fifth segment shown in blue. In (b), the left hand side, or start, of the phase-time relationship from the fifth segment is shown in blue, and has clearly deviated away from the correct phase-time relationship from the overlapped fourth segment shown in red.

By using the same phase offsets as with Figure 5.48, the segments for the different signals can be similarly aligned and overlaid.

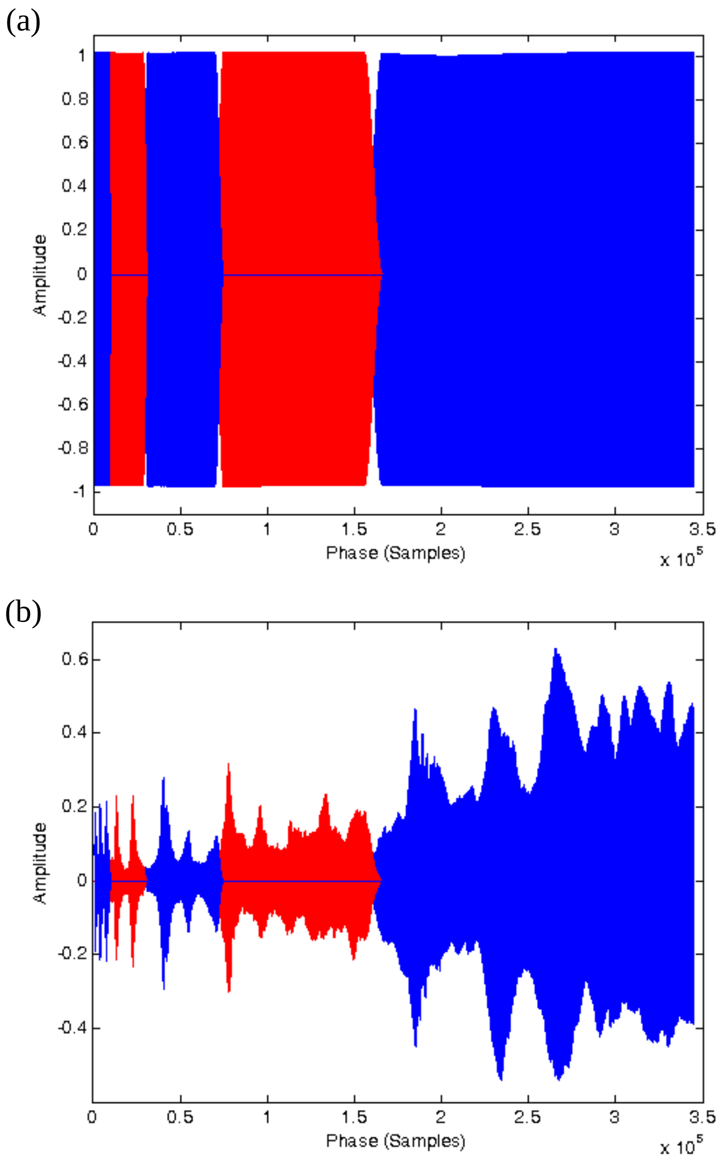


Figure 5.50 – Overlayed order-tracked segments
(a) Duplicate force signal segments (b) acceleration signal segments

In a similar fashion to Figure 5.43 on page 300, Figure 5.50 shows an overlay of the different order-tracked segments, correctly offset in phase. In each figure the segments are shown in alternating red and blue. Figure 5.50 (a) shows the order-tracked segments from the duplicated force signal, which are correctly aligned in the angle domain. As with Figure 5.43 (a), the windowing of each segment in Figure 5.50 (a) can clearly be seen, and it can also be seen that the windowing of neighbouring segments has remained complementary after the order-tracking process. Figure 5.50 (b) shows the order-tracked segments from the acceleration signal, which are also correctly aligned in the angle domain.

With these windowed and order-tracked segments correctly aligned in the angle domain, they can simply be summed together to form the order-tracked versions corresponding to the original signals (minus the section below the low frequency cut-off).

(a)

(b)

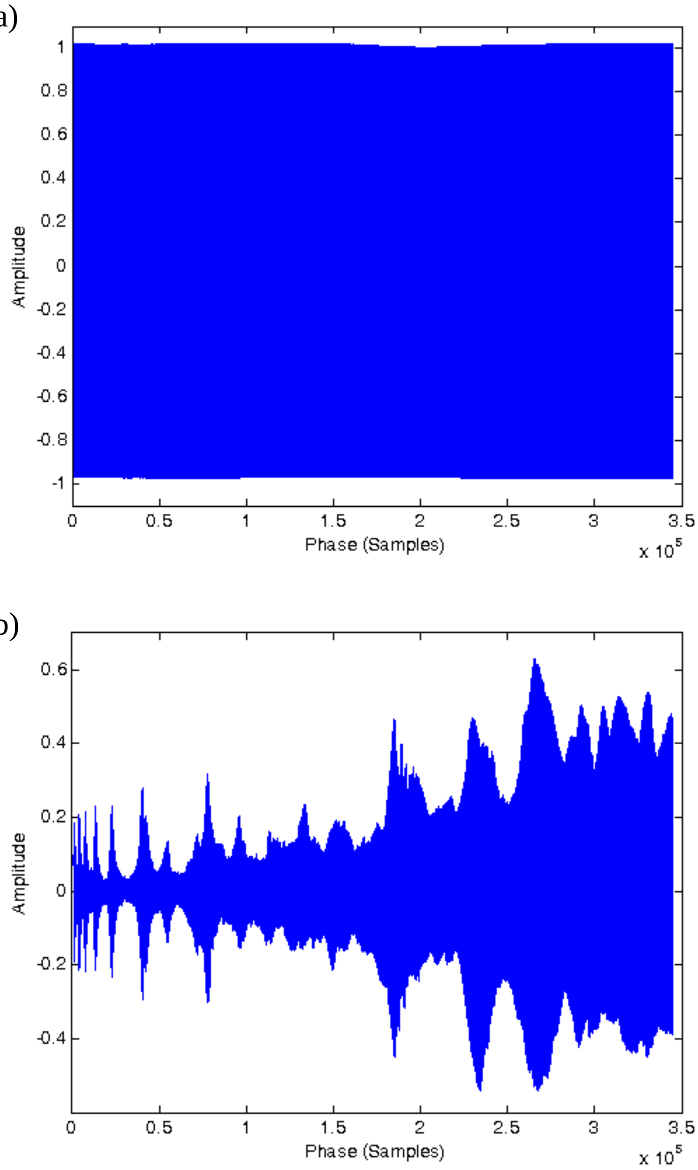


Figure 5.51 – Recombined order-tracked signals

(a) Duplicated force signal (b) Acceleration signal

Figure 5.51 shows the recombined order-tracked signals, with (a) being the recombined duplicate force signal, and (b) being the recombined acceleration signal. As can be seen in both figures, the recombined signals have been correctly joined together and contain no discontinuities. In both cases, the overlapped and complementary regions between segments have been correctly summed together to produce the correct final signal. It should be noted that the slight variation visible in the envelope of the recombined duplicate force signal is visual aliasing, rather than a true change in the envelope of the signal which would indicate incorrect order-tracking.

It should also be noted that the reference signal segments have not been recombined into a complete signal, as they were not windowed.

As a last step, the spectrum of the recombined signals can be checked to confirm that primarily discrete order components are present after the recombining process.

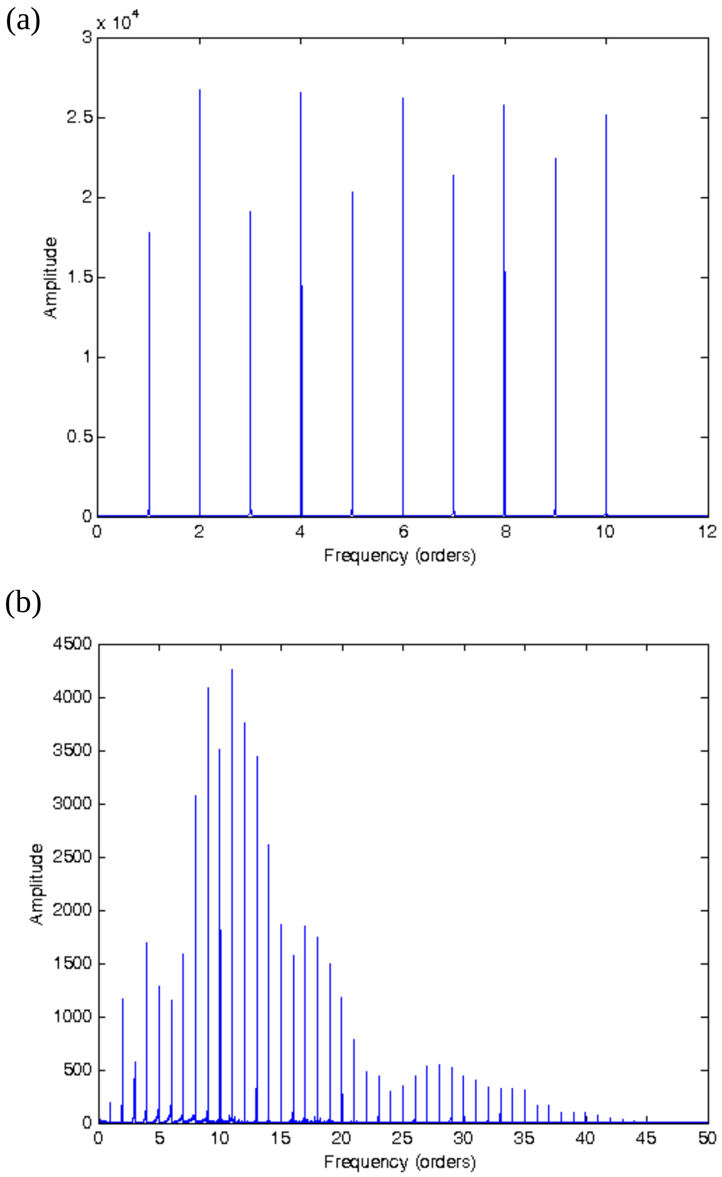


Figure 5.52 – Order spectrum for recombined order-tracked signals
(a) Duplicate force signal (b) Acceleration signal

Figure 5.52 shows the order spectra for each of the recombined order-tracked signals. As can be seen in both figures, discrete order components are again present, indicating that the recombination process has not compromised the order-tracking method. Figure 5.52 (a) shows a very similar format to the spectra shown in Figure 5.46 on page 310. Similarly, Figure 5.52 (b) shows a very similar format to the spectra shown in Figure 5.47 on page 313, except that the minor smearing due to amplitude modulation is much reduced when taken over the whole record.

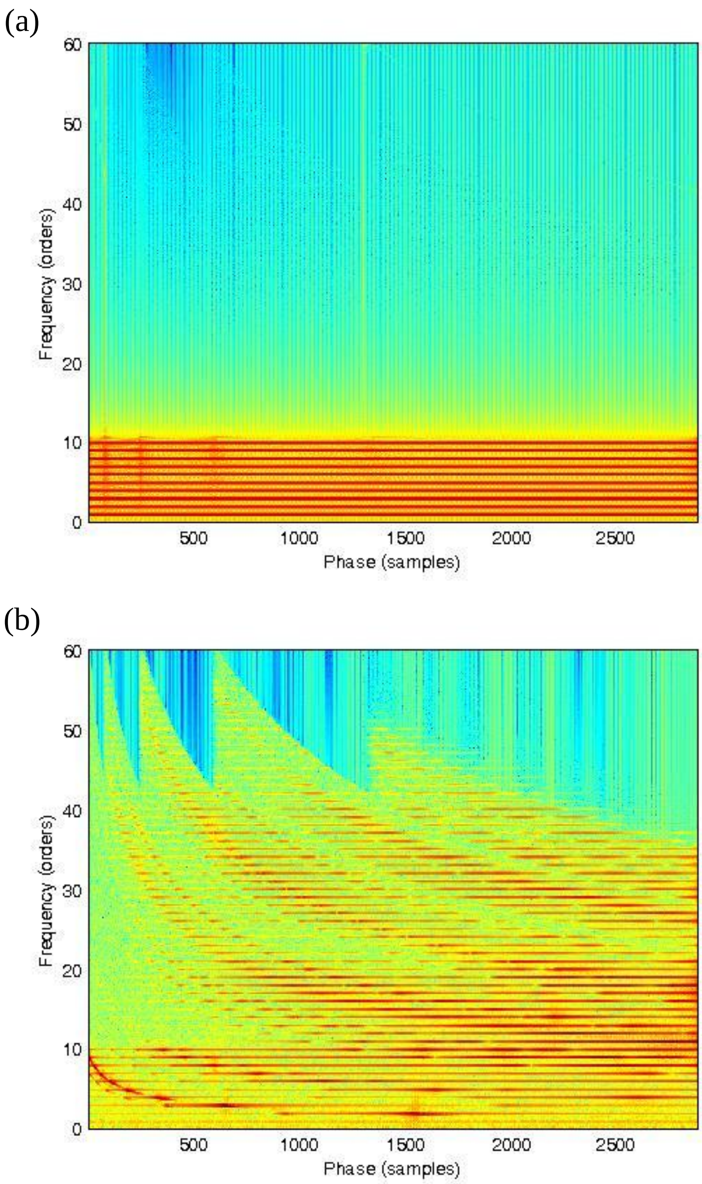


Figure 5.53 – Spectrogram plots of order-tracked signals
(a) Duplicate force signal (b) Acceleration signal

Figure 5.53 shows spectrogram plots for the recombined order-tracked signals, with the duplicate force signal in (a), and the acceleration signal in (b). In contrast to the spectrograms shown earlier for the original signals, in Figure 5.33 on page 288, and Figure 5.35 on page 289, the harmonic order components are now shown as horizontal lines, and the constant (temporal) frequency resonance bands follow a 1/speed trajectory.

In Figure 5.53 (b) the effects of the low pass filtering scheme are clearly visible in the upper frequencies. As described in section 4.4.6 on page 227, the constant (temporal) frequency low pass filter thresholds used result in a saw-tooth filter pattern in the order domain, which is clearly visible in the spectrogram. As can be seen, only partial information is present for the highest orders, with some information removed by the variable (in orders) low pass filters.

As a final step, the recombined order-tracked signals could be low pass filtered with a constant (in orders) filter threshold, before further analysis, to remove the partial order components. This was discussed at the end of section 4.4.6, which starts on page 227. In order to calculate a suitable filter threshold value, equation (4.26) on page 231 was used to calculate the highest complete order $f_{w\phi}$ present for the first four segments (numbered from the left). The value of $f_{w\phi}$ was found to be:

Segment 1: $f_{w\phi} = 41.0590$

Segment 2: $f_{w\phi} = 40.4422$

Segment 3: $f_{w\phi} = 40.4420$

Segment 4: $f_{w\phi} = 40.4421$

As seen in Figure 5.53 (b), each of these thresholds, respectively, matches the first four troughs visible in the low pass filter saw-tooth pattern in the spectrogram.

For the fifth, highest speed, segment, which was unfiltered with the applied low pass filter scheme, the low pass filter characteristics which affect this segment are the original analogue low pass filter characteristics from the original data capture. If the theoretical maximum frequency limit of half the Nyquist frequency is used, so $f_{lpf} = f_s / 2 = 16384 / 2 = 8192$ Hz, a value for $f_{w\phi}$, for the fifth segment, is calculated using equation (4.26) on page 231 as $f_{w\phi} = 41.0683$. However, as can be seen in the spectrogram in Figure 5.53 (b), the fifth and last trough of the low pass filter saw-tooth pattern drops noticeably below 41 orders. This is because the frequency information for the unfiltered fifth segment does not extend all the way to the Nyquist frequency. As was seen in Figure 5.36 on page 290, the analogue low pass filter begins to partially filter the signal above approximately 7kHz. Instead, the rough rule-of-thumb threshold value for an analogue filter cut-off threshold of $f_{lpf} = 0.4 f_s = 6553.6$ Hz, as described in section 4.4.6 on page 227, can be used with equation (4.26) on page 231, to give $f_{w\phi} = 32.768$ orders for the complete orders threshold for the fifth segment. As can be seen in Figure 5.53 (b), this value for $f_{w\phi}$ is now a slightly conservative estimate of the highest non-partial order present for the fifth segment, and can be correctly used.

So, from all five segments, the lowest value of $f_{w\phi}$ is 32.768 orders with the fifth segment, and this value is appropriate to be used as the constant-order filter threshold to low pass filter the entire recombined order-tracked signals to remove partial order components. Note that as the effects of analogue low pass filtration are gradual, which will affect at least one segment from the entire signal with the described filtration scheme, it is suggested that a constant-order filter threshold level is not directly estimated from a spectrogram plot, such as Figure 5.53 (b), by visual inspection, and the calculations just shown here be employed.

Please note that the results of employing a constant order low pass filter, as described above, to remove partial order components are not illustrated here, as they would be visually almost identical to the results already shown.

Chapter 6 Experimental testing of PDOT, for reversible order-tracking

Chapter Overview

The following chapter presents a novel method whereby a variable signal can be pre-processed in order to improve subsequent OMA analysis.

In this method, a signal is order tracked, and then edited using Cepstrum techniques to remove order based components, and then finally reversed back to the time domain before subsequent OMA analysis.

This demonstrates the success of using the PDOT in a reversible fashion.

In addition, while the presented editing scheme is presented in conjunction with PDOT, it could also be employed with any other suitable reversible order-tracking method, and is novel in its own right.

6.1 Reversible order-tracking for Operational Modal Analysis

The following example shows how a reversible approach to using PDOT can be used for pre-processing of variable speed vibration signals, for the application of Operational Modal Analysis (OMA). Note that the example presented here will be limited solely to the pre-processing step, which will produce signals which should be suitable for use with any subsequent (general) OMA method.

To begin, a brief overview of the application is needed to give context to the work.

6.2 Overview of Modal analysis, including both experimental and operational modal analysis

Firstly, Modal analysis describes methods which focus on determining the dynamic structural properties of a structure excited by an input, for example where a structure is vibrating. Modal analysis has been extensively described in literature, such as in [46], and so only a brief description is presented here.

System vibration, as a function of frequency, can be represented as:

$$F(f)\times H(f)=X(f) \quad \text{...} \quad (6.1)$$

where $F(f)$ is the input, corresponding to the force input to the system, commonly called the forcing function. $X(f)$ is the response output, corresponding to an acceleration output. Finally, $H(f)$ is the system structural properties, and is called a Frequency Response Function (FRF) when expressed as a function of frequency.

With this definition, modal analysis firstly focuses on obtaining the FRF $H(f)$. The FRF can then be further separated into a series of components, known as resonance frequencies, corresponding to individual modes (of vibration).

Experimental Modal Analysis (EMA) refers to modal analysis methods where both the forcing function $F(f)$ and the response function $X(f)$ are directly measured, allowing $H(f)$ to be found directly using equation (6.1). $H(f)$ can then be decomposed to give the modes of vibration of the structure. It should be noted that different transmission paths on a structure result in different FRFs, and the presence and level of individual modes of vibration varies with the distribution of the excitation. To determine all the modes of vibration for a given structure, it is common to make an array of measurements from different transmission paths on a structure, to obtain multiple FRFs. The multiple FRFs can then be analysed to find all the modes of vibration of the structure.

In order to conduct EMA, it is necessary to use known forcing functions which can be directly measured. This is commonly accomplished using one of two alternative methods. For the first method, the input is provided by an impact with an impact hammer. An impact hammer has a force transducer fitted at the tip of the hammer, so the forcing function provided by the impact can be directly measured. For the second method, the input is provided by a shaker, which is a device directly coupled to the structure which provides an input force. As with the impact hammer, a shaker has a force transducer fitted at the coupling point with the structure, so the input force can be directly measured. Note that both these methods give an input force at a single location (and direction). In order to ensure that the structure is not excited by any additional unknown forces, and give known boundary conditions, it is preferable to elastically isolate the structure, which is commonly done by suspending the structure with elastic cords. The force applied by the shaker should also be in one direction only, and with no accompanying moments.

To obtain an array of different transmission path measurements, in order to obtain an array of FRFs, two alternative approaches are commonly used. The first approach uses a fixed common response measurement location (fixed location for $X(f)$), and then the location of input force is varied (varied location for $F(f)$). The second approach does the opposite, where the input force location is fixed, and the response measurement location is varied.

EMA has the advantage that calculations are relatively simple, as the input force is measured. Applying the input force has a possible disadvantage however, as the structure needs to be excited by known forces in an experimental fashion. This can be unrealistic, as the structure may be excited by a force quite different from that encountered in operation. This may also be impractical for structures where it is difficult to apply externally generated forces or implement elastic isolation, which could be due to space considerations where external excitation and elastic suspension cannot be easily implemented, or due to size considerations where a structure is simply too large to be practically excited at a point source, or be suspended elastically, such as with a bridge or building.

Note that EMA can be made more robust by applying multiple forces simultaneously and using mathematical methods to separate the responses to each input.

An alternative approach to modal analysis is Operational Modal Analysis (OMA). In OMA, a structure is excited during normal operation by the unknown operating forces, rather than being subject to known experimentally applied forces as with EMA. OMA has the advantage that a structure is measured in its operating condition, and so determined structural properties should not differ from the operating case, as they can with EMA. In addition it is not necessary to apply external forces (requiring less equipment), or to elastically isolate the structure (preferable with EMA).

The disadvantage of OMA is that the input forces are not measured, as typically a structure is not excited by single point sources which can be measured by a transducer. As with EMA, it is necessary to measure an array of transmission paths; however, as the input force is unknown, unlike with EMA, it is only possible to vary the measurement locations for the response measurement for OMA.

When considering equation (6.1) on page 327, for OMA only the response signal $X(f)$ is measured. As a result, to determine the structural properties it is necessary to estimate the value for the FRF $H(f)$, and hence the structural modes, from the combined values of $F(f)\times H(f)$. As a result, the unknown (not directly measured) input force can have a significant impact on the success of estimating the value of $H(f)$ with OMA methods. In some cases an input force component can mask a resonance, so it is missed during the OMA process. Conversely, the discrete frequency input force components can sometimes be mistaken for a resonance, and false modes can be identified during the OMA process.

6.3 Pre-processing to improve operational modal analysis, for variable speed situations

One branch of research to improve OMA methods, which is not covered here, has focused on developing improved OMA methods which are insensitive to the presence of discrete frequency input forces.

A different approach to improve the implementation of OMA is to pre-process the response signal, in order to filter out discrete frequency components in the input $F(f)$, before OMA is then implemented on the remaining filtered signal, excited by broadband excitation. A number of filtering approaches have been developed to filter a significant portion of the discrete frequencies from a response signal $X(f)$, for situations where the input signal has many discrete components, such as occurs during the constant speed operating state with machinery. Some examples of filtering methods include frequency bin editing, Time Synchronous Averaging (TSA), Linear prediction, Self-adaptive noise cancellation (SANC), Discrete/random separation (DRS), and Cepstrum editing, most of which are summarised in [2].

For cases where machinery is operating with variable speed, an input signal is no longer dominated by discrete frequency components, as was discussed in section 1.1.1 on page 3. In this instance filtering methods used for constant speed situations cannot be immediately applied. As discussed in this work, with variable speed cases, a response signal can first be order-tracked. A filtration method suitable for constant speed signals can then be utilised on the order-tracked response signal, which now contains discrete components. While this is suitable for separating shaft order components from the rest of the response signal, it is not then appropriate to conduct

OMA on the filtered order-tracked signal, as OMA can only correctly be applied to response signals in the time domain (excepting the case of order-tracking with nominally constant speed, where the time domain and order domain signals would be almost identical). In order to conduct OMA analysis on the filtered order-tracked response signal, it is necessary to first reverse the order-tracking and transform the filtered signal back to the time domain.

6.3.1 Summary of prior works, with reversible order-tracking

It is not a new concept to use a reversible order-tracking method to facilitate pre-processing of response signals, for variable speed cases, before subsequently applying OMA. The concept has been specifically presented in some new works, and older works have also presented a significant portion of the concept.

In 1996 Lembregts et al. [15-18] (note all four papers are practically identical) first presented the concept of employing order-tracking in a reversible fashion, though it should be noted that only a cursory methodology was given for the reversing process, and no information was given in regards to what order-tracking method was reversed. While these authors did not present the concept of using a reversible order-tracking process for the application of pre-filtering a signal to better observe structural components, they made two independent statements, which if considered together would give rise to the concept. The authors firstly identified that in a variable speed vibration response signal, order-based and fixed-frequency-based components can more clearly be seen in the order-domain, and frequency domain, respectively, and so a reversible order-tracking process allows a signal to be interchangeably transferred between both domains in order to observe each type of component more clearly in the appropriate domain, in a complementary fashion. The authors specifically give resonance frequencies as an example of a fixed-frequency-based component which could be observed in this fashion. Separately, the authors identify that order-based frequency components can be modified in the order-domain, and then the modified signal transformed back to the time domain for further analysis. Logically, combining these two statements gives rise to the concept of editing order-domain components in order to better observe fixed-frequency resonances in the time domain. It should be specifically noted that Lembregts et al. did not make this link in their work.

In 2005 Groover et al. [14] presented a method which employed a reversible approach to order-tracking, using a time domain based order-tracking method, in order to remove order-based frequency components using a frequency bin editing method applied in the order domain, so that fixed frequency resonances could then be analysed in the time domain without masking by order components. This was presented for the application of monitoring the fixed-frequency blade natural frequencies of turbine blades as a structural health diagnostic feature, where the blade natural frequencies were masked by order-based components which needed to be removed. While it was not identified in this work, the final filtered signals produced by the described method could, in principle, then be used to conduct OMA for the structure (in this instance a turbine). It should be noted that Groover et al. [14] presented the concept of using a reversible order-tracking approach as being novel, and did not identify, or make reference to, the earlier work by Lembregts et al. [15-18] which first proposed the concept, at least in a cursory fashion. The precedence of the basic concept being first presented by Lembregts et al. was only later identified by Peeters et al. [47], whose work is discussed below.

In 2007 Peeters et al. [47] presented a method to use reversible order-tracking to remove order-based components from a varying speed response signal, in order to pre-filter the signal, before then using OMA. While this is very similar to the work proposed by Groover et al. [14], it represents the first instance of reversible order-tracking being specifically and explicitly employed for the application of OMA. The method presented by Peeters et al. [47], while not specifically stated, based on the descriptions in the paper appears to employ a time domain based order-tracking method, similarly to the method employed by Groover et al. [14]. In order to filter out order-based components, Peeters et al. make use of time synchronous averaging (TSA), applied in the order domain, which is in contrast to the work by Groover et al., which used frequency bin editing.

6.3.2 Evaluating prior works, for the application of pre-processing a variable speed signal before OMA

For the work presented here, in addition to utilising PDOT instead of a time domain based order-tracking method, it was also desired to improve any other steps in the process for the application to OMA, if possible. The prior works by Groover at al. [14], and Peeters et al. [47], were evaluated for the OMA application, and both methods were identified as using sub-optimum filtering methods, which could be improved upon.

In both cases, amplitude modulation had not been sufficiently taken into consideration. In general, for machines which operate with varying speed, the vibration response signal contains an amplitude modulation component, in addition to the frequency modulation component from the varying speed, so the combined effect from the input signal in the response is given by a general modulation of a harmonic series carrier, as discussed in section 2.5.6 on page 114, rather than solely by frequency modulation of a harmonic series carrier. Typically, the amplitude modulation component is periodic, and has the same period as the reciprocal of the modulating frequency of the frequency modulation component, and the magnitude of the variation of the amplitude modulation component increases as the speed variation increases (typically by traversing resonances). This can be seen in the experimental results shown in section 5.3, which starts on page 255. Specifically, amplitude modulation is visible in the response signal in Figure 5.9 (d) on page 256 for a 10% speed variation, with the same period (inverse of modulating frequency) as the frequency modulation component. Additionally, a larger amplitude variation is visible in Figure 5.15 (d) on page 264 for the larger 25% speed variation case. When the general modulated input signal components are order-tracked, the frequency modulation component is changed to being constant, and the resulting input signal component in the order domain becomes an amplitude modulated signal. The spectral components for the input signal, in the order-tracked response signal, will consist of a harmonic series of discrete ‘carrier frequency’ components, and each carrier frequency is surrounded by a family of sidebands, which in general would be smeared since the reason for the amplitude modulation is not necessarily periodic in the order domain. In order to successfully remove the order related signal components from the order-tracked signal, it is necessary to completely remove the amplitude modulated harmonic series carrier component.

In Peeters et al. [47], TSA applied in the angular domain was used as the filtration method. TSA has the property that it is tuned to a single period of repetition over which the average is conducted, corresponding to a single discrete frequency. The averaging process then filters out the single discrete frequency, as well as any exact integer multiples of the single discrete frequency, e.g. a single harmonic series is filtered, when the TSA is tuned to the fundamental harmonic. When TSA is applied to an amplitude modulated harmonic series carrier in the angular domain, tuned to the carrier frequency, only the ‘average’ amplitude is filtered from the signal (in the angle domain), and all amplitude variations around the average, which should be filtered out for correct filtering, remain. When viewed in the order spectrum, TSA filters out the carrier frequency from each harmonic order, but all the amplitude modulation sidebands remain. As a result, TSA is not suitable for completely filtering an amplitude modulated harmonic series carrier component from the rest of a response signal (in the angle domain), which generally is the input signal type present in the angle domain from a varying speed machine. Hence, TSA is not appropriate to use for the application of pre-filtering response signals before using OMA, for typical varying speed situations.

It should be noted that in [47], Peeters et al. do show the successful implementation of TSA in the angle domain to filter the input signal for three examples, however none of the examples contains the commonly present amplitude modulated harmonic series carrier component as an input in the angular domain, and so the examples do not prove the applicability of the use of TSA as a filtering method in the general case for varying speed. In the first example, a simulated 6 Degree of Freedom (DOF) system was excited by both white noise, and a frequency modulated harmonic series carrier with +/- 4% frequency deviation, to produce a simulated response signal. Once order-tracked, the frequency modulated harmonic series carrier component in the response signal became a discrete harmonic series in the angular domain, which could be successfully filtered using TSA due to the absence of amplitude modulation. In the second example, a response signal is recorded from a helicopter in-flight. In this example no value is given for the amount of speed variation. A plot is given showing an estimation of the instantaneous frequency for the helicopter (Figure 10 in [47]), however this plot has no values or zero level indicated on the amplitude axis, and so the magnitude of the speed variation cannot be determined from this plot. A plot of the corresponding estimated phase is also given (Figure 8 in [47]), and while this plot also has no amplitude values indicated, the phase variation is seen around the linear phase progression of the ‘carrier frequency’, and the variation has a comparatively tiny magnitude, indicating that this example likely has a speed variation of jitter only, and is for nominally constant speed. As amplitude modulation generally is greater for larger amounts of speed variation, as noted earlier, no amplitude modulation is expected when a machine is operating with nominally constant speed. The interpretation that this case is for nominally constant speed is reinforced by a plot of the response signal spectrum (Figure 11 from [47]), which is dominated by a single harmonic series with slight smearing, with no sidebands present. While it is not clarified if this plot of the response is either a spectrum or order spectrum (before or after order-tracking), for the case of a non-trivial speed variation, sidebands would be expected in either plot type (either from amplitude modulation in the order spectrum, or from general modulation in the temporal spectrum), and the absence of sidebands reinforces that this example is for a nominally constant speed case. For the third example, a response signal was taken from a large diesel engine running at a stationary state, so again a case with nominally constant speed, where TSA would be expected to work correctly. Note that for this third example, no results are presented, and only a qualitative statement is given stating that the TSA filtering process and subsequent OMA analysis worked. As such, while TSA was successfully employed in all three example cases, the absence of amplitude modulation from all cases means these successes using TSA cannot be extrapolated to conclude that TSA is applicable for general speed variation cases, as implied by Peeters et al. [47].

In 2005 Groover et al. [14], frequency bin editing in the order domain was used as the filtration method to remove order based components. For the described method, the order based components are assumed to be a discrete frequency for each harmonic order, and so the possibility of accompanying amplitude modulation of the varying frequency component was not considered. With the method, either one, or multiple, harmonic orders of the order-based component are removed from the signal by editing their frequency value down to the local noise level. Two alternative editing processes are described for the frequency editing step. For the first alternative, the replacement edited value for each order is calculated using linear interpolation between the immediately adjacent samples (e.g., for a sample N being edited, the linear interpolation is conducted between values from samples $(N-1)$ and $(N+1)$). The interpolation is conducted separately for the real and imaginary components, so the final edited value is the combination of the calculated real and imaginary interpolated values. The authors found this first alternative method caused incorrect editing for signals with well-defined phase shifts, so a second alternative method was given where the edited value was simply taken as the minimum of the two adjacent (complex) samples. In [14], both a simulated and experimental example are given showing the successful removal of order components before transformation back to the time domain. However, as with [47], both examples in [14] did not contain amplitude modulated order components. For the simulated example, a single constant (temporal) frequency component was added to a single FM sinusoid which has what appears to be a ramp up frequency variation. In this instance, the order component resolved into a single discrete peak after order-tracking, which was successfully removed by editing a single frequency bin in the order spectrum. For the experimental example case, a torsional vibration signal was captured from an experimental rig. This case was conducted for nominally constant speed, and so as described above for the experimental case for [47], as amplitude modulation of an order component is typically greater for greater frequency variation, a nominally constant speed signal would normally not have amplitude modulated order components. The spectrum for this experimental case shows only discrete order components with smearing due to jitter, with the corresponding order spectrum containing a harmonic series of discrete order components which were successfully removed with frequency bin editing. As with the TSA filtering method employed in [47], the frequency bin editing scheme described in [14] would not be successful in completely filtering an amplitude modulated order domain component from a signal, which is commonly present with typical frequency variation of machinery. In a similar fashion to TSA, the described frequency bin editing method would only remove the ‘carrier frequency’ component and incorrectly leave amplitude modulation sidebands in the signal.

However, unlike with TSA, the described frequency bin editing scheme in [14] could be extended to remove amplitude modulation sidebands, however this extension would very likely result in significant disruption of the spectra. For order spectra which contain multiple discrete amplitude modulation sidebands, the sidebands could be edited in the same fashion as the discrete ‘carrier frequency’ components. This however would be a laborious process to correctly target the bin of every single sideband surrounding every harmonic ‘carrier frequency’, as the sidebands typically significantly outnumber the discrete carrier frequencies. The amount of editing locations required in the order spectrum would also cause increased disruption of the signal, when compared to the editing of only discrete components as in [14]. For more general amplitude modulation, where the amplitude modulation component consists of a contiguous block (of unspecified frequency content) surrounding the carrier frequency, as described in 2.5.2 starting on page 77, the entire sideband block could theoretically be edited down to the surrounding noise level, however this would most likely result in significant disruption of the order spectrum with large swathes of signal being replaced by a constant patch of spectrum at the noise-level.

As stated previously, for the method presented here it was decided to use an alternative editing method, rather than a method used by the previous works, as both of previous filtering schemes would not be ideal when applied to amplitude modulated order components. The TSA method used in [47] would simply not work for amplitude modulated components, as described above. The frequency bin editing method used in [14] could be extended to work with amplitude modulated order components, as described above, however this was expected to result in significant spectral disruption and so would not be ideal.

6.4 Cepstrum domain editing

For the method presented here, a Cepstrum editing technique was chosen to remove order-domain content, in order to pre-filter a signal before application of OMA. The chosen cepstrum editing technique had been used previously, in 2012, by Randall et al. [48] to pre-process a signal before using OMA, for constant speed applications where order-tracking was not used. The cepstrum editing technique was modified in this work to be applicable for use with varying speed signals.

6.4.1 Overview of the Cepstrum

The most general definition of the Cepstrum, is that it is the inverse Fourier transform of logarithmic spectral information. This is expressed as:

$$C(\tau) = \mathcal{F}^{-1}\ \log\left[X\ \ f\ \right] \quad \text{...} \ (6.2)$$

where $C(\tau)$ is the Cepstrum, $\mathcal{F}^{-1}$ is the inverse FFT ($\mathcal{F}$ being the forward FFT), and $X(f)$ is a function containing spectral information, in terms of frequency.

The primary advantage of the Cepstrum, is that components which are a multiplication in the frequency domain become additive in both the log spectrum and the cepstrum. The result is that components that are naturally coupled together in the spectrum, by multiplication, are separated from each other in the Cepstrum. For this application, it is desired to separate out input signals from structural information, which as per equation (6.1) on page 327 are a multiplication in the spectrum, and so are additive (and often separate) in the cepstrum, and hence are much easier to separate. In the cepstrum domain, at least for a single forcing function, the modal properties tend to be concentrated at very low “quefrency” (defined below). Frequency modulated forcing functions will be smeared in the time domain, but discrete in the order domain, but in any case generally at higher quefrencies than the modal information.

Additionally, for this application, it is desired to filter out amplitude modulated input signals (in the order spectrum). This is made easier with the cepstrum, because all related periodic structures in the spectrum, are separate in the cepstrum. For modulated structures from the spectrum, which consist of two related periodic structures (carrier frequency harmonic series, and surrounding sidebands), these periodic structures are separated from each other into unrelated structures in the cepstrum, which allows for much easier editing.

Due to these two advantages, the use of Cepstrum editing was evaluated as being preferable to using a modified version of the frequency bin editing method (based on the spectrum), based on the method used in [14].

As a result of the structures in a Cepstrum quite closely corresponding to structures seen in the spectrum, and that the very earliest definition (no longer used) for the Cepstrum was as a "spectrum of a spectrum", a quirky naming convention was utilised in the first publication of the cepstrum [49], which is still employed almost identically in modern works and will be used in this work. With the naming convention, structures in the cepstrum, which are similar to those seen in the frequency spectrum, are named by reversing the first syllable of the corresponding frequency domain term to form the cepstral term. For example, the cepstral structure corresponding to frequency is titled "quefrency", with the first syllable reversed, and Cepstrum itself is formed by reversing the first syllable of "spectrum". Table 6.1 shows a summary of the cepstrum terms used in this work, matched to the corresponding spectrum term. Note that the corresponding term to the frequency domain was originally defined as "quefrency domain", but is more commonly referred to as the "cepstral domain" in later works.

Frequency Spectrum term	Corresponding Cepstrum term
spectrum plural, spectra	cepstrum plural, cepstra
frequency	quefrency
frequency domain	quefrency domain (original definition [49]) cepstral domain (modern , common, name)
harmonic (s)	rahmonic (s)
filter	lifter
low pass filter	short pass lifter

Table 6.1 – Cepstrum terms corresponding to spectrum terms with first syllable reversed

The Cepstrum domain has properties and structures which are similar to the time domain. The Cepstrum is gained by an inverse Fourier transform from the frequency domain, so similarly to the time axis in the time domain, the quefrency axis of the cepstrum has units of seconds. Similarly to the time domain, where the period in the time domain is the reciprocal of the corresponding frequency from the frequency domain, in the Cepstrum quefrency is the reciprocal of the corresponding frequency from the frequency domain. A similarity that the Cepstrum shares with the frequency domain is that the Cepstrum also contains positive and negative halves.

In the Cepstrum domain, as stated previously, components such as the ‘carrier frequencies’ and sidebands from amplitude modulation, which are coupled in the frequency domain, are separated out in the cepstrum. In the Cepstrum, any periodically repeated components in the frequency domain (regardless of whether they are harmonic, i.e. have an integer multiple coincident with zero frequency, such as a harmonic series, or are non-harmonic, i.e. do not have an integer multiple coincident with zero frequency, such as sidebands) manifest as a rahmonic series, which is a series of repeating peaks in the Cepstrum, where the zero'th peak of the series corresponds to zero quefrency, so of similar form to a harmonic series in the frequency domain. As the Cepstrum is analogous with the time domain, the rahmonic spacing in the Cepstrum is analogous to period (or delay time, as for the autocorrelation function) in the time domain, and so rahmonic spacing is the reciprocal of the spacing between components in the spectrum. For a harmonic series, the spacing between components is typically the fundamental frequency, and so the rahmonic spacing of the corresponding rahmonic series would be the reciprocal of the fundamental frequency. For non-harmonically repeated components, such as sidebands, the spacing between the components in the spectrum (e.g. sideband spacing) is inverted to give the corresponding rahmonic spacing in the Cepstrum.

6.4.2 Real and Complex Cepstrum

Two common versions of the Cepstrum exist in the literature. The first version is titled the ‘Complex Cepstrum’, and is the inverse Fourier transform of the log of the complex spectrum. If the basic definition of the Cepstrum is taken as equation (6.2) on page 344, then the complex cepstrum $C^{cp}(\tau)$ is given by equation (6.2), when $X(f)$ is defined as the full complex spectrum, such that:

$$X(f) = \mathcal{F}\left[y(t)\right] = Y(f) \quad \text{...} (6.3)$$

where $y(t)$ is a function in time, and $Y(f)$ is the corresponding complex function in frequency. As $Y(f)$ (and hence $X(f)$ with this definition) is complex, it can be expanded in polar form, such that:

$$X(f) = Y(f) = A(f)\exp\left(j\phi(f)\right) \quad \text{...................................} (6.4)$$

where $A(f)$ is the amplitude of $Y(f)$, and $\phi(f)$ is the phase of $Y(f)$.

Substituting equation (6.4) into (6.2) gives the complex cepstrum $C^{cp}(\tau)$ as:

$$C^{cp}(\tau) = \mathcal{F}^{-1}\left\{\log\left(A(f)\exp\left[j\phi(f)\right]\right)\right\} \quad \text{...............................} (6.5)$$

Equation (6.5) can additionally be further expanded (or simplified, depending on perspective) to:

$$C^{cp}(\tau) = \mathcal{F}^{-1}\left\{\ln\left[A(f)\right] + j\phi(f)\right\} \quad \text{..................................} (6.6)$$

Note that even though the spectral information is complex, the resulting complex cepstrum is actually real, since the log amplitude spectrum is even and the phase spectrum odd.

The second common version of the Cepstrum is titled the 'Real Cepstrum', and is the inverse Fourier transform of the log of the amplitude of the spectrum. Hence, the real cepstrum $C^r(\tau)$ is given by equation (6.2) on page 344, when the spectral information term $x(f)$ is given by the amplitude $A(f)$, such that:

$$|X(f)| = A(f) \quad \text{...} (6.7)$$

where $A(f)$ is the amplitude of $Y(f) = \mathcal{F}[y(t)]$. Substituting equation (6.7) into (6.2) gives the real cepstrum $C^r(\tau)$ as:

$$C^r(\tau) = \mathcal{F}^{-1}(\log[A(f)]) = \mathcal{F}^{-1}(\ln[A(f)]) \quad \text{............................} (6.8)$$

Note that as the spectral information $|X(f)|$ consists of only (real) amplitude values (hence the name), and is also even, all information is contained in the positive quefrency part of the real cepstrum.

It should be noted that the real cepstrum in equation (6.8) is equivalently gained by substituting $\phi(f) = 0$ into calculations for the complex cepstrum (substituting $\phi(f) = 0$ into equation (6.4), and subsequently equations (6.5) and (6.6)).

The primary differences between the two alternative versions are based on two attributes.

For the first attribute, the complex cepstrum contains the phase information from the spectrum, and so can be reversed back to the time domain from the cepstrum. The real cepstrum, however, can only be reversed back to an amplitude spectrum, and cannot subsequently be reversed back to the time domain as the spectral information is not complex. Thus, to edit signals in the cepstrum, which are then returned to the time domain, based on this attribute, only the complex cepstrum is suitable for use.

For the second attribute, the complex cepstrum requires that the phase of the signal can be unwrapped to a continuous function of frequency (e.g. single transients), which is not possible with stationary signals (both forcing functions and response signals). The real cepstrum on the other hand can be employed with stationary signals. Thus for the application to stationary response signals, only the real cepstrum is suitable for use.

Based on these two attributes, neither the complex nor the real cepstrum are directly suitable for this application, where the goal is to edit a response signal (stationary signal) in the cepstrum, and then ultimately transform the edited signal back to the time domain. Based on this fact, it was formerly believed that cepstrum editing was not possible for this application.

6.4.3 Cepstrum editing method for time domain signals

In 2011 Sawalhi and Randall [50] identified a novel extension to the real cepstrum which allowed for editing in the cepstrum and then subsequently reversing the edited signal back to the time domain. This is the basic cepstrum editing technique which was subsequently used in [48], and will be used in this work.

For this extension to the real cepstrum, the traditional real cepstrum process is firstly used to edit the log amplitude spectrum, via the real cepstrum, and then the edited spectral amplitude information is recombined with the original phase of the signal, by taking the edited log spectrum as real component and the original phase as imaginary component (of form similar to the right hand side of equation (6.6) on page 348), which can be exponentiated to give complex frequency values, which in turn can then be transformed back to the time domain.

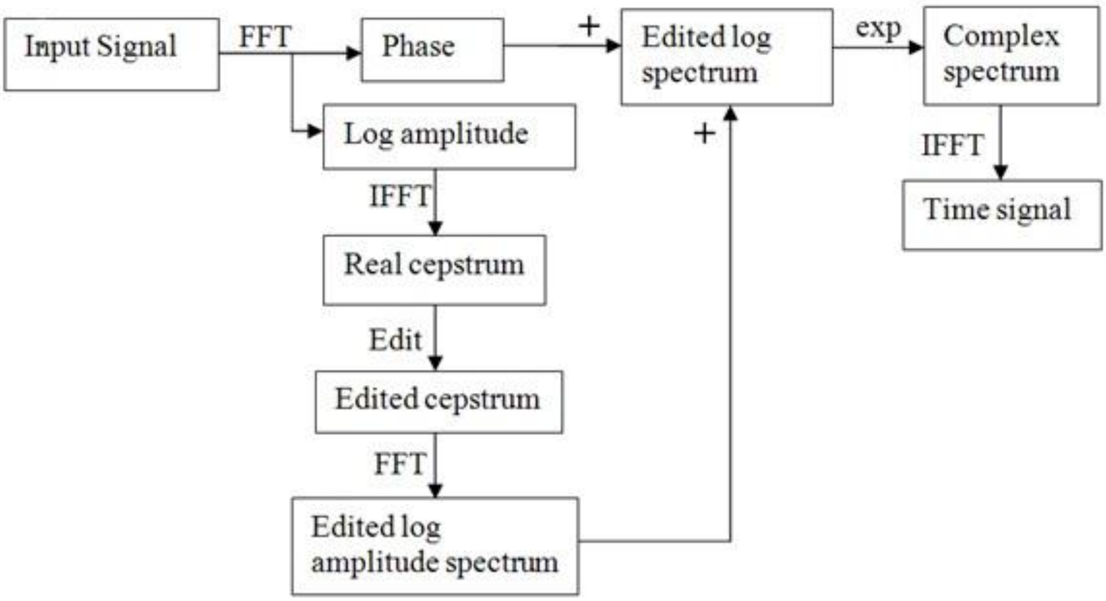


Figure 6.1 – Schematic diagram of the cepstral editing method, based on an extension to the real cepstrum, which allows for an edited signal to be returned to the time domain. [48, 50]

Figure 6.1 shows a schematic diagram showing the editing process for a signal, using the real cepstral editing method.

For a mathematical representation of the above description, and Figure 6.1, for the reversal of edited cepstrum values back to the time domain, firstly real cepstrum values $C^r(\tau)$ are obtained as per equation (6.8) on page 349. These cepstrum values are then edited, resulting in edited real cepstrum information $C^r_{ed}(\tau)$. The edited cepstrum values can then be transformed back to the frequency domain using an FFT, giving (logarithmic) edited frequency information $L_{ed}(f)$, such that:

$$L_{ed}(f) = \mathcal{F}\left[C^r_{ed}(\tau)\right] \quad (6.9)$$

Edited complex logarithmic spectral information $Z_{ed}(f)$ is then formed by combining the edited frequency information $L_{ed}(f)$ as real component, with the original (unedited) phase information $\phi(f)$ (from $Y(f) = A(f)\exp(j\phi(f))$) as imaginary component, so that:

$$Z_{ed}(f) = L_{ed}(f) + j\phi(f) \quad \text{......................................} (6.10)$$

The edited complex (non-logarithmic) frequency domain information $Y_{ed}(f)$ is then formed by exponentiating $Z_{ed}(f)$, so that:

$$Y_{ed}(f) = \exp[Z_{ed}(f)] = \exp[L_{ed}(f) + j\phi(f)] = \exp[L_{ed}(f)]\exp[j\phi(f)] \text{ .... } (6.11)$$

From equation (6.11), the amplitude of $Y_{ed}(f)$ is $\exp[L_{ed}(f)]$ (based on edited cepstral values), and the phase is $\phi(f)$ (original phase values).

The edited time domain function $y_{ed}(t)$ is finally obtained by taking the inverse FFT of $Y_{ed}(f)$, such that:

$$y_{ed}(t) = \mathcal{F}^{-1}[Y_{ed}(f)] = \mathcal{F}^{-1}\{\exp[L_{ed}(f)]\exp[j\phi(f)]\} \text{ .... } (6.12)$$

It should be noted that in effect, this cepstral editing process only edits the amplitude values of a spectrum, and so the phase of an edited spectrum will not be correct at locations where the amplitude has been modified. However, as the goal is to reduce the amplitude of certain components down to the noise level (i.e. remove them), the effect on the overall signal due to the minimised amplitude components is often negligible [51].

One important thing to note for editing is that the real cepstrum has positive and negative sides, which are mirrored around the middle quefrency value along the record, similarly to the frequency (amplitude) spectrum being a mirror around the Nyquist frequency for the same situation. As with the frequency domain, due to the mirroring around the centre of the record, it is normally necessary to lifter both the positive and negative sides of the cepstrum in a mirrored fashion, to correctly remove components.

6.4.4 Lifter types for cepstrum domain editing

In [48], Randall et al. present two different editing methods for the Cepstrum domain, these being a notch lifter (which could alternatively be called a comb lifter), and an exponential short pass lifter. In [51], Randall and Sawalhi presented a second type of notch lifter, and introduced the naming convention, of the type 1 notch lifter as referring to the notch lifter from [48], and the type 2 notch lifter referring to the newer of the two notch lifters, from [51].

The two notch lifter types are both designed to remove a rahmonic series from the real cepstrum (note the term 'real cepstrum' will henceforth be referred to as simply the 'cepstrum' in this work, for the sake of brevity, as the complex cepstrum is not used again in this work). Both lifters consist of a unity pass band over the entire cepstrum, with notches down to zero amplitude centred on each order of the rahmonic series which is targeted for removal. Note that multiple individually tailored notch lifters could be applied consecutively in order to remove multiple rahmonic series from a single cepstrum. Setting a rahmonic value in the cepstrum to zero amplitude, has the result of automatically smoothing the amplitude of the corresponding components in the spectrum to the amplitude of adjacent noise in the frequency domain, rather than leaving notches in the (log) spectrum (which would give non-zero components in the cepstrum). While it is not possible to use a custom smoothing technique in the spectrum, when employing cepstrum editing, such as the two alternative smoothing methods used with frequency bin editing in [14], the natural amplitude smoothing inherent with setting a rahmonic to zero amplitude in the cepstrum is ideal for the application of (multiple) component removal. Note that, as discussed earlier, this smoothing only occurs between the amplitude values in the frequency domain, and the phase remains unaffected.

The primary difference between the two notch lifters presented in [51], is that the type 1 notch lifter uses constant width notches, and the type 2 notch lifter uses notches which increase in proportion to rahmonic order.

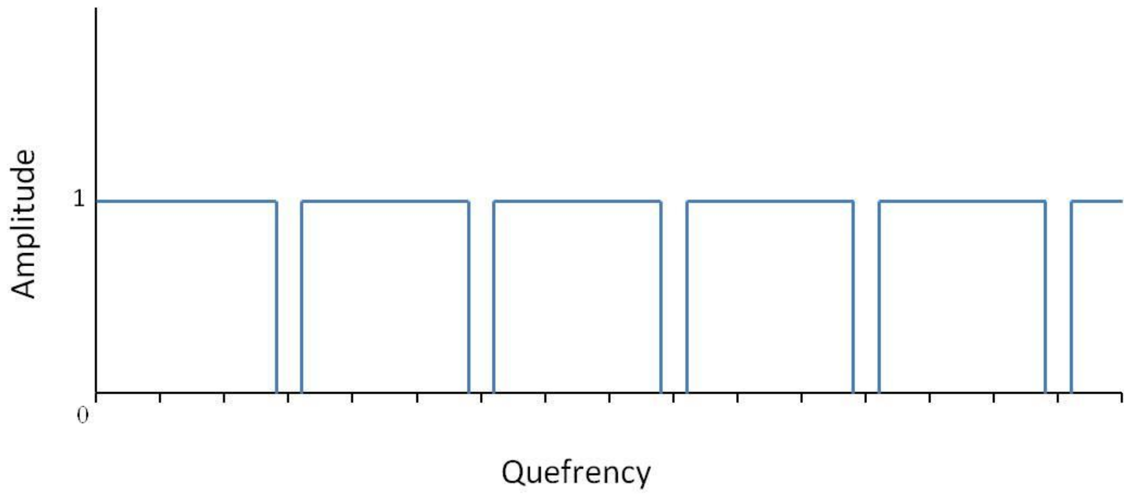


Figure 6.2 – Type 1 notch lifter, for removal of a constant-width rahmonic series, in the cepstrum domain

Figure 6.2 shows the lifter characteristic for the type 1 notch lifter, which has constant width notches down to zero amplitude. Note that each notch in the lifter would be centred over a rahmonic order to be removed. This lifter is best suited for use with removing a rahmonic series which remains constant width, and does not smear out at higher rahmonic orders (characteristically similar to the spectrum for an AM signal, as shown in Figure 2.21 on page 78, where each harmonic order is the same width). Note that this depicted notch lifter is suitable for use for the positive side of the cepstrum (from the left hand side to the centre), and that the positive side notch lifter needs to be mirrored about the centre quefrency, in order to correctly lifter the negative side of the cepstrum.

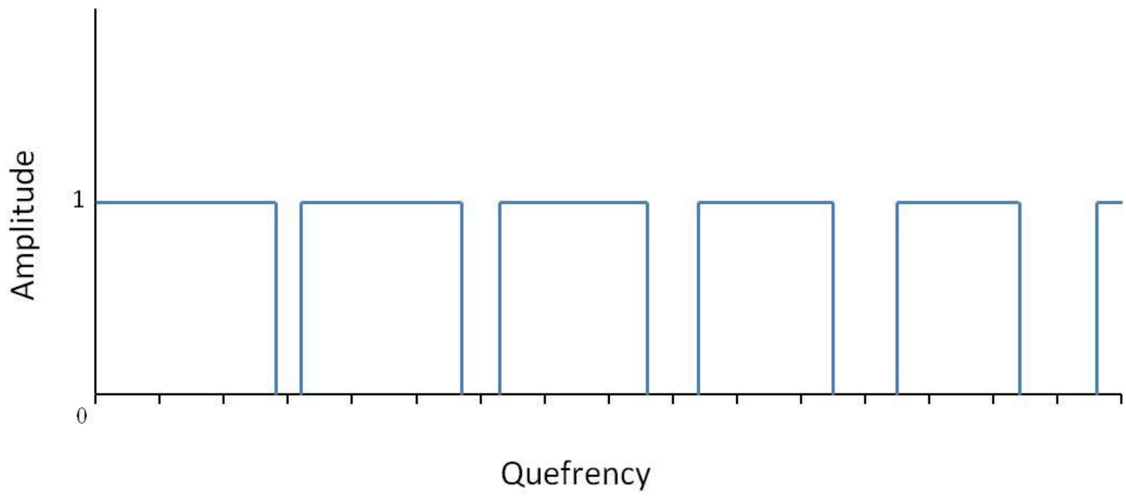


Figure 6.3 – Type 2 notch lifter, for removal of an increasing-width rahmonic series, in the cepstrum domain

Figure 6.3 shows the lifter characteristic for the type 2 notch lifter, where each notch is increasing in width with rahmonic order. This lifter is best suited for removing a rahmonic series which becomes smeared with higher rahmonic order (characteristically similar to the spectrum for an FM signal, as shown in Figure 2.27 on page 106, where the bandwidth of each harmonic order increases (smears) with harmonic order). One feature to note is that notches for higher rahmonic orders will overlap once they become broad enough, which would result in a brick-wall short-pass lifter effect at higher quefrency (analogous to a low pass filter in the frequency domain).

The third type of lifter, used in [48], is the exponential short pass lifter. A short pass lifter is one which removes high-quefrency components, and is analogous to a low pass filter in the frequency domain. This type of lifter is specifically useful for OMA applications, as all modal information for a structure is present at low quefrency values, and so high quefrency information can be removed en masse when the goal is to retain only modal information. The use of this type of lifter, in effect, de-noises modal information by removing non-modal information in a blanket fashion.

This can be demonstrated, by making use of the fact that the cepstrum of an impulse response function (IRF) has the same mathematical form as the IRF (a complex exponential) further damped by multiplication by $1/n$, where n is the (discrete) time (quefrency) sample number (and where the time axis of the cepstrum is known as “quefrency”). Thus, as shown in [52], the complex cepstrum of a minimum phase function can be expressed as:

$$C^{cp}(n) = 2\sum_{j=1}^{N_p/2}\frac{A_j^n}{n}\cos(\omega_j n) - 2\sum_{k=1}^{N_p/2}\frac{A_k^n}{n}\cos(\omega_k n) \quad \text{.....................} (6.13)$$

where A_j are the moduli and ω_j are the frequencies of the poles in the z-plane; A_k and ω_k are the same for the zeroes, all inside the unit circle. Note that A_j^n and A_k^n are exponential decays which can be written as $\exp\left(-\sigma_i n \Delta t\right)$, where σ_i represents in rad/s half the 3 dB bandwidth of the corresponding resonance peak, and Δt is the time sample spacing. Thus, if a cepstrum is multiplied by an exponential window $\exp\left(-\sigma_0 n \Delta t\right)$, the only effect on the modal properties is to add the constant damping value σ_0 to the damping of all poles and zeroes in the FRF, and this can in principle be compensated for after modal analysis has been implemented.

As a result, an exponential short pass lifter is preferable to use for modal analysis applications, and for OMA specifically.

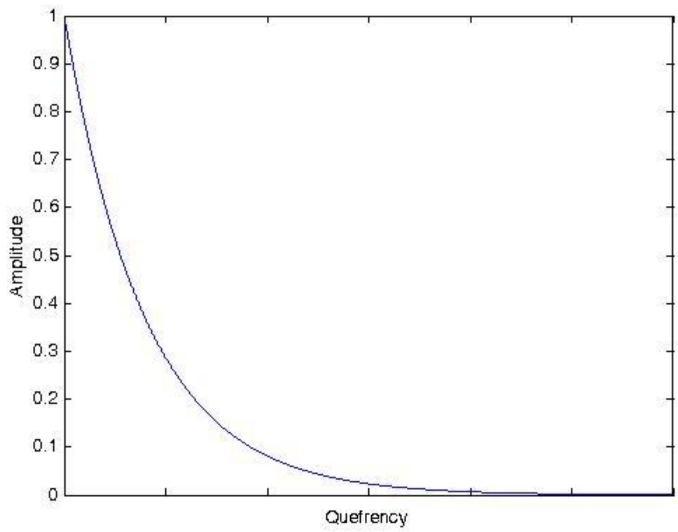


Figure 6.4 – Exponential short pass lifter, for removal of high quefrency components, in the cepstrum domain

Figure 6.4 shows the lifter characteristics for a decaying exponential short pass lifter. Note that, as with the previous lifters, this lifter characteristic would be applied to the positive side of the cepstrum, and would be mirrored around the centre quefrency to give the filter characteristic which is applied to the negative side of the cepstrum, to perform correct editing.

6.4.5 Cepstrum as applied to order-tracked signals

One important factor to note, is that the Cepstrum, as defined in previous works, is specifically based on the time and frequency domains. In this work, it is specifically desired to use ‘Cepstrum editing’ with order-tracked signals. As discussed in 1.1.1 starting on page 3, the primary usage of order-tracking is to allow MCM signal processing methods which are applicable for constant-frequency signals, to be applied to varying speed signals. This is made possible by converting a variable speed signal to a constant rotational phase axis, so the basic signal is in the angle domain rather than the time domain, and has constant ‘frequency’ representing harmonic orders of the rotational speed. The ‘spectrum’ of a phase domain signal is gained by taking the FFT of the phase angle signal, which gives the order spectrum. The order spectrum, for a variable speed signal, is functionally equivalent to the spectrum from a constant speed signal. In an identical fashion to the real cepstrum, it is possible to take the ‘real cepstrum’ of order domain spectral information, based on the rotation angle domain of a variable speed signal, which results in a ‘cepstrum’ which is functionally equivalent to the real cepstrum which would be gained from a time signal for a constant speed machine. As such, identical processes can be applied in the ‘cepstrum’, regardless of whether it is ultimately based on a time domain signal, or a phase domain signal.

As a result of this functional equivalence, to avoid confusion the term ‘cepstrum’, as well as associated cepstral terms such as shown in Table 6.1, will be used in this work to refer to both the traditional cepstrum, based on the frequency and time domain, as well as the ‘cepstrum’, when based on the order and phase domain. Where appropriate, it will be identified if the cepstrum is referring the traditional time domain based definition, or the phase domain based definition. Furthermore, exactly which cepstrum version is being shown in results can be inferred from the dimensions and units of the x-axis of the cepstrum and associated domains, e.g. frequency (Hz) and quefrency (time, in seconds) for the time domain based domains, or order and quefrency (rotations) for the angle domain based domains.

Temporal based domains		
Domain	Dimension	Units
time domain	time	seconds (sec)
frequency domain	frequency	Hertz (Hz)
Cepstrum domain	quefrency	seconds (sec)

Phase based domains		
Domain	Dimension	Units
phase angle domain	phase	revolutions
order domain	dimensionless	orders
Cepstrum domain	quefrency	revolutions

Table 6.2 – Units and dimensions of phase and time based cepstrum, and associated domains

Table 6.2 shows the differences in dimensions and units for the two cepstral domain versions used in this work. If unclear in the text descriptions, the dimensions and units of results shown in this work can additionally be used to infer if the described cepstrum results are referring to those based on a time domain signal (either the original or reversed signals), or to cepstrum results based on a phase domain signal (an order-tracked signal, with speed variations compensated for).

6.5 Experimental setup

In order to demonstrate the new method presented here, to pre-process a variable speed response signal using cepstrum editing, to remove order-related components so that OMA could subsequently be performed, the processing of an array of recorded measurements from an experimental test will be conducted.

For the experimental test, an array of response signals were captured from a gearbox casing (with no internals) which was externally excited by an attached shaker. This is the same gearbox casing which was used to capture data used in section 5.5, starting on page 284, to illustrate segmented order tracking. For this experimental test, the excitation shaker was mounted horizontally, as opposed to the vertical mounting used in section 5.5. Both the shaker and casing were independently elastically suspended, in order to isolate the system. Figure 6.5 shows a photo of the experimental setup, and the shaker position is indicated by the red arrow. Note that the separate elastic suspension of the casing and shaker are visible, with the casing elastic suspension extending upwards outside the photo area.

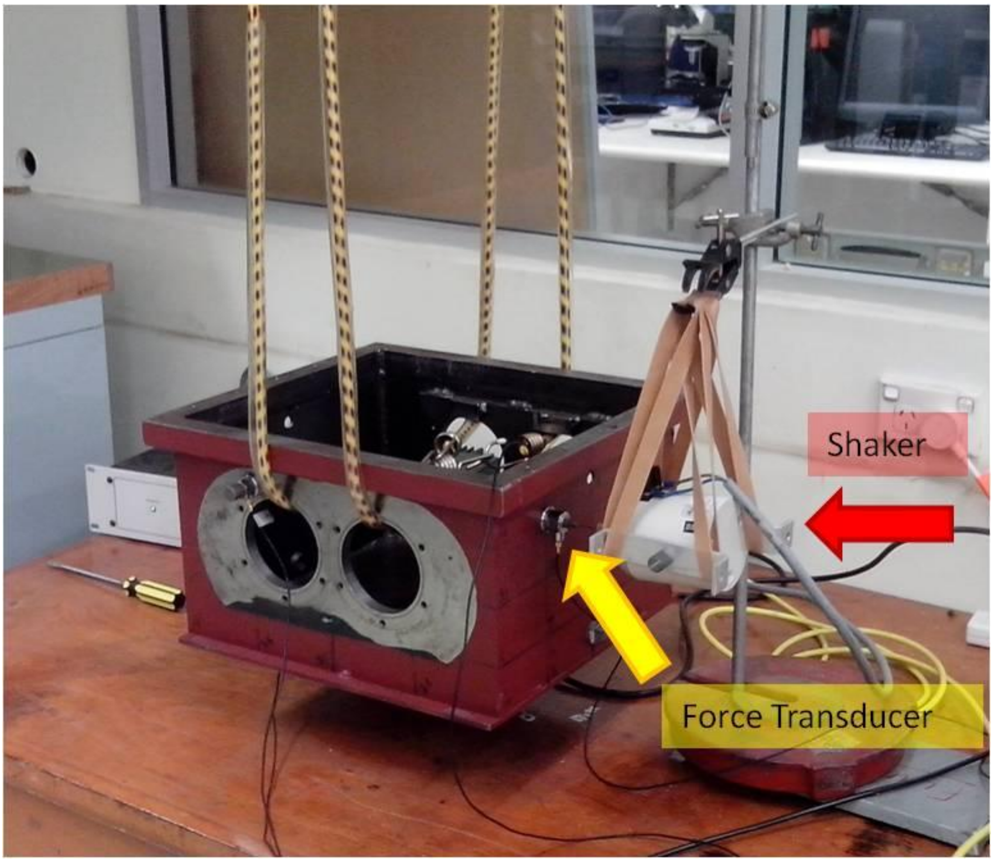


Figure 6.5 – Experimental setup, indicating shaker mounting position and force transducer location

For this experimental test, an array of acceleration (response) measurements was captured from the casing, from the four side panels and the base, which would both be suitable for OMA analysis. Each panel consisted of five equally spaced measurement grids along each axis, with measurements at each edge, the centre line, and then at the midpoints between each edge and centre line. Note that the array locations can be seen faintly, marked in back, on the visible side panels of the casing in Figure 6.5. With five measurement grids on each axis, this resulted in $5 \times 5 = 25$ response measurement locations on each side, and the base. Each response measurement was taken perpendicular to a particular side, and multiple measurements were taken with different axes at the common edges between panels. This array scheme resulted in 125 response measurement positions, split over the three cardinal axes (positive and negative directions on some axes). This measurement array scheme is the same used by Deshpande et al. [53] (additionally in [54]), who conducted modal analysis using an impact hammer with the same casing, and was chosen in the hopes of directly comparing their EMA results to those in this work. However, the modal test conducted in [53] was conducted using a different suspension scheme, and ultimately the results were found to be not directly comparable to those captured during the experimental test conducted in this work.

In addition to response measurements, a force measurement was recorded at the interface point where the shaker was attached to the gearbox casing. The force measurement was intended to allow for experimental modal analysis to be conducted using the same data, to compare to OMA results, however faults were discovered with the force measurements recorded which prevented this parallel analysis to be conducted. This experiment will be re-conducted at a later date, to gain results suitable for use with EMA to provide a direct comparison, however this will be conducted as future work, and not included in this thesis.

Experimental data was recorded in multiple data sets, with each set containing four data measurements. For each set, a force measurement was captured, a (common to all sets) driving point acceleration signal was taken at the shaker interface (inside of the casing wall, opposite the shaker mounting), and then two roving acceleration signals were recorded for each data set. This resulted in a total of 62 separate data sets being recorded, in order to span the 125 response measurement locations.

6.5.1 Input signal for experimental test

The input signal to drive the shaker, and excite the system, was constructed by combining a frequency modulated periodic pulse signal (to represent a varying speed machine component) with an additional white noise component which spanned the full valid frequency range (to ensure the entire spectrum was excited). These signals were recorded with a signal length of 120 seconds, with a sampling rate of 4096 Hz.

For the periodic part of the input excitation signal, a pseudo-pulse signal was constructed containing 22 harmonics, by summing 22 FM cosines together, in a similar fashion to the process used in section 5.5 on page 284, to produce an input signal with 10 harmonics. It was specifically selected to use 22 harmonics, as the 23rd harmonic was found to be the first which extended above the Nyquist frequency of the response signal sampling rate (4096 Hz). An amplitude decay factor was introduced to lower the amplitude of higher harmonics. The amplitude decay factor used was $0.98^{(n-1)}$, where n is the harmonic order, so for the highest order, the amplitude was $0.98^{21} = 65.4\%$ of the amplitude of the fundamental harmonic.

The fundamental ‘carrier frequency’ used for the periodic part of the input signal was $f_c = 76\text{ Hz}$, with the frequency variation (deviation) of $\pm 15\%$ around the carrier frequency, and modulating frequency of 2 Hz (0.5s period, resulting in 240 modulation periods over the 120 second long response signals). This signal was constructed with a sampling rate of 25 kHz, and with a length of 30 seconds, which was the maximum length of signal supported by the signal analyser being used to generate the shaker input signal. In order to reach the desired time length of 120 seconds, the 30 second periodic part of the input signal was repeated four times by the signal analyser.

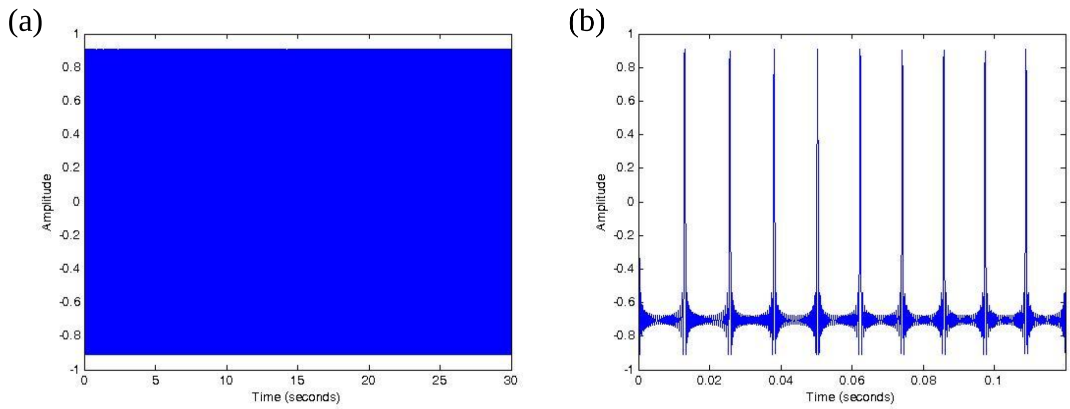


Figure 6.6 – Periodic FM component of Input signal
(a) 30 second signal which is repeated (b) zoom showing characteristic pulses

Figure 6.6 (a) shows the time domain plot of the periodic part of the input signal, with 30 second record length, and it can be seen that the envelope of this signal is constant over the length. Figure 6.6 (b) shows a close zoom of (a), in order to show the characteristic pulses of the signal which are formed by adding the 22 cosines together.

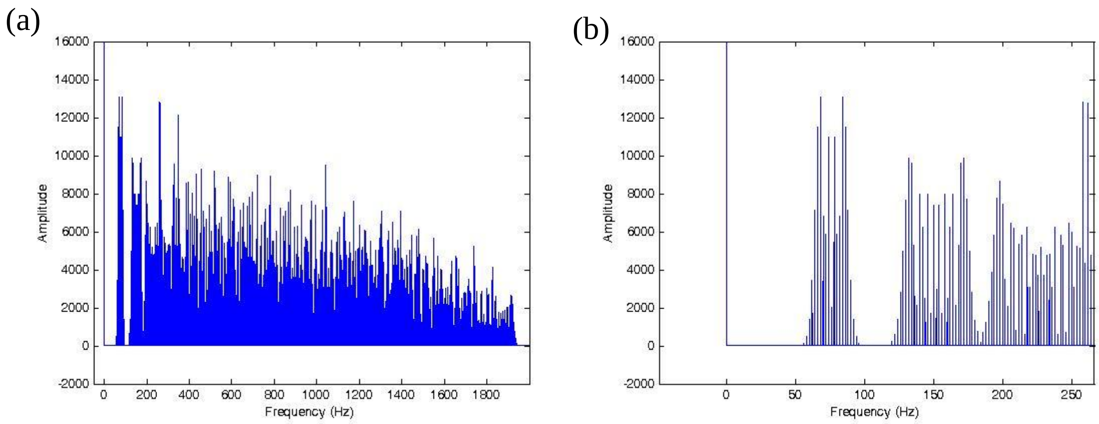


Figure 6.7 – Spectrum of periodic FM component of input signal
(a) Full spectrum (b) zoom showing first harmonic

Figure 6.7 (a) shows the spectrum of the periodic part of the input signal, and the frequency variation is clearly seen from the smearing of the 22 harmonics. Additionally, it can be seen that 22 harmonics extend almost up to the Nyquist frequency (2048 Hz) of the response signal sampling rate to be used (anti-aliasing filter in response signal is at 1600 Hz). Figure 6.7 (b) shows a zoom of the spectrum over the first three harmonics, and it can be seen that with a $\pm 15\%$ speed variation the first harmonic is separable from the second.

It should be noted that, as the first harmonic is separable, the order-tracking step can be conducted with the basic PDOT method in this instance. If the speed range was large enough that the first harmonic was not separable, analysis could still be conducted by combining reversible order-tracking and segmented order-tracking in a modular fashion, as all the different modules to the generalised PDOT methodology can be employed together as needed, as discussed in Chapter 4.

For the noise component of the input signal, randomly generated white noise was used with a frequency span of 10 kHz, corresponding to 40% of the 25 kHz sampling frequency of the input signal. The level of the white noise, relative to the periodic FM component, was tuned so that the fundamental harmonic would have a DR larger than 40 dB, ensuring that order-tracking could be conducted using the fundamental harmonic. Note that the noise component was randomly generated for each data set, and so each of the 64 data sets recorded were excited by a different noise component.

Both the FM periodic component, and the white noise component (generated on different channels of the same signal analyser) were combined to produce the input signal used to excite the shaker.

Unfortunately, due to a problem with the signal generator, the periodic component of the input signal was not repeatably generated by the signal analyser. The main problem appears to be that the 30 second periodic component was not correctly repeated four times, and rather an amplitude modulated version was produced for each 30 second section, with the amplitude modulation being variable for each section, and so each section is not repeated identically within an individual signal, and additionally matching segments between data sets are also not equal.

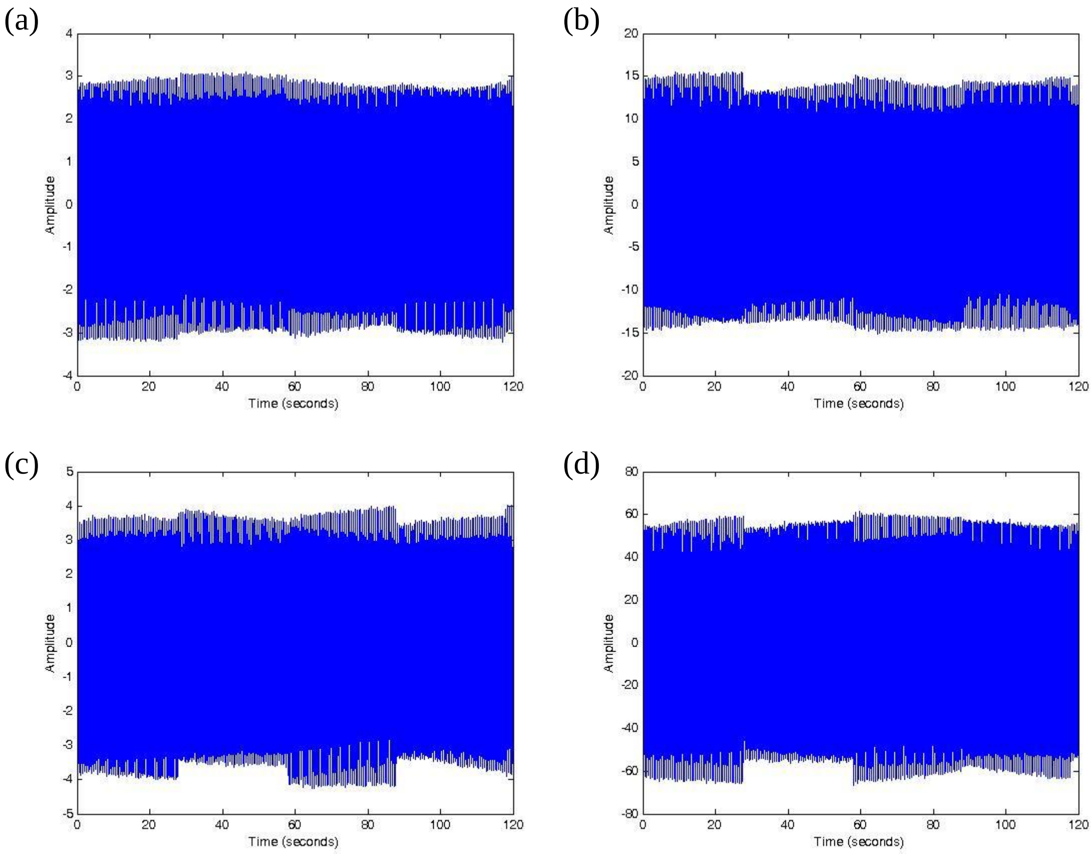


Figure 6.8 – Time domain plots of response signals from different data sets showing ‘incorrect’ envelope

From data set: (a) 1 (b) 2 (c) 3 (d) 4

Figure 6.8 (a) to (c) show the time domain plots of response signals from different data sets. It can be clearly seen that in each plot, there are distinct joins every 30 seconds, where the signal analyser has generated a new input signal section which does not match the previous. Secondly, it can be seen, by comparing envelopes, that none of the response signals correctly match the originally constructed signal envelope, from Figure 6.6 (a) on page 367, and that all have a slight amplitude modulation. Finally, it can also be seen by comparing envelopes between signals in Figure 6.8 (a) to (c), that each data set has been excited by different input sections, even over a common

30 second section, and so the input signal has not been repeatable over the different data sets.

While it is not ideal that the input signal was not repeatable between data sets, it is perfectly acceptable for OMA to be conducted using these results. For many OMA applications, the input signal is not repeatable, and analysis still needs to be conducted. So, this experimental case, with a non-repeatable excitation, simply represents a more challenging OMA instance, rather than preventing OMA due to the issues.

6.6 Pre-processing method to remove order components, illustrated with example response signal

6.6.1 Forward order-tracking using PDOT

While it had been intended to conduct order-tracking by using the periodic component of the input signal as a tacho signal, this was not possible with the issues with the input signal, as described in the previous section.

In essence, the signal generator created an ‘incorrect’ input signal (output from generator) which didn’t match the constructed input signal (input to generator) sent to the signal generator, and so while the response measurements correspond to the ‘incorrect’ input signal generated, which was not recorded, they do not match the constructed input signal. As a result, the constructed input signal cannot be used as a tacho.

Instead of using a ‘tacho’ for this experimental case as originally planned, PDOT was instead performed using the response signals themselves as the reference signal, as described in section 4.2.3 on page 183.

In order to demonstrate the pre-processing method developed here, the pre-processing of a single response signal is firstly shown for all stages of the method. The first roving response signal from the first data set will be used as the example, where the time domain plot of this signal was shown in Figure 6.8 (a) on page 369.

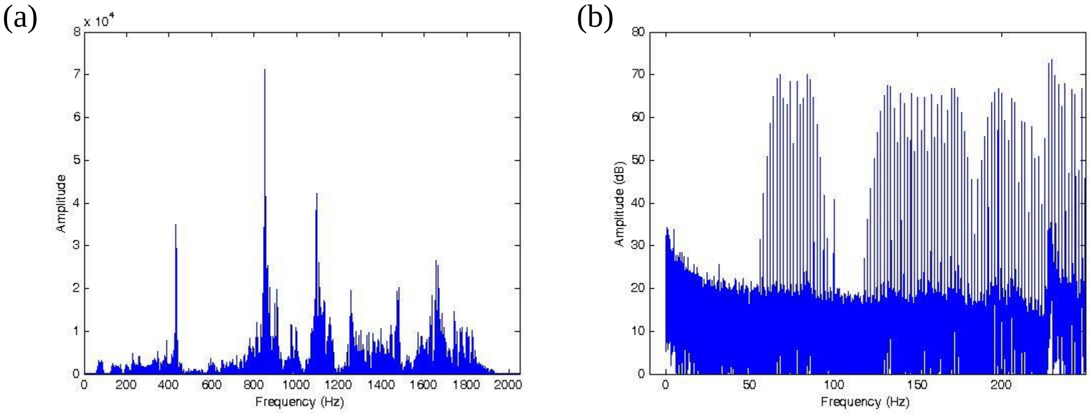


Figure 6.9 – Response spectra for example case, from data set 1

(a) Linear amplitude (b) dB amplitude, zoomed on lowest three orders

Figure 6.9 (a) shows the full spectrum for the response signal being used as the example case, and it can be seen that frequencies have been excited over the full frequency range. In addition, the smearing of components due to the frequency variation is seen throughout the spectrum.

Figure 6.9 (b) shows a zoom over the first harmonics on a dB scale, and it can be clearly seen that the first harmonic is separable from noise and adjacent harmonics with a 40dB dynamic range, and so is suitable for use with the PDOT method to conduct order tracking. As a result, the first harmonic of the response signal was used to conduct the PDOT order tracking.

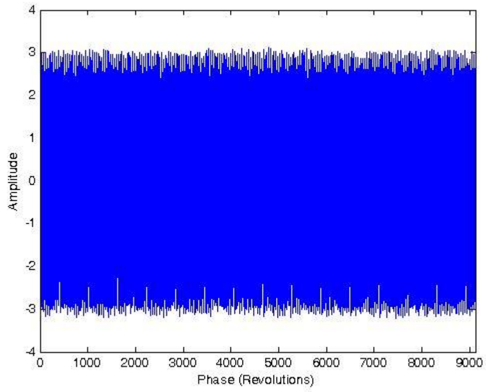


Figure 6.10 – Order-tracked response signal – time domain

Figure 6.10 shows the time domain plot for the order-tracked signal, and it can be seen that the envelope of the signal has been correctly maintained during the order-tracking process, indicating successful order-tracking.

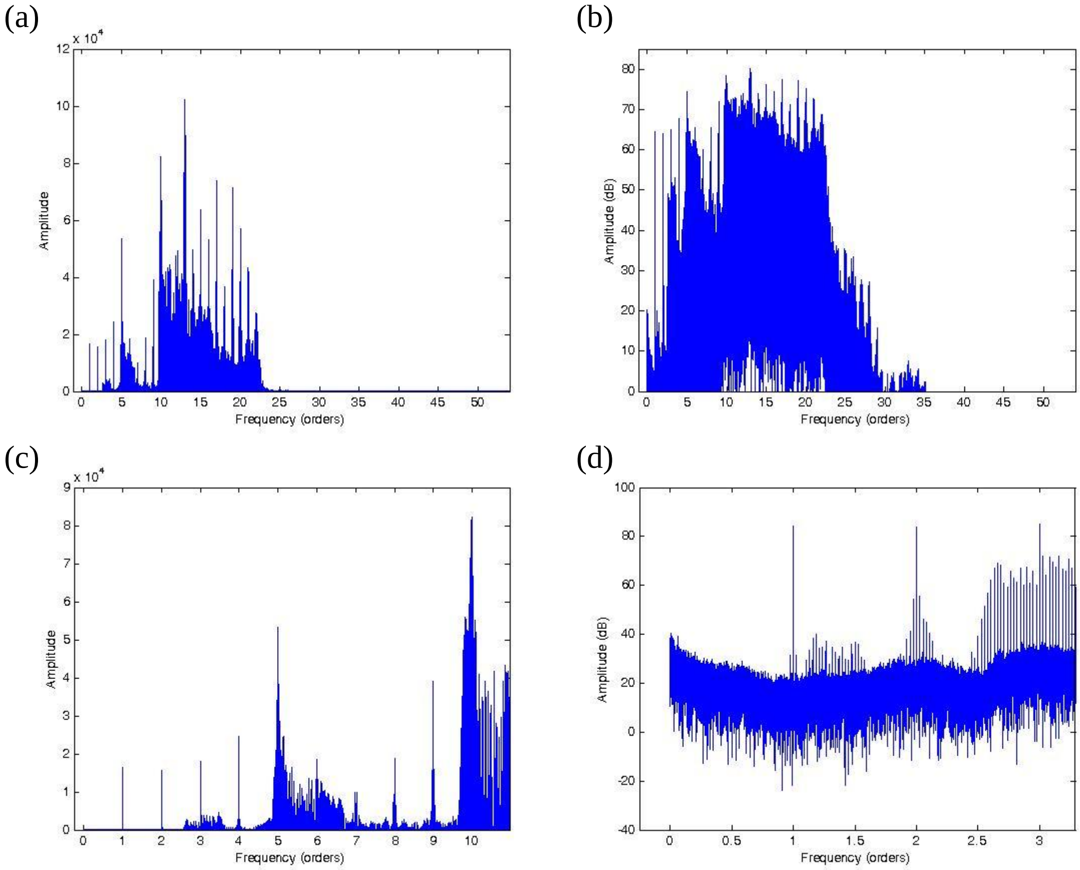


Figure 6.11 – Order spectra of order-tracked signal
(a) full spectrum, linear scale (b) full spectrum, dB scale
(c) zoom on first 10 orders, linear scale (d) zoom on first 3 orders, dB scale

Figure 6.11 shows the order spectra from the order-tracked signal. Figure 6.11 (a) shows the full order spectrum on a linear scale, and sharp discrete peaks are seen for the first 22 orders, corresponding to 22 harmonics in the excitation signal, indicating that order-tracking was successful. Figure 6.11 (b) again shows the full spectrum, but on a dB scale, and in this plot it can be seen that frequency information extends across the entire order spectrum. Figure 6.11 (c) shows a zoom of the spectrum, over the first 10 orders, on a linear scale, and from this plot the discrete peaks at each order value are even more clearly seen, though sidebands are now visible at some orders. Figure 6.11 (d) shows a zoom of the spectrum, for the first three orders, on a dB scale, and in this plot the amplitude modulation sidebands surrounding the orders are now clearly visible. These sidebands have primarily been caused as a natural consequence of the structure being excited with a variable frequency input, and would be expected to be present with most variable speed cases. It is the presence of these sidebands, caused by amplitude modulation, which was not adequately addressed in the earlier works [14, 47] aimed at removing order-based content from a signal, as discussed in section 6.3.2 on page 336. For correct removal of order-based components for pre-filtration, both the discrete peaks at each order, as well as the surrounding amplitude modulation sidebands need to be removed from the signal.

6.6.2 First stage of Cepstrum editing using notch lifter, of order-tracked signal

To begin the editing process of the order-tracked signal, the order spectrum is transformed to the cepstrum domain, using the ‘real cepstrum’ method. At this time it is also suitable to record the phase of the original order spectrum, which will be recombined with the edited ‘real cepstrum’ values after editing has been conducted.

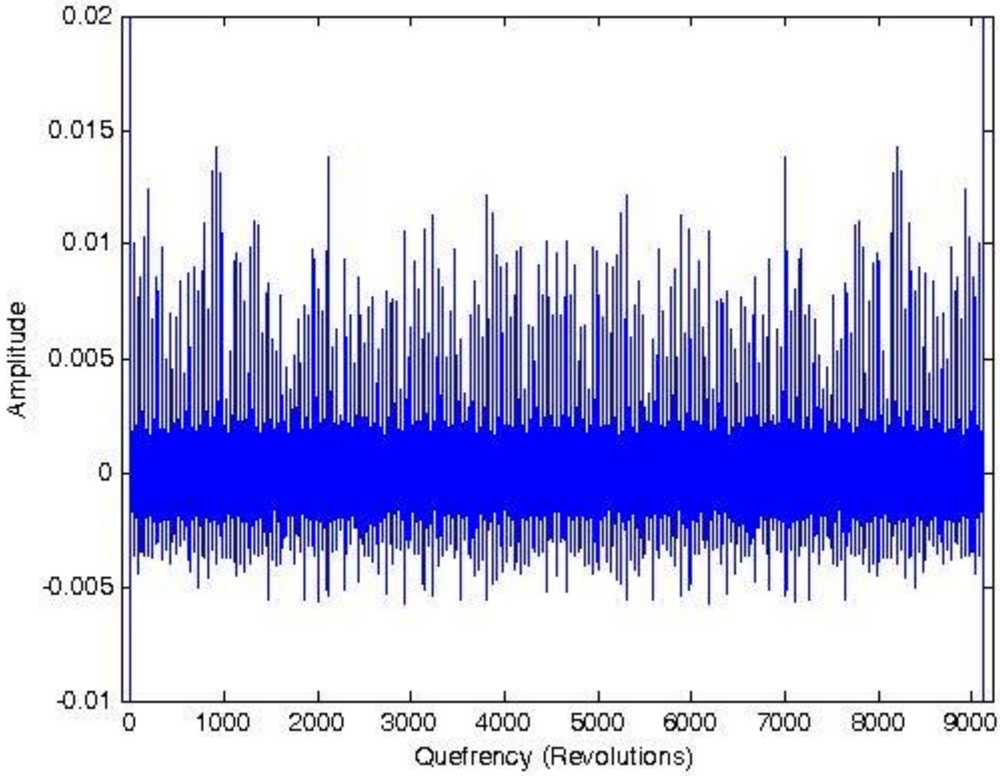


Figure 6.12 – Real cepstrum before editing, two sided, based on order spectrum

Figure 6.12 shows the real cepstrum, based on the order spectrum. This is the entire cepstrum, and as described earlier, this consists of both a positive and negative side, which is mirrored around the centre quefrency. Note that for all subsequent plots, the negative side of the cepstrum will not be shown for convenience, similar to the third general form of a frequency spectrum plot, as described in section 2.1 starting on page 34, and shown by Figure 2.4 on page 42.

For the first stage of editing of the real cepstrum, a notch lifter is applied to remove the discrete ‘carrier frequency’ rahmonic series.

As the frequency spacing of the carrier frequency components in the spectrum is 1 order, the spacing in quefrency of the corresponding rahmonic series, is given by the inverse of the frequency spacing, which in this case is simply 1 revolution. So, in the cepstrum, the components to be removed consist of a discrete series of rahmonic peaks, with one peak at every integer revolution quefrency value.

Examination of the rahmonic series, which has 1 revolution spacing, showed that the discrete peaks had very little smearing over the entire cepstrum. This was expected, as the underlying signal has been correctly order-tracked, and the carrier frequency harmonic series in the order spectrum was discrete, which would also give discrete values in the cepstrum.

As the series is not increasing in width at higher quefrency, a type 1 notch lifter with constant width, as described in section 6.4.4 on page 355, and used in [48] (among other works as detailed earlier), was chosen to remove the carrier frequency rahmonic series component. For this lifter, a constant notch width of 7 samples was chosen, which was centred over each integer value of revolutions. As described previously, this notch lifter was extended up to the middle quefrency value to span the positive side of the cepstrum, and then mirrored and applied to the negative quefrency values.

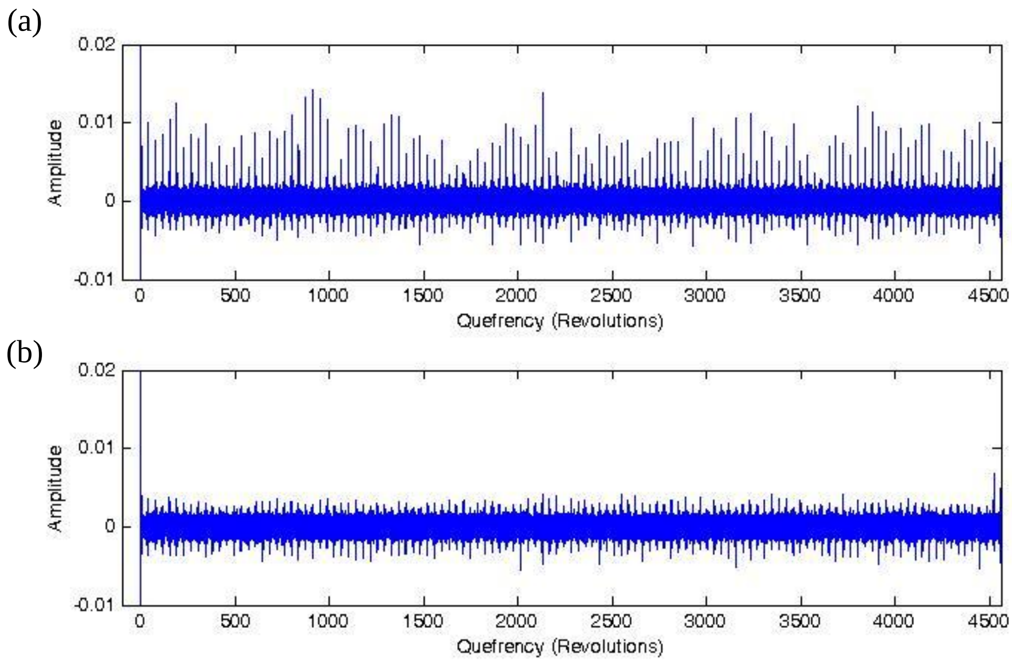


Figure 6.13 – Cepstrum, before and after notch editing, positive side shown
(a) before notch editing (b) after notch editing

Figure 6.13 shows the cepstrum plots, both before and after editing. Figure 6.13 (a) shows the cepstrum before editing, and is the same plot as Figure 6.12, except only the positive side of the cepstrum is shown. Figure 6.13 (b) shows the cepstrum after the notch lifter has been applied, and it can be clearly seen that a significant amount of information has been removed from the cepstrum.

In order to more clearly illustrate the application of the notch lifter, a zoom of these plots over just a few revolutions is shown below.

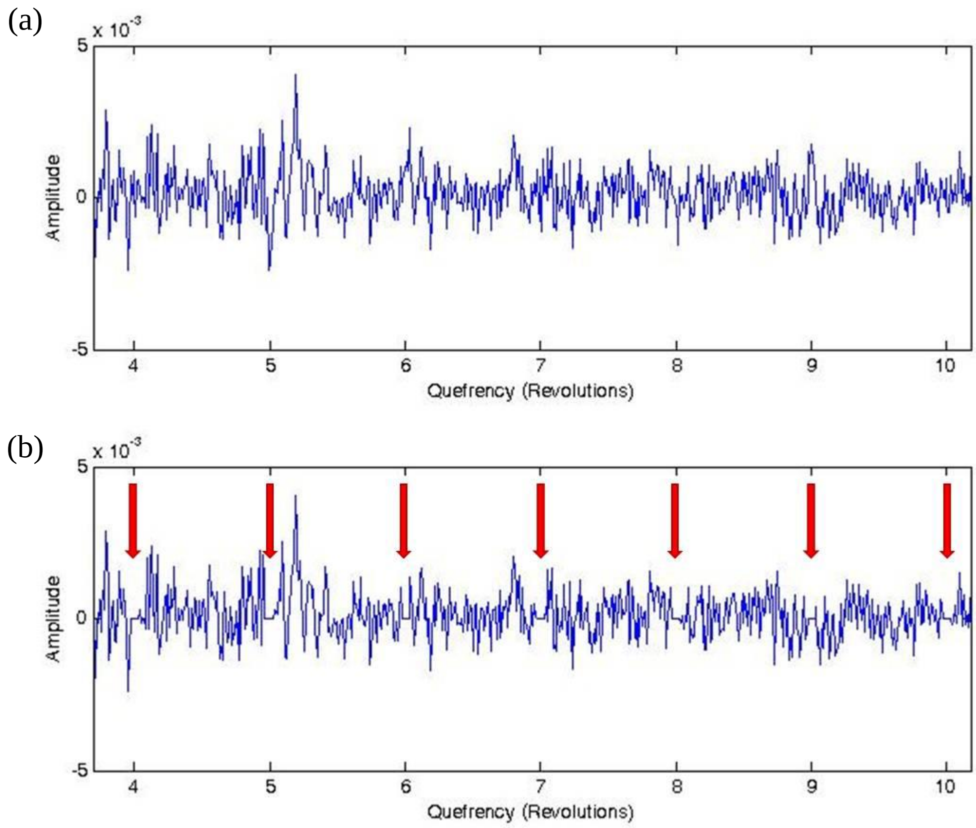


Figure 6.14 – Cepstrum, before and after notch editing, with editing locations marked by red arrows
(a) before editing (b) after editing

Figure 6.14 again shows the before and after cepstra from Figure 6.13, only this time zoomed in to show quefrency values between 4 and 10 revolutions. The locations where the notch editing were applied are marked by red arrows in Figure 6.14 (b), and by comparing to Figure 6.14 (a) the effects of the narrow notches used can just be seen.

To evaluate the success of this editing process, this edited signal was transformed back to the order domain, by combining with the original phase, as per equation (6.11) on page 353.

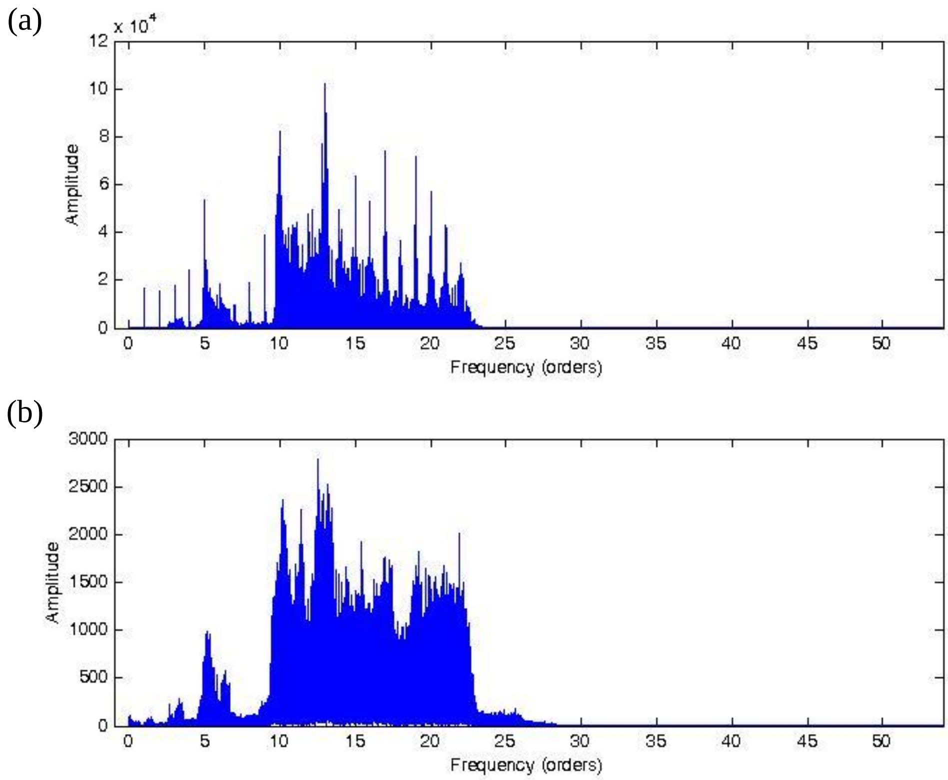
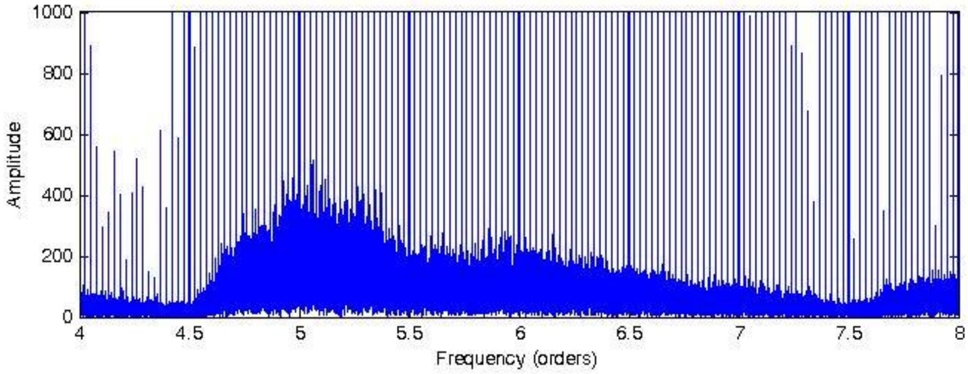


Figure 6.15 – Order spectra, before and after notch editing
(a) before editing (repeat of Figure 6.11(a))   (b) after editing

Figure 6.15 (a) shows a repeat of the original spectrum from Figure 6.11 (a), on page 374, which can be compared to the new order spectrum in (b) gained after the notch editing applied in the cepstrum. As can be clearly seen in Figure 6.15 (b), all of the discrete peaks from the 'carrier frequency' harmonic series, present at each integer order, have been successfully removed.

However, on comparing the plots of Figure 6.15 (a) and (b), it can be seen that the shape of the 'noise floor' in each plot does not exactly match, which would not be expected if information had only been removed from the integer order values. Furthermore, the discrepancy in amplitude ranges of both plots indicates that a significantly larger amount of signal has been removed by application of the cepstrum notch liftering process. By examining a zoom of both plots over the same identical amplitude range, the impact of the cepstrum notch lifter can be clarified.

(a)



(b)

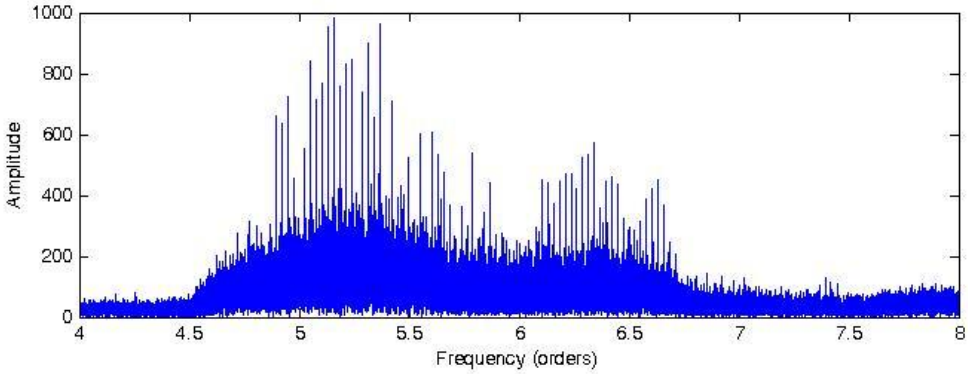


Figure 6.16 – Zoom of order spectra, before and after notch editing
(a) before editing (b) after editing

Figure 6.16 shows a zoom of the order spectra from Figure 6.15, between the order values of 4 and 8 orders, now with a common amplitude range. With this closer zoom, it can now be seen in Figure 6.16 (a) that the ‘noise floor’ seen in the envelope of Figure 6.15 (a), is not the true noise floor of the spectrum, and this envelope is caused by a field of very closely spaced discrete sidebands. By comparing Figure 6.16 (b) with (a), it can be seen that in addition to removing the targeted ‘carrier frequency’ harmonic series, which was the target of the notch lifter, the lifter has also removed an extensive number of sidebands from the signal.

As these sidebands are also order-dependent, they were to be removed in subsequent stages, so it is not a problem that they were inadvertently removed by an earlier editing stage. Upon examination, the cause of the removal of these sidebands by the notch filter was found to be a result of the rahmonic series, corresponding to many of the sidebands in the order spectrum, being coincident with higher rahmonic orders from the ‘carrier frequency’ series in the cepstrum, and so both the ‘carrier frequency’ rahmonics and the coincident ‘sideband’ rahmonics were removed by the same notches.

This occurrence of sideband rahmonics being coincident with their carrier frequency counterparts is very situational, and would depend on the particulars of a machine in question, and so is not expected to always occur whenever this editing method is employed.

6.6.2.1 Determining optimum notch width

As stated previously, for the notch lifter, a notch width of 7 samples was selected. This notch width was chosen by trial and error, by testing the two roving response signals from the first data set, with the success of each notch width being evaluated by determining the amplitude of the first order in the order spectrum, after editing. With better removal, the amplitude of the component at the first order should be minimised.

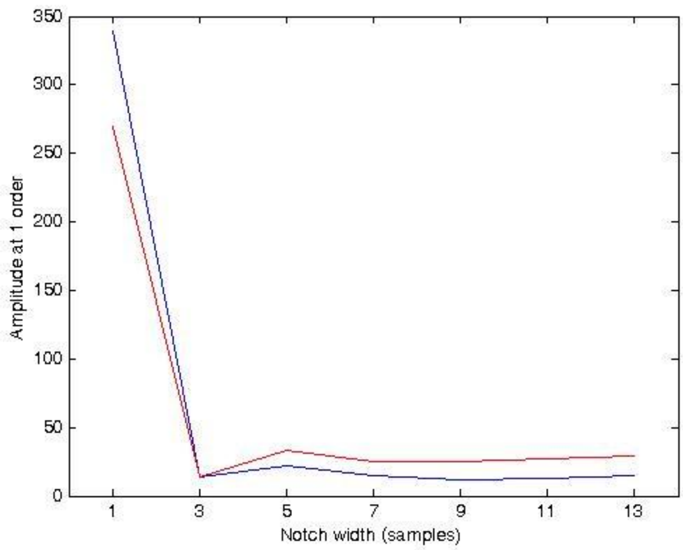


Figure 6.17 – Success of different notch widths at removing frequency at 1 order, for two response signals.

Figure 6.17 shows the effects of using a notch filter, with widths ranging from 1 sample to 13. Note that only oddly numbered widths are possible, for a notch centred over a specific sample. As can be seen in Figure 6.17, while only using a 1 sample notch is not ideal, using a notch width of 3 or higher gives approximately comparable success. A notch width of 7 was chosen to use for this experimental case, to allow for some leeway with smearing in other data sets, given only results from one data set were checked here.

It is recommended, by this author, to undertake a similar check to determine the optimum notch width to use for a given editing situation. The above evaluation method could be improved in multiple ways, if required for a given application. For example, the amplitudes at every harmonic order of the series being targeted for removal could be checked, rather than just one order; A number of response signals could be checked from multiple data sets, or even from every data set, rather than from just a single data set; The total (or average) power of the edited signal could be recorded as an alternative metric, as improved editing should lower the total power of the signal.

6.6.3 Second stage of Cepstrum editing using decaying exponential short pass lifter, of order-tracked signal

Looking back at the editing results, it is seen in Figure 6.16 (b) that some AM sidebands still remain in the edited spectrum, and so additional editing is needed to remove these.

In order to remove further order-dependent signal components from the signal, a second liftering stage was used in the cepstrum, in series with the first notch lifter. Note that the first and second liftering stages are done consecutively in the cepstrum, and the signal is not returned to the order domain between stages. The intermediate transformation back to the order domain above, after the first liftering stage, was done solely for evaluation purposes, and does not form part of the editing process.

At this stage, a second notch lifter could be employed in order to directly target the rahmonic series, corresponding to the sidebands still remaining in the edited order spectrum. However, on checking, the rahmonic spacing of the remaining sidebands was determined to have a high quefrency value, and so the lowest rahmonic of the series did not occur until relatively high quefrency.

As this editing is being performed to enhance OMA, it is suitable to employ an exponential short pass lifter. As described earlier, modal information is confined to low quefrency, and so for the application of pre-filtering a signal for OMA, it is appropriate to simply remove all high quefrency information which cannot be mode related. On checking the appropriate ”time” constant of an exponential lifter with this signal, the liftering position of the exponential lifter was found to be significantly lower than the first rahmonic from the remaining sidebands. As a result, an exponential short pass lifter will remove all high quefrency information, including the residual sidebands, and so an additional notch filter specifically targeting the sideband rahmonics was unneeded.

One significant issue with employing a decaying exponential lifter in the cepstrum based on the order domain, as opposed to the time domain, is that the lifter will apply a constant damping effect in the order domain. Once the signal is reversed back to the time domain, the constant damping in the order domain becomes a variable damping effect (variable damping constant) in the temporal spectrum. The variation is identical, in percentage, to the speed variation of the original signal, which in this case is $\pm 15\%$. As a result, the effects of applying an exponential lifter in the cepstrum, based on the order domain, cannot be easily compensated for in the final signal, as it can be with the comparable lifter applied in the time domain. This issue cannot be directly addressed by modifying the exponential short pass lifter applied in the order domain based cepstrum. However, it can be compensated for by applying an additional dominant low pass filter in the (time based) cepstrum, after the edited signal has been reversed to the time domain.

The absolute worst-case speed variation possible with the basic PDOT method is a 2:1 change. As a result, by using an additional exponential lifter in the cepstrum, based on the temporal spectrum, which has half the time constant (in samples, as both the order-tracked and reverse-order-tracked signal have approximately the same length in samples) as the exponential lifter applied in the cepstrum, based on the order spectrum, the combined effect of both lifters in the final time domain signal will be dominated by the effect from the lifter applied to the temporal cepstrum.

So, for the second stage of liftering, a decaying exponential short pass lifter is applied in the cepstrum, based on the order spectrum. This lifter has a ‘time constant’ of 800 samples, and then the third stage of liftering, in the temporal cepstrum, which is discussed later in section 6.6.5 on page 393, will have a time constant of 400 samples.

For the cepstrum, based on the order spectrum, the ‘time constant’ of 800 samples for the decaying exponential short pass lifter corresponds to 7.3982 rotations.

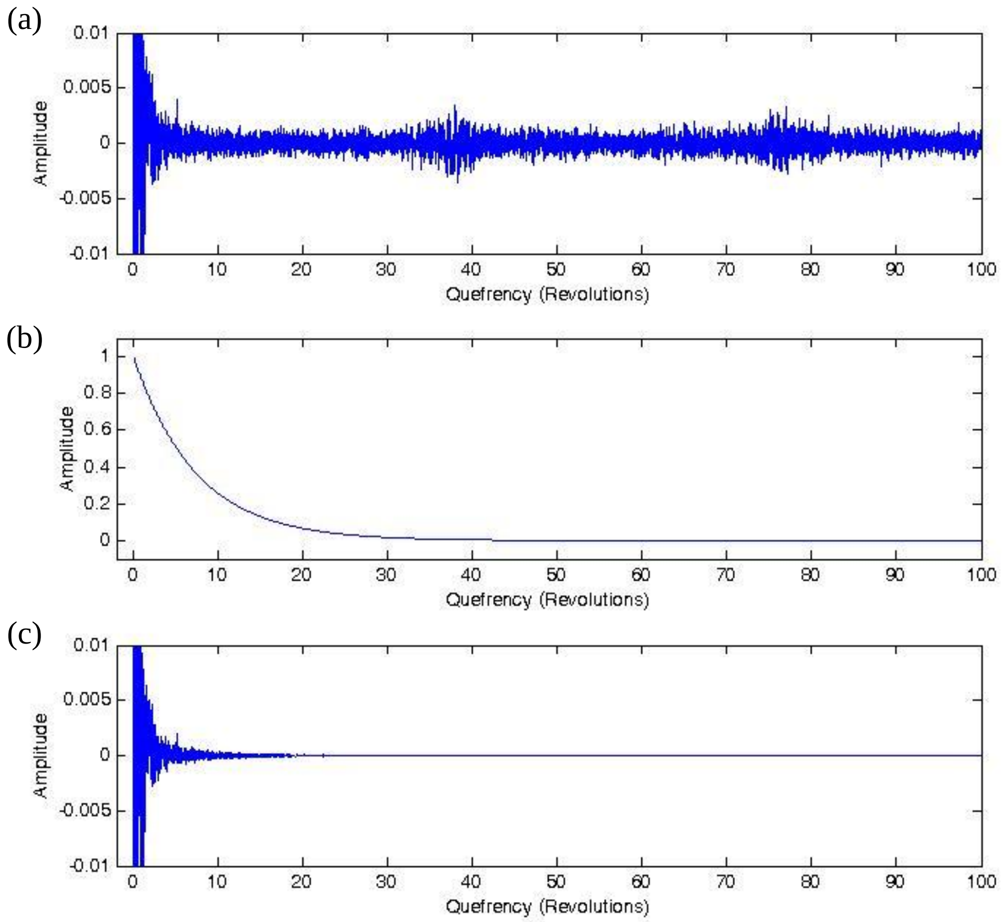


Figure 6.18 – Cepstra, based on order spectrum, before and after using decaying exponential short pass lifter
(a) before using lifter (b) decaying exponential short pass lifter used
(c) after using lifter

Figure 6.18 (b) shows the decaying exponential short pass lifter, with ‘time constant’ of 800 samples, or 7.3982 revolutions. Figure 6.18 (a) shows the cepstrum before the second stage of liftering is applied, which corresponds to the earlier Figure 6.13 (b) on page 378. Figure 6.18 (c) shows the cepstrum after the exponential short pass lifter has been applied, and by comparing with (a) it is clear that all high quefrency information has now been removed from the twice-edited signal in (c), while leaving the modal information, present at low quefrency, though smeared, in the twice-edited signal.

In regards to the sidebands which were remaining in the signal before the second stage of editing, the first rahmonic of these sidebands is visible in Figure 6.18 (a) at approximately 38 rotations. As can be seen in Figure 6.18 (b), the applied exponential lifter, as a second stage of liftering, takes effect well below 38 quefrency, and in (c) all rahmonics corresponding to the residual sidebands have been removed. However, it should be specifically noted that if sidebands are present in a signal at quefrency low enough to remain after exponential short pass liftering, then a further specific stage of notch liftering might be necessary to remove the low quefrency sideband information.

In order to evaluate the success of this second stage of liftering, the now twice-edited signal is again transformed back to the order spectrum, by combining with the original phase, as per equation (6.11) on page 353.

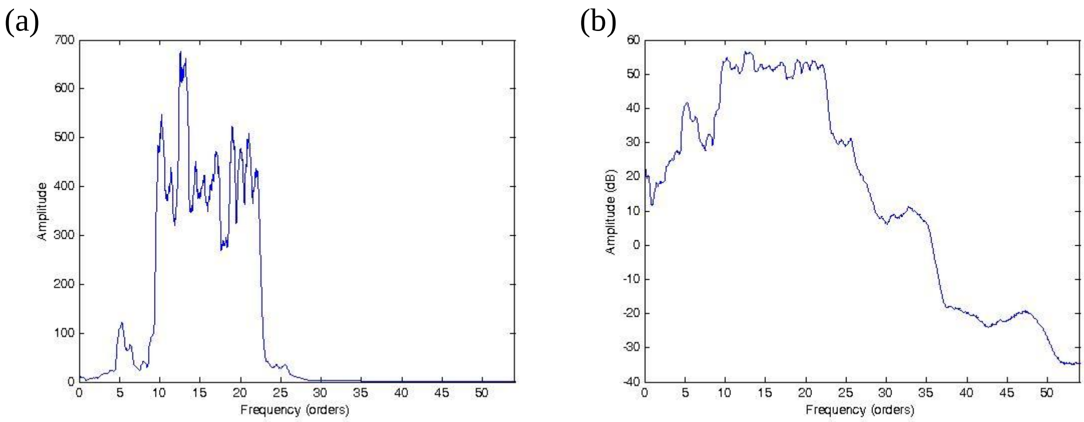


Figure 6.19 – Order spectrum after second stage of liftering
(a) linear scale (b) dB scale

Figure 6.19 (a) and (b) show the order spectrum after the exponential lifter has been applied as a second stage of liftering, with (a) on a linear scale and (b) the same plot on a dB scale. It can be seen that all high quefrency information has been removed from the signal, and the final edited signal contains no discrete ‘carrier frequency’ components at integer order value, or any modulation sidebands around the ‘carrier frequency’ order values. In essence, the decaying exponential short pass lifter has de-noised the entire order spectrum leaving primarily modal information.

At this stage, it can be noted that the edited order spectrum now superficially resembles a typical FRF which would be obtained with EMA through averaging, with the edited order spectrum now being quite smooth. However, it should be specifically emphasised that no averaging has been used to arrive at this edited order spectrum, so the edited order spectrum retains the original fine frequency resolution, despite now being relatively smooth. On the other hand, an FRF is typically averaged tens or hundreds of times to arrive at a comparably smooth FRF, and has a correspondingly lower frequency resolution given by the averaging fraction (e.g. averaging 100 times with no overlap results in a 100 times lower frequency resolution (less fine), and signal overlapping decreases the frequency resolution further), so this OMA editing process is preferable, in that it is able to retain the original full frequency resolution.

6.6.4 Reversing order-tracking, using PDOT

Now that all order related input signal components have been removed from the order-tracked signal, the signal can be reversed back to the time domain, using the reversible PDOT method described in 4.5, which starts on page 236. The signal is first transformed back to the phase angle domain, from the cepstrum, and then the order-tracking is reversed, changing the signal to the time domain.

Note the edited signal is reversed back to the same sampling rate as the oversampled original signal. As one of the first steps of PDOT, the original signal is oversampled by a factor of 2, so the edited signal reversed back to the time domain has twice the sampling rate of the original recorded response signal, so $f_s = 4096 \times 2 = 8192$ Hz. However, it should be noted that any frequency information above the original Nyquist frequency of 2048 Hz in the new edited and reversed signal is simply noise, introduced by the editing process, as the original signal contained no information above this frequency.

Note that the signal reversed back to the time domain is typically a few samples shorter than the original oversampled signal. This is a result of not using extrapolated values with the phase-time curve when reversing the order-tracking process. Typically, when going forward, the phase sampling has a slightly shorter span (in samples) than the original time span (in samples) to avoid extrapolating values. Then, when reversing the order tracking, the ‘new’ time span (in samples) must then be slightly shorter than the phase span (in samples), again to avoid extrapolating values, with the net result that the reversed time signal is multiple samples shorter than the original oversampled signal. Typically, the loss of samples is a fraction of one revolution of data (often only

1-2 samples in total), and has no meaningful impact on analysis. However, the shortening means that it will often be advisable to truncate back to the next lower integer number of periods, so as to give order spectra with discrete lines.

After reversing the edited signal back to the time domain (not pictured), the signal can be transformed to the frequency domain.

(a) (b)

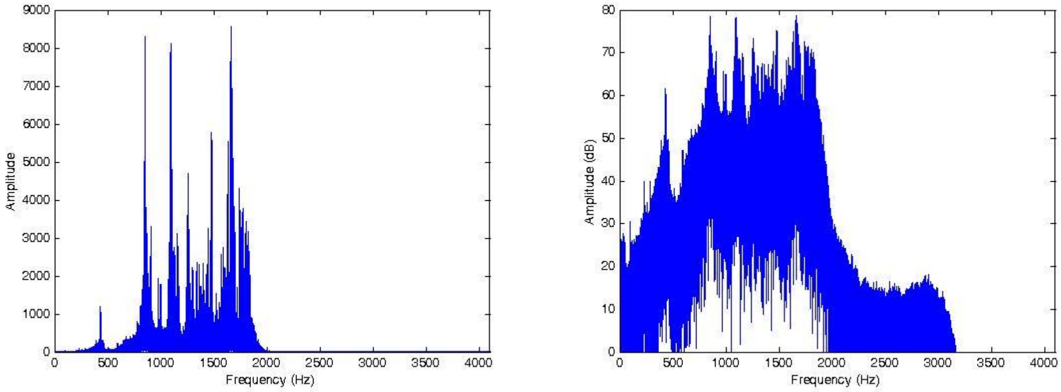


Figure 6.20 – Spectrum of edited signal, reversed back to the time domain
(a) linear scale (b) dB scale

Figure 6.20 (a) and (b) show the spectrum of the edited signal, which has been reversed back to the time domain. In summary, this signal has had two liftering stages applied in the cepstrum, based on the order domain, and then been transformed to the frequency domain. As can be seen in Figure 6.20 (a) and (b), the spectrum no longer contains any order-based components, when compared to the original spectrum plots in Figure 6.9 on page 372. However, when compared to Figure 6.19 on page 389, the spectrum shown in Figure 6.20 is no longer smooth, and has effectively been re-noised across the entire spectrum by the reverse order-tracking process.

6.6.5 Third stage of Cepstrum editing using decaying exponential short pass lifter, of time based signal

The re-noising of the signal is also evident in the cepstrum, corresponding to the spectrum.

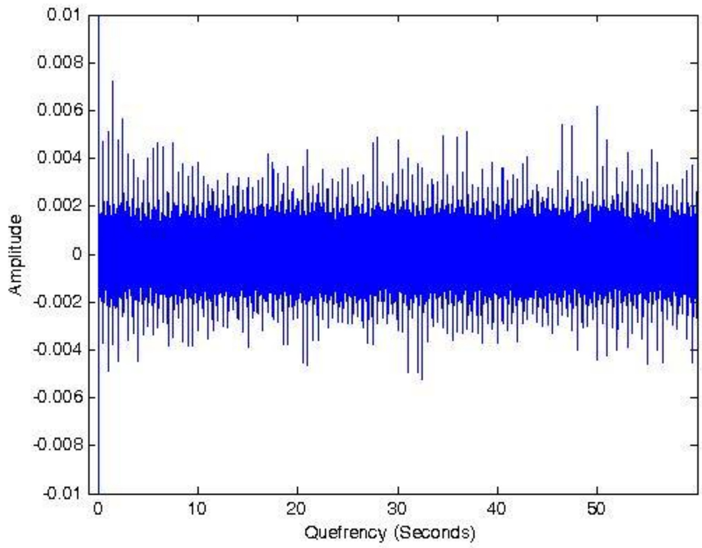


Figure 6.21 – Cepstrum of edited signal, based on spectrum, after 2 stages of editing, showing positive quefrencies

Figure 6.21 shows the entire positive side of the cepstrum, based on the spectrum, of the edited signal (after 2 stages of editing), and it can be seen that the entire cepstrum has been re-populated with ‘noise’. Note the change of units of the cepstrum plot, when compared to earlier cepstrum plots, now the cepstrum is based on the spectrum rather than the order spectrum.

As stated in section 6.6.3 on page 385, a third stage of editing is now employed. For this stage, a second decaying exponential short pass lifter is applied to the cepstrum. As stated previously, this exponential lifter has a time constant of half the second stage (in samples), so this lifter will dominate the final change to damping, resulting in a constant damping effect (in time) being the dominant effect applied to the final signal. As stated previously, the time constant of the decaying exponential lifter is 400 samples, which in the time domain based cepstrum corresponds to a quefrency value of:

$$\tau = \frac{1}{f_s} \times S_\tau = \frac{1}{8192} \times 400 = 0.0488 \text{ s} = 48.8 \text{ ms} \quad \text{......................} (6.14)$$

where τ is the quefrency value of the time constant, f_s is the sampling rate of the reversed signal, and S_τ is the number of samples for the time constant. The damping factor σ introduced by the decaying exponential lifter with time constant τ is given by:

$$\sigma = \frac{1}{\tau} = \frac{1}{0.0488} = 20.48 \text{ rad/s} \quad \text{or} \quad \frac{20.48}{2\pi} = 3.26 \text{ Hz} \quad \text{...................} (6.15)$$

and this added damping adds $2\sigma = 6.52$ Hz to the 3 dB bandwidth of all resonances (and antiresonances) in the final spectrum, which can in principle be exactly compensated for after OMA analysis has been completed.

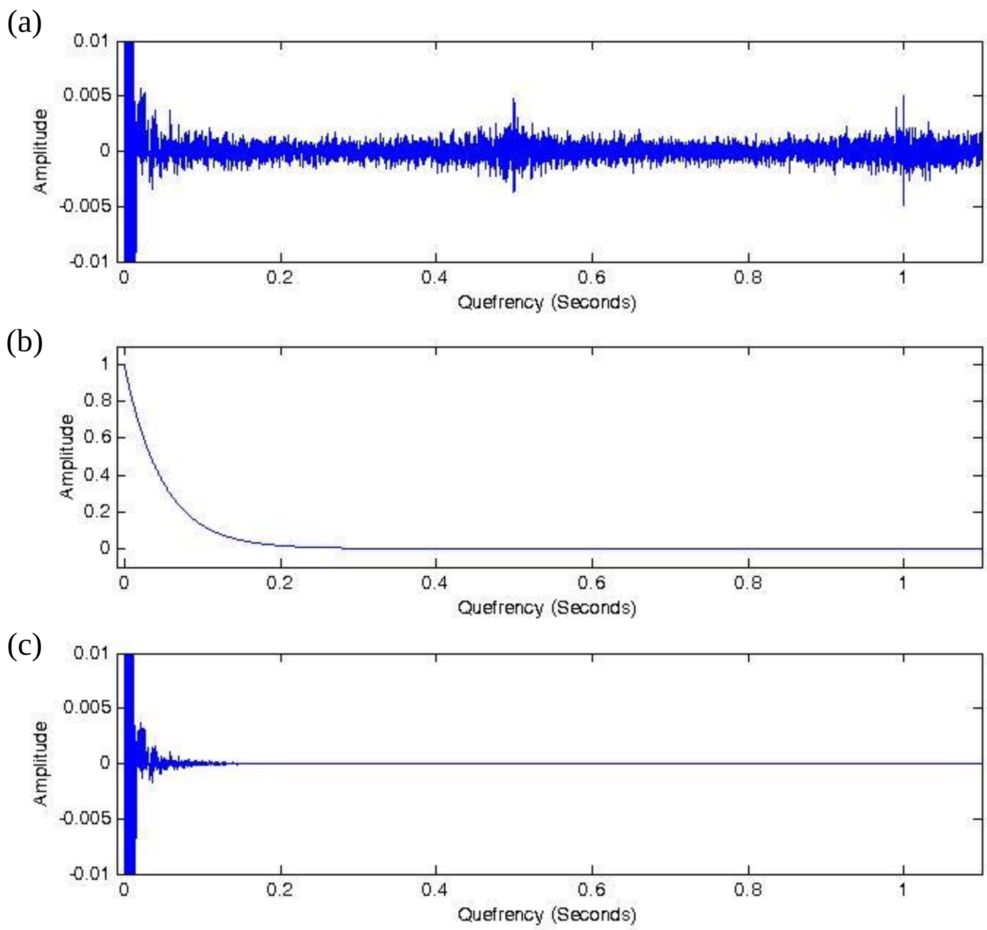


Figure 6.22 – Cepstra, based on time domain, before and after using second decaying exponential short pass lifter

(a) before using lifter (b) decaying exponential short pass lifter used

(c) after using lifter

Figure 6.22 (b) shows the decaying exponential short pass lifter, with time constant of 400 samples, or 48.8 ms, which is used to lifter the cepstrum as the third stage. Figure 6.22 (a) shows the cepstrum before the second stage of liftering is applied, which corresponds to the earlier Figure 6.21. Figure 6.22 (c) shows the cepstrum after the exponential short pass lifter has been applied, and by comparing with (a) it is clear that all high quefrency information has again been removed from the edited signal in (c) by the third stage of liftering, while leaving the modal information, present at low quefrency, in the three stage edited signal.

One interesting feature to note in the zoomed cepstrum in Figure 6.22 (a), before the third stage of editing is applied, is that a noticeable rahmonic series is present with quefrency spacing of 0.5 s. This quefrency spacing corresponds to a 2 Hz spacing in the frequency domain, which is equal to the modulating frequency of the original speed varying component, from the input signal. At this stage it has not been identified if this rahmonic series either remained during the editing process conducted on the order-tracked signal, or was introduced by reversing the order tracking process, as the reverse-order-tracking effectively re-modulates the signal by the same speed variation parameters originally compensated for with the forward order-tracking. Investigating the source of this rahmonic series, in a signal which has been order tracked, edited, and then reversed back to the time domain, will be a topic for future work.

As with previous editing stages, the edited cepstrum signal can be reversed back to the spectrum by combining with the original phase, as per equation (6.11) on page 353.

(a) (a)

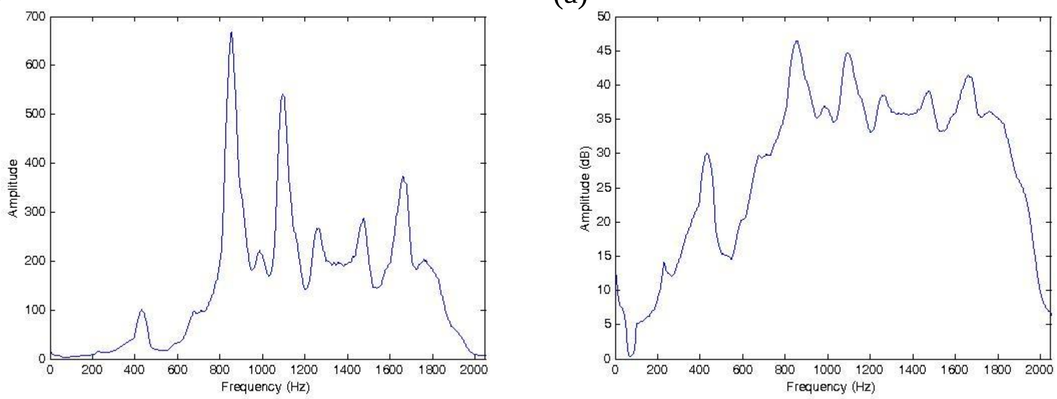


Figure 6.23 – Spectrum after third stage of liftering, showing 0 to 2048 Hz
(a) linear scale (b) dB scale

Figure 6.23 shows the spectrum after the decaying exponential short pass lifter has been applied to the cepstrum, as the third stage of liftering, where (a) shows the spectrum on a linear scale, and (b) shows the same spectrum on a dB scale. Note the second half of the positive frequencies above the original sampling rate, which only contain noise, have not been shown in these plots. When compared to Figure 6.20, on page 392, it is clear that Figure 6.23 has again been de-noised by the application of the exponential short pass lifter, only this time in the time based signal. As the effects of the input signal have been edited from the signal, the final edited signal now consists of almost entirely modal information.

This edited signal can be transformed back to the time domain, as per equation (6.12) on page 353, and can then be subsequently used with any appropriate OMA technique.

Unfortunately due to time constraints, the signals recorded for this experimental test have not been processed with an OMA technique to obtain final modal information. Additionally, due to the issue with the force signals being faulty, no comparable FRF’s were calculated using EMA techniques with this same data, for comparison to the spectrum produced here.

However, as mentioned previously, an experimental modal study was conducted by Deshpande et al. [53] (also [54]) using the same gearbox casing, which used a roving impact hammer method, rather than the shaker used here. While the EMA test had a different mounting method, along with the different excitation method and transducer positions, and so was not directly comparable, the FRFs gained in their study should have similar modal frequencies.

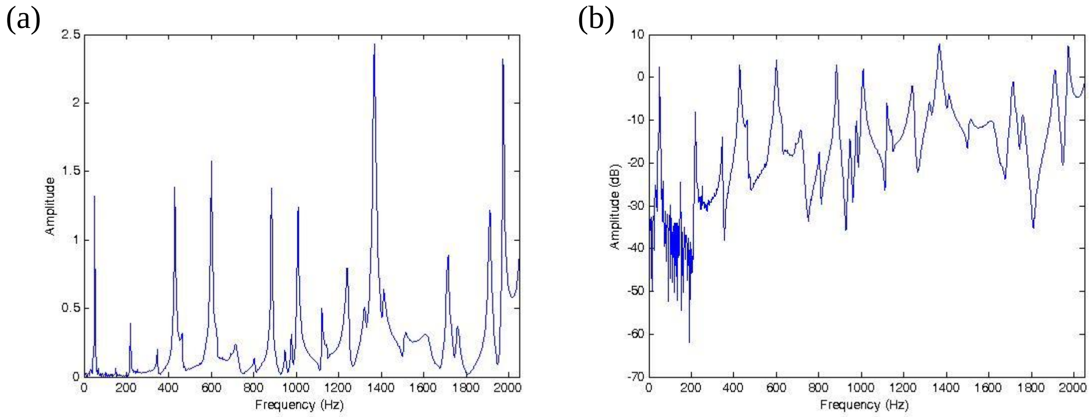


Figure 6.24 – Similar FRF from EMA test by impact testing
(a) linear scale (b) dB scale

Figure 6.24 shows a typical FRF from the EMA test conducted in [53] (note [53] and [54] do not contain these FRF plots, however these plots were made from data used in those works). Figure 6.24 (a) shows the FRF on a linear scale, and (b) shows the same FRF on a dB scale. By comparing Figure 6.24 with Figure 6.23, it can be seen that the edited response signals, from this work, contain many of the same, or similar, modal frequencies to those present in the FRF shown. The modal peaks are broader in Figure 6.23, indicating the presence of higher damping, but as discussed previously this is expected, as the exponential short pass lifter stages increase the damping of the modes in the edited response signal in Figure 6.23. Note that the fixed location reference points for the edited response signal (shaker attachment point) and FRF (response accelerometer position) were not the same, and so the FRF is not for the same transfer function as the edited response signal, and so both are not expected to be identical.

6.7 Pre-processing of the array of experimental signals, to remove order components

The above example illustrates the editing process, which was then applied to all the response signals recorded for the experimental test. As discussed previously, data was recorded in 62 separate data sets, with three response signals (one fixed) per data set.

To summarise the editing method described above, the editing process involves:

1. Order-track response signals, to phase angle domain
2. Apply stage 1 editing in cepstrum (based on order domain), which was a notch lifter, targeting ‘carrier harmonics’
3. Apply stage 2 editing in cepstrum (based on order domain), which was a decaying exponential short pass lifter, removing all high quefrency information including sidebands
4. Reverse order-tracking, back to time domain
5. Apply stage 3 editing in cepstrum (based on spectrum), which was a shorter decaying exponential short pass lifter, removing reintroduced noise, and dominating the time constant.

While this basic process was successful in editing many of the response signals, a portion could not be initially edited. The problem was that some of the response signals could not be initially order-tracked using the basic PDOT method, when using the individual signal itself as the reference signal.

As listed in section 3.2.1, starting on page 131, it is a requirement that the (time domain) amplitude within a selected demodulation band to be used for PDOT must not be equal to zero at any point, or incorrect order-tracking occurs. As discussed in the section on the basic phase demodulation method, in section 2.6.1 starting on page 120, an amplitude value of zero within a demodulation band causes incorrect discontinuities in the associated demodulated phase values.

On investigating the signals which initially were incorrectly order-tracked, by examining their demodulated amplitude, as discussed in section 2.6.2, starting on page 125, the demodulation bands from these signals were found to contain amplitude values going down to zero, which was introducing phase discontinuities in the corresponding demodulated phase, causing incorrect order tracking. Such signals should not be used as references for PDOT.

In all previously processed results with the PDOT method, the occurrence of zero amplitudes preventing order-tracking with a given signal has been exceedingly rare. For reasons which have not yet been determined, an abnormally large portion of the experimental response signals recorded with this specific study exhibited this phenomenon.

However, by realising that all the signals within a data set are recorded simultaneously, all can be order-tracked by using a common reference signal, the best in each group.

For each data set, it is only required that one response signal be suitable for use as a reference signal, for all three response signals.

To highlight the frequency and impact of the presence of these zero amplitude components in the recorded experimental response signals, every response signal from every data set was tested for suitability for use as a reference signal, and Table 6.3 below shows the success of using each response signal as a reference signal, with successful order-tracking indicated by and unsuccessful order-tracking indicated by . As previously stated, each data set contained three response signals, and these have been given the abbreviations of 'CA' (Common Accelerometer), for the driving point response accelerations signal which is common for every data set, 'RA 1' (Roving Accelerometer) for the first roving accelerometer signal in a given data set, and 'RA 2' for the second roving accelerometer signal. Note that the roving accelerometer locations are different with each data set.

Also note that the example response signal, used previously in this work to illustrate the developed editing method, corresponds to RA 1, from data set 1, in Table 6.3.

Data Set	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
RA 1				X			X							X		
RA 2				X		X	X		X		X	X				X
CA	X	X	X		X	X	X	X	X	X	X	X	X	X	X	

Data Set	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
RA 1			X					X								X
RA 2		X		X	X								X		X	
CA	X	X	X	X	X	X	X		X	X	X	X		X	X	X

Data Set	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
RA 1							X		X		X	X		X	X	
RA 2	X								X	X		X		X	X	
CA	X	X	X	X	X	X	X	X		X	X		X	X	X	X

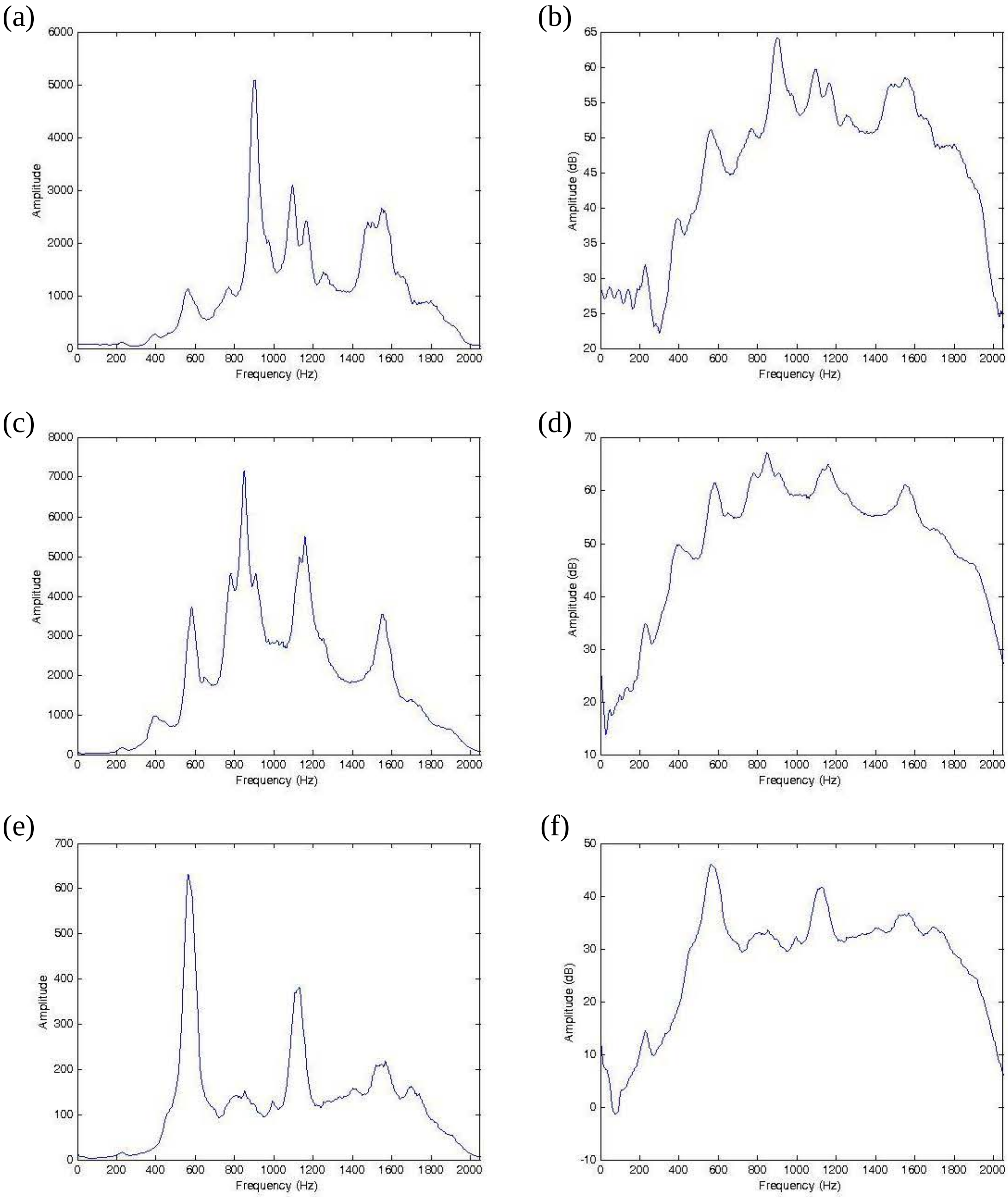
Data Set	49	50	51	52	53	54	55	56	57	58	59	60	61	62
RA 1					X			X	X		X			X
RA 2					X	X		X	X	X	X			
CA	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Table 6.3 – Success of initial order-tracking using different response signals

order-tracking successful   X order-tracking unsuccessful

As can be seen from Table 6.3, a significant number of response signals were unsuitable for use as the reference signal for order-tracking, with 97 out of 186 individual response signals, or 52.15 %, being unsuitable for use. However, this is mitigated by the fact that only one response signal out of each data set needs to be usable for the whole data set, and as seen in Table 6.3, only 7 out of 62 data sets, or 10.3 %, have all three response signals as unsuitable, and so could not be order-tracked.

So, by using one of the valid response signals as reference signal, a total of 165 response signals from 55 data sets were pre-processed using the steps described in this work. This pre-processing was successful in every instance, including the reverse PDOT order-tracking stage, and in each case an edited time signal was produced, which has a spectrum comparable to an FRF obtained using EMA, and is suitable for use with subsequent OMA analysis.


(a)
Amplitude
Frequency (Hz)
(b)
Amplitude (dB)
Frequency (Hz)
(c)
Amplitude
Frequency (Hz)
(d)
Amplitude (dB)
Frequency (Hz)
(e)
Amplitude
Frequency (Hz)
(f)
Amplitude (dB)
Frequency (Hz)

(g) (h)

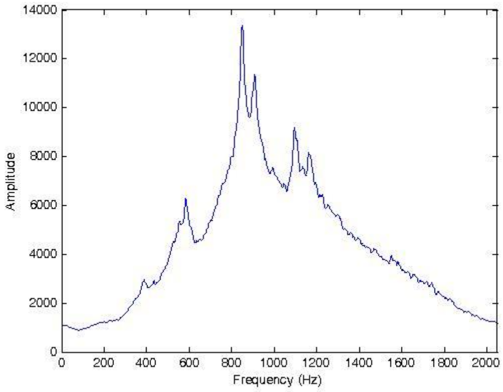


Figure 6.25 – Spectra of different edited response signals
(a, c, e, g) linear scale (b, d, f, h) dB scale
(a, b) Data set 8, RA 1 (c, d) Data set 19, RA 2
(e, f) Data set 38, RA 1 (g, h) Data set 44, CA

Figure 6.25 shows the response spectra for a selection of pre-processed signals from different data sets, and from a mixture of roving and fixed accelerometer positions, and it can be seen that in each case a spectrum comparable to Figure 6.23, on page 396, and Figure 6.24, on page 398, has been produced, indicating correct pre-processing has been performed in each instance.

So, the editing process described here was successful in pre-processing every response signal, which could be initially order-tracked, by removing order based components from a variable speed signal. These signals should all be suitable to use with subsequent OMA analysis.

While some signals could not be processed, this was a result of the signals not being suitable for use as a reference signal, with the PDOT method. This was a result of a known limitation with phase demodulation, upon which the PDOT method is built.

It can also be concluded that the reversible approach to applying PDOT was successful in every instance, where the forward PDOT process could also be employed.

At this stage the demonstration of this pre-processing method is only indicative, as subsequent OMA analysis has not been applied, and directly comparable EMA results were not available. However, it is believed that this editing scheme is suitable for the application, and fully validating this editing scheme will be a task for future work.

Chapter 7 Conclusions and Future Work

Chapter Overview

The following chapter lists the conclusions from this thesis work, and highlights the successful development of the generalised methodology for PDOT in this work.

Subsequent avenues for continuing research with PDOT, as future work, are then identified.

7.1 Conclusions

As stated in the Aims section 1.1.4 on page 9, the basic aim of this doctoral research work was to expand greatly upon the single record order tracking method I developed for my BE thesis in [9], which has been successfully accomplished with this thesis work.

What began as a basic concept to use phase demodulation based order-tracking, suitable only for use with a tacho signal, and presented only with limited simulated results, with no experimental confirmation, in my BE thesis, has been dramatically expanded into a generalised methodology for Phase Demodulation based Order-Tracking (PDOT), which can be implemented using a vast number of different options in a modular fashion, culminating with this thesis work.

The basic mathematical model, upon which the BE thesis work was based, has been progressively updated through the course of multiple peer reviewed academic publications, resulting in the comprehensive mathematical model presented here in Chapter 3 which successfully addresses many, if not all, of the factors relevant to variable speed situations which had not been addressed in prior works. This comprehensive mathematical model has allowed the formulation of the detailed conditions required for the successful implementation of PDOT, including the limits to permissible speed variations possible with the basic PDOT method, and for the development of relevant practical guidelines to determine if this suitability is met with a given real-world signal, as presented in this work.

The different modular components which can be used with the PDOT method were explained in detail in Chapter 4.

To summarise, the PDOT can be employed with a variety of reference signal types, including using no separate reference signal, and the differences between the results produced using these different reference signals were explained in detail in this work.

The PDOT method can be employed with a single stage of order-tracking, or be used with the novel multiple-iteration approach developed in this work, which allows for further refining of the order-tracking process for higher harmonics, even in the presence of large speed variations.

This thesis contains the first publication of a new segmented approach to PDOT, which extends PDOT so it can be employed with almost any speed variation.

This thesis also contains the first publication of a reversible approach to PDOT, which allows an edited order-tracked signal to be re-transformed back to the original time domain, allowing for joint time and order based analysis of the same signal with minimal masking.

Multiple experimental examples are presented for different applications of PDOT, highlighting the suitability of the PDOT method for many variable-frequency applications.

These examples include:

- bearing diagnostics and gear diagnostics, in the presence of small to large speed variations, showing successful detection of faults.
- Using multiple stages of order tracking, to significantly refine order-tracked results.

- Conducting a segmented approach to PDOT, which allowed for the successful order-tracking of a run-up signal.
- Successful pre-processing of signals suitable for use with subsequent OMA methods, where order-based components were extracted from a signal using cepstrum editing, which was then returned to the time domain, which would allow for subsequent analysis of modal information without the presence of contaminating signal components.

The above experimental example, of successfully pre-processing of signals before OMA, is also the first publication of a novel approach to use cepstrum editing on an order-tracked signal, before subsequent analysis in the time domain. While this novel approach to using cepstrum editing is presented in conjunction with PDOT, it could feasibly be employed with any reversible order-tracking method, and is novel in its own right.

7.2 Future work

This work on PDOT additionally leads to an extensive number of subequent topics which could be researched to improve the PDOT methodology. These topics both consist of logical progressions to those presented in this work, which will result in additional new order-tracking processes, as well as topics which have been touched on briefly in this work, which could be investigated more fully.

It is the intent of this Author to investigate some of these topics as future work, however this is not the case for every identified topic, and many might be suitable for investigation as part of an undergraduate, masters, or similar research topic.

The following is a summary of many of the additional subsequent topics which have been identified to improve the PDOT method. However, given the volume of material in this work, it is expected that there would be additional sunsequent topics beyond those summarised here.

In regards to the basic PDOT method, as discussed in Chapter 2:

- In [9], a preliminary study was made evaluating the effects of different parameter selections on the accuracy of phase demodulation based order-tracking, using simulated signals which allowed for a quantitative evaluation of the success of order-tracking. These parameters included the effects of sample rate, signal length (frequency resolution), carrier frequency, in addition to the modulating frequency and frequency deviation parameters which are also investigated in this work. A more comprehensive investigation of the effects of chosen parameters on the basic PDOT method could be undertaken, following the updated work in this thesis.

- A large part of this thesis work, for the basic PDOT method, has focused on the impact on order-tracking results when a higher order contaminates the harmonic order to be used for order tracking, establishing the importance of minimising the contamination to the demodulation bandwidth used to conduct order-tracking. A number of processes, including acceptability charts, have been dedicated to allow an application to be evaluated to determine if contamination from higher orders is small enough to allow accurate order-tracking. One thing which has not been established, is if unrelated (random) noise has the same impact on order-tracking results, as related contamination from higher orders.

  If unrelated (random) noise has a lesser impact, this would simplify the application of the PDOT method, and allow for it to be employed in instances which are currently excluded under the current PDOT guidelines as presented in this work.
- Develop automated methods to determine when a zero amplitude phenomenon has occurred, causing incorrect phase demodulation, and extraction of the incorrect phase-time map, which was seen to be affecting results in section 6.7 on page 399.
  - Investigate techniques which may allow for a signal which contains a zero amplitude phenomenon to be correctly order-tracked using PDOT techniques, by either bypassing or compensating for the phenomenon.
- In regards to an acceptable reference signal for PDOT, it has been identified that constant time width pulses from a tachometer are not suitable to be used,

and only constant aspect ratio based pulses are acceptable.

However, pulse type tachometers and encoders with constant aspect ratio pulses have a widely dispersed quality, in regards to how accurately the pulses correspond to the machinery being measured.

It would be of interest to investigate how sensitive/insensitive the PDOT method is to different qualities of constant aspect ratio reference signals, from different quality tachometers and encoders. Ideally, this study would be conducted with a variety of different quality tachometers and encoders for comparison; however it is possible this might be able to be evaluated in a simulated fashion.

This topic was not investigated during this thesis work, as only a single physical tachometer type was available with each experimental setup, and so comparison measurements could not be undertaken.

In regards to the modular extensions to the PDOT method:

- A variety of modular options which can be employed with the PDOT method are presented in Chapter 4. These modular options are all suitable to be utilised together, in a variety of combinations for different applications.

  In the following application example chapters, only a select number of examples are presented, which employ only a small number of the total possible modular combinations.

  It would be of interest to identify additional mechanical applications which require modular combinations of the PDOT which have not been investigated in this work, in order to evaluate the success of the modular PDOT approach with different modular combinations.

In regards to the specific modular option of using mult-stage PDOT:

- Develop methods to improve the implementation of the multi-stage approach when using the response signal as the reference, by removing the effects from amplitude modulation, which impacts the successful application of multiple stages. The effects of this impact were seen in results in section 5.3.2 on page 273

In regards to the specific modular option of using segmented based PDOT:

- For the segmented based PDOT method, two alternative philosophies to undertaking the segmenting process were dicussed in this work. The first was the optomised approach, which requires a recursive approach to calculations, as discussed in section 4.4.3 on page 212. The second is a simplified approach, where a common conservative percentage speed range is selected which is suitable for every segment, as discussed in section. The example studied in this work using a segmented approach only employed the simplified method to conducting segmenting. It would be of interest to directly compare the differences (if any) between using either an optomised, or a simplified, approach to determining the segment locations.

In regards to the application examples for PDOT:

- All of the application examples presented in this work, both for the basic PDOT, as well as the specific usage of the modular extensions, have only had very low percentage modulating frequency values. Primarily, this is due to most mechanical applications having, in general, low percentage modulating frequencies when compared to the corresponding frequency deviation range.

While no examples have been presented in this thesis, the PDOT method is believed to be suitable for use with a wide range of percentage modulating frequency values, and is not limited to the low cases presented in this work. Initial studies with simulated sinusoidal signals showed comparable accuracy across the full range of acceptable modulating frequencies, as indicated in the acceptability charts (section 3.6 on page 161; Appendix A on page 423).

In regards to the specific application of using reversible PDOT to pre-filter a signal before conducting OMA, as presented in Chapter 6 on page 325:

- Conducting complete experimental validation of the reversible PDOT approach to pre-process a signal using cepstrum editing, to remove order based information so subsequent OMA methods can be employed, by conducting a direct comparison to EMA, to fully confirm the method presented in Chapter 6.
- In the presented experimental work, for using a reversible PDOT approach to pre-process a signal before OMA, the random white noise component in the excitation signal was constructed to be 40dB below the level of the FM component. If the proposed study, listed above for the basic PDOT method, to determine if random noise has the same impact as contamination from higher harmonic orders from the FM component, finds that the basic PDOT method is more insensitive to random noise, then it would be suitable to re-test the reversible PDOT approach to pre-filter a signal before OMA with higher levels of random noise, to check if the method is still valid with a higher level of random noise relative to the FM component.

In regards to comparing the basic and modular extensions to PDOT to other order-tracking methods:

- Where comparable methods are available, it would be of interest to analyse a common signal/application with both a PDOT method and an alternative computed order-tracking method, to establish a quanitative measure of the differences in accuracy, to better evaluate the success of the PDOT methodology.
- More generally, it would be of interest to conduct a 'benchmarking' of all the current methods of computed order-tracking, by utilising all the methods to separately order-track a benchmark, or series of benchmark, signals, in order to gain a quantitative measure of the accuracy difference between different computed order-tracking methods.

  One complicated facet of this undertaking would be in identifying a suitable series of benchmark signals, and corresponding analysis applications for study. This is because many of the methods for computed order-tracking are optimised for a specific analysis method, or signal type, and so selecting a series of benchmark signals which both allow comparison without biasing results to a particular method would be problematic.
    - In a more expansive fashion, it would be beneficial to expand the benchmarking process beyond just computed order-tracking, and cover all current order-tracking methods.

      This would be more complicated, as some order-tracking methods output completely different format of results, which would make direct comparison more difficult.

References

1. Coats, M.D. and R.B. Randall, *Single and multi-stage phase demodulation based order-tracking.* Mechanical Systems and Signal Processing, 2014. **44**(1–2): p. 86-117.

2. Randall, R.B., N. Sawalhi, and M.D. Coats, *A comparison of methods for separation of deterministic and random signals.* The International Journal of Condition Monitoring, 2011. **1**(1): p. 11-19.

3. Randall, R.B., N. Sawalhi, and M. Coats, *Separation of Gear and Bearing Fault Signals from a Wind Turbine Transmission under Varying Speed and Load*, in *Proceedings of CMMNO'2012 – Condition Monitoring of Machinery in Non-Stationary Operations* 2012. p. 3-12.

4. Coats, M.D. and R.B. Randall, *Order-Tracking with and without a tacho signal for gear fault diagnostics*, in *Proceedings of Acoustics 2012 Fremantle, Fremantle, Australia, 21–23 November*. 2012.

5. Coats, M.D. and R.B. Randall, *Compensating for speed variation by order tracking with and without a tacho signal*, in *Proceedings of VIRM10 – Tenth International conference on Vibrations in Rotating Machinery, London, United Kingdom, 11-13 September*. 2012.

6. Coats, M.D. and R.B. Randall, *Single Record Order-Tracking*, in *In Proceedings of CM 2010 and MFPT 2010, Ettington Chase, Stratford-upon-Avon, England, 22–24 June*. 2010.

7. Coats, M.D., N. Sawalhi, and R.B. Randall, *Extraction of tacho information from a vibration signal for improved synchronous averaging*, in *Proceedings of Acoustics 2009, Adelaide, Australia, 23–25 November*. 2009.

8. Coats, M.D. and R.B. Randall, *Order tracking over a wide speed range such as run-up or run-down*, in *the 9th IFToMM International conference on Rotor Dynamics, Milan, Italy, 22-25 September*. 2014.

9. Coats, M.D., *Single Record Order-Tracking*, in *School of Mechanical & Manufacturing Engineering*. 2006, The University of New South Wales, Australia.

10. Bonnardot, F., M. El Badaoui, R.B. Randall, J. Danière, and F. Guillet, *Use of the acceleration signal of a gearbox in order to perform angular resampling (with limited speed fluctuation).* Mechanical Systems and Signal Processing, 2005. **19**(4): p. 766-785.

11. Bossley, K.M., R.J. McKendrick, C.J. Harris, and C. Mercer, *Hybrid Computed Order Tracking.* Mechanical Systems and Signal Processing 1999. **13**(4): p. 627-641.

12. Fyfe, K.R. and E.D.S. Munck, *Analysis of Computed Order Tracking.* Mechanical Systems and Signal Processing, 1997. **11**(2): p. 187-205.

13. André, H., J. Antoni, Z. Daher, and D. Rémond. *Comparison between angular sampling and angular resampling methods applied on the vibration monitoring of a gear meshing in non stationary conditions*. in *Proceedings of the International Conference on Noise and Vibration Engineering, Leuven, Belgium, September 20-22*. 2010.

14. Groover, C.L., M.W. Trethewey, K.P. Maynard, and M.S. Lebold, *Removal of order domain content in rotating equipment signals by double resampling.* Mechanical Systems and Signal Processing, 2005. **19**(3): p. 483-500.

15. Lembregts, F., J. Top, and F. Neyrinck, *Off-line synchronous resampling of vibration measurements*, in *Proceedings of IMTC-96. Instrumentation and measurement technology conference, 1996. Quality measurements: The indespensable bridge between theory and reality, IEEE*. 1996.

16. Lembregts, F., J. Top, and F. Neyrinck, *Adaptive resampling for off-line signal processing*, in *Proceedings of IMAC 14, the international modal analysis conference, Dearborn (MI), USA, 12-15 February*. 1996.

17. Lembregts, F., J. Top, and F. Neyrinck, *Adaptive resampling for off-line signal processing*, in *Proceedings of international conference on Noise and Vibration Engineering 21, Leuven, Belgium*. 1996.

18. Lembregts, F., J. Top, and F. Neyrinck, *Adaptive resampling for off-line signal processing*, in *Proceedings of SPIE the international society for optical engineering*. 1996.

19. Bonnardot, F., R.B. Randall, and J. Antoni, *Enhanced unsupervised noise cancellation using angular resampling for planetary bearing fault diagnosis.* International Journal of Acoustics and Vibrations, 2004. **9**(2): p. 51-60.

20. El Badaoui, M., J. Antoni, F. Guillet, and J. Daniére, *Rééchantillonnage angulaire par interpolation spline-cubic des signaux d'engrenages*, in *Proceedings of 6th Congrés Français d'Acoustique, Lille, France, 2002 (in French)*. 2002.

21. Combet, F. and L. Gelman, *An automated methodology for performing time synchronous averaging of a gearbox signal without speed sensor.* Mechanical Systems and Signal Processing, 2007. **21**(6): p. 2590-2606.

22. Urbanek, J., T. Barszcz, N. Sawalhi, and R.B. Randall, *Comparison of amplitude-based and phase-based methods for speed tracking in application to wind turbines.* Metrology and Measurement Systems, 2011. **18**(2): p. 295-304.

23. Combet, F. and R. Zimroz, *A new method for the estimation of the instantaneous speed relative fluctuation in a vibration signal based on the short time scale transform.* Mechanical Systems and Signal Processing, 2009. **23**(4): p. 1382-1397.

24. Zimroz, R., J. Urbanek, T. Barszcz, W. Bartelmus, F. Millioz, and N. Martin, *Measurement of instantaneous shaft speed by advanced vibration signal processing - Application to wind turbine gearbox.* Metrology and Measurement Systems, 2011. **18**(4): p. 701-712.

25. Urbanek, J., T. Barszcz, and J. Antoni, *A two-step procedure for estimation of instantaneous rotational speed with large fluctuations.* Mechanical Systems and Signal Processing, 2013. **38**(1): p. 96-102.

26. Renaudin, L., F. Bonnardot, O. Musy, J.B. Doray, and D. Rémond, *Natural roller bearning fault detection by angular measurement of true instantaneous angular speed.* Mechanical Systems and Signal Processing, 2010. **24**(7): p. 1998-2011.

27. André, H., D. Rémond, and A. Bourdon, *On the use of the instantaneous angular speed measurement in non stationary mechanism monitoring*, in *Proceedings of the ASME 2011 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference Volume 1: 23rd Biennial Conference on Mechanical Vibration and Noise, Parts A and B, Washington, D.C., U.S.A., 28-31 August*. 2011.

28. Herlufsen, H., S. Gade, H. Konstantin-Hansen, and H. Vold, *Characteristics of the Vold/Kalman Order Tracking Filter.* Sound and Vibration, 1999. (April).

29. Vold, H. and J. Leuridan, *Order Tracking at Extreme Slew Rates, Using Kalman Tracking Filters.* SAE Paper Number 931288, 1993.

30. Qian, S., *Gabor Expansion for Order Tracking.* Sound and Vibration, 2003. (June).

31. Randall, R.B., *Vibration-based condition monitoring: Industrial, aerospace and automotive applications*. 2011: John Wiley and Sons.

32. Randall, R.B., *Frequency Analysis, 3rd edition* 1987: Bruel Kjaer.

33. Norton, M. and D. Karczub, *Fundamentals of Noise and Vibration Analysis for Engineers*. 2003: Cambridge.

34. McFadden, P.D., *Interpolation techniques for time domain averaging of gear vibration.* Mechanical Systems and Signal Processing, 1989. **3**(1): p. 87-97.

35. Ziemer, R.E. and W.H. Tranter, *Principles of Communications: Systems, modulation, and Noise, 5th Edition*. 2002: John Wiley and Sons.

36. Schwartz, M., *Information Transmission, Modulation and Noise, 1st Edition*. 1959: McGraw-Hill.

37. Schwartz, M., *Information Transmission, Modulation and Noise, 4th Edition*. 1990: McGraw-Hill.

38. Rukmangadachari, E. and E. Keshava Reddy, *Engineering Mathematics, Volume III*. 2010: Pearson Education India.

39. Carson, J.R., *Notes on the Theory of Modulation.* Radio Engineers, Proceedings of the Institute of, 1922. **10**(1): p. 57-64.

40. Borghesani, P., P. Pennacchi, R.B. Randall, and R. Ricci, *Order tracking for discrete-random separation in variable speed conditions.* Mechanical Systems and Signal Processing, 2012. **30**: p. 1-22.

41. McFadden, P.D., *A revised model for the extraction of periodic waveforms by time domain averaging.* Mechanical Systems and Signal Processing, 1987. **1**(1): p. 83-95.

42. Endo, H., *Simulation of gear faults and its application to the development of differential diagnostic technique*, in *Mechanical & Manufacturing Engineering, Faculty of Engineering*. 2005, The University of New South Wales, Australia.

43. Randall, R.B. and J. Antoni, *Rolling element bearing diagnostics—A tutorial.* Mechanical Systems and Signal Processing, 2011. **25**(2): p. 485-520.

44. McFadden, P.D., *Detecting fatigue cracks in gears by amplitude and phase demodulation of the meshing vibration.* ASME Transactions Journal of vibration, acoustics, stress, and reliability in design, 1986. **108**: p. 165-170.

45. Sawalhi, N. and R.B. Randall, *Gear Parameter identification in Wind Turbines using diagnostic analysis of gearbox vibration signals*, in *Proceedings of CMMNO13 – Condition Monitoring of Machinery in Non-Stationary Operations, Ferrara, Italy, 8-10 May*. 2013.

46. Ewins, D.J., *Modal Testing: Theory, Practice and Application*. Second edition ed. 2000: Research Studies Press ltd.

47. Peeters, B., B. Cornelis, K. Janssens, and H. Van der Auweraer, *Removing disturbing harmonics in operational modal analysis*, in *Proceedings of IOMAC 2007, Copenhagen, Denmark, 1-2 May*. 2007.

48. Randall, R.B., B. Peeters, J. Antoni, and S. Manzato. *New cepstral methods of signal pre-processing for operational modal analysis*. in *ISMA2012, Leuven, Belgium, 755-764*. 2012.

49. Bogert, B.P., M.J.R. Healy, and J.W. Tukey. *The Quefrency Alanysis of Time series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking*. in *Proceedings of the Symposium on Time Series Analysis, by M. Rosenblatt (Ed.), Wiley, NY, pp. 209-243*. 1963.

50. Sawalhi, N. and R.B. Randall. *Editing Time Signals using the Real Cepstrum*. in *MFPT conference, Virginia Beach, May*. 2011.

51. Randall, R.B. and N. Sawalhi. *Cepstral Removal of Periodic Spectral Components from Time Signals*. in *CMMNO conference, Ferrara, May*. 2013.

52. Gao, Y. and R.B. Randall, *Determination of frequency response functions from response measurements. Part I: Extraction of poles and zeros from response cepstra.* Mechanical Systems and Signal Processing, 1996. **10**(3): p. 293-317.

53. Deshpande, L.G., N.M. Sawalhi, and R.B. Randall, *Application of FE model updating and reduction techniques to simulate gearbox bearing faults.* Australian Journal of Multi-disciplinary Engineering, 2013. **10**(2): p. 109-118.

54. Deshpande, L.G., *Simulation of vibrations caused by faults in bearings and gears*, in *Mechanical & Manufacturing Engineering, Faculty of Engineering*. 2014, The University of New South Wales, Australia.

Appendix A Acceptability charts for different harmonic orders

A.1 Introduction

This appendix contains acceptability charts, which can be used to determine if a specific harmonic order n, with a given combination of speed variation parameters, is suitable to be used as the reference harmonic order with the PDOT method.

These charts are constructed in the same manner described in section 3.6 on page 161, for the harmonic orders $n=1$ to $n=10$. In summary, calculations were conducted for frequency deviation percentages between 0.1 to 40%, and modulating frequency percentages between 0.1 to 50%.

Note the charts for harmonic order $n=1$ are a repeat of those in section 3.6, and are included for completeness.

Note that the acceptability chart sets for different harmonic orders have been plotted with different axis ranges, in order to better display the acceptability envelope for each set, and so are not drawn with a common scale.

An additional supplementary chart is also included for the first harmonic order $n=1$, which covers lower modulating frequency percentages down to 0.01%, which is below the lowest modulating frequency percentage of 0.1% present in the main acceptability charts.

A.2 Harmonic order $n = 1$

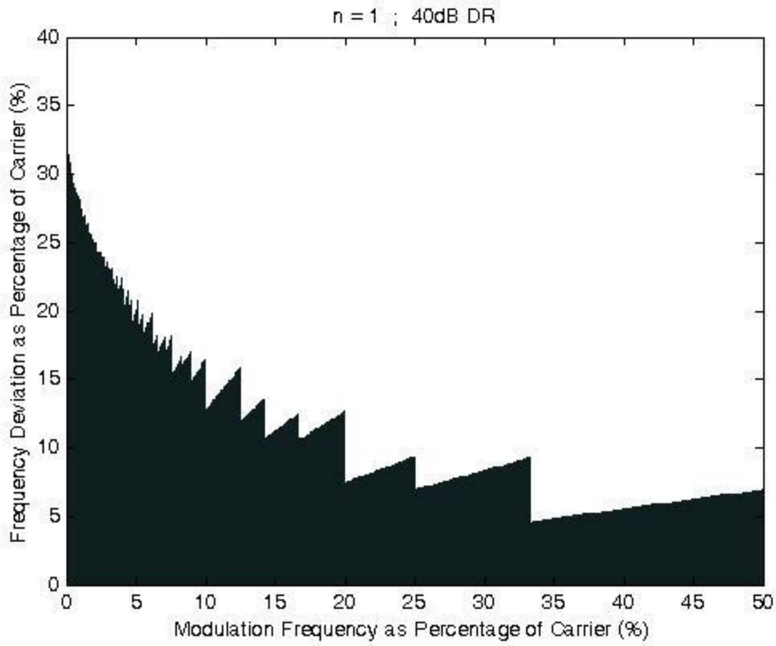


Figure A.1 – Acceptability chart, for the first harmonic order $n = 1$, using Bessel function calculations and a 40dB DR

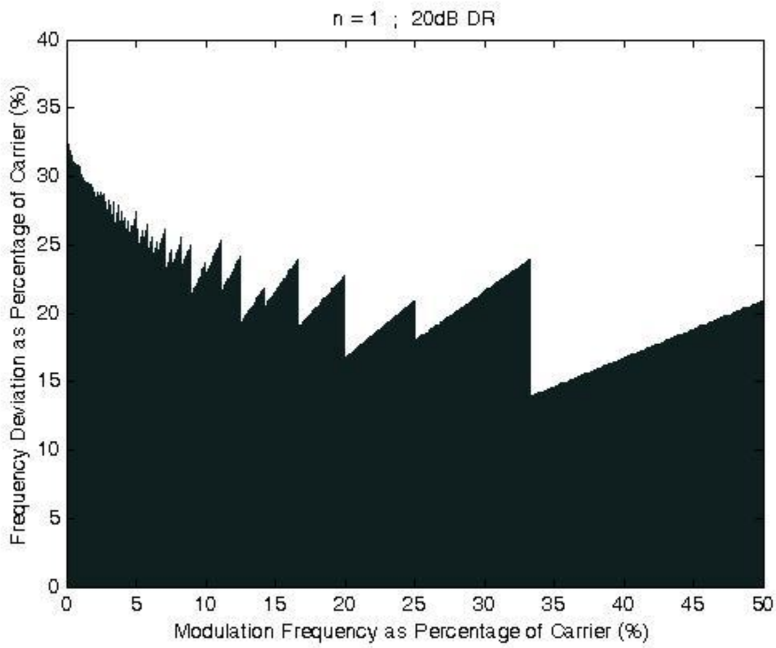


Figure A.2 – Acceptability chart, for the first harmonic order $n = 1$, using Bessel function calculations and a 20dB DR

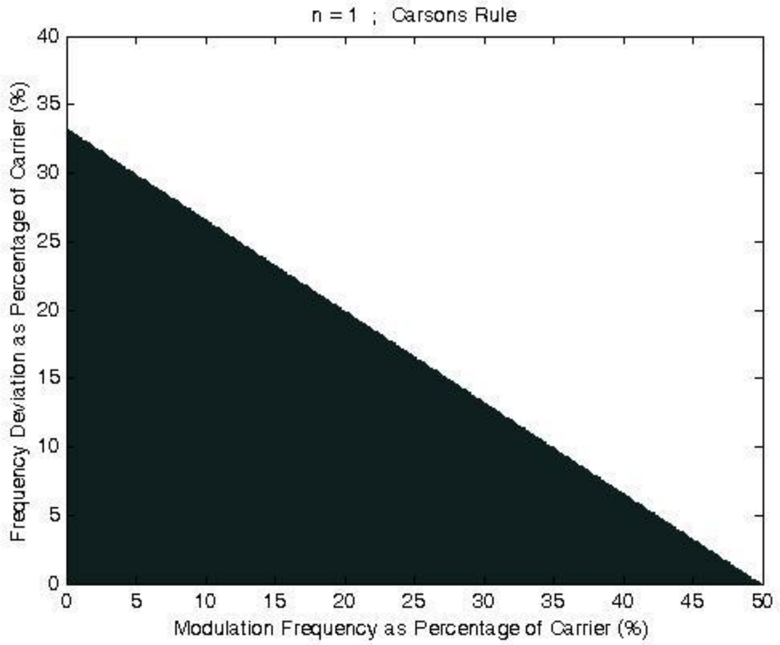


Figure A.3 – Acceptability chart, for the first harmonic order $n=1$, using Carson's Rule

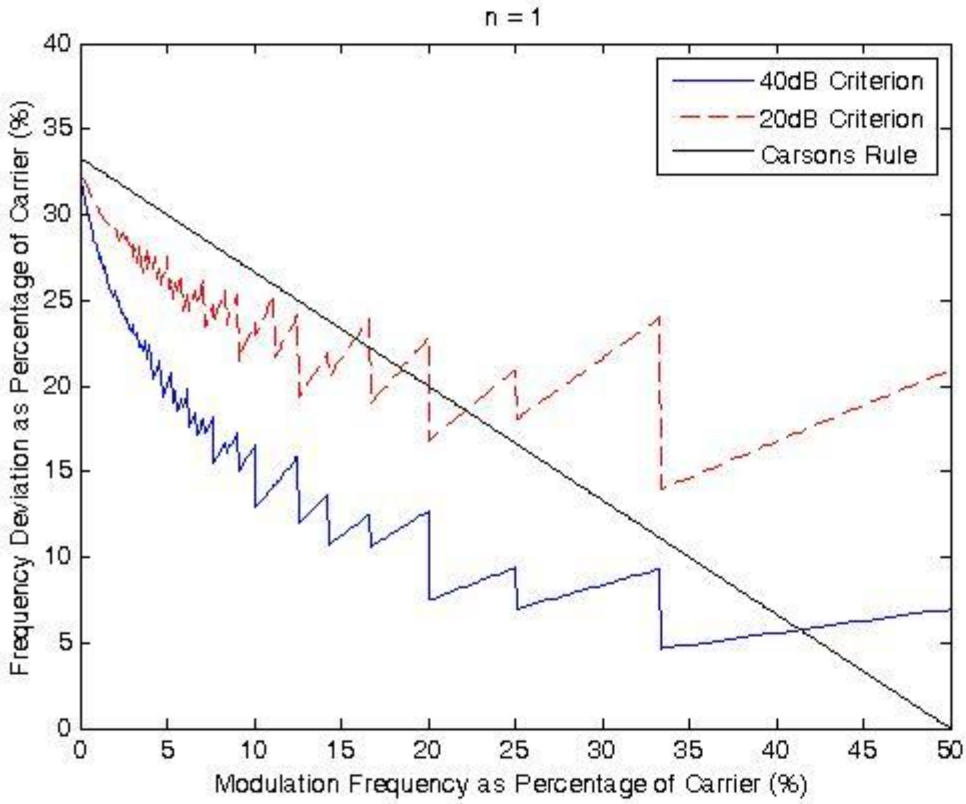


Figure A.4 – Envelope of Acceptability chart, for the first harmonic order $n=1$, using the Bessel function criterion for 40dB DR and 20dB DR and Carson's rule

A.2.1 Supplementary acceptability chart, for harmonic order $n = 1$

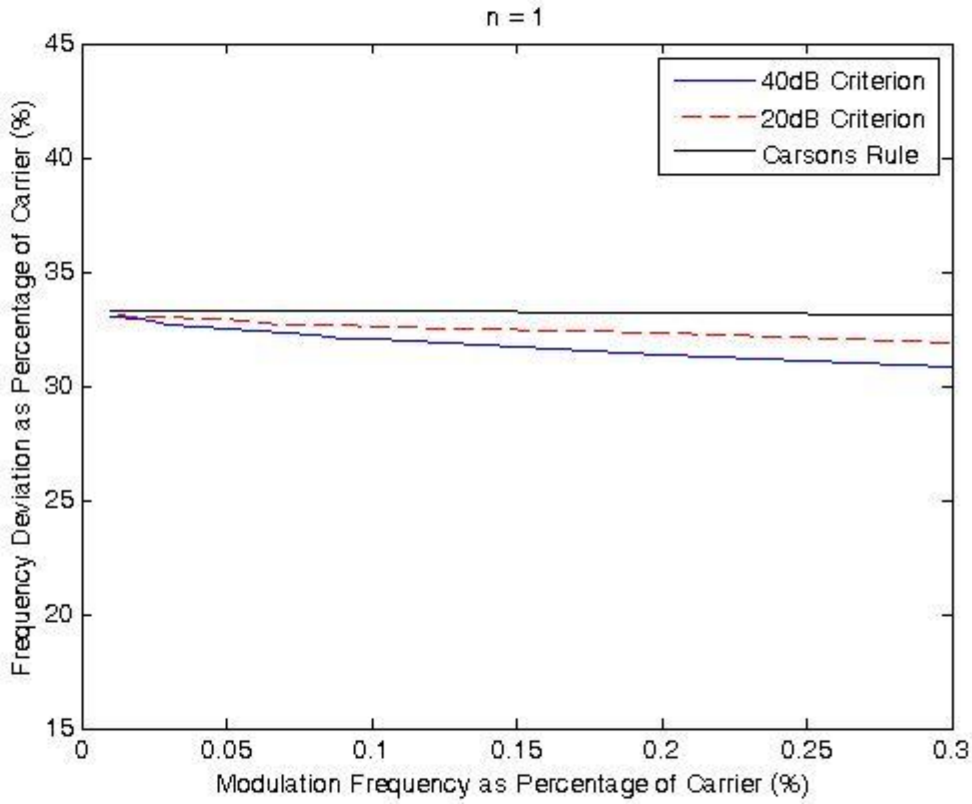


Figure A.5 – Supplementary acceptability chart, envelope, for the first harmonic order $n = 1$, using the Bessel function criterion for 40dB DR and 20dB DR and Carson's rule

A.3 Harmonic order $n = 2$

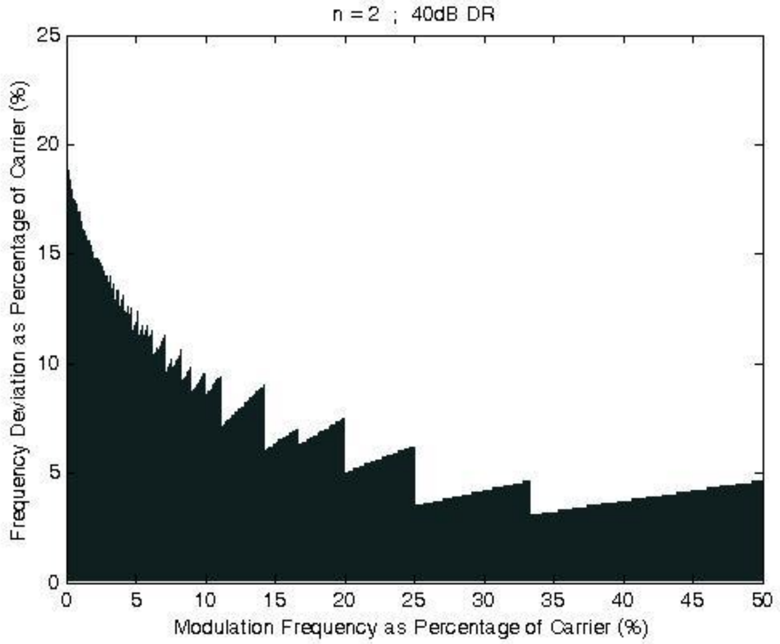


Figure A.6 – Acceptability chart, for the second harmonic order $n = 2$, using Bessel function calculations and a 40dB DR

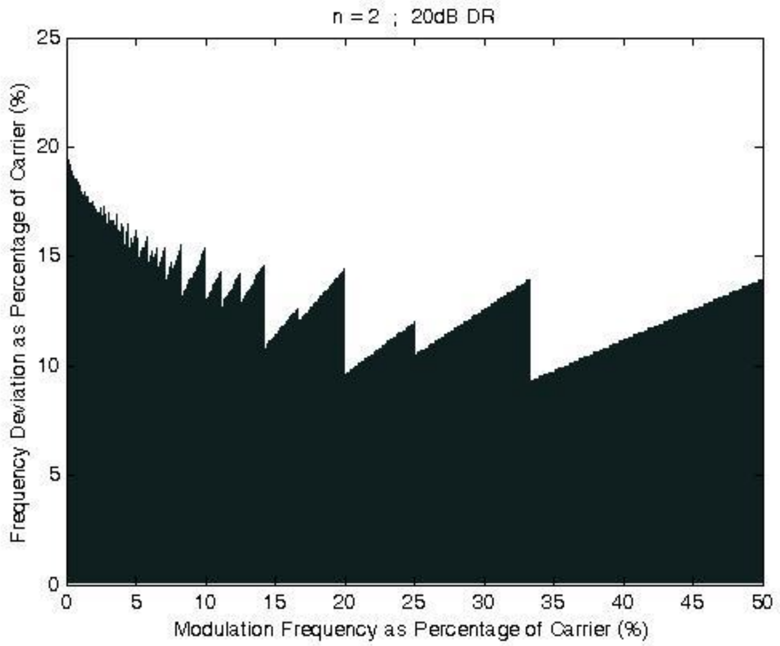


Figure A.7 – Acceptability chart, for the second harmonic order $n = 2$, using Bessel function calculations and a 20dB DR

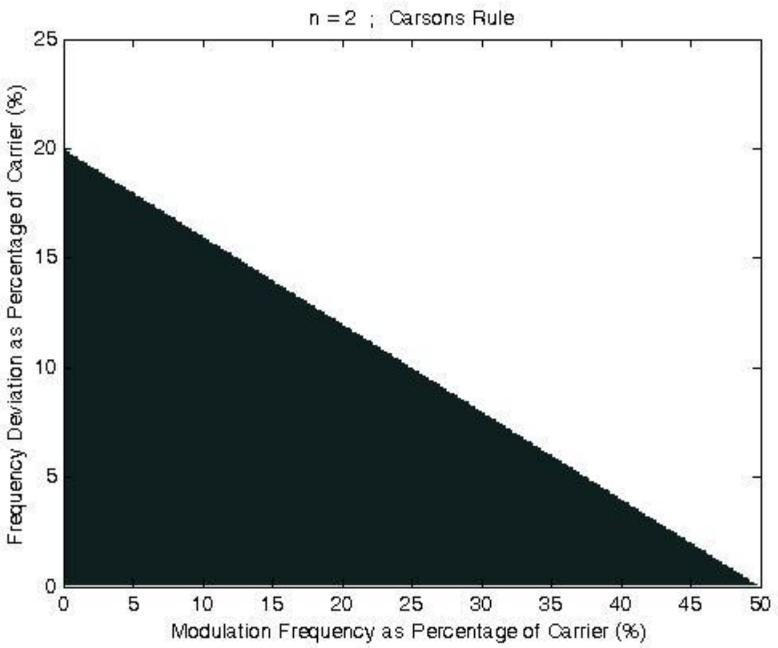


Figure A.8 – Acceptability chart, for the second harmonic order $n = 2$, using Carson's Rule

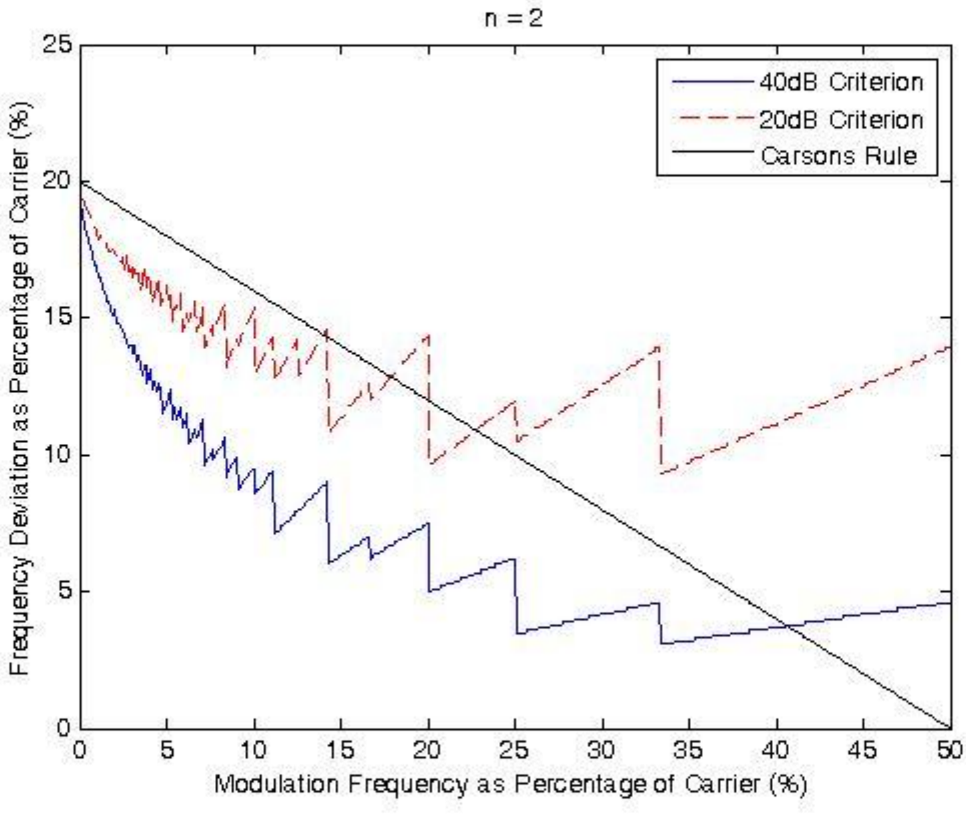


Figure A.9 – Envelope of Acceptability chart, for the second harmonic order $n = 2$, using the Bessel function criterion for 40dB DR and 20dB DR and Carson's rule

A.4 Harmonic order $n = 3$

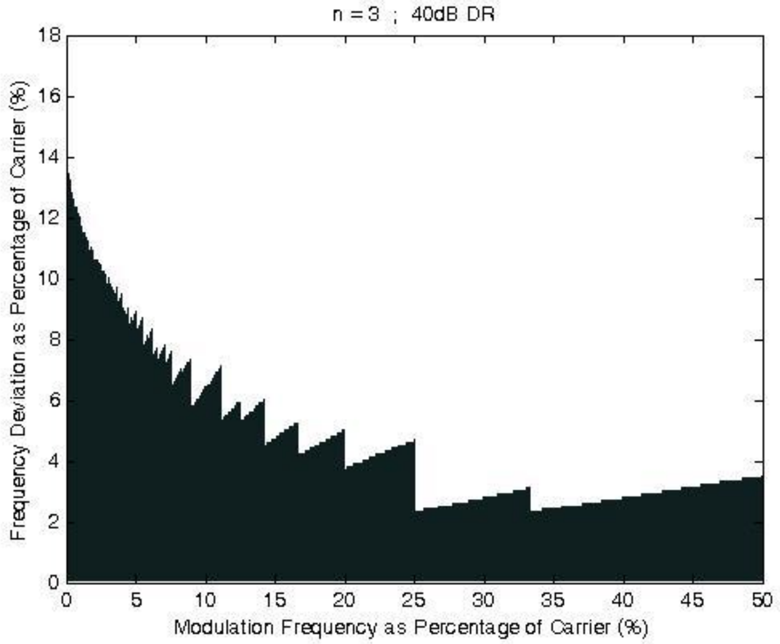


Figure A.10 – Acceptability chart, for the third harmonic order $n = 3$, using Bessel function calculations and a 40dB DR

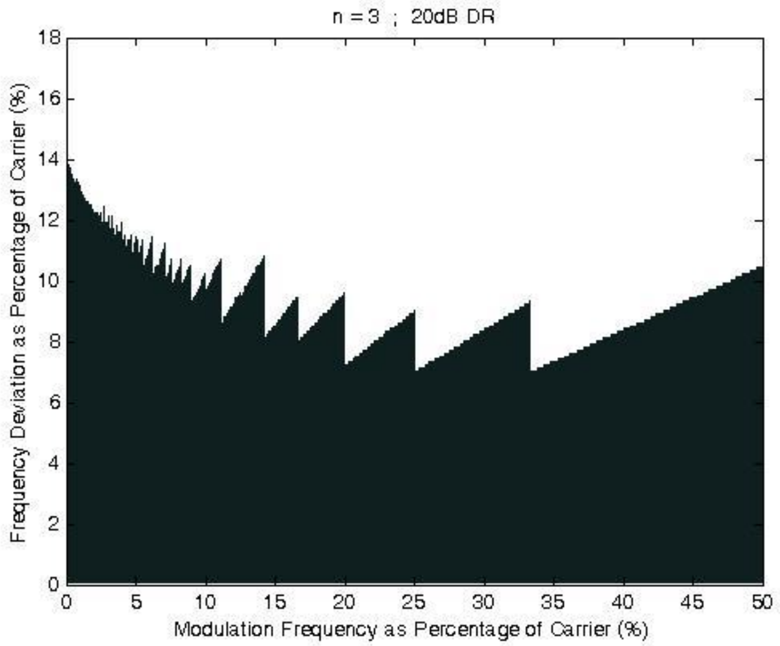


Figure A.11 – Acceptability chart, for the third harmonic order $n = 3$, using Bessel function calculations and a 20dB DR

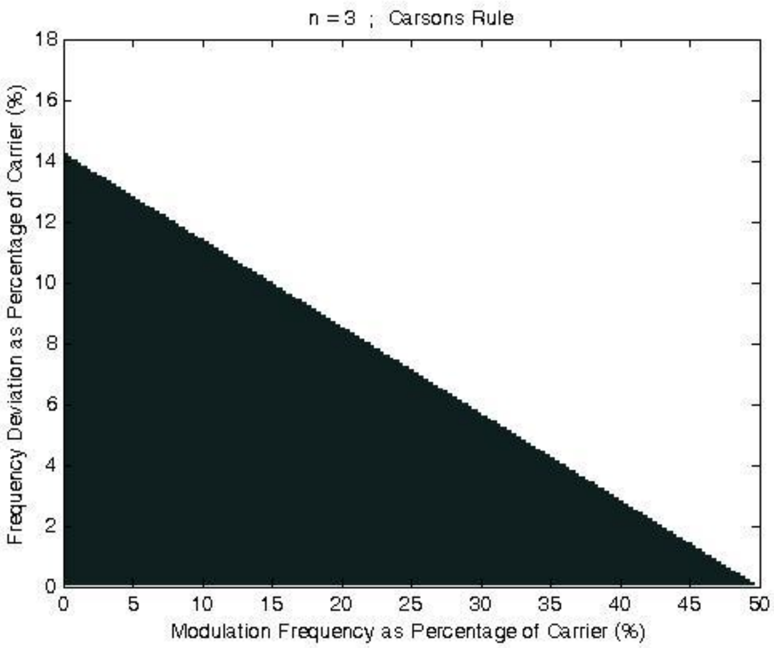


Figure A.12 – Acceptability chart, for the third harmonic order $n = 3$, using Carson's Rule

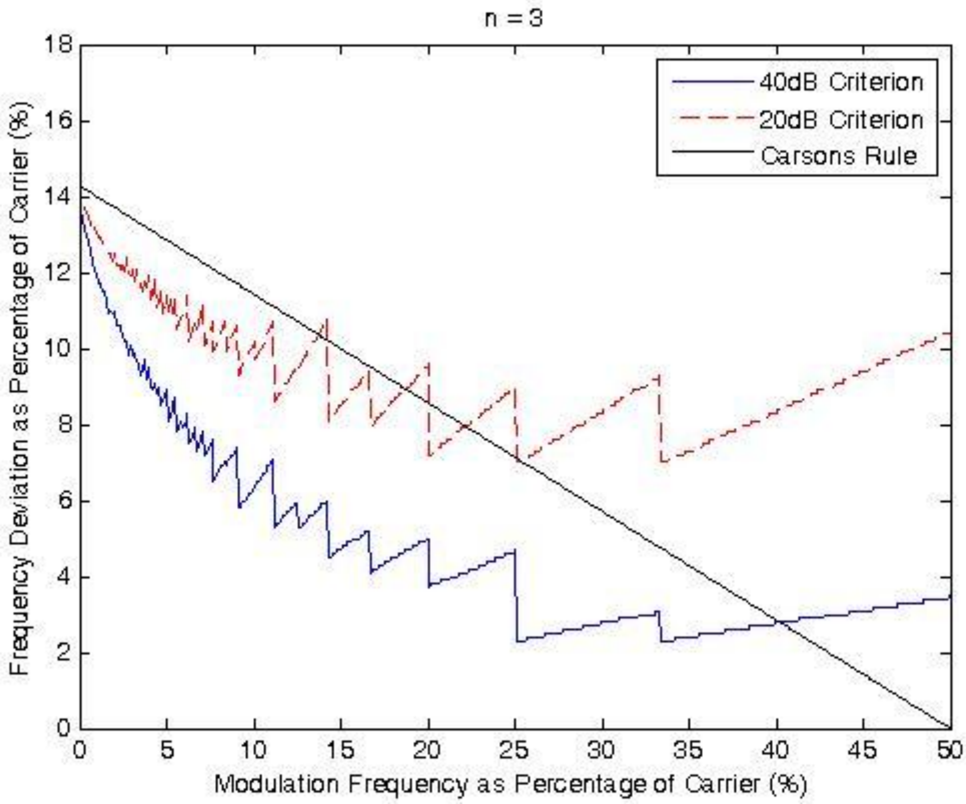


Figure A.13 – Envelope of Acceptability chart, for the third harmonic order $n = 3$, using the Bessel function criterion for 40dB DR and 20dB DR and Carson's rule

A.5 Harmonic order $n=4$

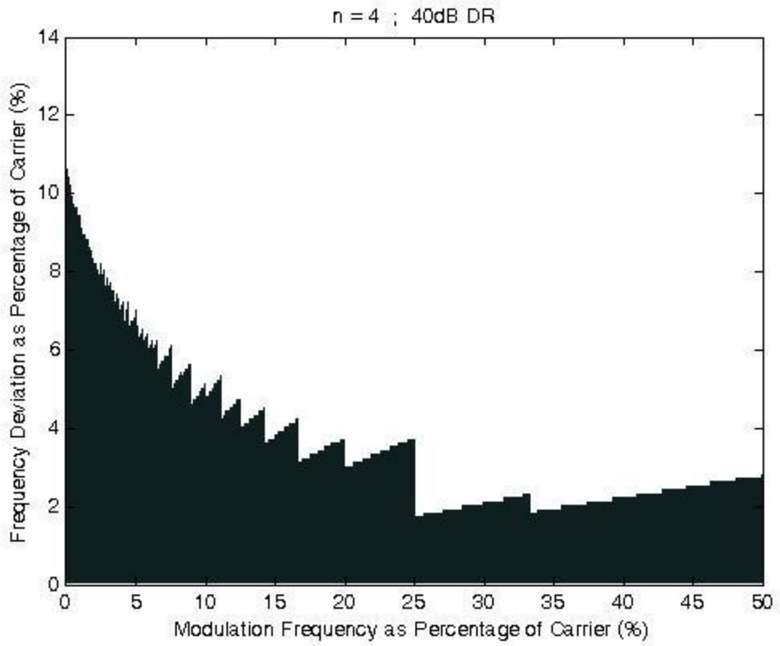


Figure A.14 – Acceptability chart, for the fourth harmonic order $n=4$, using Bessel function calculations and a 40dB DR

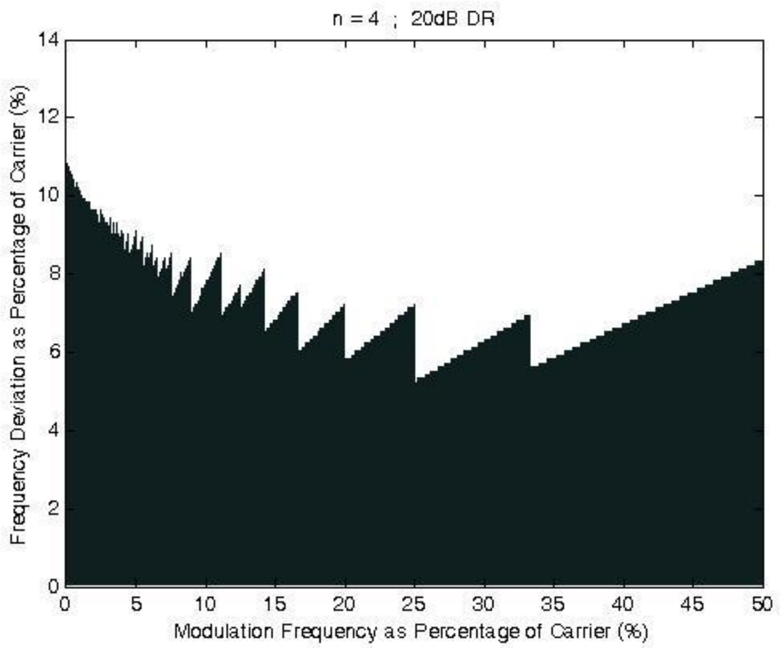


Figure A.15 – Acceptability chart, for the fourth harmonic order $n=4$, using Bessel function calculations and a 20dB DR

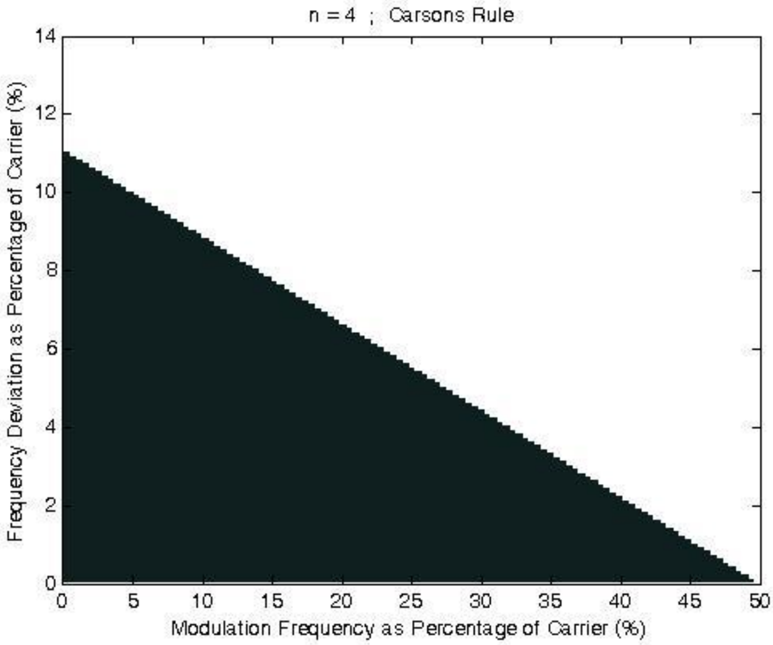


Figure A.16 – Acceptability chart, for the fourth harmonic order $n=4$, using Carson's Rule

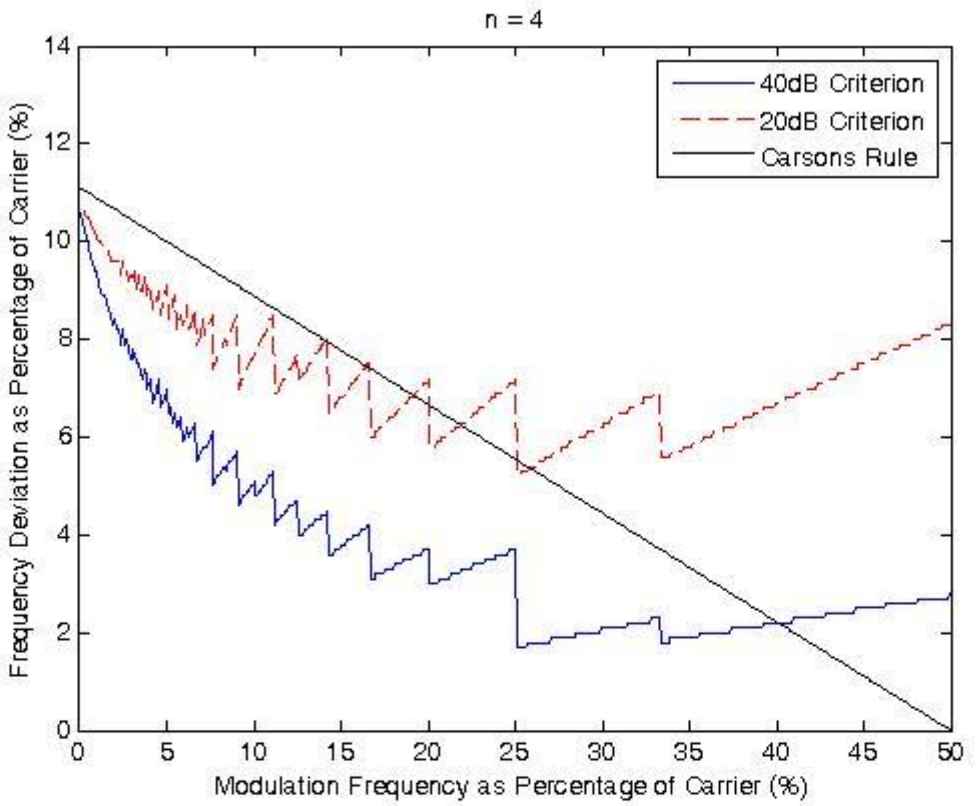


Figure A.17 – Envelope of Acceptability chart, for the fourth harmonic order $n=4$, using the Bessel function criterion for 40dB DR and 20dB DR and Carson's rule

A.6 Harmonic order $n = 5$

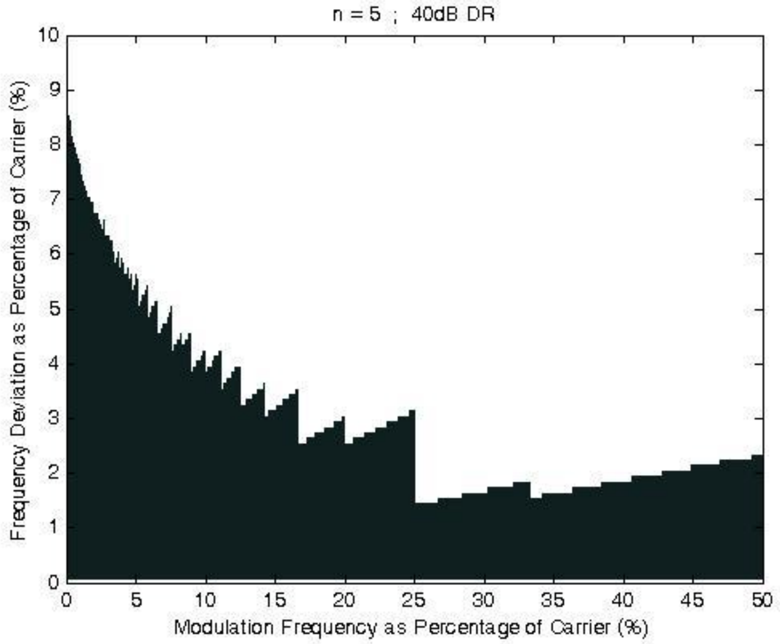


Figure A.18 – Acceptability chart, for the fifth harmonic order $n = 5$, using Bessel function calculations and a 40dB DR

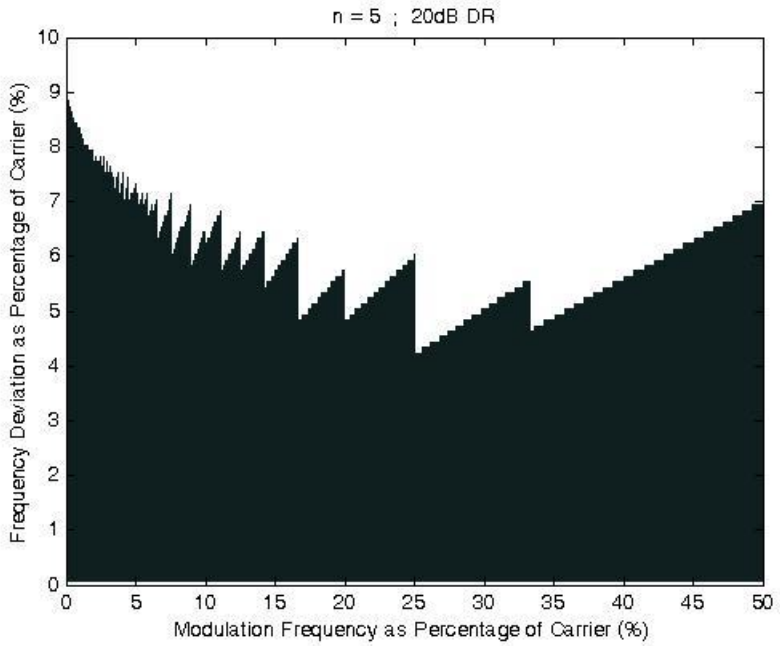


Figure A.19 – Acceptability chart, for the fifth harmonic order $n = 5$, using Bessel function calculations and a 20dB DR

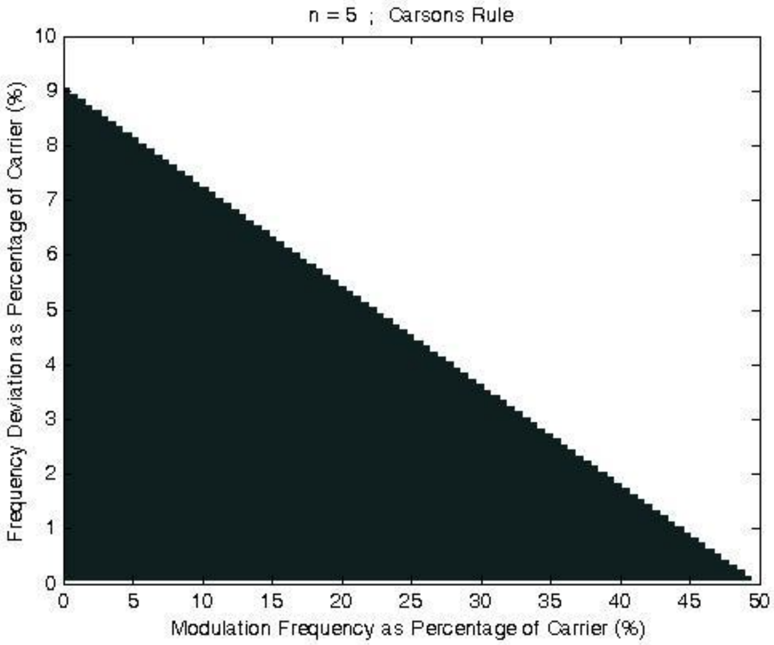


Figure A.20 – Acceptability chart, for the fifth harmonic order $n = 5$, using Carson's Rule

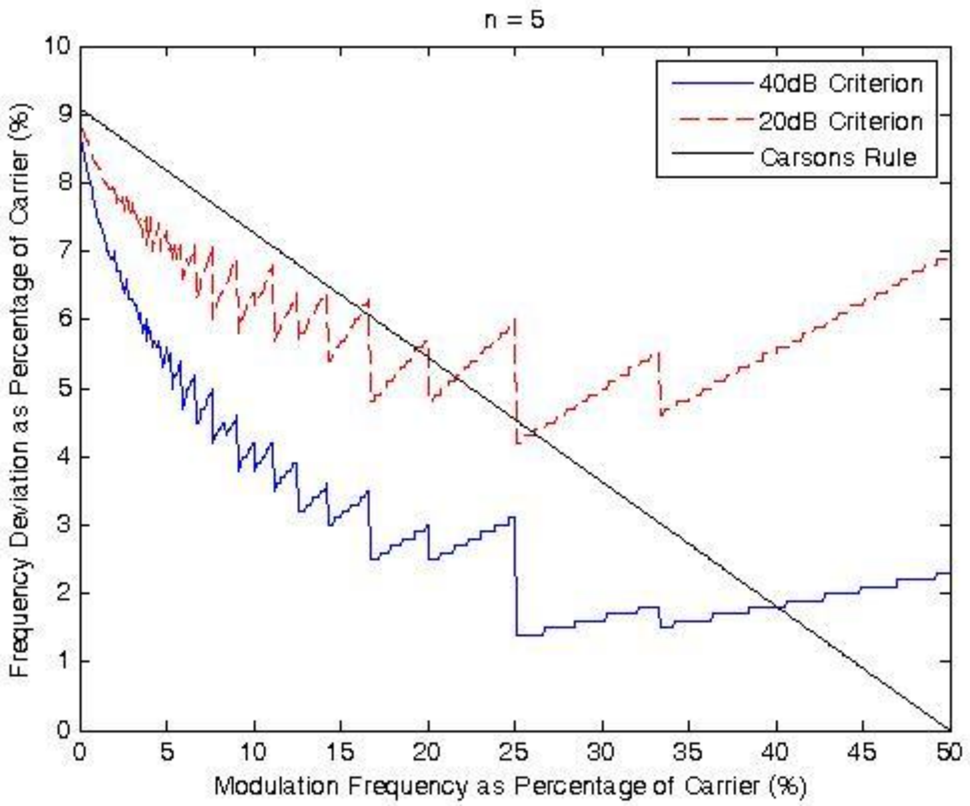


Figure A.21 – Envelope of Acceptability chart, for the fifth harmonic order $n = 5$, using the Bessel function criterion for 40dB DR and 20dB DR and Carson's rule

A.7 Harmonic order $n = 6$

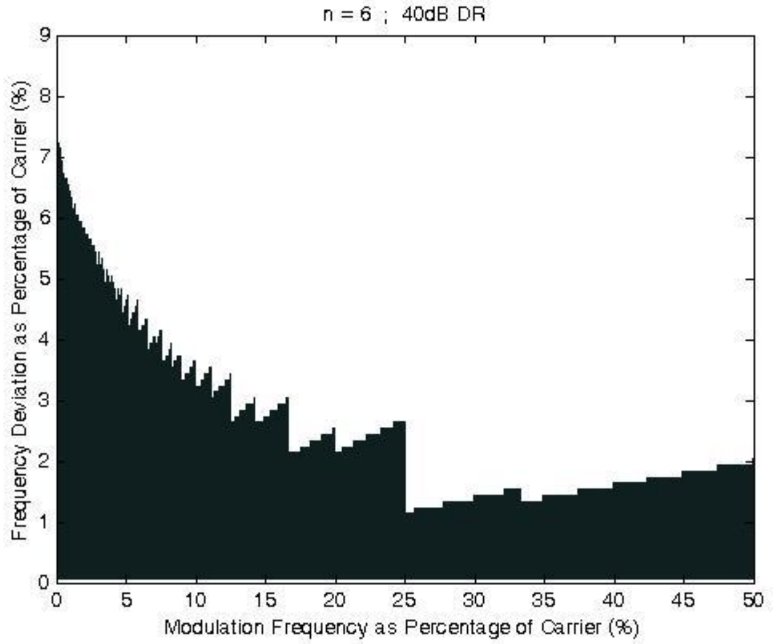


Figure A.22 – Acceptability chart, for the sixth harmonic order $n = 6$, using Bessel function calculations and a 40dB DR

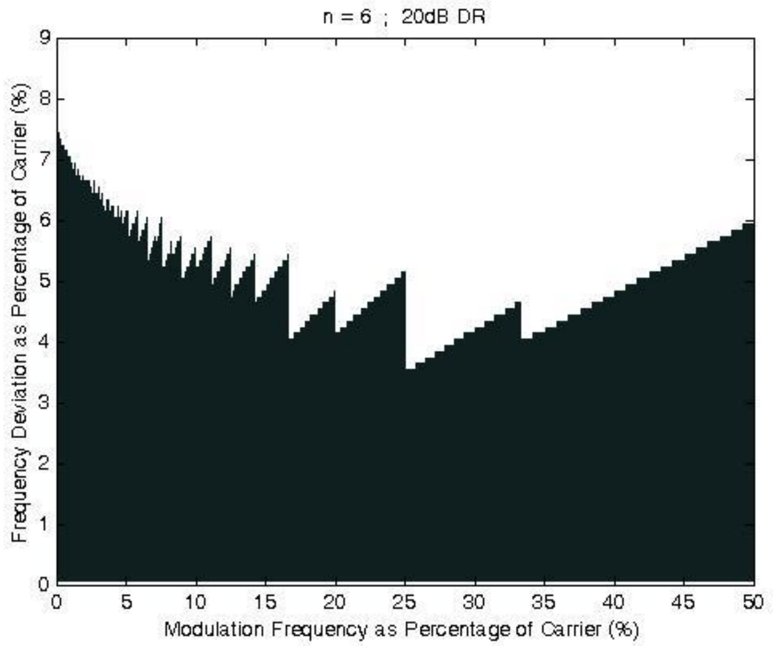


Figure A.23 – Acceptability chart, for the sixth harmonic order $n = 6$, using Bessel function calculations and a 20dB DR

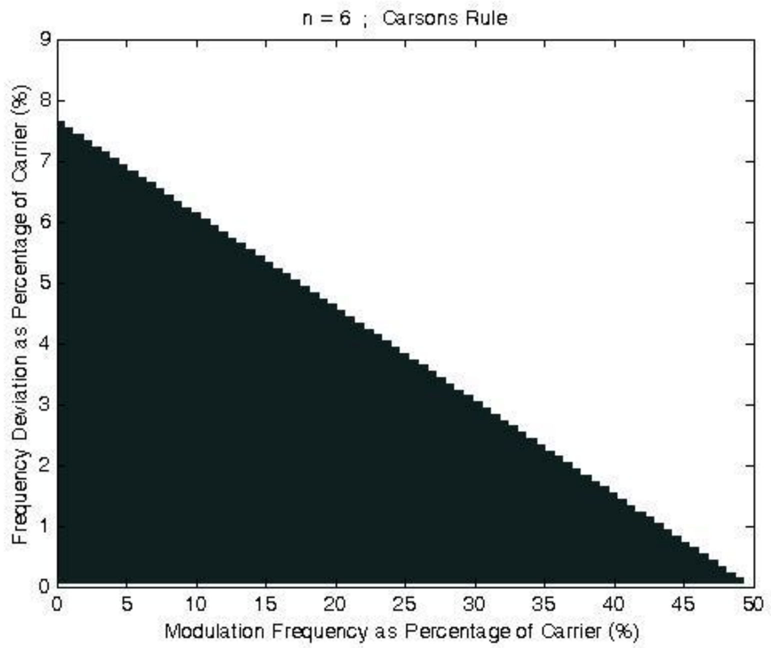


Figure A.24 – Acceptability chart, for the sixth harmonic order $n = 6$, using Carson's Rule

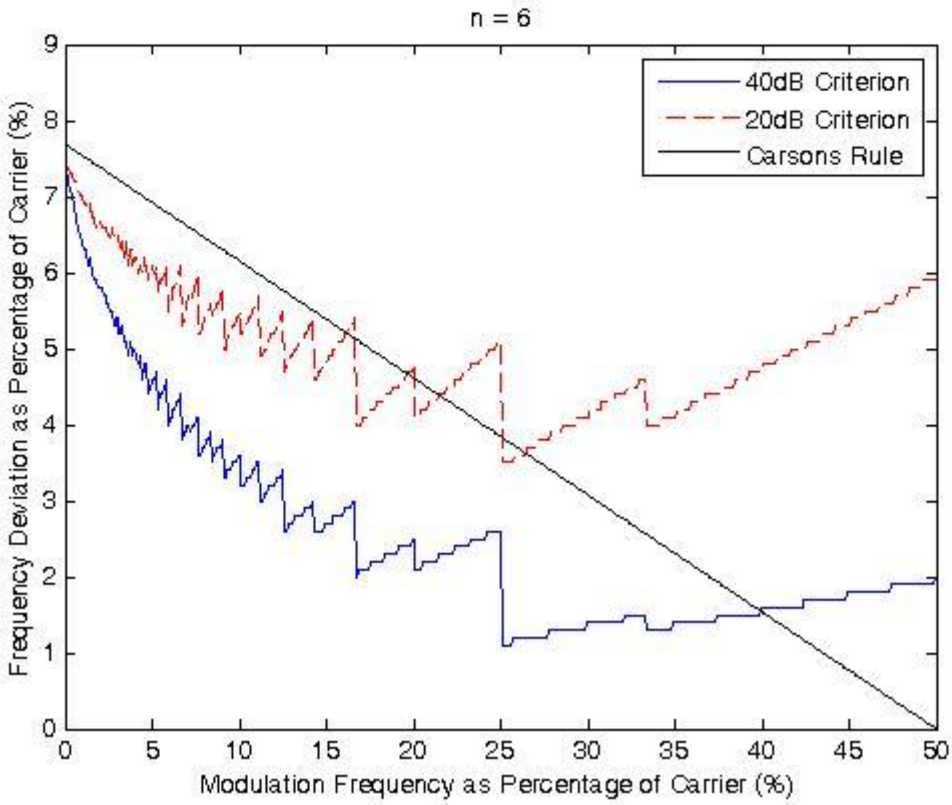


Figure A.25 – Envelope of Acceptability chart, for the sixth harmonic order $n = 6$, using the Bessel function criterion for 40dB DR and 20dB DR and Carson's rule

A.8 Harmonic order $n = 7$

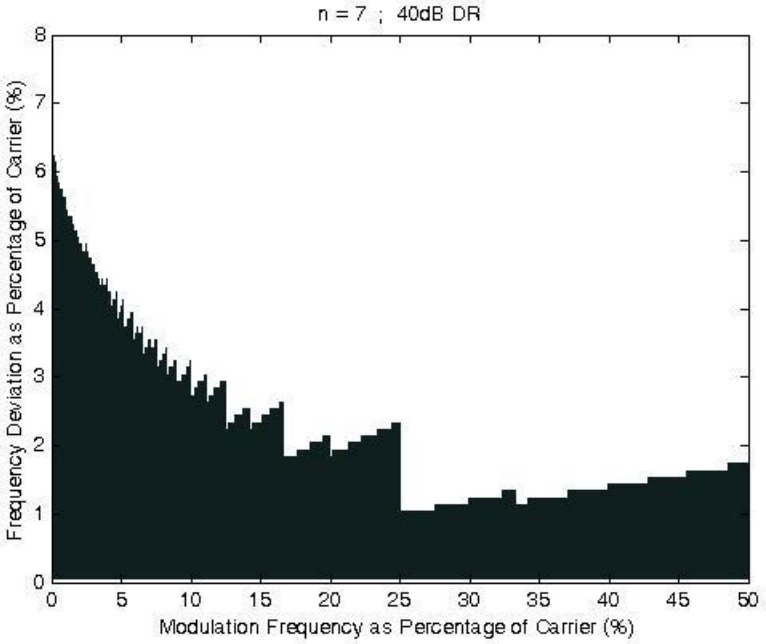


Figure A.26 – Acceptability chart, for the seventh harmonic order $n = 7$, using Bessel function calculations and a 40dB DR

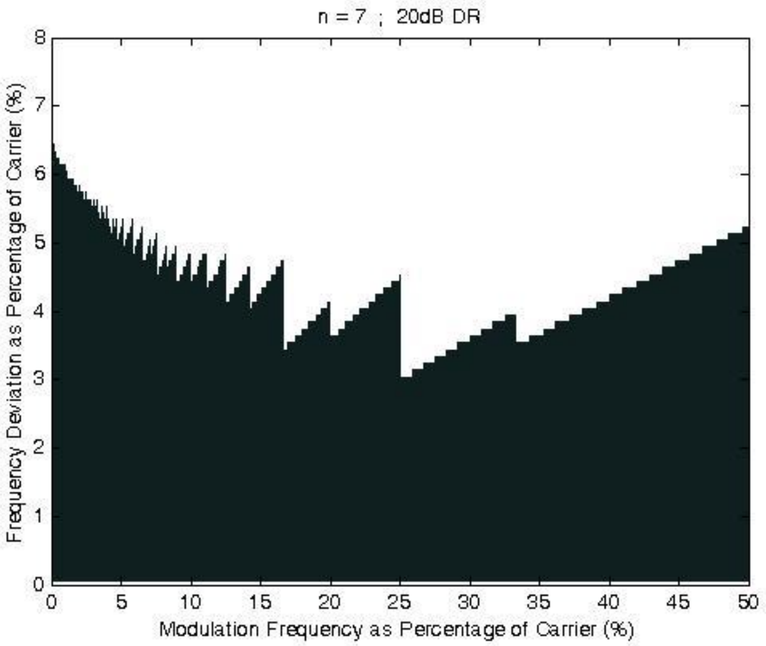


Figure A.27 – Acceptability chart, for the seventh harmonic order $n = 7$, using Bessel function calculations and a 20dB DR

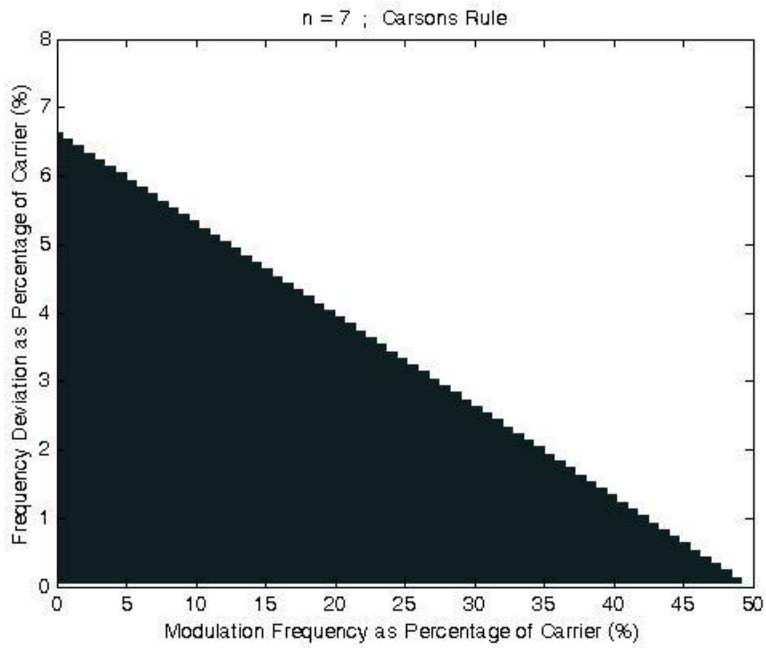


Figure A.28 – Acceptability chart, for the seventh harmonic order $n = 7$, using Carson's Rule

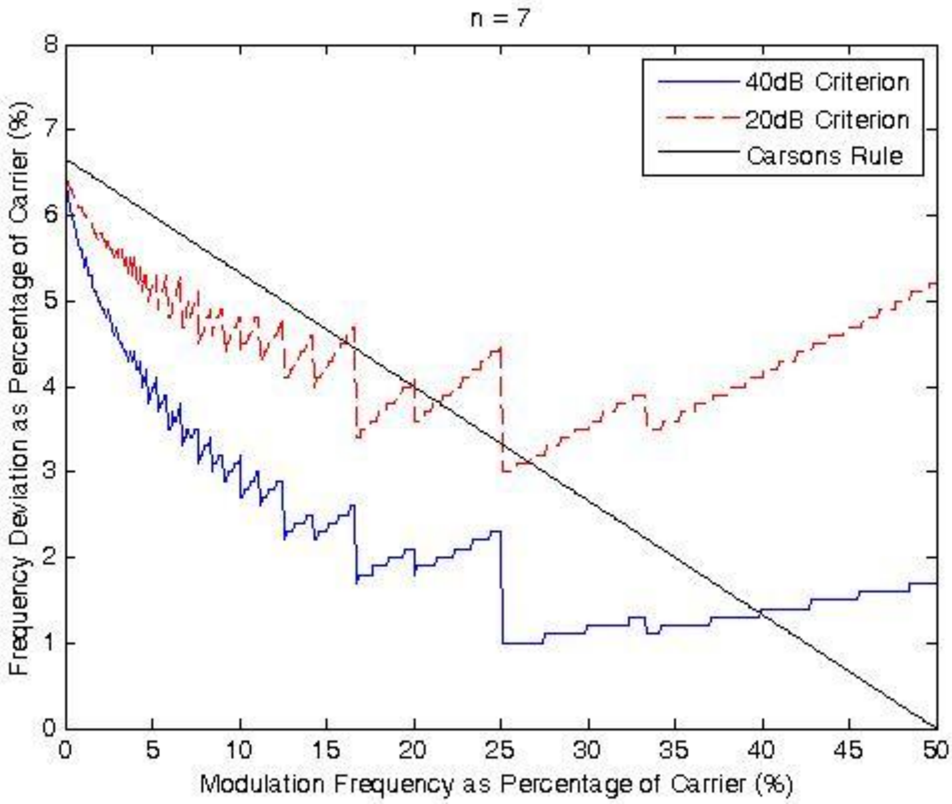


Figure A.29 – Envelope of Acceptability chart, for the seventh harmonic order $n = 7$, using the Bessel function criterion for 40dB DR and 20dB DR and Carson's rule

A.9 Harmonic order $n = 8$

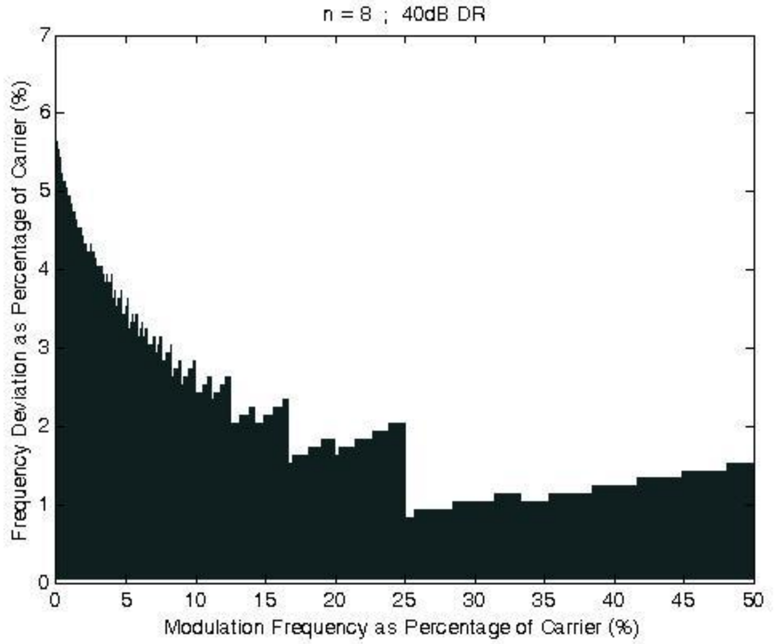


Figure A.30 – Acceptability chart, for the eighth harmonic order $n = 8$, using Bessel function calculations and a 40dB DR

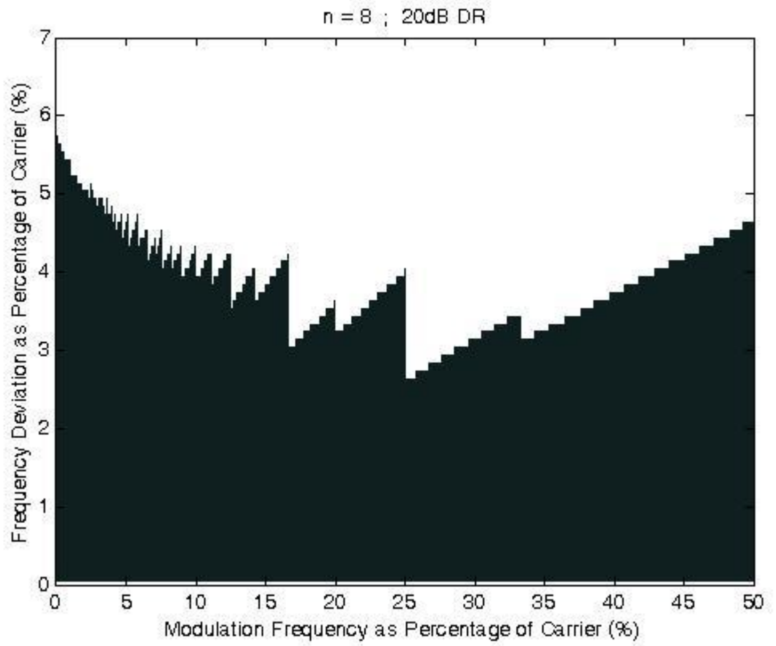


Figure A.31 – Acceptability chart, for the eighth harmonic order $n = 8$, using Bessel function calculations and a 20dB DR

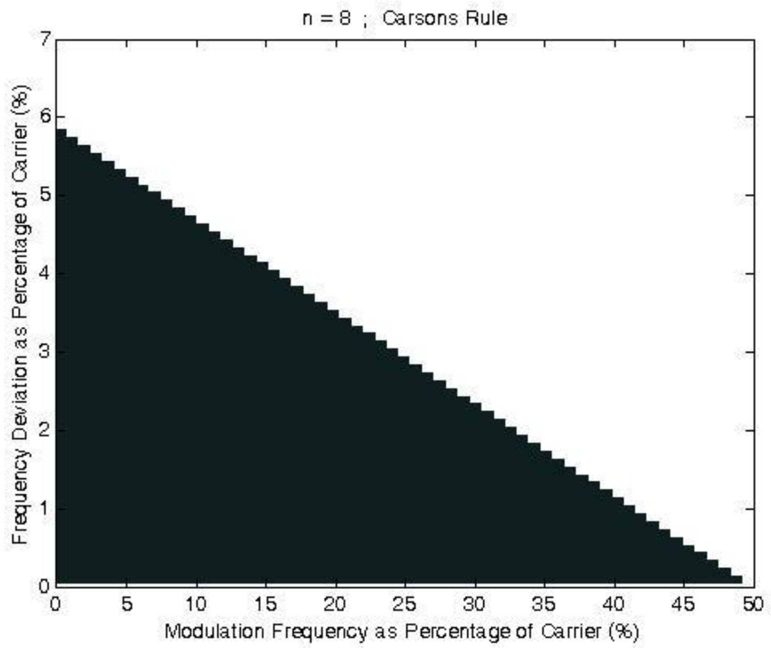


Figure A.32 – Acceptability chart, for the eighth harmonic order $n = 8$, using Carson's Rule

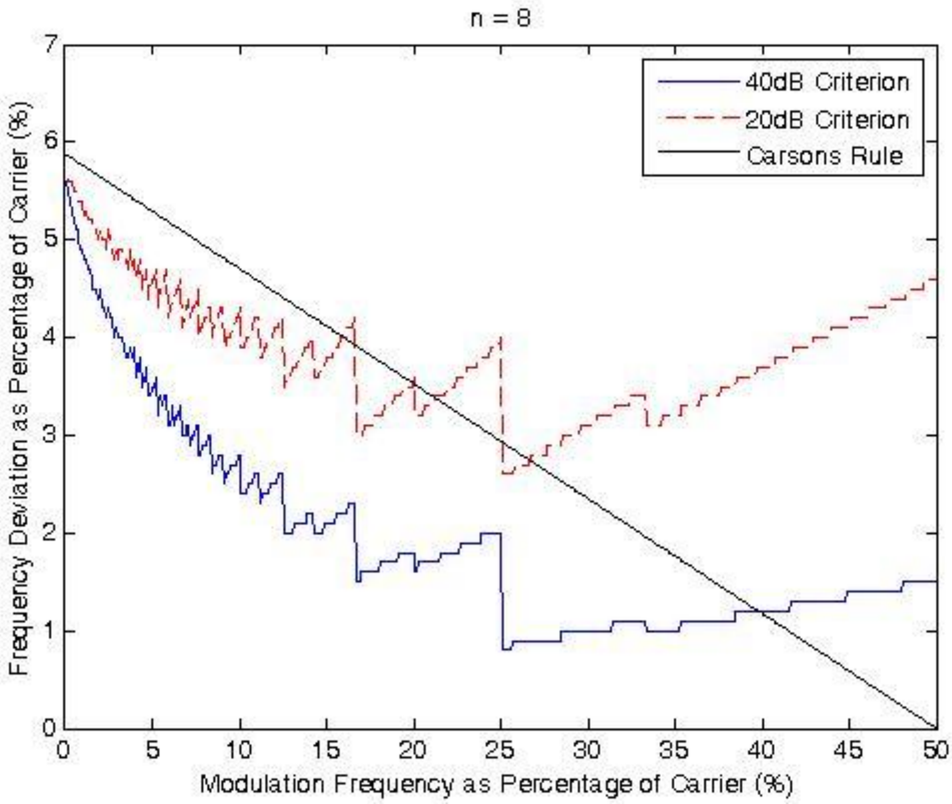


Figure A.33 – Envelope of Acceptability chart, for the eighth harmonic order $n = 8$, using the Bessel function criterion for 40dB DR and 20dB DR and Carson's rule

A.10 Harmonic order $n = 9$

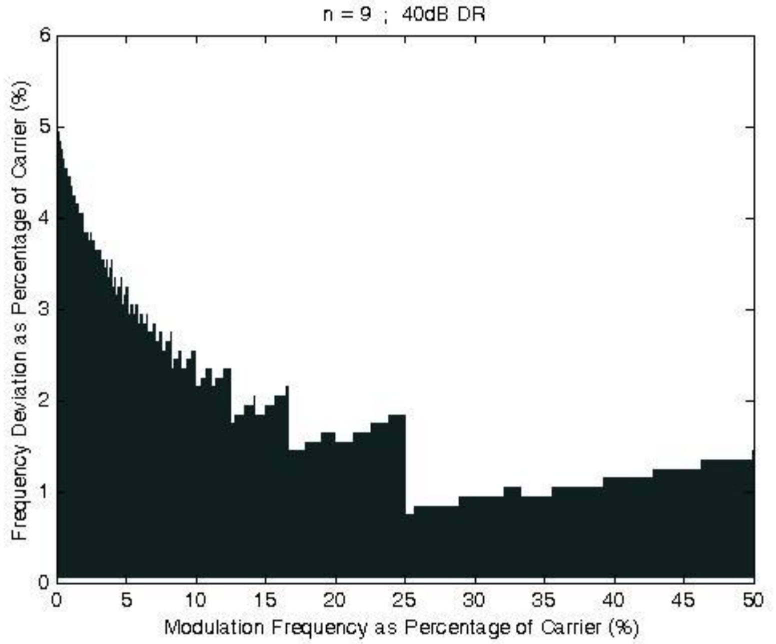


Figure A.34 – Acceptability chart, for the ninth harmonic order $n = 9$, using Bessel function calculations and a 40dB DR

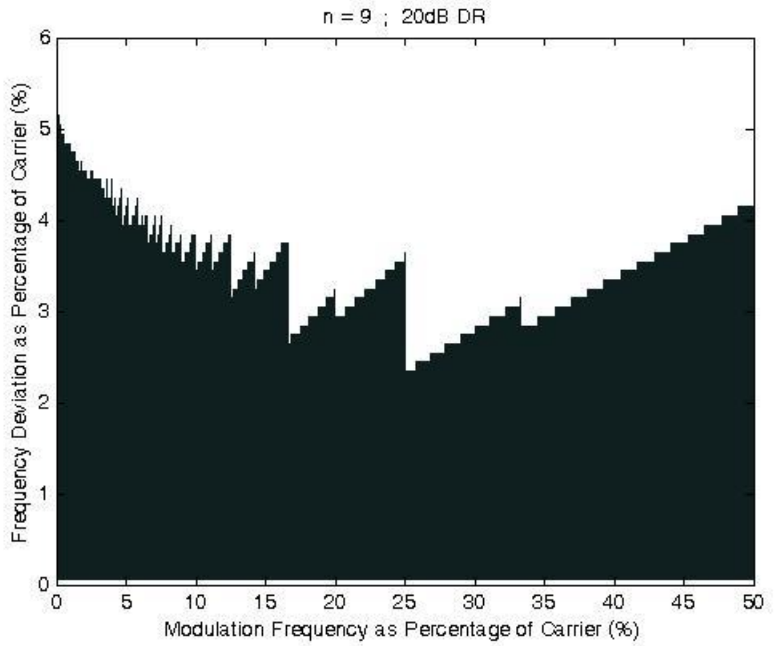


Figure A.35 – Acceptability chart, for the ninth harmonic order $n = 9$, using Bessel function calculations and a 20dB DR

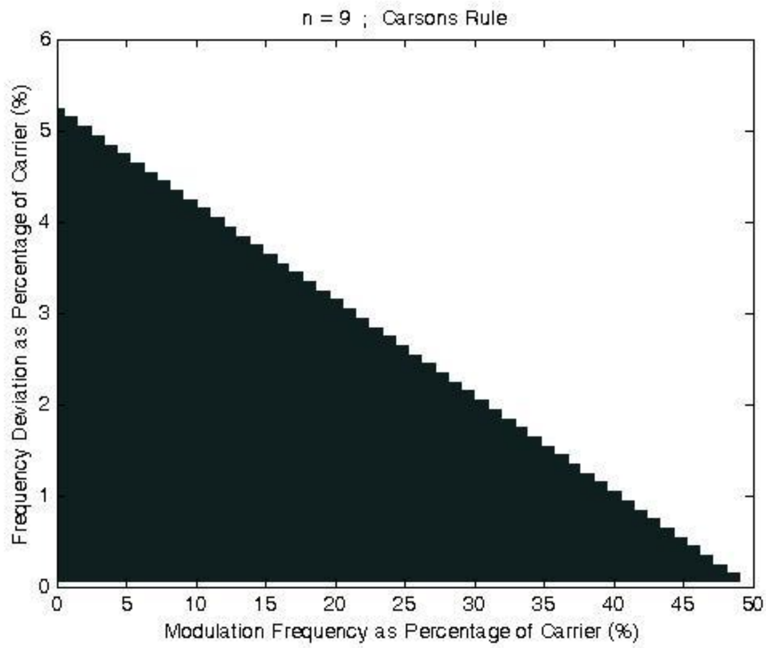


Figure A.36 – Acceptability chart, for the ninth harmonic order $n = 9$, using Carson's Rule

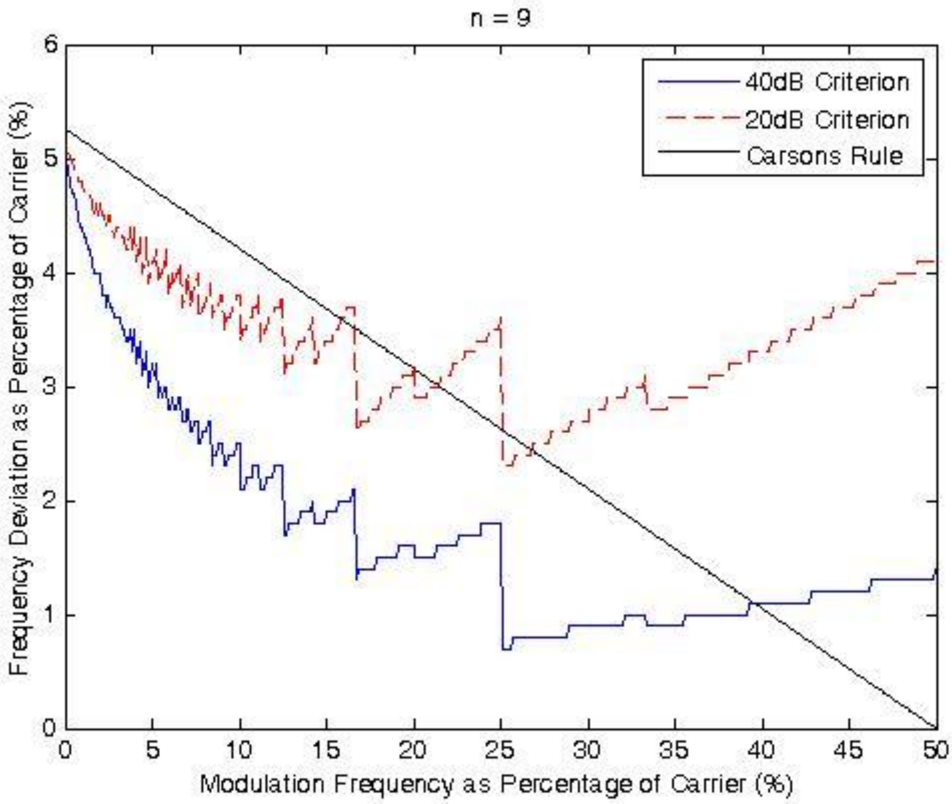


Figure A.37 – Envelope of Acceptability chart, for the ninth harmonic order $n = 9$, using the Bessel function criterion for 40dB DR and 20dB DR and Carson's rule

A.11 Harmonic order $n = 10$

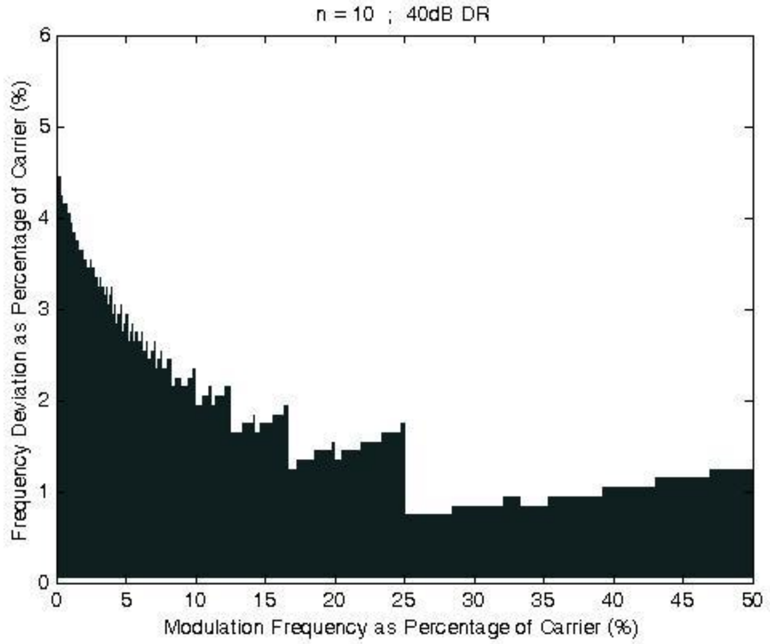


Figure A.38 – Acceptability chart, for the tenth harmonic order $n = 10$, using Bessel function calculations and a 40dB DR

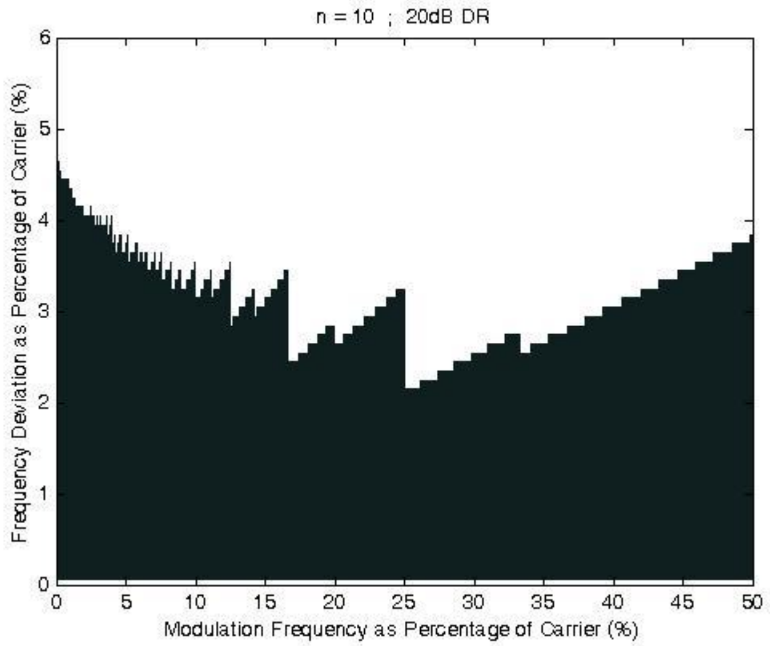


Figure A.39 – Acceptability chart, for the tenth harmonic order $n = 10$, using Bessel function calculations and a 20dB DR

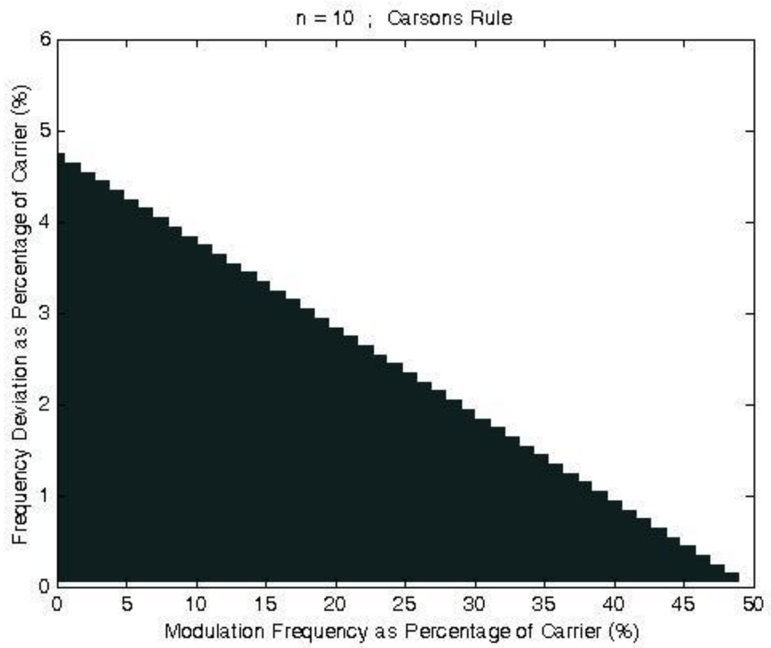


Figure A.40 – Acceptability chart, for the tenth harmonic order $n = 10$, using Carson's Rule

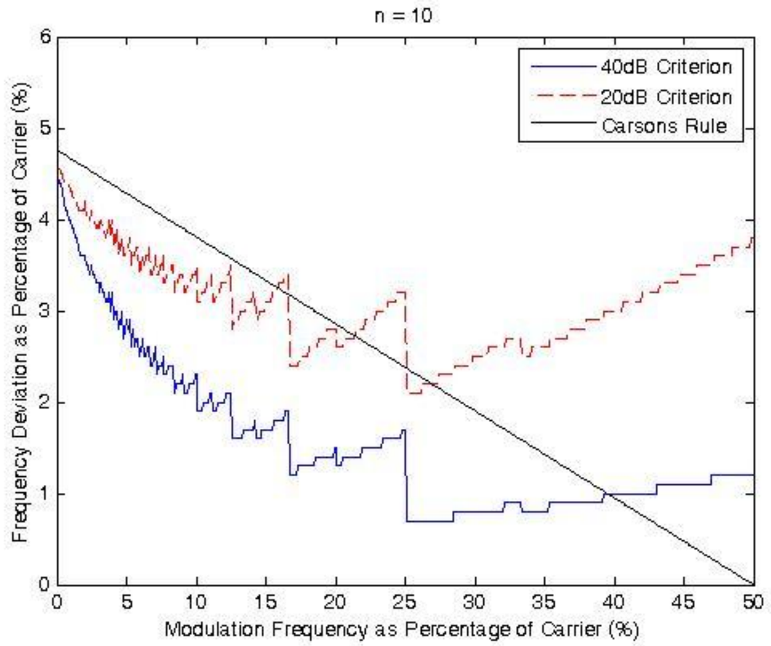


Figure A.41 – Envelope of Acceptability chart, for the tenth harmonic order $n = 10$, using the Bessel function criterion for 40dB DR and 20dB DR and Carson's rule