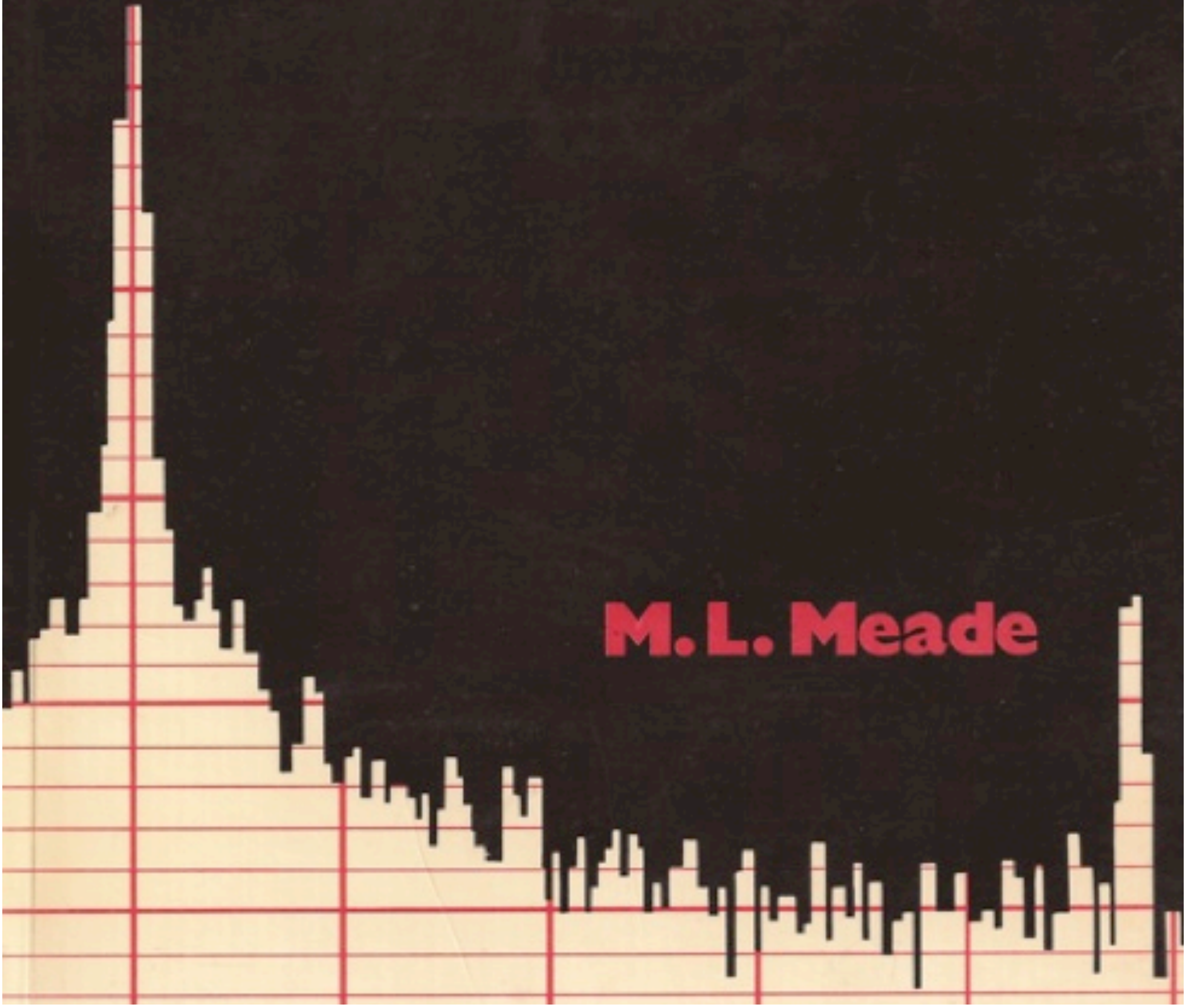
e-edition

Lock-in amplifiers: principles and applications



M. L. Meade

Lock-in amplifiers: principles and applications

M. L. Meade

e-edition

First published in Great Britain by
Peter Peregrinus Ltd, London, UK 1983

e-edition published by M.L. Meade 2013

Contents

The Contents list is for reference only.

Individual chapters can be navigated using the Table of Contents displayed in the PDF sidebar.

Preface

To the first edition

This book has been written for users of lock-in equipment and for those with an interest in the practical aspects of signal recovery and measurement using synchronous detection. The subject matter has been tackled for the most part at a systems level on the understanding that this is the approach most appreciated by research workers whose main specialization is in an area other than electronic instrumentation. Circuit designers will therefore look in vain for detailed circuit implementations; extensive mathematical developments are similarly avoided in favour of a more qualitative approach to identifying the essential features of lock-in detection systems. A basic familiarity with Fourier series and transforms provides adequate preparation for the main part of the text and it is hoped that the review of system configurations and specifications given there will prove both interesting and useful to specialists and generalists alike.

To the e-edition

It is now thirty years since *Lock-in amplifiers: principles and applications* was first published. In the intervening period it has become established as a minor classic, being amongst the most widely cited text books of its kind. Of even greater importance to me personally is that, despite a lapse of almost 25 years since the final printing, I continue to receive requests from researchers and students seeking a copy – hence my wish to create an 'authorised' electronic version and make it freely available.

Unlike the PDF held by Google Books (and numerous plagiarised versions circulating elsewhere) this e-edition has been newly compiled from the original typescript and diagrams. I make no apologies for leaving the content substantially the same as before, with no extensive revisions or additions. This reflects my view that, while there have been significant developments in the technology and implementation of lock-in systems, the greater part of the book, dealing with principles and guides to good practice, remains valid and useful. I might also confess that I am enjoying my retirement too much to spend more than the time necessary to correct obvious mistakes and to improve on the type-setting of equations which was less than satisfactory in the printed edition. There is, of course, the danger that this reworking is prone to fresh errors and, here, I am more than happy to apologise for any difficulty caused and fully prepared to amend and reissue any pages where errors are reported.

Mike Meade
Carlton
Bedford, UK

November 2013

Acknowledgements

From the first edition

The following chapters represent a greatly expanded version of a review article I prepared for the Institute of Physics in 1981[1]. A search of the literature at that time revealed that very little had been written about lock-in amplifiers beyond the technical notes published by the leading manufacturers. The task of writing this book was therefore greatly eased by the co-operation of friends and sometime colleagues at the E.G. & G. companies, Brookdeal and Princeton Applied Research, who made available a wealth of applications material and provided additional information.

At a more detailed level, the treatment in later chapters relating to heterodyne and p.w.m. systems owes much to discussions with Dr. Simon Carter of Hewlett Packard Ltd., South Queensferry. The formulation of spurious responses in these systems follows the lines developed by Dr. Carter in his Ph.D. thesis which is a prime source of reference on all aspects of lock-in systems.

I am greatly indebted to Chris Meredith of Aquarius Electronics, Beaconsfield, for providing a valuable perspective on lock-in systems and providing much painstaking, constructive, criticism on the manuscript. Also, Dr. David Crecraft and other colleagues in the Electronics Discipline of the Open University, UK, provided both assistance and encouragement throughout the writing period. Finally, I should like to make special mention of Christine Martindale and Jane Barden for managing the production of the manuscript.

For the e-edition

The e-edition was made possible by the combined efforts of the secretarial staff in the former Electronics Discipline of the Open University, UK, who undertook the scanning and hand-crafted OCR rendition of my original manuscripts during the summer of 2002. I am entirely to blame for the fact that the resulting files have only now seen the light of day.

[1] **References**

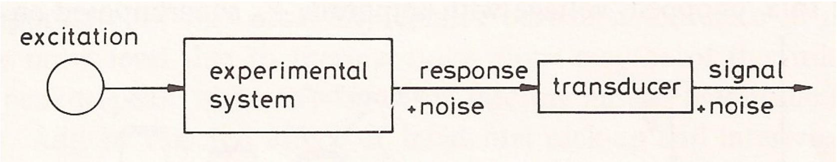
1 MEADE, M.L. (1982): 'Advances in lock-in amplifiers', J.Phys E: Sci Instrum., 15, pp. 395-403

2 CARTER, S.F. (1982): 'A systems approach to the design of lock-in amplifiers'. Ph.D.Thesis (University of Reading, England)

CHAPTER 1

Introduction

Signal recovery instruments will always be in demand as long as experiments are attempted with progressively smaller samples, weaker concentrations and fainter excitations. Nowadays it is more or less standard practice to design experiments to take advantage of signal-recovery techniques. Only then does it become possible to carry out measurements in the crucial stage of an investigation where the signals of interest become obscured by high levels of noise and interference. Of all the techniques that have been developed for signal recovery, methods based on the phase-sensitive detector and its modern counterpart, the lock-in amplifier, are by far the most widely applied in all fields of scientific research. Indeed, for many research workers, the terms 'signal recovery' and ‘lock-in recovery' are virtually interchangeable.



Fig, 1.1 A general experimental system

To see why this should be the case, let us begin with the general experimental system shown in Fig. 1.1. The system could be electrical, mechanical, optical, biological or any combination of such systems. The excitation source evokes a response in the output and the response is converted to an electrical signal in the transducer. We can suppose that the person who devised the experiment has a clear idea of how to interpret the signal.

In some experiments it may be essential to recover the *entire* output signal so that its waveform can be made available for analysis. When the signal is obscured by high levels of noise, some form of a signal averaging will be necessary and the experimenter might specify a multipoint averager or Fourier transform analyser for this purpose. Signals of a transient nature, triggered by repetitive pulses from the excitation source, are often dealt with in this fashion.

In experiments using 'static' or 'd.c.' excitations the output signal usually appears in the form of a slowly varying direct voltage proportional to an experimental parameter of interest. Severe measurement problems then result when the voltage falls to a level comparable with the error voltages due to offsets and drift in the transducer and its associated amplifier.

The temperature-dependent drift of d.c. coupled equipment is usually treated as a component in the $1/f$ noise or flicker noise that plagues low-frequency measurements. The effect of flicker noise on the determination of a fixed voltage level V_o is shown schematically in Fig. 1.2. The record shows a characteristic deviation from the initial condition established at time t = 0, resulting in an increasing measurement error as longer observation times are taken.

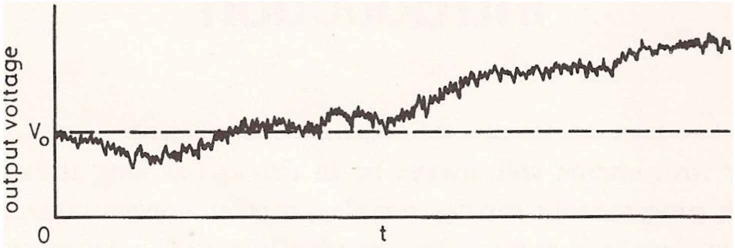


Fig. 1.2 D.C. output from an experimental system perturbed by low-frequency noise

In the light of these problems it is now established practice to interrupt normally 'static' excitations by providing some form of modulator at the input to an experiment. A typical research institute is usually a fruitful source of vibrating reeds, rotating discs and other electromechanical 'choppers' that have been engineered for this purpose. In most cases the net result is an ON/OFF modulation of the excitation source. The output voltage that originally had the form shown in Fig. 1.2 is then transformed to a 'chopped' voltage with amplitude V_o superimposed on the flicker noise voltage fluctuation. This new situation is depicted in Fig. 1.3(a).

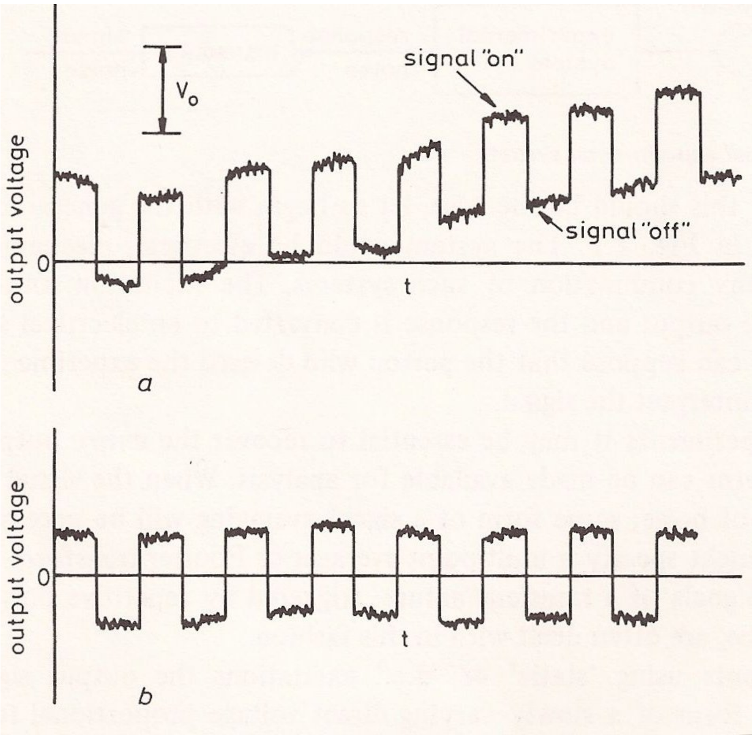


Fig 1.3 (a) Switched output voltage obtained by periodically interrupting the excitation source; (b) Output voltage with residual noise after high-pass filtering

In practice, the modulation or chopping frequency is usually made as high as possible to facilitate separation of the chopped output voltage from low-frequency noise components. This separation is achieved by using a high-pass filter and results in the *a.c.* signal shown in Fig. 1.3(b). The amplitude of this signal is proportional to the experimental parameter of interest and will usually vary in the course of an experiment. Note that in Fig. 1.3(b) the short-term noise fluctuations that appeared on the original signal have been transmitted by the high-pass filter and so appear in the final output. One way to overcome this residual fluctuation, when measuring the signal amplitude V_o, is to apply a *differential* measurement procedure to the output voltage. The idea is to measure the mean voltage *difference* between successive ON/OFF intervals and then average results over a number of modulation cycles. An improved estimate of the amplitude V_o will be obtained as the overall observation time is increased and a greater number of modulation cycles is taken into account.

In 1946, Dicke [1] showed how this type of measurement could be carried out automatically by using a *phase-sensitive detector*. A phase-sensitive detector measures the difference voltage of interest by using a synchronous *reference* voltage derived from the input modulator. We shall find that detection with respect to a synchronous reference enables the use of very long averaging times for the purpose of signal-to-noise ratio improvement and that practical systems are capable of operating with signals well below the background noise level.

The importance of this capability cannot be overstated since in many experiments the noise level due to thermal noise alone may be of the order of several millivolts peak-to-peak while the signal of interest has an amplitude measured in microvolts. Add to this the effect of incidental pick-up and interference and the result is a real signal recovery 'problem' awaiting solution.

In this context, phase-sensitive detectors offer a significant advance over alternative amplitude-demodulation schemes employing non-linear devices such as envelope detectors. The latter make no fundamental distinction between signal and noise components whereas a phase-sensitive detector is engineered to respond specifically to the information-bearing signal. If the term signal *recovery* implies that we have some prior knowledge of a signal, then the phase-sensitive detector is a true signal recovery device in that it takes account of the distinctive structure of the signal imposed by the use of a modulated excitation.

A phase-sensitive detector is responsive to the amplitude of a signal but is also sensitive to the phase difference between a signal and the derived reference. Phase-sensitive detector-based systems can therefore be devised to measure variations in both the amplitude and phase of periodic signals in the presence of noise and interference. Systems operating on the phase-sensitive detector principle are termed *lock-in systems* and the usual way of introducing a phase-sensitive detector into an experiment is to use a *lock-in amplifier.* This term has come to mean a free-standing instrument that incorporates a phase-sensitive detector, supported by preamplifiers, post-detection amplifiers and a comprehensive reference processing section.

A number of lock-in amplifier applications are listed in Appendix 1 which also serves to emphasize the widespread use of periodic excitations in experimental research. In the majority of cases the signal of interest appears in the output of the experiment at the same frequency as the fundamental excitation frequency. This certainly includes all measurements where the experimental processes are essentially *linear* insofar as no new frequencies are generated between input and output; for example, in a wide range of optical and electronic systems. To this broad class of experiments we can add examples where additional frequencies are produced, but where, once again, the frequency of interest is identified with that of the excitation source or bears a harmonic relationship to it. In any of these cases the phase-shift introduced by the experiment might be of interest, but, very often, it is sufficient to monitor changes in the magnitude of the output signal. On further examination we find that the applications can be broadly divided into two main categories. First of all, we have those many areas of activity where lock-in systems are used in their long-established role as signal-recovery tools for the measurement of modulated signals in noise. Secondly, there are many examples where lock-in systems are used for the *precision* measurement of signals, in situations where signal-recovery capability does not appear to be a prime consideration.

This essentially dual function of modern lock-in systems will be emphasized throughout the following chapters. When dealing with the principles of lock-in detection in Chapter 2 we shall begin along the lines followed by most authors in the field of telecommunications under the heading 'synchronous' or 'coherent' detection [2 ,3] Beyond this, however, we must improve upon the standard text-book treatment which is usually very disappointing to those whose interest is in experimental applications where signals are very slowly varying or may even be 'fixed' for the time available for measurement. Also, while some authors acknowledge that synchronous detectors are mathematically capable of withstanding adverse signal- to-noise ratios we find that very little attention is given to the practical aspects of demodulation under such conditions.

This is scarcely surprising since in many communication systems, signal-to-noise ratios of less than 10:1 or 20 dB would be considered quite unacceptable, while in signal-recovery work a typical starting point is with signal-to-noise ratios of less than 1:10, that is –20 dB or worse. We shall therefore find it necessary to discuss aspects of linearity and dynamic range and to make the transition from an ideal detector model to practical devices furnished with a range of specialist specifications. Most of the principles and techniques to be described could apply to lock-in amplifiers operating in almost any frequency range. Furthermore, the definitions of key specifications are independent of the particular technology used in the implementation of the lock-in amplifier. When dealing with 'typical' specifications, however, we shall take examples from commercial lock-in amplifiers optimized for the low-frequency range extending from less than 1 Hz up to a maximum of about 1 MHz. This corresponds to the frequency range in which the most significant developments in phase-sensitive detector technology have occurred and which satisfies the greatest number of applications.

The object throughout is to present information against a background of experimental work and to develop an awareness of the nature of signals and noise in experimental systems. The archetypal measurement system introduced at the beginning of this chapter will prove to be useful in this respect and is used as the basis for a general discussion about signals and noise in Chapter 2 which aims to put lock-in detection into a proper perspective. In Chapter 2, and throughout the following chapters, the treatment is mainly qualitative, with the principal mathematical developments left to the numerous Appendices.

Discussion on lock-in systems in the early chapters is confined to the so-called 'traditional' or 'conventional' variety of lock-in amplifier where the overall handling characteristics are essentially those of the switching phase-sensitive detector. These include some undesirable features, notably the susceptibility of the phase- sensitive detector to interference signals at the odd harmonics of the reference frequency.

The drawbacks of conventional systems are briefly reviewed in Chapter 6 in preparation for the following chapters which deal with various 'advanced' systems operating on the heterodyne and the pulse-width-modulation principles. Here we shall find that the odd harmonic responses referred to above are suppressed through more extensive processing of the signal and reference voltages, but that the switching phase-sensitive detector is retained on account of its ability to maintain linear operation under the most adverse noise conditions. At the same time, the more complex system configurations are found to be characterized by a number of additional spurious responses which must be minimized at the design stage. There are also certain inevitable trade-offs, notably with respect to dynamic range and frequency coverage, which must be taken into consideration, and we shall take account of these where appropriate.

Heterodyne and pulse-width-modulation lock-in amplifiers are almost invariably supplied and used as self-contained units. In Chapter 7, however, we find that the topic of phase-locking bridges the gap between 'conventional' and 'advanced' systems in that reference processing can be arranged by using standard modules such as phase-sensitive detectors and voltage-controlled oscillators. Readers familiar with the literature on this subject will find the approach here heavily biased towards locking with noisy signals, the object being to derive a ‘local’ reference voltage when this is otherwise unavailable. The treatment is unavoidably mathematical in this case, but it is nevertheless intended that non-specialist readers will benefit from a review of the problems of locking in noisy conditions. The optimization procedures described represent but one way of approaching the phase- lock problem. They do, however, take account of long-term variations in signal amplitude, a feature which is noticeably lacking in the general literature.

Chapter 10 deals with some of the problems inherent in bringing signal-recovery equipment under computer control. Here, as elsewhere, it is hoped that the development of ideas will be accessible to readers with an interest in lock-in systems as measurement tools but who are otherwise non-specialists in the general areas of electronics and telecommunications.

The same remarks apply to Appendices 5 and 6 which give an appraisal of noise in amplifiers and the problems associated with signal connections, in particular the avoidance of ground loops. It is noted there that lock-in recovery can be a more-or-less straightforward business provided that proper attention is paid to signal handling. While acknowledging that familiarity with specifications and the basic rules of instrument management are best learned 'at the bench', it is hoped that the guidelines established in this book are of the sort which make lock-in recovery a reasonably exact science and that they will ensure that many common pitfalls leading to erroneous or misleading results will be avoided.

1.1 References

1 DICKE, R.H. (1946): "The measurement of thermal radiation at microwave frequencies', *Rev. Sci. Instrum.,* 17, (7), pp. 268-275

2 TAUB, H., and. SCHILLING, D.L. (1971): 'Principles of Communication Systems' (New York, McGraw Hill)

3 BETTS, J.A. (1970): 'Signal processing, modulation and noise' (London, English Univ. Press)

CHAPTER 2

Basic concepts in lock-in recovery

2.1 Introduction

Lock-in amplifiers are characterized by a wide dynamic range which gives the ability to measure signals accompanied by relatively high levels of noise and interference. It is appropriate therefore to begin with an examination of the signal and noise voltages which might appear at the output of an experimental system of the general type introduced in Chapter I and, in particular, to identify the combinations of signals and noise which give rise to a signal recovery 'problem'. As we shall see, a fairly close examination is necessary in order to give substance to the vague, if popular, notion of signals 'buried in noise'. We shall then turn our attention to methods of signal measurement and outline the principles of synchronous detection which underlie the operation of lock-in amplifiers.

It was shown in the introductory chapter that signal recovery applications of lock-in amplifiers usually involve the measurement of amplitude variations and - to a lesser extent - phase variations of periodic signals. In support of this, Appendix 2 gives some attention to the structure of modulated signals and gives methods of estimating signal bandwidth. In many signal recovery applications, however, the signal modulations are relatively slowly varying, not just with respect to the excitation frequency but also with respect to observation intervals ranging from several seconds up to many hours. Indeed, in many circumstances the signal may have fixed characteristics throughout the time available for measurement. We shall find in this chapter that many of the important differences between synchronous detection methods and other, non-linear, detection methods can be readily demonstrated on this assumption.

This chapter includes an introduction to basic lock-in amplifiers. A brief appraisal of the various system components highlights some of the main topics for discussion in the chapters to follow.

2.2 Evaluating the signal recovery 'problem'

In the following we shall suppose that signals are obtained in the form of a voltage variation either directly from an electrical transducer or from a suitable combination of transducer and preamplifier. Although of extreme importance, the technicalities of amplifier selection and the practicalities of signal connections will not be a major concern at this stage; our objective will be to examine the general characteristics of signals and noise which might be encountered in the course of a typical measurement.

We define ‘noise’ as all unwanted signals, so all sources of random disturbance must be included in its description, together with the effect of more-or-less well-defined interference originating from neighbouring experiments and installations. The broad spectrum of interference phenomena is charted in Appendix 2 and includes several well-known - if not notorious - sources such as mains-frequency ‘hum’, r.f. breakthrough from pulsed experiments and pick-up from t.v. and radio transmissions. Interference phenomena in general have a different status to the noise sources that are inherent in the experimental processes under investigation, in the amplifying devices employed and in the transducers used to provide the output signals. For example, interference effects can often be suppressed - if not entirely eliminated - by careful experiment design, so it is useful to make a distinction between noise which is fundamental to the system and noise of external origin.

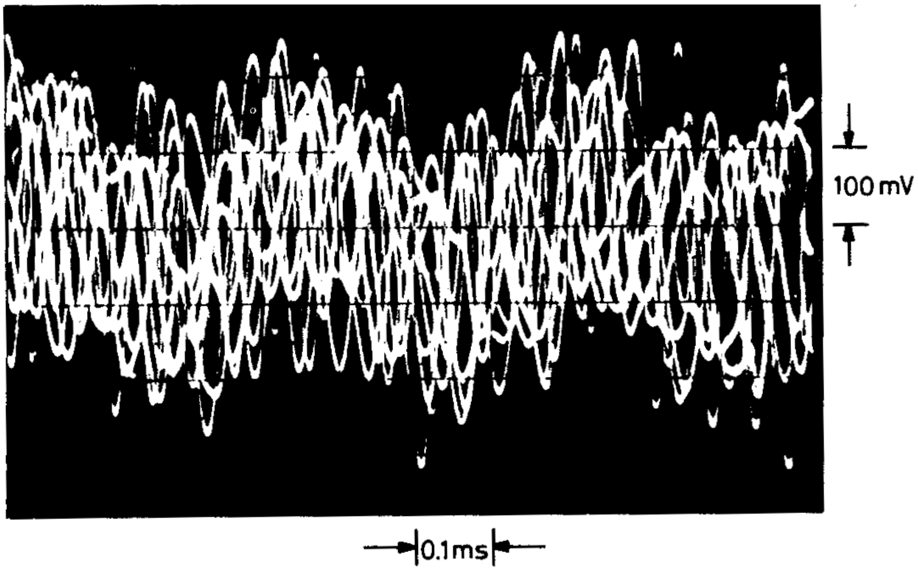


Fig. 2.1 Typical oscilloscope view of a signal contaminated by noise and discrete interference

At the same time we should not make the mistake of supposing that the system noise is so 'fundamental' that it cannot be reduced further. In practice there is often scope for improvement through selection of amplifiers, a topic which is treated at some length in Appendix 5.

The first direct contact with signals in a laboratory is usually made via an oscilloscope. In the case of noisy signals the resulting display is often a confused jumble for example. Fig. 2.1 which shows the effect of noise, including discrete interference, on a low-level signal.

The term 'buried in noise' seems to be most appropriate here, but in many cases this conclusion tells us more about the method of display than it does about the signal. Let us look therefore at an alternative way of displaying the characteristic of the signal, by using a spectrum analyser.

The spectrum-analyser approach gives a more explicit and graphical interpretation of the relationships between signals and noise in a system. We find that by making the transformation from time domain to frequency domain we can often sort the signal and noise into their respective categories, a process which overcomes the confusing superposition of the time domain picture.

Fig. 2.2 is such an example and uses the same input voltage as displayed on the oscilloscope. The signal is now very much in evidence and well separated in terms of frequency from sources of interference such as high-frequency breakthrough from a neighbouring experiment. The noise which appeared to dominate the oscilloscope display is now seen to be spread quite 'thinly' over a wide range of frequencies.

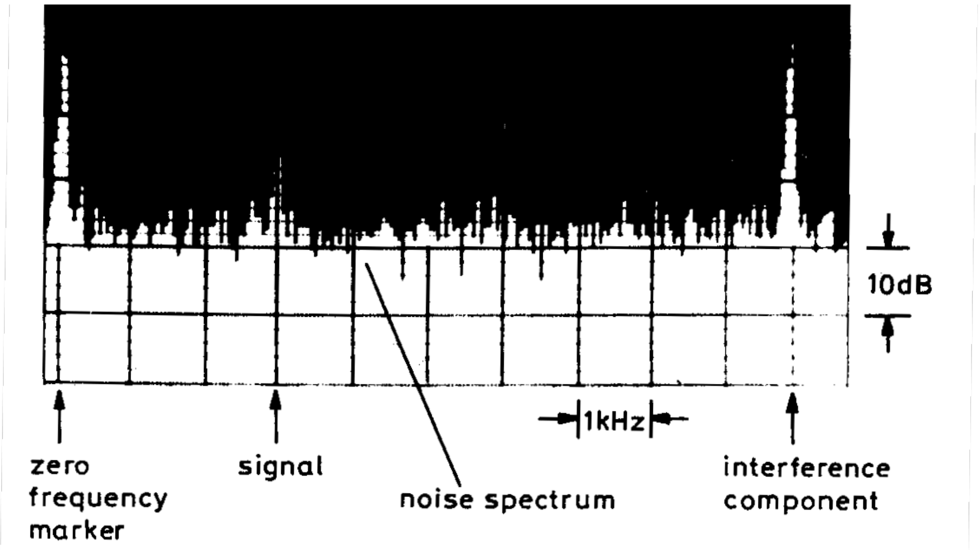


Fig. 2.2 Spectrum-analyser display of the signal shown in Fig. 2.1. On this scale the mains frequency pick-up region is immediately adjacent to the zero frequency marker

The spectrum analyser picture also serves to remind us that many sources of interference can be described quite adequately in terms of a 'line' spectrum, representing the concentration of power at discrete frequencies while the fundamental noise and other interference sources give rise to a continuous spectrum of noise up to the upper cut-off frequency of the output transducer or output amplifier. Appendix 2 gives a breakdown of the principal interference mechanisms according to their characteristic frequency ranges and spectral signatures, and then gives attention to providing simple mathematical models for the various kinds of spectrum which might be encountered in practice. These include broadband spectra resulting from thermal noise and shot noise, narrowband spectra and the ubiquitous flicker-noise or $1/f$ noise. The latter is associated with a rise in noise power density at low frequencies. These models provide the means of estimating the contribution to an observed fluctuation which emanates from different parts of the noise spectrum in preparation for the first step in signal recovery; this is the elimination of unwanted noise by filtering. Thus, given a combination of signal and noise spectra such as that shown in Fig. 2.2, we find that there is ample scope for improvement by introducing filters to 'clean-up' the signal *prior* to detection. The elimination of unwanted components in this way is an important aspect of *signal conditioning* and is often effective in bringing about a substantial increase in signal-to-noise ratio. In this example, the use of a bandpass filter centred on the signal component indicated in Fig. 2.2 results in the displays shown in Fig. 2.3. The signal of interest is now relatively noise-free so that variations in, say, signal amplitude could be measured using a conventional a.c. voltmeter without incurring a serious noise penalty.

The idea of noise reduction by bandwidth reduction is of course central to an; discussion of signal recovery. If, however, we restrict ourselves to the question of signal-to-noise improvement *before* demodulation, then we find that the benefit obtained by signal conditioning can be very limited.

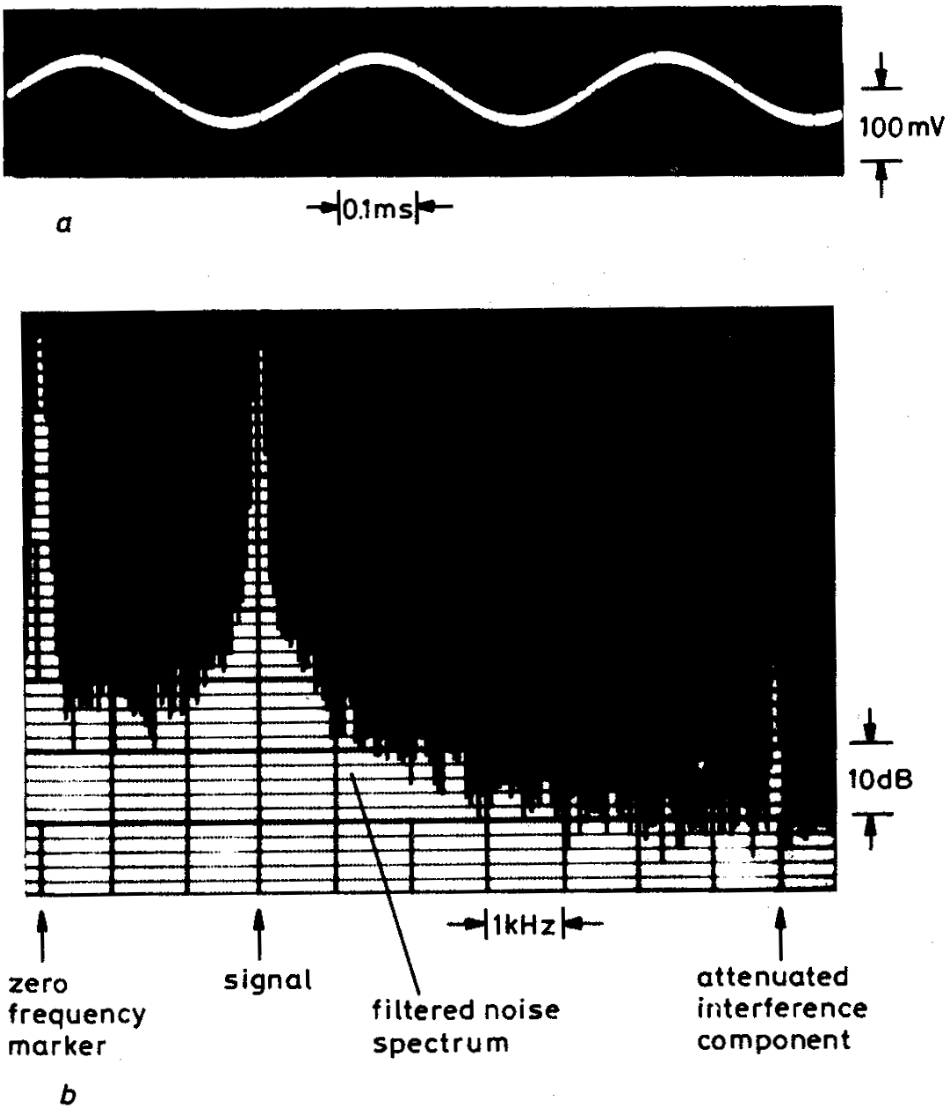


Fig. 2.3 (a) Signal of Fig. 2.2 following the introduction of a bandpass filter centred on the signal frequency; (b) corresponding frequency-domain picture

To illustrate this let us look at the idealized case shown in Fig. 2.4 where a bandpass filter with bandwidth B_0 is used to reduce the noise power appearing with the signal, leaving the signal unaffected. It is shown in Appendix 2 that the noise power is reduced by a fraction B_0/B_I where B_I is the 'input' noise bandwidth to the filter. If, at some later stage in the measurement, the signal amplitude should be reduced, then the output signal-to-noise ratio can only be maintained by a further reduction in filter bandwidth. For example, if the signal amplitude falls by a factor of 10 (signal power reduced by 100) the filter bandwidth must be reduced to 1/100 of its former value in order to restore the signal-to-noise ratio.

Clearly, this process cannot be repeated indefinitely. First of all, if the signal is carrying modulation it will occupy a finite bandwidth, and this, in turn, will determine the smallest filter bandwidth that can be used. Secondly, there are practical limits to designing filters with very high selectivity, which, in any case, results in a tightly 'tuned' measurement system. This, almost invariably, gives rise to additional problems such as susceptibility to drift and the inability to follow even small variations in signal frequency.

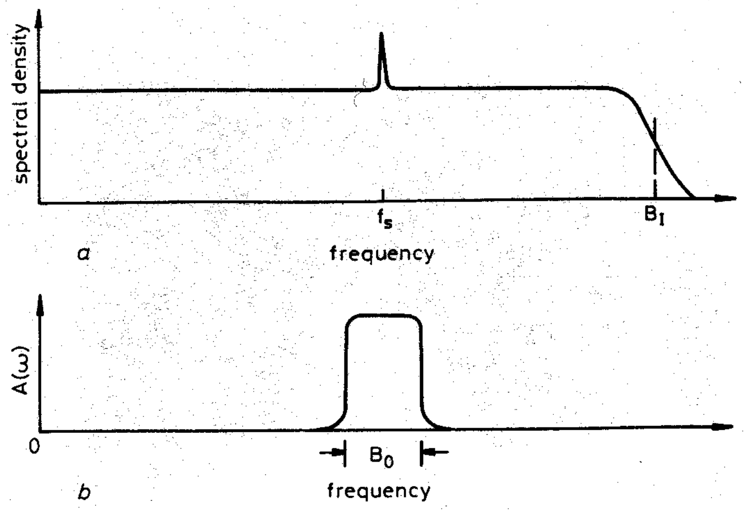


Fig 2.4 **(a) Spectrum of signal and noise; (b) amplitude response of bandpass filter used for noise reduction. In this ideal case, the filter signal bandwidth, B_0 is equal to the noise equivalent bandwidth**

Bearing these points in mind we can make a clear distinction between our original example and the case shown in Fig. 2.5. The noise power is now concentrated with high density in the region of the signal frequency and the signal is over-shadowed by a massive interference component nearby. We are now faced with a signal recovery problem of quite a different order, where adequate separation of a signal and noise cannot be achieved by filtering and a substantial fraction of the noise and interference must filter through to the equipment used for demodulation.

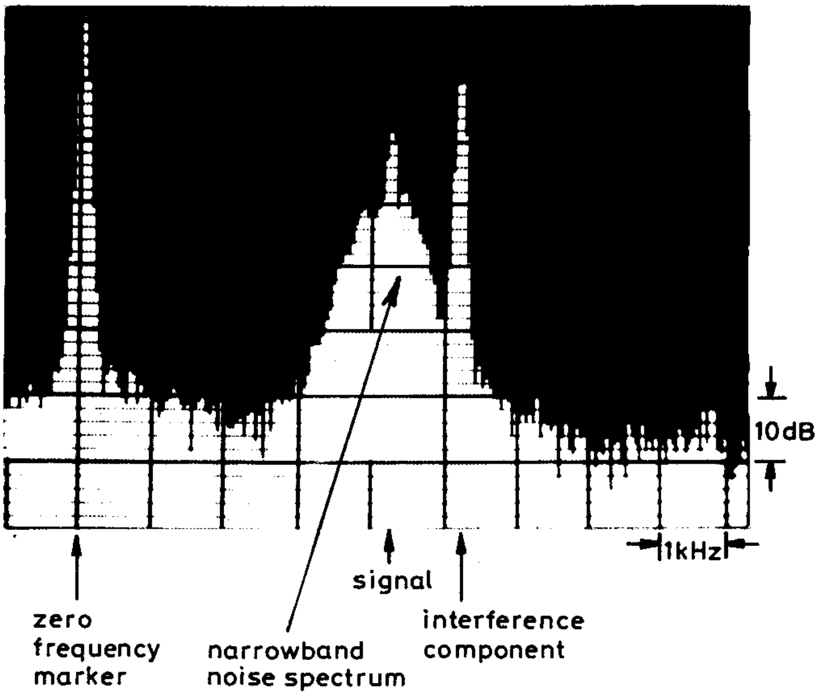


Fig. 2.5 **Frequency-domain view of a signal recovery 'problem'**

In evaluating the effect of noise on a signal, therefore, we are interested more in the distribution of the noise components with frequency rather than in the *total* noise power which accompanies the signal. Thus, the total peak-to-peak fluctuation which can so seriously obscure an oscilloscope display is not in itself sufficient to prevent measurement of a signal, provided that adequate noise reduction can be achieved through signal conditioning. If this approach fails, then the final outcome of an experiment will depend entirely on the ability of the demodulator to function in the presence of noise.

2.3 Demodulators for signal recovery

It was suggested in Chapter 1 that the term 'signal recovery' implies some prior knowledge of a signal and that this knowledge is used to advantage *at the point of detection.*

The use of signal conditioning filters is an example of how 'prior knowledge' of the signal can aid the process of detection. Let us now turn our attention to the provision of a demodulator and ask: are there any techniques available which take specific account of the 'character' of the signal imposed by the choice of excitation source? In answering this question we would separate the 'true' signal recovery techniques from those which make no fundamental distinction between signals and noise. We would find an essential difference in that demodulators for signal recovery are almost invariably supplied with a *reference* signal which is precisely synchronized with the signal of interest.

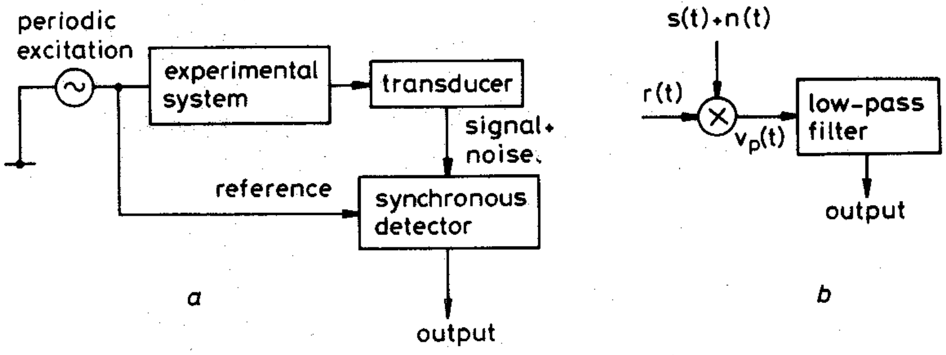


Fig. 2.6 (a) Incorporation of a synchronous detector in an experimental system; (b) multiplier model for a synchronous detector

It is the availability of the reference which enables us to exploit the principle of synchronous detection referred to in Chapter 1. The use of the term 'lock-in systems' in this context reminds us that we are dealing with demodulators which are 'locked' to a signal of interest by virtue of a synchronous or coherent reference voltage. Fig. 2.6(a) is fairly typical in that the reference has been obtained directly from the excitation voltage at a fixed level.

The response of a synchronous detector to variations in the amplitude and phase of a synchronous signal will be considered in the next section. In general terms, operation depends on the high degree of correlation which is known to exist between a periodic signal of interest, $s(t)$, and the reference $r(t)$. The presence of correlation is tested using the scheme outlined in Fig. 2.6(b), by first multiplying the two inputs to form the product

$$v_p(t) = r(t)\,[s(t) + n(t)]$$

where $n(t)$ represents the entire disturbance due to additive noise and interference.

When $r(t)$ and $s(t)$ are closely correlated, their product gives rise to a distinctive response which depends upon the amplitude of the signal and its phase relative to the reference. For example, in the important case of practical interest where $s(t)$ is a sinewave having fixed characteristics during some stage of a measurement, the product term $r(t)s(t)$ will give rise to a voltage which includes a constant term proportional to signal level. The purpose of the low-pass filter shown in Fig. 2.6(b) is to separate this voltage from the higher-order products of multiplication and allow it to filter to the final output. Regarding the noise, there should be no correlation with the reference in this case and the average value of the noise product $r(t)n(t)$ is always zero in the final output. The response to a fixed signal is quite unambiguous because there is no error due to rectified noise components: any residual fluctuations due to noise appear as an *a.c.* variation which does not affect the average value of the 'true' output voltage due to the signal. In principle, these residual fluctuations can always be attenuated to an acceptable level by reducing the bandwidth of the output low-pass filter. This represents the major mechanism for signal-to-noise improvement in synchronous detection systems.

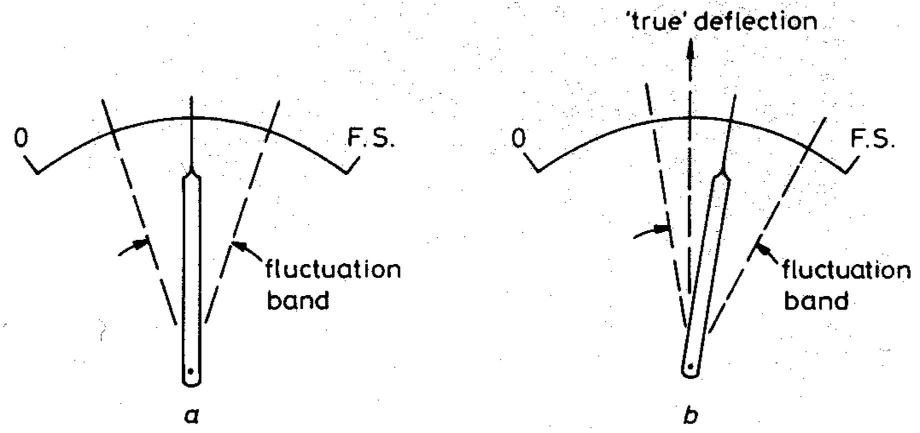


Fig. 2.7 Meter indications at the output of: (a) a synchronous detection system; (b) an envelope detector. The 'true' deflection due to signal corresponds to half-scale in each case

The response to a signal with changing amplitude will be a varying output voltage with additive noise. When the amplitude is changing very slowly we can interpret the response as a slowly varying 'd.c.' level. In this case an output filter with very small bandwidth can be used to smooth the fluctuation due to noise. The problem is then to estimate the value of the d.c. component. This might be achieved by observing the output on a voltmeter as shown in Fig. 2.7(a). With a synchronous detector there is no residual deflection due to rectified noise and the desired d.c. response is given by observing the average deflection of the meter. Other types of detector, for example envelope detectors, give a meter indication characterized by Fig. 2.7(b). Here the response is subject to an error due to rectified noise components which contribute to the net deflection of the meter. This source of error cannot be removed by using a low-pass filter. The only recourse in this case is to eliminate noise components *before* detection by the use of filters tuned to the signal frequency.

A further point to be noted is that rectifier-detectors are subject to *intermodulation* effects whereby signals and noise become multiplied together. Intermodulation is inevitable in any process which depends on a non-linear operation such as squaring or envelope detection. The effect can be safely ignored at high input signal-to-noise ratios, but when the signal is reduced the intermodulation products dominate and give rise to the phenomenon of *threshold* or *signal suppression.* This generally occurs at input signal-to-noise ratios below about 1:1 and corresponds to a loss of information in that the detector output no longer contains a term which is simply proportional to the desired modulation.

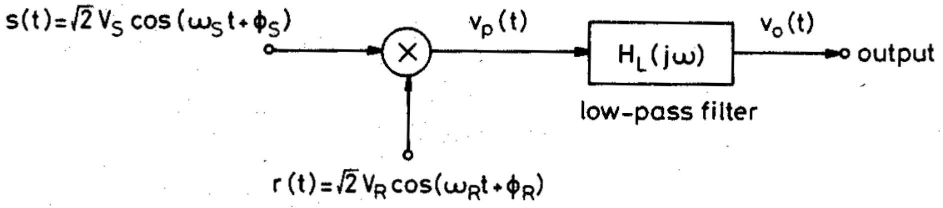
In the light of this general discussion it is evident that 'ideal' multiplier synchronous detectors do not generate products of the type $s(t)n(t)$; a distinct separation between signal and noise is maintained throughout. As a result, the residual noise output *adds* to the desired response. We can say that synchronous detectors are *linear* insofar as the principle of superposition can be applied in order to combine the responses due to individual components in the signal and noise voltages.

Let us now turn from these general considerations and look at some specific relationships between the reference, signal and noise in a synchronous detection system.

2.4 Operation of synchronous detectors

2.4.1 Introduction

We shall begin by summarizing basic mathematical relationships which apply to the ideal multiplier followed by a low-pass filter shown in Fig. 2.8.



Fig, 2.8 Synchronous detector with sinewave signal and reference

A sinewave reference will be used and we shall calculate the response to a single sinewave component in the signal path. Signal and reference are conveniently expressed in terms of their r.m.s. values V_s and V_R:

$$s(t) = \surd 2\ V_\text{s} \cos\left[\omega_\text{S} t + \phi_\text{s}\right]$$

$$r(t) = \surd 2\ V_\text{R} \cos\left[\omega_\text{R} t + \phi_\text{R}\right]$$

If we now form the product of signal and reference we can separate the result into sum and difference components:

$$\begin{aligned} v_\text{p}(t) &= V_\text{s}V_\text{R} \cos\left[(\omega_\text{s} + \omega_\text{R})t + \phi_\text{s} + \phi_\text{R}\right] \\ &+ V_\text{s}V_\text{R} \cos\left[(\omega_\text{s} - \omega_\text{R})t + \phi_\text{s} - \phi_\text{R}\right] \end{aligned}$$

This is an operation which will recur in succeeding chapters and which is almost invariably linked to the assumption that the low-pass filter cuts off at a frequency much less than ω_R. In this case the sum-frequency component is effectively eliminated from the final output. The fate of the other component will depend on the magnitude of the difference frequency $\Delta\omega = |\,\omega_\text{S} - \omega_\text{R}\,|$. If this is less than or comparable with the bandwidth of the low-pass filter we find that the output appears in the form of an alternating or 'beat' response at frequency $\Delta\omega$. To calculate the magnitude of this response we require the frequency-response function $H_L(j\omega)$ of the low-pass filter. In general, the beat' component will have amplitude

$$|v_o| = V_S V_R A_L(\Delta\omega)$$

where

$$A_L(\omega) = |H_L(j\omega)|$$

Since $A_L(\omega)$ has a cut-off well below ω_R we find that the system is able to accept only those signal components which lie very close to the reference frequency. This can be described in terms of a *transmission window* centered on the reference frequency with a characteristic dependent on $A_L(\omega)$ as shown in Fig. 2.9. We can thus argue that the combination of reference, multiplier and low-pass filter functions as a band-pass system giving a response only to signals in the vicinity of the reference frequency.

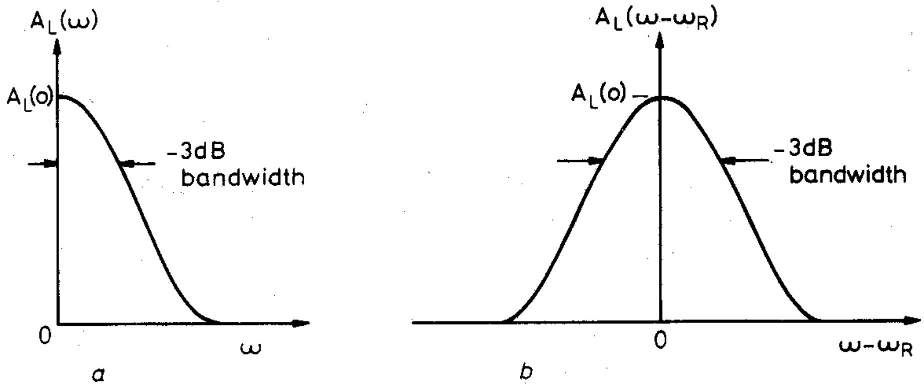
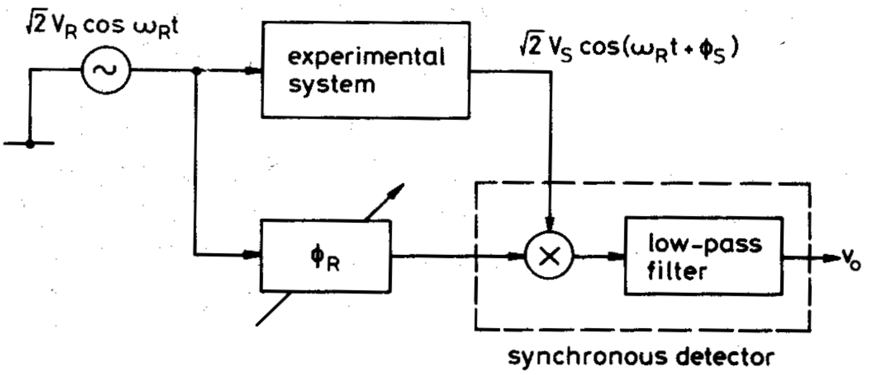


Fig. 2.9 (a) Frequency-response magnitude of low-pass filter; (b) transmission window derived from the characteristics of the low-pass filter and centred on the reference frequency, ω_R

2.4.2 Demodulation with a synchronous reference

Let us now turn to the case of greatest practical significance, where the signal and reference are derived from the same source, for example in the experimental scheme shown in Fig. 2.10 below.

The excitation of the experiment is sinusoidal at frequency ω_R and provides a reference which is to be used for detection of the output signal. The output signal appears at the same frequency as the excitation and suffers a phase-shift ϕ_S in the experiment. The reference is applied to the multiplier via a variable phase-shift network.



$\sqrt{2} V_R \cos \omega_R t$
experimental system
$\sqrt{2} V_S \cos(\omega_R t + \phi_S)$
ϕ_R
×
low-pass filter
V_o
synchronous detector

Fig. 2.10 Using a synchronous detector with a variable phase-shift network in the reference path

At this stage let us assume that the signal is noise-free. The output from the low-pass filter can then be calculated by putting $\omega_s = \omega_R$ in the results derived earlier. For signals with fixed amplitude and phase we obtain the classic phase-sensitive response in the form of a d.c. indication:

$$v_o = k_R V_s \cos\phi$$

where

$$\phi = \phi_s - \phi_R$$

$$k_R = V_R A_L(0)$$

Usually the reference amplitude is fixed so that k_R is a constant scaling factor. The response is then simply proportional to the signal amplitude and exhibits a phase dependence through the cosine term. Let us now see how the reference phase- shifter can be used to measure some specific modulations on the signal.

2.4.3 Amplitude demodulation

The reference phase is adjusted to bring the signal and reference in phase at the multiplier to give an output:

$$V_o = k_R V_s$$

for

$$\phi_R = \phi_s$$

The output will follow variations in signal amplitude provided that the low-pass filter has a bandwidth wide enough to transmit the modulation signal without distortion. For example, when the signal has the form:

$$v_s(t) = m(t) \cos \omega_o t$$

the output voltage will be

$$v_o(t) = k_R m_F(t)$$

where $m_F(t)$ is a low-pass filtered version of the modulation signal $m(t)$.[1]

In this case the output voltage of the synchronous detector will have a spectrum;

$$V_o(j\omega) = k_R M(j\omega) H_L(j\omega)$$

$$= M(j\omega) H_D(j\omega)$$

Here, $M(j\omega)$ is the Fourier transform of the modulation signal and we identify

$$H_D(j\omega) = k_R H_L(j\omega)$$

as the frequency-response function of the synchronous detector.

2.4.4 Phase demodulation

To use the synchronous detector as a phase demodulator we must assume that the signal amplitude is constant. The first step is to *null* the output of the detector by bringing the signal and reference into *quadrature* at the multiplier, giving;

[1] If we denote the impulse-response of the low-pass filter by h_L(t), the filtered version of the modulation function will be given by:

$$m_F(t) = m(t) \otimes h_L(t)$$

where ⊗ denotes convolution.

$$\phi_R = \phi_s - \pi/2$$

If the signal phase subsequently changes by an amount ϕ_m the response of the detector will be

$$v_o = k_R v_s \sin\phi_m$$

The synchronous detector will operate as a linear phase detector only for small phase variations. In this case we restrict the magnitude of ϕ_m, so that $\sin\phi_m \cong \phi_m$ and obtain the approximately linear response:

$$v_o \cong k_R v_s \phi_m$$

Let us now take the general case where ϕ_m is a time-varying phase-shift $\phi_m(t)$. When $|\phi_m(t)| << 1$ this corresponds to low-index phase modulation. We must now consider the effect of the low-pass filter on the frequency components *of* $\phi_m(t)$. If the phase modulation has a Fourier transform $\Phi_m(j\omega)$ the required relationship is

$$\begin{aligned} V_o(j\omega) &= k_R V_s \Phi_m(j\omega) H_L(j\omega) \\ &= V_s\ \Phi_m(j\omega) H_D(j\omega) \end{aligned}$$

or, in terms of a time variation;

$$v_o(t) = k_R V_s \phi_m(t) \otimes h_L(t)$$

The sensitivity of the system to phase variations is thus proportional to the amplitude of the signal. Otherwise, the detector frequency-response function $H_D(j\omega)$ plays the same role in amplitude and phase detection.

2.4.5 Mixed modulations

In general, we must expect that the signal of interest appears with both amplitude and phase modulations, in the form

$$s(t) = m(t) \cos\left[\omega t + \phi_s + \phi_m(t)\right]$$

If the reference phase is adjusted to bring signal and reference in phase at the multiplier such that $\phi_s = \phi_R$ the system response becomes:

$$\begin{aligned} V_o(t) &= k_R m(t) \cos\phi_m(t) \otimes h_L(t) \\ &= k_R m(t) \otimes h_L(t) \end{aligned}$$

for $|\phi_m(t)| << 1$ radian

We see that amplitude detection is first-order independent of phase variations on the signal. The implication is that small errors in setting-up the ‘in-phase’ condition do not seriously affect accuracy when signal amplitude is to be measured. Indeed an error of ±10° in the in-phase condition leads to an error of only ±1.5 % when measuring the amplitude of a fixed signal.

It is worth noting that the operation of a synchronous amplitude detector does not depend on $m(t)$ being constrained to take only positive values. Negative values of $m(t)$ correspond to a phase reversal of the carrier signal which will be faithfully reproduced as negative voltages in the output of the detection system.

Turning now to operation as a phase detector, we have noted that the overall response is proportional to signal amplitude. This simple system is therefore unsuited to operation as a phase detector when the signal carries amplitude modulation.

More complex systems which involve two synchronous detectors operated in quadrature will be considered in later chapters. These systems allow amplitude *and* phase variations to be measured simultaneously without restrictions on the maximum allowable phase shift. The key to this mode of operation lies in the phasor representation of the signal and reference voltages given in Fig. 2.11, drawn with respect to the reference phase. The output of a synchronous detector is now seen to be proportional to the *in-phase* component of the signal, while changing the reference phase by π/2 gives us a measure of the *quadrature* component of the signal. With the possibility of generating these two pieces of information we begin to see the potential of synchronous detection systems for phasor analysis and signal characterization beyond their traditional role in signal recovery.

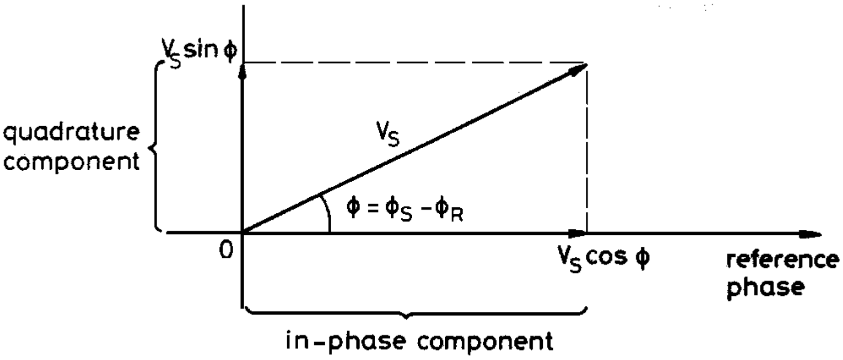


Fig. 2.11 Phasor representation of a sinusoidal signal drawn with respect to reference phase

2.4.6 Noise rejection

The response of a synchronous detector to a synchronous signal of fixed amplitude and phase is always obtained in the form of a *d.c.* indication. In contrast, asynchronous signals such as noise and discrete interference components always give rise to an alternating response in the form of 'beat' frequency products as discussed in Section 2.4.1. Furthermore, we have seen that the only unwanted components which can give rise to spurious outputs are those which originate in the immediate vicinity of the reference frequency, confined to the transmission 'window' defined by the characteristics of the low-pass filter. We can conclude, therefore, that the rejection of a large part of the background noise spectrum is inherent in the operation of a synchronous detector and that the frequency selectivity of the detector is governed by the properties of the low-pass filter. In practice, the experimental system itself will set limits to the maximum rate of change in the amplitude and phase of the signal of interest. This in turn will determine the minimum bandwidth which can be tolerated in the low-pass filter if the recovered information is to be transmitted without distortion. In many applications this maximum rate of change is deliberately restricted in order to achieve the smallest possible value of output bandwidth and so enhance the noise rejection properties of the detector (for example, by reducing the scan rate of spectrometers and swept-response analysers). The case of a signal accompanied by a wide band of white noise is given consideration in Appendix 3. Because of the inherent linearity of synchronous detectors we find that the classic noise reduction formula can be applied to calculate the signal-to-noise improvement obtained by demodulation. This is defined in terms of the 'output' signal-to-noise ratio (SNR_{o}) of the recovered information signal and the 'input' signal-to-noise ratio (SNR_{I}) of the modulated signal. The latter is assumed to appear in white noise with bandwidth B_{I}. For the recovery of amplitude modulation we have

$$\text{Improvement factor} = \frac{SNR_{\mathrm{o}}}{SNR_{\mathrm{I}}} = \frac{B_{\mathrm{I}}}{B_{\mathrm{o}}}$$

where B_0 is the (noise) bandwidth of the low-pass filter, set to a value just wide enough to pass the modulation signal.

The use of the filter noise bandwidth ensures that noise transmitted in the 'tails' of the filter beyond the signal cut-off frequency is accounted for in the calculation. The noise bandwidths of several important filter types are given in Appendix 4.

To conclude this section let us review some of the practical advantages of synchronous detection schemes compared with methods of non-linear detection. We can begin with Fig. 2.12 which compares the signal-to-noise improvement obtained for amplitude demodulation using a synchronous demodulator and an envelope detector.

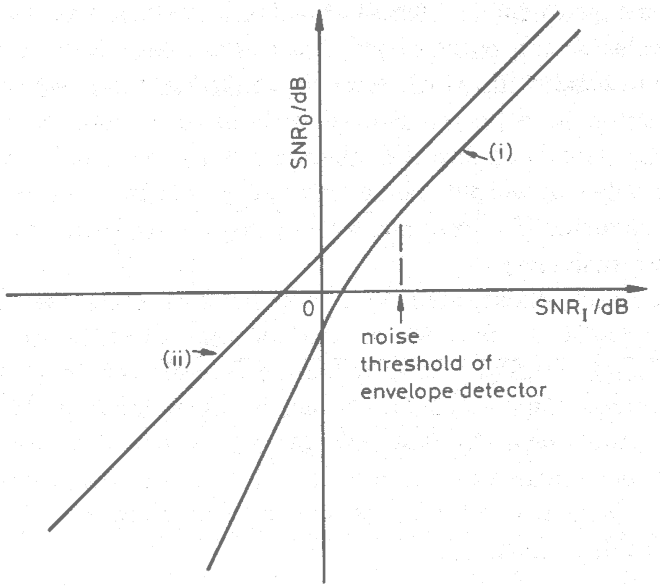


Fig. 2.12 Comparison of input and output signal-to-noise ratios for (i) an envelope detector and (ii) a synchronous detector

For a valid comparison to be made we must ensure that the systems are identical with respect to the noise-rejection filters used before and after demodulation. We find that for strong signal conditions the two methods are comparable. The improvement factor is constant here, with SNR_0 strictly proportional to SNR_I (a discrepancy of a few dB is barely significant in signal recovery terms). The situation is quite different, however, when the input signal-to-noise ratio is low. The synchronous demodulator is capable of maintaining a constant improvement factor for all levels of SNR_I whereas the envelope detector deteriorates rapidly when the input ratio falls to about 1:1. In the region below threshold the output signal-to-noise ratio of the envelope detector falls faster than SNR_I, and the demodulated output becomes grossly distorted.

In addition, the performance of the envelope detector is degraded for signal-to-noise ratios which would normally be considered quite favourable in a signal recovery context. We therefore find it necessary to provide noise suppression filters centred on the signal frequency to ensure a high signal-to-noise ratio *before* detection. In difficult conditions this probably means that we require highly selective tuned circuits which are subject to drift and other temperature effects and which render the system incapable of operation if the signal frequency is changing by accident or design.

Synchronous demodulators, and hence lock-in systems, are normally operated without front-end filters and the final noise rejection takes place by averaging in a *low-pass* filter which follows the multiplier. This final bandwidth need be no wider than to transmit the information signal without distortion. In many applications the signal variations are of such a long-term nature that a final bandwidth of a few hertz may be quite adequate. The practical significance of defining such a narrow bandwidth by means of a low-pass filter, rather than by using a highly selective bandpass filter at the signal frequency, cannot be overstated. As a direct result we find that lock-in systems are ideally suited to swept-response measurements. By using a swept reference, it is possible to track signals over many decades of frequency.

A plot similar to Fig. 2.12 can be drawn to show the occurrence of a noise threshold in systems used for phase demodulation. Lock-in systems are free of this effect for small index phase variations but this aspect is rather overshadowed by the growth in importance of lock-in systems used for *precision* measurements of amplitude and phase on relatively 'clean' signals.

This is a relatively new development brought about by improvements in electronic circuit and system design during the last decade and which will be given due attention in the following chapters.

2.5 Basic lock-in amplifiers

2.5.1 Introduction

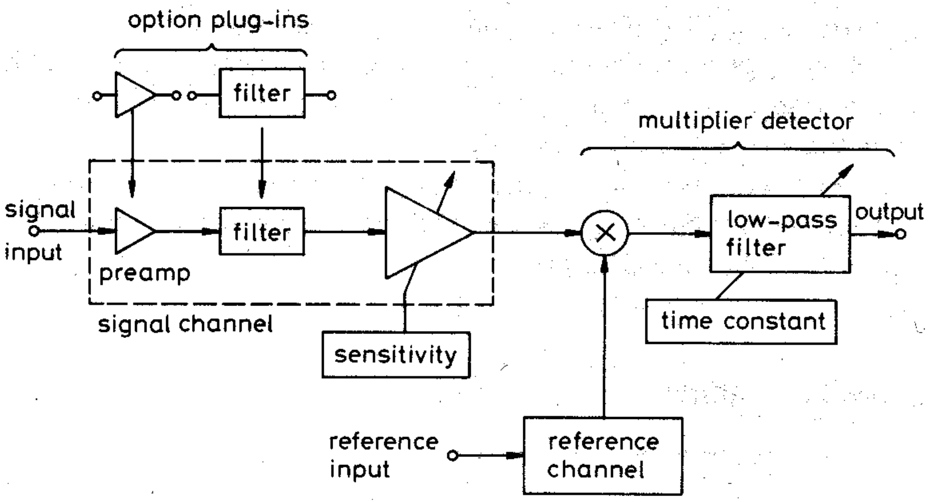


Fig. 2.13 A basic lock-in amplifier

The requirements of a basic lock-in system are shown in Fig. 2.13. Although such a system could be built up by interconnecting individual units, the purchase of a lock-in amplifier usually represents a more cost-effective approach. Lock-in amplifiers incorporate all the features of Fig. 2.13 in a single unit which is optimized for operation over a range of frequencies. The major advantages of using an integrated system of this type are that the controls are calibrated directly in terms of full-scale sensitivity for a synchronous signal and that the relative phase of the signal and reference channels can be maintained to within close limits over the recommended frequency range.

Let us now identify the essential requirements of each main block in the lock-in amplifier and take the opportunity to review some practical aspects of lock-in operation.

2.5.2 The signal channel

An amplifier is necessary to bring the signal to a level sufficient to overcome the self-noise of the multiplier, and the provision of switched gain permits the sensitivity of the system to be varied. As shown in Fig. 2.13 lock-in amplifiers are usually provided with an optional range of preamplifiers. The objective here is not merely to boost the gain but to provide an optimum *noise match* to the signal source. In this way we can ensure that the spectrum of noise in the vicinity of the signal frequency is not enhanced at the expense of the signal due to an excessive noise contribution from the amplifier. The main considerations in choosing a suitably 'low-noise' amplifier are reviewed in Appendix 5. An associated problem is the occurrence of ground loops and pick-up in a complex experimental system which serve to degrade the signal-to-noise ratio even further. This topic is discussed further in Appendix 6 which outlines a basic stratagem for making interconnections between instruments.

2.5.3 Signal conditioning

As we have seen, there is no fundamental requirement to 'clean up' the signal prior to detection in a synchronous system. It might seem surprising, therefore, that many lock-in amplifiers are provided with an array of filters for this very purpose.

The reason is that signal conditioning is often an essential step when a high amplification factor is required to obtain the required system sensitivity. If a substantial fraction of the noise and interference were not eliminated in this way the amplifier could be driven into saturation at an undesirably low gain factor. The resulting distortion and intermodulation would then seriously degrade system performance.

Practical multiplier circuits are similarly designed to cope with a specified range of signal and noise voltages, so that a degree of signal filtering might be necessary to protect the vital detection process against overload. In either case, the main target in providing filters is often the spectrum of discrete interference components. If an electronic circuit is to saturate on noise it will most likely be the discrete frequency components occurring at large peak-to-peak values which make the greatest contribution.

A review of the main types of signal-conditioning filter is given in Appendix 4, which describes their frequency-response characteristics and catalogues their noise bandwidths. Of these, one of the most useful is a sharply tuned 'notch' filter which can be used to suppress a dominant interference component, say at mains frequency. The remaining components can then be reduced along with the out-of-band noise by the use of high- and low-pass filters as indicated in Fig. 2.14.

The attenuation of the remaining discrete frequency components will be obtainable from the amplitude responses of the filters which combine to form a band-pass response extending from f_L to f_H. In an extreme case a resonant filter tuned to the signal frequency might be used, but, as we have seen, this would introduce undesirable features which we have been so anxious to avoid in electing to use a lock-in system. The effects of introducing a tuned filter in the signal channel are discussed further in Chapter 4. For the moment it is sufficient to note that the inclusion of filters in the signal channel introduces undesirable restrictions on a signal recovery system and even the most simple high-pass or low-pass filter introduces a phase shift which must be compensated in the reference path. Fortunately, the capability of modern systems is such that filtering can be kept to a minimum in all but the most demanding of applications.

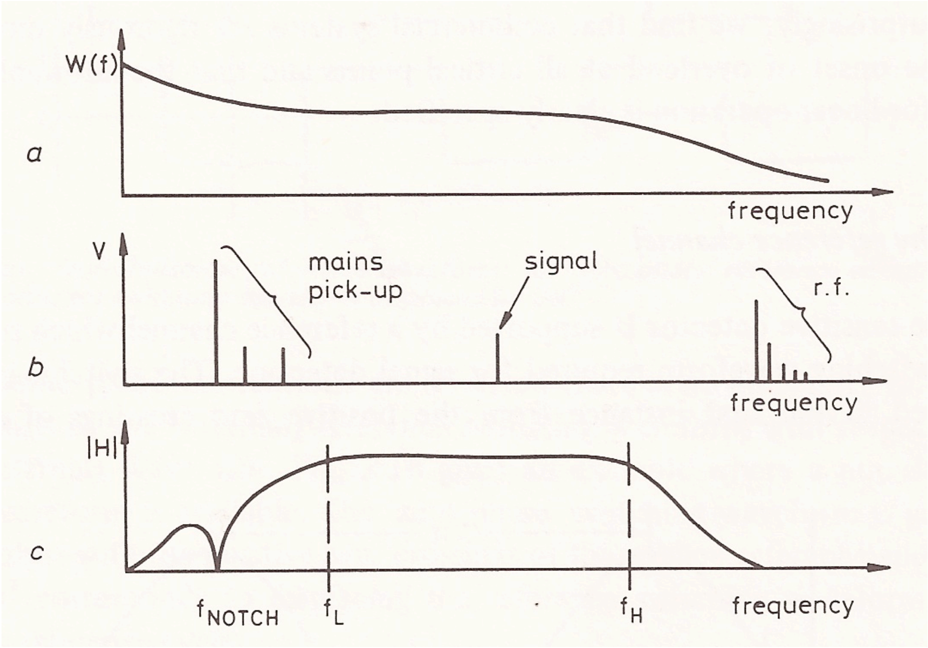


Fig 2.14 **(a) Spectrum of fundamental noise; (b) amplitude spectrum of signal and discrete interference; (c) combined amplitude-response characteristic of a mains-frequency notch filter, a high-pass filter cutting off at f_L and a low-pass filter with cut-off frequency f_H**

2.5.4 The multiplier

The basic principles of synchronous detection have been established in terms of an 'ideal' multiplier which can maintain its performance under all applied levels of signals and noise. If the advantages cited in Section 2.4 are to be realized, therefore, practical multipliers must possess exceptional dynamic range. Otherwise, departures from linearity under conditions of even moderate signal-to-noise ratio will give rise to intermodulation effects and the eventual suppression of the desired response.

A major objective in the past has been to identify those circuit configurations which offer the widest dynamic range in order to maximize the signal recovery capability of practical systems. These efforts have resulted in the almost universal adoption of the switching multiplier or phase-sensitive detector which is found at the heart of all lock-in recovery systems. The essential difference in this case is that the multiplier is now driven by a squarewave switching waveform that is precisely synchronized to the applied reference waveform as described in the next section.

The operating characteristics of phase-sensitive detectors are dealt with in some detail in the next chapter and it is shown there how the properties of a synchronous detection system are modified by the inclusion of a switching multiplier. On the basis of our discussions so far we can emphasize the need for good dynamic performance and note that linearity at the point of detection must be supported by linearity elsewhere in the system. As we have seen, this applies mainly to the signal-channel amplifier, which is expected to handle small signals accompanied by relatively large noise and interference voltages. Not surprisingly, we find that commercial systems are rigorously monitored to detect the onset of overload at all critical points and that the allowable range of voltages for linear operation is clearly specified.

2.5.5 The reference channel

The phase-sensitive detector is supported by a reference channel which supplies the precise switching waveform required for signal detection. The switching waveform is triggered in the first instance from the positive zero crossings of an applied reference waveform as shown in Fig. 2.15, and is always arranged to be precisely symmetrical irrespective of the symmetry of the original reference input. The displacement ϕ_R is usually introduced by means of a calibrated phase control. In a 'broadband' reference channel this phase-shift can be maintained to a high degree of precision over a wide range of reference frequencies and the symmetry of the switching waveform is rigidly controlled.

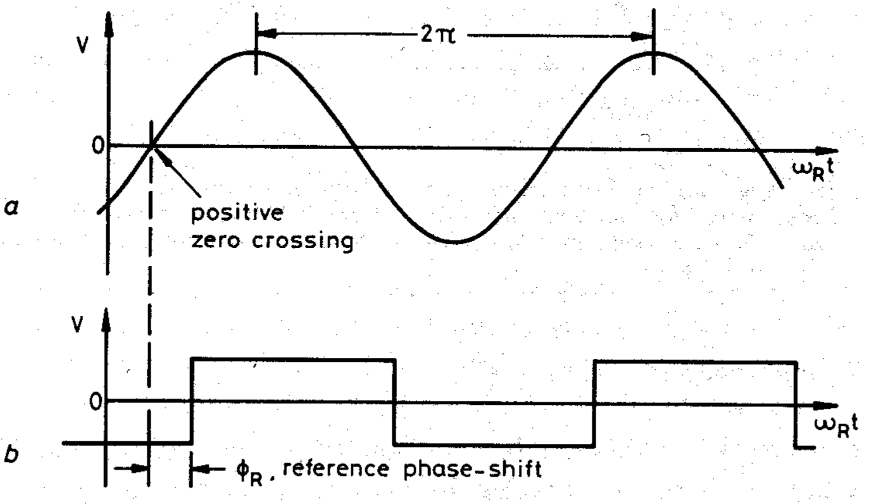


Fig. 2.15 (a) Reference input; (b) Symmetrical switching waveform triggered from (a) and phase-shifted ϕ_R

Provided the phase conventions are observed there is no ambiguity in defining the phase shift of the (internal) reference switching waveform with respect to any applied (external) waveform. Fig. 2.16 gives an example where a non-sinusoidal reference waveform is available. The 'zero phase' switching waveform is generated in synchronism with the positive zero crossings of the applied reference and a phase shift of 90° corresponds to displacing the reference

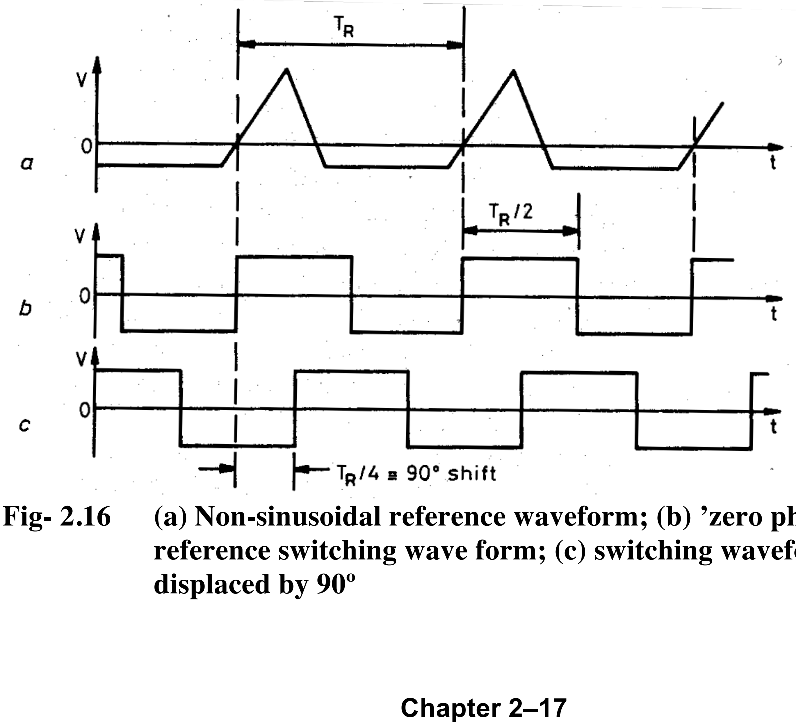


Fig- 2.16 (a) Non-sinusoidal reference waveform; (b) 'zero phase' reference switching wave form; (c) switching waveform displaced by 90°

switching waveform by one quarter of a reference cycle. Phase-shift controls are usually provided in the form of a continuously variable adjustment covering the range 0°–100°, together with 90° and 180° pushbuttons for quadrant selection.

In addition, even the most basic lock-in amplifier is usually provided with a '2*f*' facility whereby the system becomes synchronized at the second harmonic of the applied reference waveform. The 2*f* mode is normally made available by pushbutton selection up to reference frequencies of one half the maximum value allowed in normal operation

2.5.6 The low-pass filter

The low-pass filter provided with the majority of lock-in amplifiers is based on either a single-section or two-section *RC* filter giving a roll-off of 6 dB or 12 dB per octave beyond the cut-off frequency. A range of bandwidths is supplied and the range switch is almost invariably presented as a *time-constant* control, that is in terms of the characteristic time $T_0 = RC$ of the filter. The characteristics of these filters are presented in Appendix 4. In calculations involving noise, the noise bandwidth of the filter must be used, the appropriate values being $1/(4T_0)$ Hz for a single section filter and $1/(8T_0)$ Hz for a two-section filter. The smaller value of noise bandwidth for the filter with sharper roll-off reminds us that it is more effective in suppressing noise at frequencies beyond cut-off. Also, it should not be forgotten that the signal bandwidth of a two-section filter is smaller than that of the corresponding single-section filter; the bandwidth $B = 1/(2\pi T_0)$ gives the –6 dB frequency in the first case and the –3 dB frequency in the second.

In the majority of experiments it is advisable to use the smallest possible value of time constant to ensure that the response of the lock-in amplifier is not too sluggish. The *settling time* of the filter is important in this context since it takes account of the recovery of the filter following a 'step' change in signal level. This could also apply to an increment in sensitivity caused by range switching or to switching a phase increment in the reference channel. The step responses of the two filter types are shown in Fig. 2.17.

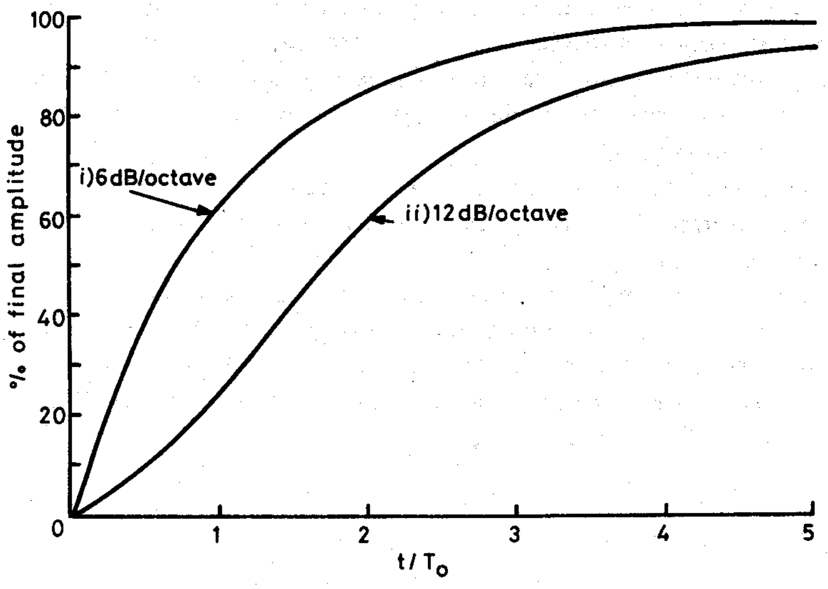


Fig. 2.17 Step responses of R-C low-pass filters.

We find that a settling time of four time constants brings the output of the single-section filter to within 2% of its final value while the two-section filter output is in error by 10%. For a given value of time constant, therefore, we must be prepared to trade noise rejection for the ability to follow a changing signal. Clearly, for time constants of 10 s or greater, the settling time can become a major factor.

In this respect it should be noted that both the settling time and the *mean-square* fluctuation due to residual noise in the filter output are reduced in proportion to the time constant setting. Using the results of Appendix 3 with the appropriate value of noise bandwidth for the low-pass filter, we obtain improvement factors:

$I_1 = 4B_1T_0$, 6dB/octave filter

$I_2 = 8B_1T_0$, 12dB/octave filter

where B_1 is the input noise bandwidth.

The improvement factors refer to the mean-square fluctuations attending the signal before and after detection. Since, in most practical cases, the r.m.s value of a signal is of interest, it is usual to consider the *r.m.s.* value of the output fluctuation. Thus, increasing the time constant by a factor x increases the settling time by the same amount but brings a reduction of only $\sqrt{x}$ in the r.m.s. value of the noise measured on an output meter or chart recorder.

2.6 Signal recovery ‘capability’

The word ‘capability’ is used throughout the literature supplied by lock-in amplifier manufacturers. ‘Capable’ systems are those which can withstand the effect of massive levels of noise while maintaining a linear response to a synchronous signal. When dealing with theoretical models we can assume infinite ‘capability’. Also we find that the signal-to-noise improvement obtained through synchronous detection is limited only by the minimum allowable output bandwidth. As we have seen, this depends upon the nature of the recovered information signal: if the signal parameters are effectively fixed for the duration of a measurement the output bandwidth can be made very small and the potential for signal-to-noise improvement is correspondingly high.

In a practical experiment the improvement factor must bridge the gap between the input signal-to-noise ratio, SNR_I and the desired output signal-to-noise ratio, SNR_o. SNR_I must, of course, be restricted in order to preserve linear operation while the improvement factor should be achievable *within the time set aside for the experiment.* If an arbitrarily long time is not available for measurement the desired improvement factor might not be attainable even though the input signal-to-noise ratio lies within the ‘capability’ of the lock-in system. The following discussion will provide a valuable perspective on system performance when we consider aspects of specification at a later stage.

Let us begin by supposing that we have a sinusoidal signal appearing against a background of white noise with bandwidth B_I. In order to maximize the prospects of recovery we shall assume that the amplitude and phase of the signal are fixed. We can then choose whatever output bandwidth we wish without being constrained to respond to variations in signal level.

The improvement factor brought about by synchronous detection is B_1/B_o, where B_o is the noise bandwidth of the output filter. The output signal-to-noise ratio is therefore

$$SNR_{\text{o}} = \frac{B_1}{B_{\text{o}}} \times SNR_1$$

The normal experimental procedure is to select B_0 to give an acceptable value of SNR_0 for a given value of input signal-to-noise ratio. The problem arises when the required value of B_0 is so small that it results in an excessively sluggish response and inconveniently long settling time in the output circuit. Under these circumstances we must either accept a lower value for SNR_0 or conclude reluctantly that the input signal-to-noise ratio is too low to permit measurement to the required precision in the available time.

At a time constant of 100 s the response of a synchronous detector is such that a settling time approaching 10 minutes is required to recover from the slightest disturbance in the signal. In many circumstances therefore, an output signal-to-noise ratio of about 1:1 would represent a reasonable limit to detection in view of the length of time required to average the response from the residual noise background. Let us therefore limit the maximum time constant to 10 s and demand that the output indication appears with a signal-to-noise ratio of about 10:1 (that is about 3:1 in terms of r.m.s. fluctuation). Putting $B_0 = 1/(8T_0)$ with $T_0 = 10$ s, we obtain the following bound on the input signal-to-noise ratio:

$$SNR_1 > 1/(8B_1) \quad (B_1 \text{ given in hertz})$$

For detection of audio-frequency signals in an input bandwidth of about 10 kHz we find that for reasonable precision at a moderate observation time the input signal- to-noise ratio must be better than 1/80 000 (−50 dB). If the noise appears in a wider bandwidth (with a correspondingly lower noise density in the vicinity of the signal) or if a larger time constant can be tolerated, the limit could be relaxed to about −60 dB. This corresponds to measuring the amplitude of a 100 μV signal in a noise background of 100 mV r.m.s. Even at this level, the achievable performance falls well short of the popular notion of recovering signals from 100 dB of noise. What is important here is that we have reached this conclusion without referring to linearity or the ability of electronic circuits to function correctly with noisy inputs. It would appear that a lock-in amplifier capable of handling signals in the presence of 60 dB of wideband noise would be able to fulfil all but the most demanding of measurement tasks. Unfortunately, the simple calculations given here refer only to disturbance by *white* noise, whereas this idealized situation is rarely observed in practice. It can safely be assumed that the most spectacular claims in respect of lock-in amplifier ‘capability’ refer to disturbance by narrowband noise or to large-scale interference components appearing at frequencies well removed from any transmission ‘windows’ associated with the phase-sensitive detector.

CHAPTER 3

Phase-sensitive detectors

3.1 Introduction

The wide dynamic range of modern lock-in amplifiers results from the use of a switching multiplier as a synchronous detector. The adoption of a switching circuit leads to a degree of precision which cannot be matched by 'true' multipliers and, moreover, has the added advantage of operational simplicity. This is evident from the block diagram given in Fig. 3.1.

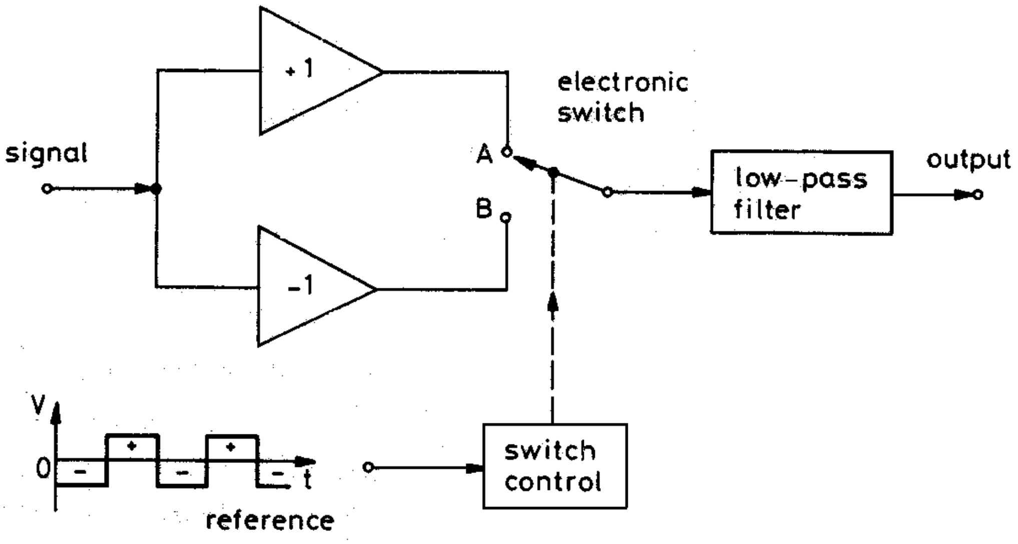


Fig 3.1 Phase-sensitive detector: principles of operation

We shall find it convenient at this introductory stage to maintain a distinction between the switching network and the low-pass filter in the output circuit. For the most part, however, we will conform with the usual practice of referring to the entire combination as a phase-sensitive detector[1]. Thus, when we come to consider the specification of phase-sensitive detectors it will be the behaviour of the switch/filter combination which is of interest.

In line with comments made in the Preface we shall be concentrating on the systems aspects of phase-sensitive detectors rather than on detailed circuit implementations. Those with an interest in circuit techniques are recommended to read the paper by Carter and Faulkner [1] which contains several circuit configurations of phase-sensitive detectors and examines sources of error in practical designs. This paper is one of the very few published accounts where the level of treatment is appropriate to the performance of commercial systems.

[1] The alternative forms, 'mixer' and 'demodulator', are widely used.

3.2 Principles of operation

The key to the operation of the phase-sensitive detector is the two-state switch which is controlled electronically from the reference voltage. The switch changes position between points A and B as the reference changes polarity. This action gives a systematic change of gain between +1 and −1 in the signal path.

We shall be considering the classic operation in which the phase-sensitive detector spends equal times in its two states, an arrangement which gives rise to the well known waveforms of Fig. 3.2

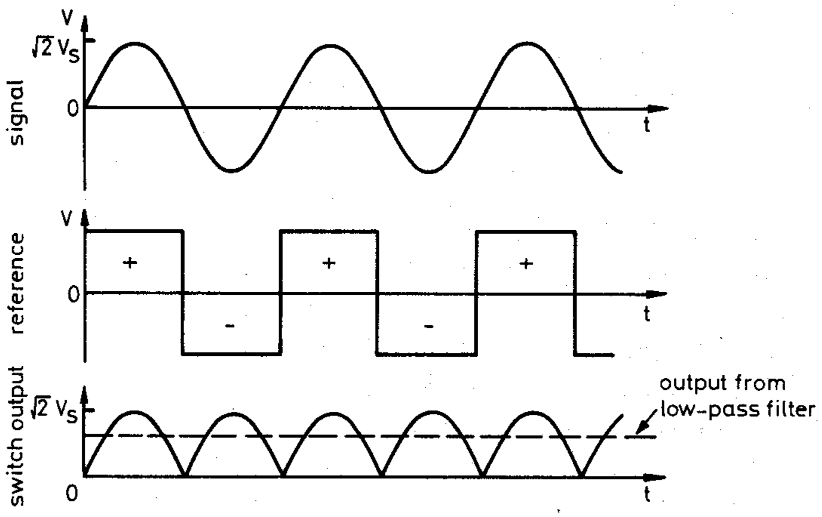


Fig 3.2 Waveforms in a phase-sensitive detector operating with an in-phase sinewave signal

In Fig 3.2, the sinewave signal and applied reference are precisely in phase. The reference changes polarity in a symmetrical fashion, in step with the signal, and so causes full-wave rectification of the signal at the switch output.

The output of the switch is then applied to the low-pass filter which smooths out the ripple component and delivers a d.c. voltage which is proportional to the amplitude of the signal.

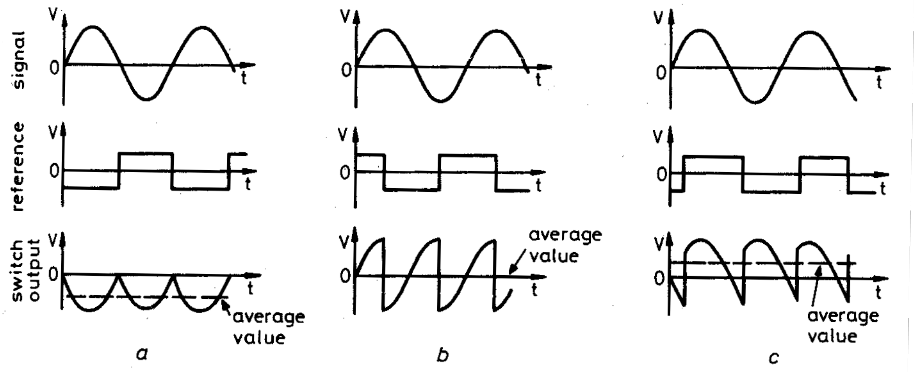


Fig. 3.3 Waveforms in a phase-sensitive detector for different phase conditions: (a) ϕ = 180°; (b) ϕ = 90°; (c) arbitrary phase

In most applications the signal and reference will not arrive at the phase-sensitive detector exactly in phase; hence, Fig. 3.3(a) which shows the effect of a phase reversal, giving a negative d.c. level, and Fig. 3.3(b) which shows the output when the phase displacement is 90°. In this case the output from the switching stage is a bipolar waveform which averages to zero and gives no net response from the low-pass filter. Finally, Fig. 3.3(c) shows the output at some intermediate value of phase shift, giving a d.c. level somewhere between the positive and negative maxima obtained with $\phi = 0°$ and $\phi = 180°$.

To determine the exact relationship between signal and reference, we recognize that the switching operation is equivalent to multiplication of a signal by a squarewave taking values of +1 and –1. We can therefore use the ideal multiplier model shown in fig. 3.4 where the reference waveform has the Fourier series representation

$$r(t) = \frac{4}{\pi}[\cos(\omega_R t + \phi_R) - \frac{1}{3}\cos 3(\omega_R t + \phi_R) + \frac{1}{5}\cos 5(\omega_R t + \phi_R) - ...]$$

Fig 3.4(b) shows the switch output for the case where the signal and reference are asynchronous. The switched signal has zero average value and its general form can be obtained by forming the product:

$$v_p(t) = r(t)s(t)$$

where

$$s(t) = \sqrt{2}V_s \cos(\omega_s t + \phi_s)$$

Multiplying term by term and separating into 'sum' and 'difference' components we obtain

$$v_p(t) = \frac{2\sqrt{2}V_s}{\pi}[\cos(\omega_R t \pm \omega_s t + \phi_R \pm \phi_s) - \frac{1}{3}\cos(3\omega_R t \pm \omega_s t + 3\phi_R \pm \phi_s)$$
$$+ \frac{1}{5}\cos(5\omega_R t \pm \omega_s t + 5\phi_R \pm \phi_s) - ...]$$

The development is thus similar to the case of the 'ideal' synchronous detector discussed in Section 2.4. For synchronous operation we put $\omega_s = \omega_R$ and, as before, we assume that the low-pass filter cuts off well below the reference frequency. This eliminates multiplier products at frequencies $2\omega_R$, $4\omega_R$, $6\omega_R$, etc. from the final output which contains only the phase-sensitive d.c. component:

$$V_o = \frac{2\sqrt{2}}{\pi} V_s A_L(0)\cos(\phi_R - \phi_s)$$

Here, $A_L(0)$ gives the magnitude of the filter response at zero frequency.

Apart from a scaling factor, the response of a phase-sensitive detector to a synchronous sinewave is identical to that of an ideal multiplier-detector operating with a sinewave reference. The essential difference in this case is that the phase-sensitive detector will also give a phase-sensitive d.c. output in response to signals at frequencies $3\omega_R$, $5\omega_R$ etc. This is shown in Fig. 3.5 for a sinewave signal at the third harmonic of the reference frequency. The relative sensitivity of the detection system at these additional frequencies is 1/3, 1/5 and so on, corresponding to the relative magnitudes of the reference Fourier components.

A detection system with this property is said to be harmonically responding. We shall find that there is minimal practical advantage in having a harmonically responding system; indeed, the additional responses are frequently dismissed as 'anomalous' or 'spurious'. In later chapters, we shall be investigating ways in which the harmonic responses can be suppressed by improved system design. It is significant, however, that all these improved systems continue to rely on a switching phase-sensitive detector to provide the dynamic range essential for signal recovery operation.

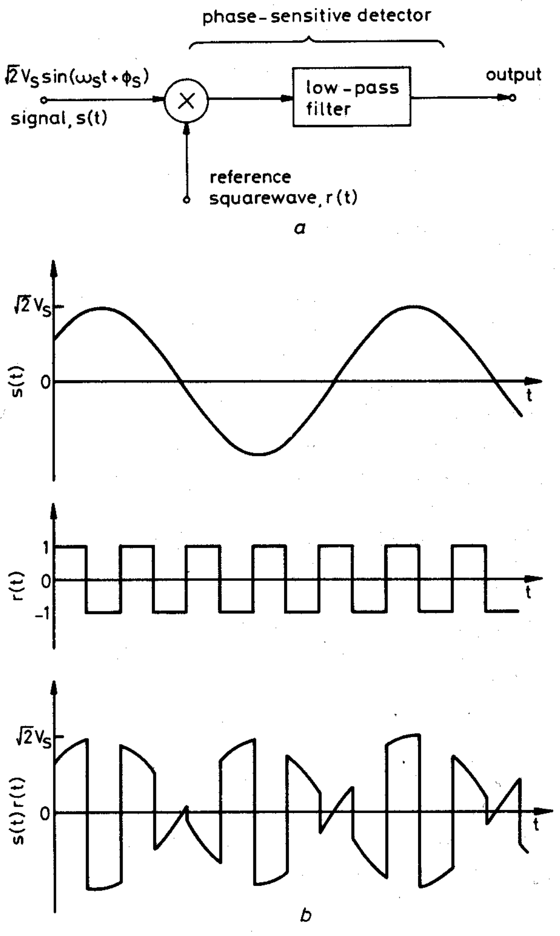


Fig 3.4 (a) Ideal multiplier model for a phase-sensitive detector; (b) output from switch, before low-pass filter, for a sinusoidal signal with $\omega_s < \omega_R$

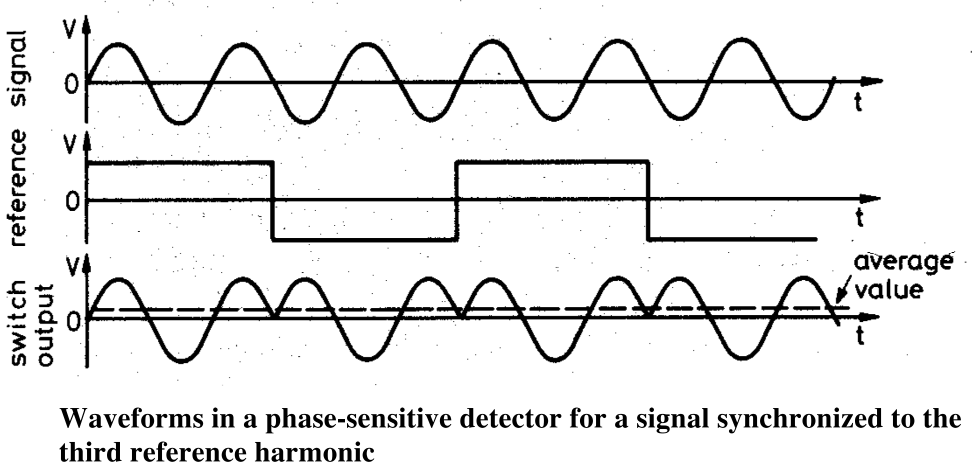


Fig. 3.5 Waveforms in a phase-sensitive detector for a signal synchronized to the third reference harmonic

3.3 Harmonic transmission windows

The idea of a transmission 'window' was introduced in Chapter 2 as a convenient way of representing the response of a synchronous detector to signals with frequencies close to the reference frequency. If we now consider the effect of using a squarewave reference waveform we find that the additional harmonic components lead to the set of transmission windows illustrated in Fig. 3.6.

The transmission windows are centered on the odd harmonics of the reference frequency and the maximum magnitude of each window is weighted by the magnitude of its associated reference Fourier component. Before a signal can produce a response at the output of the phase-sensitive detector it must lie within one of the transmission windows. In order to produce a 'true' d.c. response, a signal must be synchronous with one or more of the reference Fourier components as was shown in Fig. 3.5. Otherwise, the output will appear as an alternating 'beat note' at the difference frequency between the signal and the centre frequency of the transmission window.

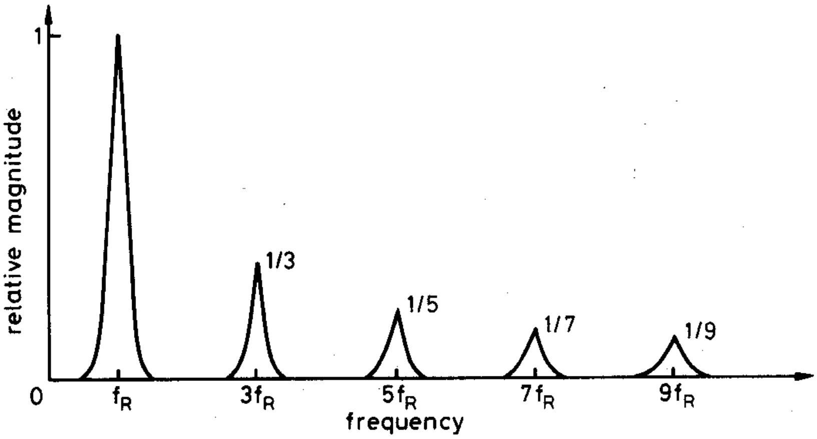


Fig. 3.6 The first five harmonic transmission windows of a switching phase-sensitive detector

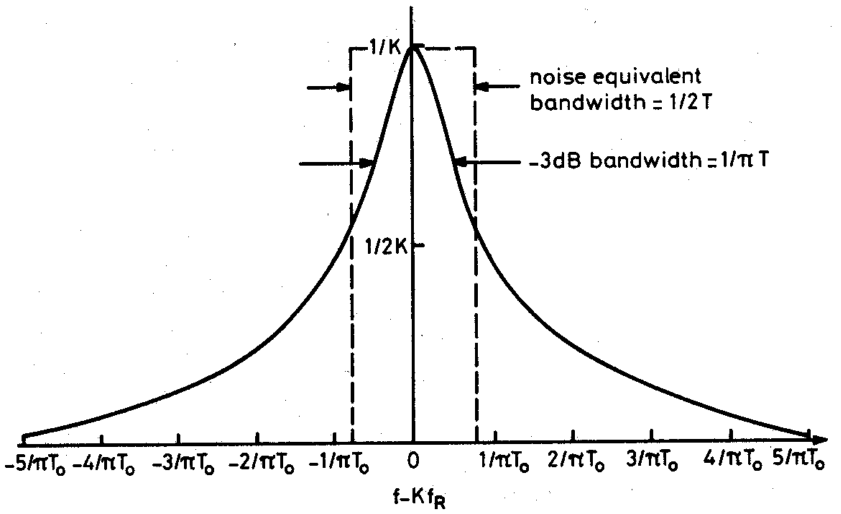


Fig. 3.7 The *K*th transmission window of a phase-sensitive detector corresponding to a 6 dB/octave low-pass filter with time constant T_0
The amplitude response of the noise equivalent filter is shown as a dashed characteristic

Fig. 3.7 shows the form of one of these windows obtained when a low-pass filter having a roll-off of 6dB/octave is used to follow the switching stage. It is centered on the harmonic frequency Kf_R where K takes values 1, 3, 5 etc. The –3 dB and noise bandwidth are always independent of the centre frequency.

The practical importance of the transmission windows stems from the fact that they represent frequency regions where the phase-sensitive detector is susceptible to large-scale discrete interference components. 'Ideal' synchronous detectors operating with a *sinewave* reference are, of course, relatively immune to such components unless they originate very close to the reference frequency. The additional susceptibility of switching phase-sensitive detectors to interference above the reference frequency, together with possible measurement ambiguities resulting from a harmonically responding system, can be a severe limitation in some circumstances. Some of the problems associated with the harmonic responses are discussed in Chapter 6.

3.4 Noise bandwidth of phase-sensitive detector

When a signal in broadband noise is measured using a phase-sensitive detector, the post-detection noise output is increased over the level which would be calculated for a multiplier-detector using a sinewave reference. This is because of the effect of the additional transmission windows which are able to 'leak' noises through to the final output.

The total effect for a band of white noise can be calculated with the help of Figs. 3.7 and 3.8.

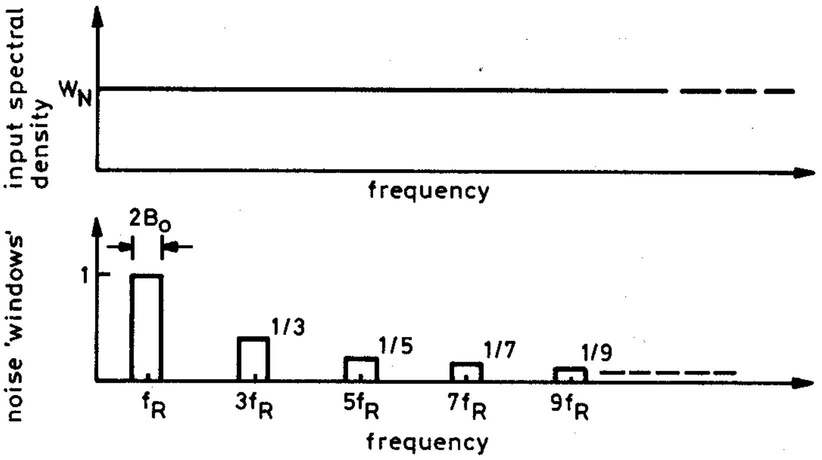


Fig. 3.8 Calculation of noise output from a phase-sensitive detector for a white-noise output. Each transmission window is replaced by a rectangular noise equivalent window with Bandwidth $2B_o$

The noise bandwidth of each transmission window is given by $2B_0$ where B_0 is the noise bandwidth of the low-pass filter. The noise outputs due to components which fall within each window will have mean-square values proportional to W_N B_0, where W_N is the spectral density of the white noise.

The mean-square value of the noise transmitted by the fundamental window can be written as

$$\overline{n_1^2} = a_N W_N B_0$$

where the constant a_N takes account of the scaling factor of the phase-sensitive detector and any associated amplifiers (see Appendix 3).

To calculate the noise output from any other window we must include a weighting factor which depends on the magnitude of the associated Fourier component. The noise output from a window from the *K*th harmonic is therefore

$$\overline{n_{\mathrm{K}}^2} = a_{\mathrm{N}} W_{\mathrm{N}} B_0 / K^2, \quad K = 3, 5, 7 \ldots$$

There is no coherence in the noise contributions from the individual transmission windows. This means that the total noise output can be obtained directly by summing mean-square values. We obtain

$$\overline{n_{\mathrm{T}}^2} = a_{\mathrm{N}} W_{\mathrm{N}} B_0 \left(1 + 1/9 + 1/25 + \ldots\right)$$

Noting that

$$\sum_{n=0}^{\infty} 1/(2n+1)^2 = \pi^2/8$$

We obtain a total mean-square noise voltage:

$$\overline{n_{\mathrm{T}}^2} = \frac{\pi^2}{8} a_{\mathrm{N}} W_{\mathrm{N}} B_0 \cong 1.23 \times a_{\mathrm{N}} W_{\mathrm{N}} B_0$$

This result shows an increase of only 23% over the mean-square noise transmitted by the fundamental window and represents an increase of around 11% in the r.m.s. output fluctuation. We can conclude that for *white* noise the effect of noise in the higher-order windows is negligible in practical terms. However, for exact calculations, the standard noise reduction formula derived in Appendix 3 for an ideal synchronous detector can be modified to give a less optimistic improvement factor.

$$\frac{SNR_{\mathrm{I}}}{SNR_{\mathrm{o}}} = \frac{B_{\mathrm{I}}}{B_0} \times 8/\pi^2$$

Here, B_{I} is the input noise bandwidth and $\pi^2 B_0/8$ is the exact noise bandwidth of the phase-sensitive detector.

3.5 Non-sinusoidal signals

3.5.1 Introduction

So far, the properties of synchronous detection systems have been discussed solely in terms of sinusoidal signals. We shall now extend the discussion to include all types of periodic signal and pay particular attention to some waveforms which have special practical importance.

We shall assume that a synchronous reference squarewave is available for detection of the signal. In practice, this would usually be generated by a reference unit of the type described in Section 2.5.5, triggered in the first instance by the positive zero crossings of an applied reference voltage.

A problem which is common to many measurements involving phase-sensitive detectors is to introduce the reference switching waveform in the correct phase to maximize the d.c. response to a given signal. This must often be achieved under noisy conditions for a signal where the d.c. response varies with the reference phase setting in a complicated way.

In the special case of sinewave signals we have seen that the switching phase-sensitive detector response follows a $\cos\phi$ law, where ϕ represents the relative phase of the signal and switching waveform measured at the phase-sensitive detector. The following procedure can therefore be adopted to adjust the phase of the detection system, starting from an arbitrary initial phase condition. First, the phase-sensitive detector output is ‘nulled’ by adjusting the reference phase. This enables a quadrature condition at the phase-sensitive detector. The desired ‘in-phase’ condition is then obtained by shifting the set phase by 90°. For a sinewave signal the ‘in-phase’ reference setting arrived at through this procedure maximizes the response of the detection system. In practice, the ‘null’ point can be determined with high accuracy even when the signal is extremely noisy and the procedure defines a precise and repeatable reference condition for subsequent measurements.

This approach to setting the phase of a phase-sensitive detector is widely applied to periodic signals *of all shapes and forms.* This procedure is often justified on the assumption that a phase displacement of 90° from the null-point automatically maximizes the response of the phase sensitive detector to *all types* of periodic signal. Unfortunately, this is not necessarily the case as we shall see in the following section.

3.5.2 General considerations

Let us take a periodic signal with fundamental frequency ω_0 having a Fourier description

$$s(t) = \sum_{n=1}^{\infty} \alpha_{\mathrm{n}} \cos n\omega_0 t + \sum_{n=1}^{\infty} \beta_{\mathrm{n}} \sin n\omega_0 t$$

In describing $s(t)$ we have omitted a d.c. component since, in practice, signals are invariably a.c. coupled prior to phase-sensitive detection.

The signal is switched by a synchronous squarewave with the Fourier series:

$$r(t) = \frac{4}{\pi} \sum_{n=0}^{\infty} (-1)^n \frac{\cos\left[(2n+1)(\omega_0 t + \phi_{\mathrm{R}})\right]}{(2n+1)}$$

where ϕ_{R} is the phase of the reference unit defined with respect to an externally applied reference waveform.

The d.c. response of the phase-sensitive detector is obtained by forming the product $s(t)r(t)$ and extracting the difference-frequency terms at zero frequency. Since the reference comprises only odd harmonic components it is only the odd harmonics of the signal which enter the calculation. We obtain an output voltage:[·]

$$V_{\mathrm{o}} = \frac{2}{\pi} \sum_{n=0}^{\infty} (-1)^n \alpha_{2\mathrm{n}+1} \frac{\cos(2n+1)\phi_{\mathrm{R}}}{(2n+1)} - \frac{2}{\pi} \sum_{n=0}^{\infty} (-1)^n \beta_{2\mathrm{n}+1} \frac{\sin(2n+1)\phi_{\mathrm{R}}}{(2n+1)}$$

In general, for any set of α_{n} and β_{n}, there will be a value of ϕ_{R} which leads to $V_{\mathrm{o}} = 0$ and which therefore corresponds to the ‘null’ point referred to earlier. We shall denote this value by ϕ_{q}.

At some other value of ϕ_{R}, denoted by ϕ_{i}, the magnitude of V_{o} will be a maximum. This value is obtained by solving

$$\mathrm{d}V_{\mathrm{o}} / \mathrm{d}\phi_{\mathrm{R}} = 0$$

where

$$\mathrm{d}V_{\mathrm{o}} / \mathrm{d}\phi_{\mathrm{R}} = \frac{-2}{\pi} \sum_{n+0}^{\infty} (-1)^n \alpha_{2\mathrm{n}+1} \sin(2n+1)\phi_{\mathrm{R}} - \frac{2}{\pi} \sum_{n+0}^{\infty} (-1)^n \beta_{2\mathrm{n}+1} \cos(2n+1)\phi_{\mathrm{R}}$$

[·] We assume that $A_{\mathrm{L}}(0) = 1$ in this case

We shall not attempt a general solution in either case. It is sufficient to note that if the value of ϕ_q is determined, either by calculation or experiment, the value of $\phi_q \pm 90^\circ$ will not satisfy the condition for maximum response unless the signal waveform is subject to certain constraints. These will now be investigated.

3.5.3 Symmetrical periodic signals

Many non-sinusoidal waveforms of practical importance such as squarewaves, triangle and rectangular pulse waveforms, possess a high degree of symmetry. In all these cases the time origin can be chosen to give a waveform which is either an even or an odd function of time. The corresponding Fourier series then consists only of cosine terms in the first case ($\beta_n = 0$) and sine terms in the second ($\alpha_n = 0$). Inspection of the expressions for V_o and $dV_o/d\phi_R$ shows that for waveforms with this essential symmetry we always arrive at values of ϕ_q and ϕ_i which satisfy

$$\phi_i = \phi_q \pm 90^\circ$$

Relationships between the reference switching waveform and a number of non-sinusoidal but symmetrical signals are shown in Fig. 3.9. We can make the following observations about the types of signal chosen.

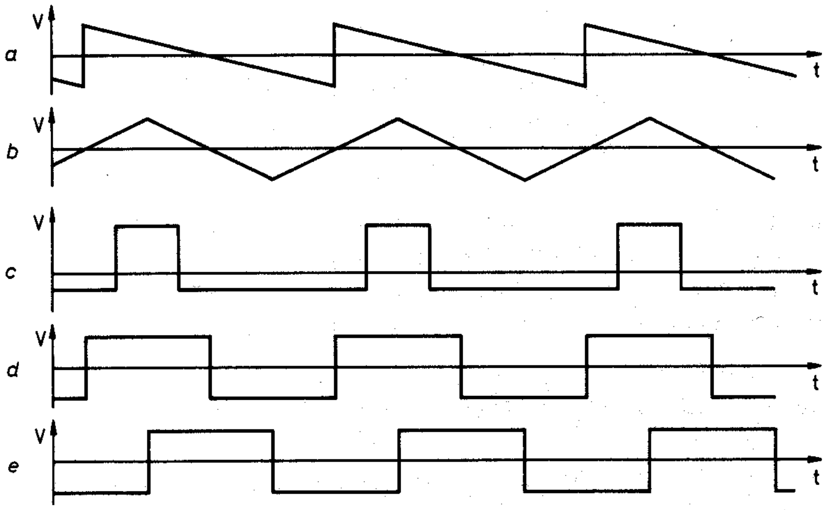


Fig. 3.9 (a) - (c) Symmetrical waveforms; (d), (e) reference switching waveforms introduced at $\phi_R = \phi_i$ and $\phi_R = \phi_q$ respectively

i) Setting the reference channel phase by first nulling the phase-sensitive detector output and then shifting the phase by 90° automatically maximizes the d.c. output of the phase-sensitive detector. We shall refer to this as the 'null-shift' procedure.

ii) The conditions for zero output and maximum response correspond to bringing the Fourier components of the reference switching waveform first in quadrature and then in phase with the corresponding components of the signal.

iii) The null-shift procedure summarized in (i) does not depend on absolute phase calibration. For recovery work, a variable phase-shifter and a calibrated phase increment of 90° are sufficient to achieve optimum detection of symmetrical signals.

iv) When the reference channel is calibrated according to a convention such as that described in Section 2.5.5 the null-shift procedure provides a basis for measuring the phase-shift of a sinewave signal relative to that of an applied reference voltage. It is now apparent that the validity of such a measurement will be in doubt if the sinewave signal is subject to harmonic distortion. This aspect is discussed further in Chapter 4 in relation to precision phase measurement.

3.5.4 Asymmetrical periodic signals

In the present context, 'asymmetrical' signals are those for which it is impossible to choose a time origin such that $s(t) = \pm s(-t)$. In practice signals in this category are most likely to occur when a normally symmetrical signal is subject to linear filtering giving rise to waveforms such as those shown in Fig 3.10. The waveforms in Figs. 3.10(b) and (c) in particular will often be encountered in experiments using 'chopped' excitation where a squarewave signal has been transmitted by a low-pass or a high-pass signal conditioning filter. To this general class of signals we can add the important case of sinewaves subjected to arbitrary harmonic distortion.

In all these cases it is possible to determine a value of reference phase-shift which nulls the phase-sensitive detector output. In general, however, changing the phase by 90° from the null point fails to maximize the phase-sensitive detector output. Fortunately, in all but the most extreme cases, the resulting output is normally within 10% to 20% of its maximum possible value. In addition, the null output obtained for $\phi_R = \phi_q$ provides an ideal datum point for setting the phase which can be reproduced on future occasions even under noisy signal conditions.

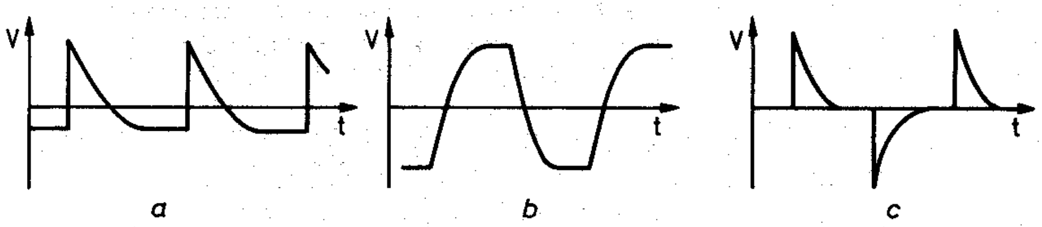


Fig. 3.10 Examples of asymmetrical waveforms

In view of these remarks, the procedure of nulling followed by introducing a phase offset of 90° has much in its favour even when the resulting response is less than optimum. As a result, this approach is almost universally adopted in signal recovery work where it is sufficient to obtain a consistent measure of signal magnitude in the presence of noise. It is, nevertheless, worth bearing the following points in mind.

i) With symmetrical signals the derivative $dV_o/d\phi_R$ is zero for $\phi_R = \phi_q \pm 90°$. The phase-sensitive detector output is then maximized and becomes first-order independent of small phase adjustments in the reference channel. This advantage is lost when the null-shift procedure is applied to asymmetrical signals.

ii) In the two-phase lock-in systems described in Chapter 5, an automatic phase control loop is sometimes used to determine the null point $\phi_R = \phi_q$. Detection is subsequently carried out at relative phase $\phi_R = \phi_q - 90°$. In the light of our discussions we must expect that the overall response of such a system to an asymmetrical signal will have less than its maximum value.

iii) We can conclude that the only way in which the null-shift procedure can be applied to give a maximum response to all forms of periodic signal is to use a detection system with 'fundamental-only' response. Some methods of suppressing the harmonic responses of phase-sensitive detectors are discussed in Chapters 8 and 9.

3.5.5 Squarewave signals: a special case

Fig. 3.11 illustrates the waveforms which result when a squarewave of V_S volts peak is applied to a switching phase-sensitive detector with an in-phase reference. The switch output is a constant level of V_S volts, the 'ripple' component which is characteristic of operation with sinewaves being absent in this case (compare Fig. 3.2).

The net response is larger than that obtained with a sinewave of the same r.m.s. value because there is now a contribution from each of the fully synchronous Fourier components of the signal and reference voltages.

Thus, if we begin with a sinewave signal and then convert it to a squarewave by adding Fourier components of the correct magnitude and phase, the d.c. output voltage will increase by a factor

$$\left[1+1/9+1/25+...\right]=\pi^2/8$$

which, from Section 3.4, is the factor by which the *mean-square* noise is increased when white noise is 'leaked' through the harmonic transmission windows. The effect of adding the additional Fourier components to the sinewave signal is, therefore, to boost the signal indication at the expense of the noise. The response is increased by 23%, which is more than compensates for the 11% increase in the r.m.s. noise fluctuation. The reason is, of course, that the signal contributions add coherently while the individual noise outputs must be added in a mean-square sense.

In terms of an ideal multiplier model we could reach similar conclusions for any signal waveform which is perfectly 'matched' to the reference. In a wider sense, this topic belongs to the realm of matched filter theory which aims to devise schemes for the optimum detection of signals in white noise.There are more complex lock-in systems to be discussed in Chapter 9 which are capable of approaching this type of operation; however, as in the case of squarewave signal and reference, the noise benefits are often marginal and usually outweighed by benefits in other areas.

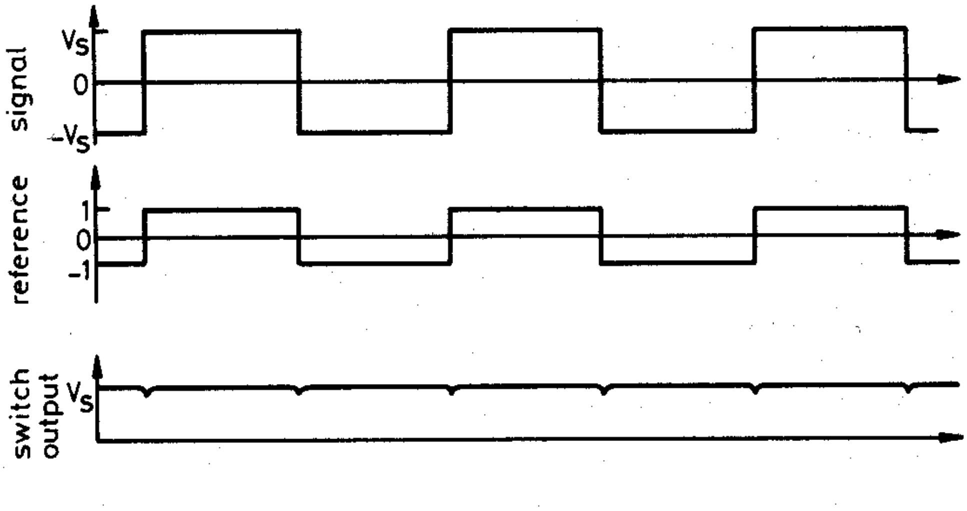


Fig. 3.11 Waveforms in a phase-sensitive detector with in-phase squarewave signal

For example, when operating a switching phase-sensitive detector with an in-phase squarewave signal, the principal practical advantage arises because of the absence of components at $2f_R$, $4f_R$ and so on in the output of the multiplier. When using sinewave signals, the time constant of the output signal should always be sufficiently long to suppress these 'ripple' components, irrespective of whether the signal is noisy or clean. This can often lead to inconveniently long response times when measuring with low-frequency signals (say, below 10 Hz). When squarewave signals are used in this regime it may be possible to relax requirements on the low-pass filter and so obtain results more quickly. At the same time, the effects of opening wide the transmission windows should be taken into consideration. As we saw in Section 3.3 the susceptibility to interference is then greatly increased even when wideband noise is not identified as a major problem.

Finally, let us examine the phase relationships in squarewave operation using Fig. 3.12, which gives us the waveforms at an arbitrary phase-shift ϕ. The phase dependence could be determined by multiplying together the Fourier series for signal and reference and extracting the low-frequency output, but it is more convenient to deduce the form of the dependence directly from Fig. 3.12(a). At arbitrary phase, a low-pass filter is obviously necessary to 'smooth' the products of multiplication. The d.c. component remaining in the final output is then

$$V_o = V_s[1-\phi/90], \ 0 \le \phi \le 90°$$

with ϕ expressed in degrees.

In lock-in amplifier systems the phase shift introduced in the reference channel can be varied in the full range 0° – 360°, which gives rise to a characteristic piecewise-linear variation in the d.c. output as drawn in Fig. 3.12(b). The classic response to a synchronous sinewave is also shown for comparison, to remind us how the handling characteristics of a phase-sensitive detector can change with different types of signal.

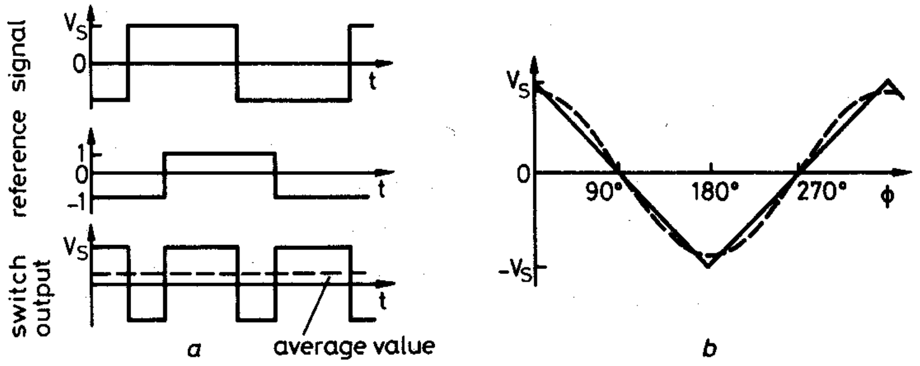


Fig. 3.12 **(a) Waveforms in a phase-sensitive detector with a squarewave signal and reference at arbitrary phase shift; (b) phase response for a squarewave signal. The $\cos\phi$ dependence is shown for a sinewave signal with the same r.m.s. value**

3.6 Phase-sensitive detector specifications

3.6.1 Introduction

At this stage we shall be concerned only with identifying some key features of the specifications of phase-sensitive detectors. The extension to more comprehensive systems will be left until the next chapter. It has been remarked elsewhere that the major developments in lock-in amplifier design have been achieved for instruments operating in the frequency range up to 1 MHz and the 'typical' specifications given in the following sections will be for phase-sensitive detectors of this type. In recent years it has become common practice for manufacturers to quote some specification figures at 'midband' or at a specified operating frequency, say 1 kHz. The deterioration of such 'typical' specifications towards lower or higher operating frequencies is not always given. At best it may be possible to obtain the required information by carefully reading the data sheet. At worst it must be assumed that some deterioration will occur.

3.6.2 Full-scale sensitivity

Phase-sensitive detectors are 'general purpose' in the sense that they are capable of operating with a wide range of signal types. However, we have seen that the magnitude of the response depends on the Fourier composition of the signal and so must differ in each case.

To enable a sensible comparison to be made between alternative designs the scaling factor of phase-sensitive detectors is almost invariably specified for a synchronous sinewave signal. A simple scaling factor relates the d.c. output voltage to the r.m.s. value of an *in-phase* synchronous sinewave at the phase-sensitive detector input. In practical systems the d.c. output must be limited to some maximum value. This is the *full-scale output* which is usually ±10 V in modern systems. The *full-scale sensitivity* of a phase-sensitive detector is defined as the r.m.s. value of an in-phase synchronous sinewave which gives a full-scale d.c. output, and is expressed in volts, millivolts or even microvolts.

If the full-scale sensitivity S_D is known, the d.c. response to a synchronous sinewave with r.m.s. value V_s and relative phase-shift ϕ is

$$V_o = V_F(V_s / S_D)\cos\phi$$

where V_F is the full-scale output voltage.

As is usual with this sort of specification, a phase-sensitive detector with 'high' sensitivity has a low value of S_D and is able to give a full-scale response to signals at 'low' level.

3.6.3 Linearity and out-of-phase rejection

We have seen in Chapter 2 that an 'ideal' synchronous detector is inherently free from non-linear effects. Thus, in principle, a synchronous signal can be measured in the presence of noise without incurring errors due to intermodulation and offsets due to rectified noise components.

In making the transition to practical devices we must first recognize that there is a maximum level of noise voltage - asynchronous voltage - that a phase-sensitive detector can withstand. This is determined by the level of asynchronous input that gives rise to gross detection errors due to distortion and 'clipping' in the electronic circuits. Unfortunately, in practical devices, the effects of non-linearity, and the resulting detection errors are likely to be discernible when the signal of interest is accompanied by asynchronous voltages well below the clipping level. In an attempt to reflect these limitations it used to be common practice for manufacturers to quote a specification known as *out-of-phase rejection*. This measure was widely used a number of years ago for comparing competing instruments.

Out-of-phase rejection is specified and measured with a synchronous sinewave signal adjusted to give an output at or near full scale. A second, asynchronous, sinewave at a frequency well removed from any transmission windows is added to the first and increased until a pre-determined change in the output occurs. This output change is generally attributed to non-linearity in the phase-sensitive detector transfer function although the precise nature of the non-linearity is not usually specified. The standard approach is to calculate the out-of-phase rejection using:

$$\text{Out - of - phase rejection} = \frac{\text{r.m.s. value of asynchronous sinewave}}{\text{r.m.s. value of synchronous signal} \times \text{fractional output change}}$$

For example, if we choose a fractional change of 1% and this change corresponds to an asynchronous/synchronous voltage ratio of 1000:1 the out-of-phase rejection will be ×100 000 or 100 dB. The high value of voltage ratio obtained emphasizes that to be of use in signal recovery applications a phase-sensitive detector should be able to sustain the effect of asynchronous voltages far in excess of a full-scale synchronous signal.

This fractional change in a near full-scale output increases with increasing asynchronous voltage in a highly non-linear fashion. Thus, if out-of-phase rejection were to be measured for a 0.1% or 10% change in output, entirely different values would be obtained.

Users of phase-sensitive detector-based equipment should therefore be aware that there is a trade-off between linearity and the level of asynchronous voltage permitted at the phase-sensitive detector. The overall behaviour is such that, for good linearity, the asynchronous voltage should be maintained at a relatively low level in order to restrict the total voltage swing at the phase-sensitive detector. In cases where a signal is very noisy, the user might permit a larger voltage swing due to asynchronous components but would then suffer a penalty in the form of reduced linearity at the point of detection. Clearly, the ‘maximum allowed’ asynchronous input to the phase-sensitive detector would be quite different in these two extreme examples, determined in each case by the errors due to non-linearity that could be tolerated by the user. We shall return to this topic when discussing the linearity of different system configurations in Chapter 4.

3.6.4 Dynamic reserve

When considering the specification of out-of-phase rejection in the last section it became evident that a small fraction of the total input range of a phase-sensitive detector is in fact ‘used’ by synchronous signals while a much larger fraction is held ‘in reserve’ for noise and interference. This idea leads to the definition of a prime phase-sensitive detector specification; this is *input dynamic reserve*·, given by the ratio:

$$r_{\mathrm{i}} = \frac{\text{max. allowed p - p level of asynchronous sinewave}}{\text{p - p value of full - scale synchronous signal}}$$

Dynamic reserve thus gives a direct measure of the worst-case signal-to-noise ratio that can be tolerated at the input to a phase-sensitive detector consistent with maintaining a full-scale output. The term ‘overload capability’ is also used in this context, but we shall be using this description in a more specialized sense in Chapter 4.

Dynamic reserve is a useful concept but only serves as a basis for comparing different systems when there is general agreement on defining ‘maximum allowable’ asynchronous input.

One way to do this is to identify the level of asynchronous input that causes a change of 5% in the response to a full-scale signal. This approach has its roots in the specification of out-of-phase rejection and is a way of reaching a compromise between two conflicting requirements: first of all, to achieve the largest possible input voltage swing, thus maximising the signal recovery ‘capability’ of the phase-sensitive detector; secondly, to restrict the total input to a level where errors due to non-linearity are - for most practical purposes - just discernible.

· Usually known simply as "dynamic reserve"

An alternative approach is simply to equate the maximum allowed level with a value marginally less than the input clipping level. This is the *input overload level* quoted by manufacturers beyond which gross measurement errors will be incurred. Circuit designers find this definition attractive since the input clipping level is usually a well-defined circuit parameter, unlike the 5% error limit which can only be determined by painstaking observations at the phase-sensitive detector output.

In many cases, these two approaches give rise to very similar results for dynamic reserve. It might be thought, therefore, that the first approach is by far the most satisfactory because it incorporates the twin ideals of 'capability' and linearity. However, in practice, the second approach is almost universally adopted since it establishes a predictable relationship between the internal gain of a phase-sensitive detector and dynamic reserve.

To see this, let us consider some specification figures for a modern phase-sensitive detector. The input dynamic reserve would typically lie in the range 60 dB – 80 dB (×1000 to ×10 000) while the input overload level is of the order of several volts peak-to-peak. It follows that the full-scale sensitivity must be correspondingly high. For example, a phase-sensitive detector with a dynamic reserve of 60 dB might have a full-scale sensitivity of 1 mV r.m.s. (~ 3 mV peak-to-peak). This would allow an interference voltage to rise around 3 V peak-to-peak before the overload indicators gave warning of improper operation.

Let us suppose now that the low-frequency gain of the output low-pass filter is increased by a factor of 10. The sensitivity, and hence the signal required for a full-scale output, is now 100 μV r.m.s., while the input overload level remains at 3 V peak-to-peak.

When dynamic reserve is defined on the basis of input overload level we have no hesitation in stating that the dynamic reserve of this more sensitive phase-sensitive detector is ×10 000 or 80 dB. We thus reach the important conclusion that dynamic reserve increases in proportion to the d.c. gain of the post-detection filter. This direct relationship would clearly be lacking when the definition of dynamic reserve involved a detailed evaluation of measurement errors at different levels of applied signal and noise.

It is usually assumed that the definition of dynamic reserve is valid for a sinusoidal interference voltage and for interference from a broadband noise source. In the latter case the phase-sensitive detector output will contain a residual noise fluctuation due to noise transmitted by the fundamental and higher-order transmission windows. 'Linear' operation of the phase-sensitive detector then implies that the *mean* output voltage due to a synchronous signal is unaffected by the presence of noise at the input. From the discussion given at the end of Chapter 2, however, it is evident that, if a phase-sensitive detector was operated with a signal and broadband noise at the dynamic reserve limit, then the main limitation to precision measurement would result from the relatively large fluctuation remaining in the final output (unless an extremely long time were available for measurement). The justification for providing very high values of dynamic reserve is therefore to enable measurement to be carried out in circumstances where the dominant interference is due to narrowband noise or discrete frequency pick-up.

In view of this, manufacturers often make use of the alternative definition:

$$r_{\mathrm{i}} = \frac{\text{max. allowed r.m.s level of asynchronous voltage}}{\text{full-scale sensitivity (r.m.s. volts)}}$$

When a sinewave interference voltage is assumed, this definition is equivalent to that given at the beginning of this section.

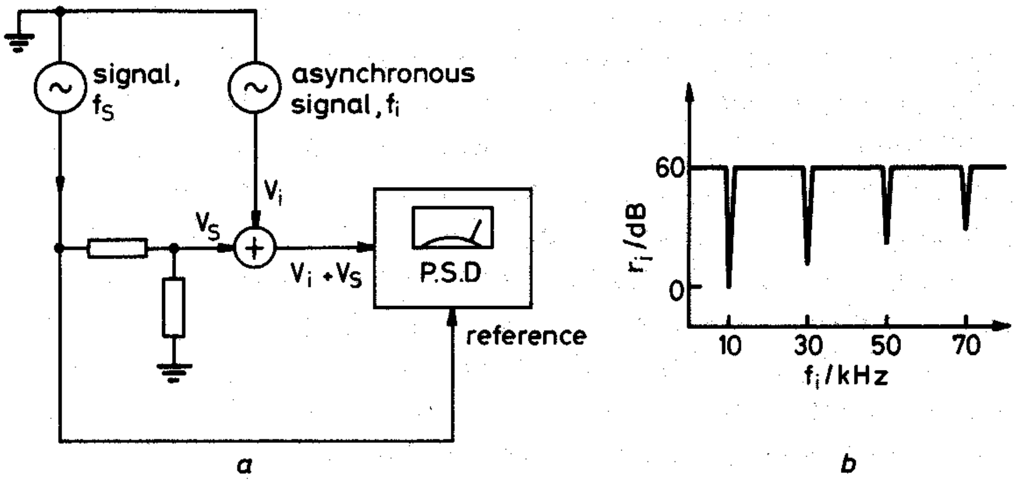


Fig. 3.13 (a) Evaluating the effect of a sinewave interference; (b) dynamic reserve characteristics of a broadband phase-sensitive detector

It should not be assumed that restriction to a sinewave interference automatically precludes difficulty in using the specification provided by a manufacturer. Consider, for example, what happens when the frequency of the interfering sinewave lies close to one of the phase-sensitive detector transmission windows. When a harmonic window is excited, the phase-sensitive detector output is perturbed by a difference-frequency 'beat' component which may be of sufficient peak-to-peak value to cause a severe output overload. When the interference frequency is fixed, the output overload can be overcome by reducing the bandwidth or the output low-pass filter. When experimental constraints determine the minimum usable bandwidth, the output overload can only be prevented by reducing the level of the interfering component. The result is a loss of dynamic reserve in certain critical frequency ranges. These effects can be evaluated using the experimental set-up shown in Fig 3.13(a). Fig 3.13(b) shows the resulting dynamic reserve characteristics of a broadband phase-sensitive detector referenced at 10 kHz.

In the light of these observations, the standard data-sheet presentation of dynamic reserve as a single unqualified figure appears to be rather inadequate. At the very least, the figure should be read as *maximum achievable* dynamic reserve. Also, wherever possible, operation at the specification limit should be avoided in the interests of maintaining good linearity. It can be argued that linearity is often of secondary importance when tackling the more fundamental problem of obtaining a measure of signal obscured by high levels of noise and interference. When extreme levels of noise cannot be avoided it is evident that a very detailed understanding of the actual dynamic characteristics of the phase-sensitive detector under the proposed measurement conditions is required.

3.6.5 Output stability and minimum detectable signal

If the total signal input to a phase-sensitive detector is removed and the reference waveform is left connected, the output should, ideally, fall to zero. The application of a millivoltmeter will, however, reveal a residual offset voltage and long-term observation will show the effect of drift in the output. The offset is, moreover, dependent upon the reference frequency and will usually increase quite markedly if the reference frequency is increased beyond about 10 kHz. This effect is known as *h.f.offset* and is due to stray reactive coupling of switching spikes from the reference input through to the final output.

Provision is usually made to trim the offset voltage to zero under a particular set of operating conditions and, if this is done, the drift component in the output represents the limiting factor to precision measurement. The drift is usually labelled as output 'stability' and quoted as a fraction of the full-scale output voltage per Kelvin.

Let us now consider the problem of measuring either very small synchronous signals or small *changes* in a synchronous signal. In these cases measurement difficulties will result whenever the corresponding change in output voltage is comparable with the drift component of the phase-sensitive detector. When operating at full-scale sensitivity S_D we accordingly define the minimum detectable signal s_{min} as

$$s_{min} = \delta \times S_D$$

where δ is the fractional drift/K.

Note that this expression serves as a *definition* of minimum detectable signal which will be more or less useful in a given practical situation. Also, when calculating s_{min}, only the numerical part of the drift specification is used. Thus, by convention, the minimum detectable signal is expressed in volts or, more usually, in microvolts.

It is worth noting that the d.c. response to a signal at the minimum detectable level can be separated from the spurious outputs due to offset and drift by introducing a phase reversal of 180° in the reference channel. The d.c. output due to the signal will then reverse its polarity while the polarity and magnitude of the offset (being phase-insensitive) will be unchanged.

This method of overcoming limitations due to offset has been exploited in systems operating on the synchronous heterodyne principle, described in Section 8.8.

3.6.6 Dynamic-reserve/output-stability trade-off

It was shown in Section 3.6.4 that both the sensitivity and the dynamic reserve of a phase-sensitive detector can be increased by increasing the d.c. gain of the post-detection filter. In practice, this additional gain can always be obtained by incorporating an output d.c. amplifier.

Unfortunately, improved dynamic reserve can only be achieved at the expense of precision since all errors due to offsets and drift in the phase-sensitive detector are enhanced by the gain of the output amplifier. If the sensitivity of a phase-sensitive detector was controlled solely by switching output gain we would find a very uncomfortable situation in which high sensitivity was linked with high reserve and low sensitivity was required for good output stability.

For maximum flexibility, a phase-sensitive detector should have switched output gain and be supported by a variable-gain amplifier in the signal path. The balance between dynamic reserve and output stability can then be adjusted at a given level of sensitivity. This aspect of system performance will be discussed in Chapter 4.

3.6.7 Dynamic range

By 'dynamic range' of a phase-sensitive detector we usually mean the *input* dynamic range. This is defined at one extreme by the maximum allowed input voltage swing and at the other by the minimum detectable signal. If we denote the maximum allowed voltage swing by Δv, the input dynamic range is simply:

$$D_I = \Delta v / s_{min}$$

It was shown in Sections 3.6.3 and 3.6.4 that the 'maximum allowed' input voltage to the phase-sensitive detector can have different values depending on the errors due to non-linearity that can be tolerated. In general, the linearity of the phase-sensitive detector is degraded as larger and larger asynchronous voltage swings are permitted, so we see that there is an important trade-off to be made between input dynamic range and linearity such that one can be improved only at the expense of the other. In order to specify input dynamic range, it is usual to equate the maximum allowed input voltage with the input overload level, that is to use the same voltage swing that appears in the specification of dynamic reserve. When this is done, both Δv and s_{min} can be expressed in terms of the full-scale sensitivity S_D of the phase-sensitive detector. Using the results of Sections 3.6.3 and 3.6.5 we obtain:

$$\Delta v = r_i \times S_D$$

and

$$s_{\text{min}} = \delta \times S_D$$

where r_i is the dynamic reserve and δ is the fractional output stability. We can therefore arrive at the following expression for input dynamic range[·]:

$$D_I = r_i / \delta$$

The result is quite properly independent of the phase-sensitive detector sensitivity and any post-detection gain stages: any attempt to increase dynamic reserve by increasing the output gain is accompanied by a deterioration in output stability. We thus see that, to achieve wide dynamic range, a phase-sensitive detector must be capable of giving good output stability at a high level of dynamic reserve and this has proved to be one of the main objectives in phase-sensitive detector design. As a rough guide, in modern systems, a combination of ×2 000 dynamic reserve and 100 p.p.m./K output stability represents 'good practice' corresponding to a dynamic range of:

$$D_I = 1000 / 10^{-4} = 10^7 \text{ (140 dB)}$$

Since the concept of dynamic range incorporates the two key specifications of phase-sensitive detectors it provides the best figure of merit to be used when comparing competing systems. It should now be clear that dynamic reserve alone is no guarantee of quality and should always be viewed in the light of the stability specification.

3.6.8 Summary of specifications

The specifications covered so far can be conveniently summarized in diagram form as shown in Fig 3.14. The diagram also serves to define two additional quantities, namely *output* dynamic reserve and the *output* dynamic range.

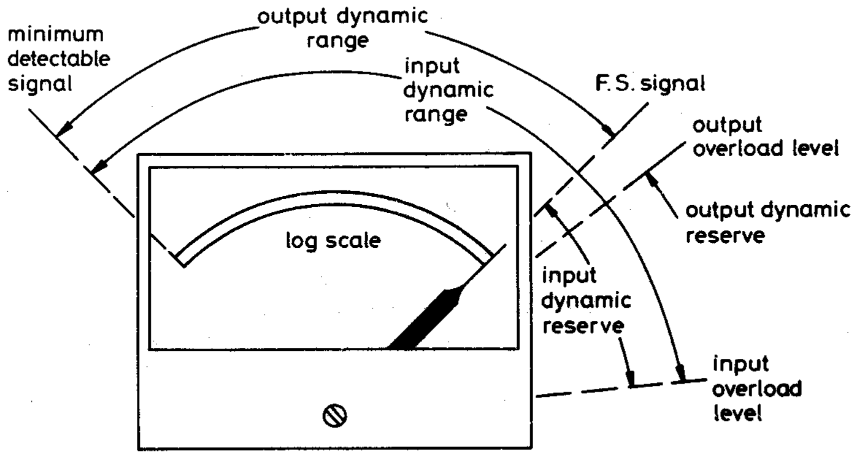


Fig. 3.14 Summary of phase-sensitive detector specifications

[·] Again, we use only the numerical part of δ, giving D_I as a voltage ratio

To be of any practical use a phase-sensitive detector must be able to give a full-scale response to a synchronous signal and yet be able to accommodate a residual output fluctuation due to noise transmitted by the low-pass filter. The residual noise will carry the output beyond full scale and there should be sufficient margin in the output circuit to allow this to happen without an overload indicator permanently flashing; hence the provision of output dynamic reserve which enables the output to exceed full scale by around 20% - 30% without suffering distortion due to ‘clipping’ in the output amplifier.

The output dynamic range is given as the ratio of the full-scale output voltage to the output drift component. If the drift is specified as a fraction of full scale- as we have assumed throughout - then the output dynamic range is simply $1/\delta$. In high-stability operation, phase-sensitive detectors are capable of operating with a relative drift of less than 10 p.p.m./K. This corresponds to an output dynamic range of ×100 000 (100 dB) and would enable a *change* of 100 µV to be observed in a full-scale output voltage of 10 V. Obviously, such a mode of operation presupposes a signal which is correspondingly free from noise.

Using the definition of output dynamic range given above, together with the results of Section 3.6.7, we see how dynamic reserve provides a link between input and output dynamic range:

Input dynamic range = Dynamic reserve × Output dynamic range

In the following chapter it is shown how the trade-offs between dynamic range and linearity, and between dynamic reserve and output stability, can be improved by incorporating filters in the signal channel of an otherwise broadband lock-in system.

3.7 References

1 CARTER, S.F., and FAULKNER, E.A. (1977): "New phase-sensitive rectifier circuit", *Electron. Lett.*, 3, pp. 339-340.

2 BLAIR, D.P., and SYDENHAM, P.H. (1975): "Phase-sensitive detection as a means to recover signals buried in noise", *J. Phys. E: Sci. Instrum.*, 8 pp. 621-627.

3 LETZTER, S.G. (1974): "Explore the lock-in amplifier", Electron. Des., 21, pp. 104-108 (also published as Technical Note 115: Princeton Applied Research Corp. Princeton, NJ).

4 "Introduction to lock-in amplifiers". Technical Note 102, Brookdeal Electronics, Bracknell, England.

CHAPTER 4

Lock-in amplifier specifications

4.1 Introduction

The basic essentials of a lock-in amplifier were introduced in Chapter 2 and a brief description was given of each of the component parts. Of these, the phase-sensitive detector has been given more detailed attention and we have seen how its performance can be evaluated in the light of some key specifications.

In this chapter we shall continue with the same theme but concentrate on the specification of lock-in amplifiers as a whole, rather than of phase-sensitive detectors in isolation. It is intended that the approach adopted will be applicable equally to lock-in amplifiers as self-contained instruments or to lock-in recovery systems plugged together from several individual units. Since we have already identified several specifications relating to phase-sensitive detectors it will be of interest to see how these contribute to the performance of a larger system and to see how systems can be set up to give an optimum solution to different types of measurement problem.

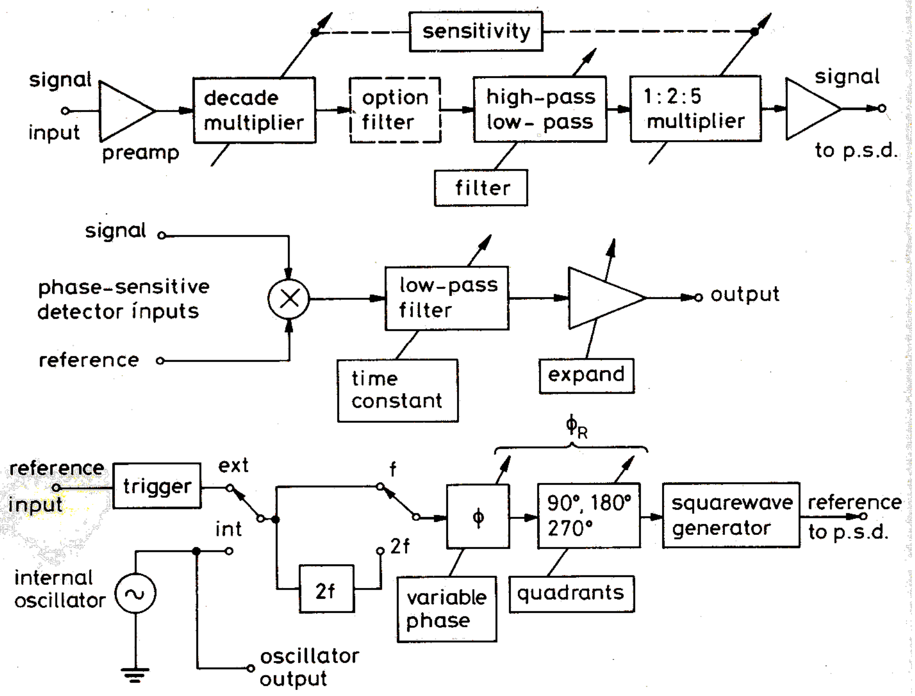


Fig. 4.1 Block diagram of a lock-in amplifier. The reference channel incorporates an optional internal oscillator and a frequency doubler to permit phase-sensitive detection at twice the applied reference frequency.

To this end, we can refer to the block diagram in Fig. 4.1 which is an extended version of the one given in Chapter 2. The signal channel is shown with variable gain which is obtained through a combination of amplifiers and attenuators. The arrangement in a commercial system gives a near-optimum noise performance in all switched positions and provision is made for signal-channel filtering, both internal and from external ‘plug-in’ modules.

We shall assume that the phase-sensitive detector has fixed sensitivity but that the output gain can be controlled by a switched d.c. amplifier. The amplifier is labelled as an ‘expand’ amplifier since it serves to expand the output from the phase-sensitive detector.

The essentials of the reference channel are also sketched in. These comprise a precision trigger circuit to respond to the positive zero crossings of an externally applied reference waveform, a broadband precision phase-shifter, and a squarewave generator to supply the final drive to the phase-sensitive detector. The phase-shifter takes the form of a control loop which sets up a phase-shift in response to a control voltage generated at the front-panel phase dial. Control over a full 360° range is usually achieved using pushbutton selection of phase quadrants in conjunction with a 0–100° continuous phase control. We shall be able to identify several critical specification features for the reference channel from a general review of system behaviour.

Note that the system shown in Fig. 4.1 incorporates two other features that are commonly found in commercial units. These are a reference-channel frequency doubler, which enables the detection system to be synchronized at the second harmonic of the applied reference frequency, and an ‘internal’ oscillator. The latter can be used as an excitation source for an experiment and is sometimes controllable from the front panel of the lock-in amplifier. When switched into operation, the internal oscillator makes a direct connection to the reference-channel phase-shifting circuits and so overcomes the need for an external cable connection.

4.2 Calibration: full-scale sensitivity

As in the case of phase-sensitive detectors, the full-scale sensitivity of a lock-in amplifier gives the r.m.s. value of an in-phase synchronous signal required to give a full-scale output. An essential difference between modular lock-in amplifiers and those supplied as integrated units is that the latter are supplied with a single sensitivity switch which enables the full-scale sensitivity to be set up immediately. In this case, the individual gains of the signal channel amplifiers and the phase-sensitive detector need not be known, whereas in a modular system these can normally be read from the front panel settings of the individual units.

In either event, the full-scale sensitivity measured from the signal channel input socket is

Full-scale sensitivity, $S_F = S_D/(G_a G_e)$

where S_D denotes the full-scale sensitivity of the phase-sensitive detector, calibrated on a sinewave, and G_a and G_e are the gains of the signal channel and expand amplifiers.

In lock-in amplifiers the usual convention is to control only the signal channel gain from the sensitivity switch which is calibrated at a given expand setting G_{CAL}. If we denote the *indicated* sensitivity on the front panel by S_I, then we have

Full-scale sensitivity, $S_F = S_I \times G_e/G_{CAL}$

When a synchronous sinewave, with r.m.s. value V_s, is applied to a lock-in amplifier having a full scale d.c. output of $\pm V_F$, the d.c. response of the system will be

$$V_o = V_F(V_S/S_F)\cos\phi$$

where ϕ is the phase difference of the waveforms appearing at the phase-sensitive detector. In most modern instruments V_F has a value of 10 V.

If the system is put into a quadrature condition to detect small phase variations (Section 2.4.4) the resulting sensitivity to *small* phase offsets is

$$V_F(V_S/S_F)\text{ volts/radian}$$

When V_F = 10 V, this corresponds to a sensitivity of 10 V/radian with a full-scale signal.

When used as an amplitude detector or as a phase detector, the response to modulations will be modified by the effect of the output low-pass filter as discussed in Sections 2.4.3 and 2.4.4.

4.3 Phase-sensitive detector related specifications

4.3.1 Introduction

It is not unusual to find that lock-in amplifiers can cater for input voltages ranging over six decades, having full-scale sensitivities calibrated in the range 1 μV to 5 V (exclusive of any additional gain provided by preamplifiers). A typical system might offer up to 12 switched ranges in the signal channel amplifier (linked to the sensitivity control) and perhaps three levels of output expansion (×1, ×10 and ×100). It follows from this that there must be a number of combinations of a.c. and d.c. gain which give the same overall sensitivity, so let us now see how the choice of gain combination influences the dynamic performance of the lock-in system.

4.3.2 System dynamic reserve

The idea of dynamic reserve was introduced in Section 3.6.3 and defined in terms of the voltages applied to the input terminals of a phase-sensitive detector. The same idea can be applied to the lock-in system as a whole except that we must transfer all measurements to the input of the signal channel amplifier since the phase-sensitive detector input is no longer directly accessible.

To illustrate this let us identify some voltages in the system using Fig. 4.2. First, the signal v_S at the input is a full-scale signal giving a full-scale indication on the output meter. Secondly, Δv_N represents the peak-to-peak value of the input noise and, finally, Δv is the maximum allowed voltage swing at the input to the phase-sensitive detector.

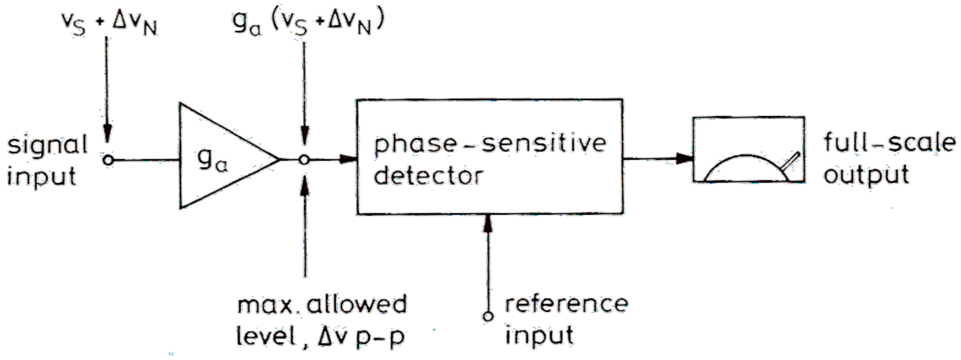


Fig. 4.2 Identification of voltages in a phase-sensitive detection system

The dynamic reserve is measured with a *broadband* signal channel so that there is no selective attenuation of the input noise components. The most convenient starting point is to assume that the lock-in amplifier is operating at its dynamic reserve limit. Referring to Fig. 4.2, the peak-to-peak voltage[*], $G_a(v_s + \Delta v_N)$, at the phase-sensitive detector input is therefore equal to the maximum allowed level, Δv. The ratio $\Delta v_N/v_s$ measured at the signal channel input is thus equal to the system dynamic reserve. If the system dynamic reserve was now measured using an interfering sinewave as described in Section 3.6.4, we would find a characteristic identical to that of the phase-sensitive detector, including the influence of the harmonic transmission windows.

It is this feature which distinguishes 'broadband' lock-in systems, where the overall response to asynchronous inputs reflects the properties of the phase-sensitive detector.

Let us now return to a discussion which was first opened in Section 3.6.6. Any increase in input noise beyond Δv_N peak-to-peak will overload the phase-sensitive detector but can be offset by reducing the signal-channel gain. The reduced response to the synchronous signal can then be 'expanded' using the d.c. output amplifier in order to restore the overall sensitivity to its original value. For a fixed level of synchronous input to the instrument we find that the system can tolerate progressively higher input noise levels as the a.c. gain G_a is reduced and the expand factor G_e is increased to maintain constant overall gain. In this way, the dynamic reserve measured at the signal channel input can be controlled by changing the internal gain distribution.

In the interests of linearity it is essential to limit the voltage swing at the input to the amplifier to a well defined maximum value. This is the input overload level which is usually of the order of a few volts peak-to-peak. If a full-scale signal of 100 mV r.m.s. is presented to the lock-in amplifier, then the input noise-to-signal ratio must be less than about 10:1 to avoid overloading the amplifier input. We thus see that the full dynamic reserve of the system can only be exploited for relatively small signal inputs where the lock-in amplifier is operated at correspondingly high sensitivity. In systems offering dynamic reserves of × 1000 (60 dB) and greater this corresponds to operating at a minimum sensitivity of around 1 mV r.m.s.

4.3.3 System overload capability

The dynamic reserve of a signal recovery system is always measured with a *broadband* signal channel and the specification is normally valid over many decades of frequency −subject to the effect of the harmonic transmission windows discussed in Section 3.6.4.

In a typical measurement the signal will be perturbed by noise which is non-white in character and a substantial fraction of the input noise voltage might be attributable to discrete interference components appearing in a well-defined frequency range. Under these circumstances, the total peak-to-peak value of the disturbance reaching the phase-sensitive detector can often be reduced significantly by using signal conditioning filters. By eliminating noise components *before* detection the dynamic reserve of the system can appear to be much greater than when operating with a broadband signal channel. We now

[*] When operated in the dynamic reserve limit the noise input to the phase-sensitive detector will usually exceed a full-scale signal by at least a factor of 10, giving $\Delta v_N >> v_s$.

refer to the *overload capability* of the system, which should always be specified with respect to a particular set of operating conditions.

Because overload capability is a narrowband specification the achievable value will depend entirely on the variety of filters available for signal conditioning and on the separation of the signal frequency from the characteristic frequencies of any associated noise and interference. At any sensitivity setting, the maximum achievable overload capability will be

$$\frac{\text{input overload level of signal channel}}{\text{full - scale sensitivity}}$$

In a lock-in amplifier which offers a maximum sensitivity of 1 μV r.m.s. and an input overload level of a few volts peak-to-peak the best achievable overload capability will be of the order × 1 000 000 (120 dB)) even though the (broadband) dynamic reserve is a maximum of × 10 000 (80 dB). It must be emphasized, however, that this order of performance is likely to be realized in practice only for high-level interference components at well-defined frequencies which can be subjected to the maximum degree of suppression by suitable choice of signal channel filters. Also, by describing the specification as essentially 'narrowband' we are reminded that high overload capability is achievable only at the expense of flexibility in the choice of operating frequency. In a basic detection system, extremely high overload capability is not generally compatible with the ability to track signals over a wide frequency range, because of the fixed characteristics of the noise-rejection filters.

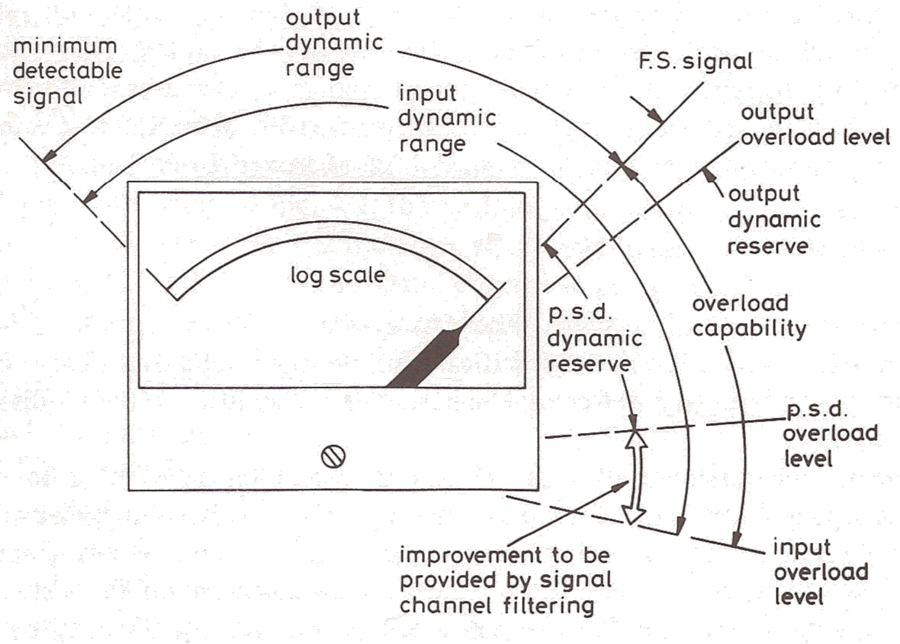


Fig. 4.3 Relationship between the specifications of a lock-in amplifier

The relationship between overload capability and dynamic reserve is brought out in the presentation of Fig. 4.3. The diagram extends the one given in Section 3.6.8 for a phase-sensitive detector by including the effect of signal channel amplification and signal conditioning filters. It is evident that if a given level of overload capability is required, the signal conditioning filters must be capable of reducing unwanted components by a factor:

$$F = \frac{\text{desired overload capability}}{\text{dynamic reserve}}$$

4.3.4 Dynamic reserve and output stability trade-off

We have seen that the effect of incorporating variable gain amplifiers before and after detection is to produce a lock-in amplifier with controllable dynamic reserve at a given sensitivity. In order to realize the full practical potential of such a system we must take into account the effect of gain selection on output stability and see how dynamic reserve and stability can be jointly controlled to the best effect.

We can formulate a basic approach to setting up a detection system as follows: first of all for very 'noisy' signals. Here, the signal channel gain G_a should be held at a sufficiently low value to avoid premature overload at the input to the phase-sensitive detector. The desired sensitivity is then obtained by increasing the gain of the expand amplifier. The output stability is degraded, but we argue that our main problem in this case is to detect the signal at all. A loss of output stability is not likely to be significant or even discernible when the final output contains a relatively large component of residual noise transmitted by the low-pass filter.

The situation is quite different for relatively 'clean' signals. These can be subjected to a large amplification factor without exceeding the maximum allowable voltage swing at the phase-sensitive detector. The minimum value of expand gain can then be used to ensure that the signal is measured with the best possible precision.

These two modes of operation can be summarized for a lock-in amplifier offering two levels of output expansion by means of the example given in Table 4.1.

Table 4.1

Signal	Sensitivity selection	Expand gain	Output stability, p.p.m./K	Dynamic reserve
'clean'	100 μV	× 1	< 10	× 100
'noisy'	1 mV	× 10	< 100	× 1000

Overall sensitivity = 100 μV f.s.

In an attempt to combine flexibility with ease of use the principal lock-in amplifier manufacturers have incorporated pushbutton selection of either 'high stability' or 'high reserve' operation. One way that this can be achieved is to retain the expand amplifier, leaving this to be controlled by the user, and to arrange for the distribution of gain between the signal channel and the phase-sensitive detector to be switched automatically as shown in Fig. 4.4.

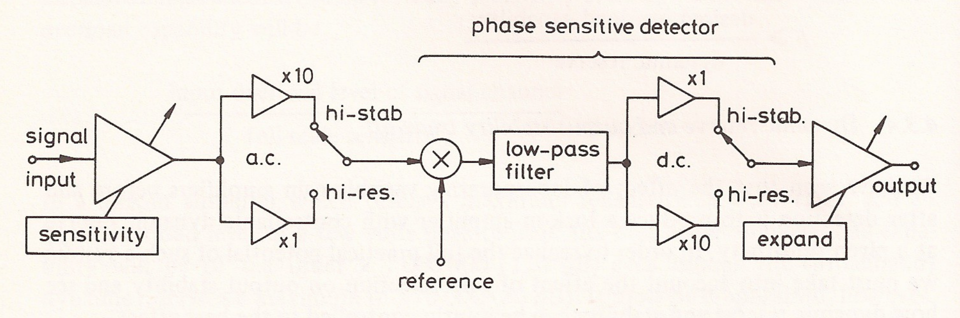


Fig. 4.4 Switch selection of high stability mode and high reserve mode in a lock-in amplifier

In 'high stability' mode the phase-sensitive detector sensitivity is reduced at the expense of signal channel gain. In 'high reserve' mode the signal channel gain is reduced while that of the phase-sensitive detector is automatically increased so as to retain the same overall sensitivity. With two levels of output expansion available the system can be adapted to meet a wide range of measurement requirements as demonstrated by the example in table 4.2.

Table 4.2

Mode selection	Sensitivity selection	Expand gain	Output stability p.p.m./K	Dynamic reserve
High Stability	100 μV	× 1	< 10	× 100
	1 mV	× 10	< 100	× 1000
High Reserve	100 μV	× 1	< 100	× 1000
	1 mV	× 10	< 1000	× 10 000

Overall sensitivity = 100 μV f.s.

4.3.5 Overload capability and output stability trade-off

Section 4.3.4 dealt only with dynamic reserve. As a result the trade-off between noise rejection and output stability has been demonstrated only for a broad-band signal channel which introduces no selective attenuation of noise components. There is, however, an aspect of noise rejection which applies to the case where a normally 'clean' signal is perturbed by large-scale interference. 'Hum' pick up on a low-frequency a.c. bridge provides a good example; the difficulty is to measure the signal to a high degree of precision.

In the absence of signal conditioning filters, the only approach to this particular problem would be to select a suitably high value of dynamic reserve. This would avoid the danger of overload on the interference component, but would automatically result in a relatively poor output stability and a consequent loss of precision.

The importance of signal conditioning becomes evident when we realize that for every 20 dB of suppression brought about by using, say, a notch filter, the system reserve can be switched down by 20 dB and the output stability improved in direct proportion. Ultimately it might be possible to achieve the desired precision with only a modest amount of suppression obtained through signal conditioning, provided that the relative magnitudes of signal and interference are brought within the capability of the phase-sensitive detector at the desired level of output stability.

In the light of earlier discussions it is evident that obtaining improved overload capability at a given level of output stability is equivalent to increasing the input dynamic range of a detection system. Furthermore, in circumstances where it is desirable to obtain high overload capability in conjunction with the best possible output stability, it is clear that the signal-channel gain should be switched to the maximum possible value consistent with avoiding overload at the phase-sensitive detector. Unfortunately, the linearity of the detection system will be degraded under these conditions, so it is important that we examine the relationship between dynamic range and linearity in practical lock-in systems.

4.3.6 Dynamic range and linearity trade-off

Suppose we wish to measure the amplitude of a 100 μV r.m.s. signal in the presence of an interference component at a level just below 10 mV r.m.s. A lock-in amplifier with the characteristics summarized in Table 4.2 would enable this signal to be measured at full-scale in a high-stability mode, using a dynamic reserve of × 100**.** The output stability would then have its best value, < 10 p.p.m., however, the linearity of the system would be degraded on account of the large asynchronous voltage swing presented to the phase-sensitive detector.

If good linearity is required, an obvious step is to reduce the pre-detection, a.c. gain by, say, a factor of 10. The total input voltage to the phase-sensitive detector is then reduced, linearity is improved and the sensitivity can be restored by using a larger 'expand' gain.

The improvement in linearity is observed as an increase in out-of-phase rejection (Section 3.6.3) since, at an input asynchronous/synchronous voltage ratio of 100, the error in the full-scale response will be reduced. However, in this new situation, with the 'expand' gain increased to × 10**,** the output stability will be degraded while the improved dynamic reserve cannot be exploited without incurring a larger voltage swing at the phase-sensitive detector. A possible increase in dynamic reserve must therefore be sacrificed in favour of an improvement in linearity. Since the output stability is worse, the net result is a loss of input dynamic range.

The situation can be improved when interference components are rejected in advance of detection by using signal-channel filters, but a trade-off exists nevertheless. We have seen that, when wide dynamic range is the main consideration, the signal-channel gain following a stage of filtering should be increased to its maximum allowed value. When good linearity is required, the signal-channel gain must be maintained at a relatively low value so as to restrict the input swing to the phase-sensitive detector. The overall sensitivity would then be restored by an increase in 'expand' gain. The linearity, measured in terms of out-of-phase rejection at the input to the signal channel, is improved while the system suffers a degradation in output stability. The dynamic range under these conditions could be greater than that obtainable without filters but will inevitably be less than the maximum achievable value.

4.4 Using a tuned filter in the signal channel of a conventional lock-in amplifier

4.4.1 Influence on overload capability

Early lock-in detection systems were traditionally operated with signal channels tightly 'tuned' to the signal and reference frequency in order to overcome the limited dynamic range of the then available phase-sensitive detectors. Although modern systems have much better performance, there are instances where they can benefit from the rejection properties of a tuned filter, and most commercial systems are supported by a range of options which includes a tuned filter.

The type of filter most often used has a symmetrical band-pass characteristic of the type described in Appendix 4. Commercial units are, moreover, almost invariably active filters which use operational amplifiers with capacitive and resistive feedback elements to simulate the circuit behaviour of inductors and so avoid the use of large inductor values at the lower frequencies required. An active filter usually has a gain greater than unity at frequencies close to the centre frequency, although this is sometimes compensated by including an output attenuator to give a net gain of unity.

When dealing with active filters we must be prepared to identify overload conditions in all the internal amplifiers, and it is usual to specify the maximum allowable input voltage swing at 'in-band' and 'out-of-band' frequencies. Fig. 4.5

shows how this can be done for a tuned filter considered in isolation using the filter frequency-response characteristic. At mid-band, interference components are subject to the full gain, G_o, of the filter, and their magnitude must be restricted to avoid driving the filter output into saturation. The increased input capability at frequencies removed from the centre frequency reflects the greater attenuation provided by the filter. The transition to a flat characteristic at larger frequency offsets marks the point where the input cannot be increased further without exceeding the maximum allowable *input* level to the filter. If a passive filter had been used, the curve would have continued along the dashed lines. In most practical active filter designs it turns out that the absolute maximum voltage swings allowed at input and output are the same, being almost equal to the power supply voltages applied to the filter.

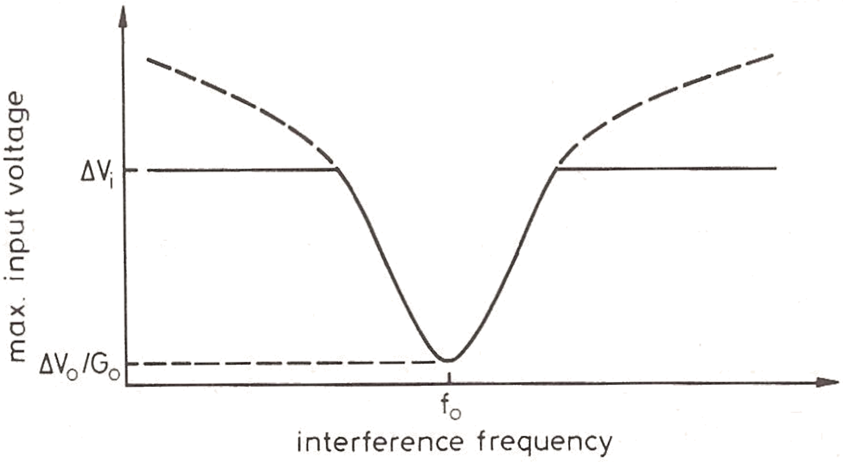


Fig. 4.5 Maximum allowable voltage swing at the input to an active tuned filter with midband gain G_o.

Δv_i and Δv_o denote the peak-to-peak overload levels at the input and output of the filter.

The dashed characteristic would be obtained with a passive filter followed by a voltage amplifier, giving a net gain of G_o at midband.

In normal operation the filter is tuned precisely to the reference frequency and inserted in the signal path. It now becomes essential to examine the relative merits of providing amplification before and after the filter in order to determine its optimum location within the signal channel. To do this, we return to the concept of dynamic reserve which can be generalized to any combination of filters and amplifiers. Following the arguments presented in Section 4.3.2 for the special case of phase-sensitive detectors, we find that the ability to handle large-scale interference components is enhanced by increasing *post-filter* gain at the expense of *pre-filter* gain. In other words: the allowable input voltage for out-of-band components reduces in proportion to the amount of gain which is introduced in front of the filter.

This conclusion is particularly relevant when the user is able to modify the configuration of the signal channel by interchanging amplifier and filter modules. The first step in obtaining high overload capability is to obtain the maximum possible voltage swing at the input to the signal channel. If a degree of flexibility is available this can be achieved by introducing the filter at the earliest possible stage. Unfortunately, the resulting improvement in overload capability is usually obtained at a cost. In this case, the penalty is an increase in system noise. Active tuned filters are generally more noisy than the high-quality amplifiers used in a lock-in signal channel. The result is that the filter makes an increasing contribution to system noise as the gain distribution is altered in favour of post-filter amplification.

The final location of the filter always reflects a compromise between system noise and overload capability. In commercial lock-in amplifiers the filter is usually introduced immediately after the low-noise input amplifier, the objective being to maintain acceptable noise performance in the high-sensitivity positions. In a modular system dedicated to measuring signals in a very noisy environment, the increase in system noise resulting from a filter directly at the signal input may well be of no significance. The incorporation of a filter in this position must then be considered as a valid means of increasing noise-handling capacity.

Whatever combination of amplifiers and filters is used it is possible to arrive at a maximum allowable input voltage swing for out-of-band frequency components. If the total available input swing is to be utilized, then the filter must supply enough attenuation to bridge the gap between the desired overload capability and the best achievable dynamic reserve at the phase-sensitive detector.

The Q-factor of a tuned filter with a centre frequency f_R is defined by

$$Q = f_R/f_B$$

Where f_B is the −3 dB bandwidth of the filter. For interference components offset by several bandwidths, the frequency response of the filter relative to its mid-band value is (Appendix 4):

$$\frac{|H(j\omega_i)|}{|H(j\omega_R)|} = \frac{f_i / f_R}{Q\left|1 - f_i^2 / f_R^{\,2}\right|}; \quad |f_i - f_R| >> f_R / Q$$

where f_i is the frequency of the interference component.

Suppose we have a lock-in system which can withstand a total voltage of 3 V peak-to-peak at its signal channel input. The phase-sensitive detector has a dynamic reserve of × 1000 (60 dB) and a signal of 10 μV r.m.s. is to be measured in the presence of 3 V peak-to-peak interference. The signal frequency is 5 kHz and the interference frequency is 1 kHz.

The interference voltage is at the maximum allowable value of 3 V peak-to-peak (≅ 1 V r.m.s.). If the signal is to be measured at full-scale we require an overload capability of $1/(10.10^{-6})$ or 10^5 while the dynamic reserve is only × 1000; the filter must therefore introduce an attenuation factor of 100. The required Q-factor can now be calculated from:

$$\frac{f_i / f_R}{Q\left|1 - f_i^2 / f_R^2\right|} \leq 1/100$$

Putting f_R = 5 kHz, f_i = 1 kHz, we obtain a value of about 20 for the filter Q-factor. Let us now look at the stability of a system incorporating a filter with this order of selectivity.

A Q-factor of 20 at a centre frequency of 5 kHz implies an effective operational bandwidth of 250 Hz. This means that the signal/reference frequency must be maintained to within ± $2^1/_2$% to stay within the filter bandwidth. If this is not achieved, the magnitude of the signal reaching the phase-sensitive detector will be in error by more than 3 dB or 30%.

The next point to be considered is the phase shift introduced by the filter. The system would normally be set up initially with the filter tuned precisely to the reference frequency. The phase-shift introduced by the filter would then, ideally, be zero, but we can suppose that any small phase error could be compensated by trimming the phase of the reference channel. The phase-shift of the filter close to the tuned condition is given in Appendix 4. We find that the change in phase which occurs when the signal drifts by a relative amount $\Delta f/f_R$ is given by:

$$\tan \Delta\theta \cong 2Q \times \Delta f/f_R$$

Using a Q-factor of 20, we calculate an incremental phase-shift of about 22° for a *one per cent* change in signal frequency. The stability of the signal source and of the filter itself must therefore be of a very high order to avoid excessive amplitude and phase modulation due to drift. It is unlikely, for example, that tuned filters with this degree of selectivity could ever be used satisfactorily with many mechanically derived excitations such as optical choppers.

In view of these constraints, most practical lock-in systems incorporating tuned filters are designed for a maximum Q-factor of around 5. Provided the signal/reference source is of reasonable stability, the overload capability of a broadband detection system can then be substantially improved without incurring excessive measurement errors. This is especially true when a tuned filter is used in the signal channel of a two-phase lock-in system. It is shown in the next chapter that two-phase systems can be used in a phase-independent mode and so overcome the excess phase-shift introduced by a tuned filter. In addition, the use of a tuned filter ensures that harmonics of the signal are greatly suppressed in advance of phase-sensitive detection. The result is a measurement system where the only significant response to either signals or noise is in the vicinity of the reference frequency. This aspect will be discussed further in the next section.

In conclusion, it should be stated that the general considerations regarding the location of a signal-channel tuned filter can be extended to any of the basic high-pass, low-pass and notch filters which are commonly used with lock-in systems. In any of these cases an overload characteristic can be drawn using the frequency response function of the filter or any desired combination of filters. In all cases the amplitude and phase responses of the filter must be considered as possible sources of error when the frequency of the signal/reference source is subject to drift. Ideally, the signal frequency should be as far as possible from the cut-off frequencies of any filters which are introduced for signal conditioning and, as we have seen, the tuned filter falls short of this ideal. In many cases, a far better approach to suppressing discrete interference components is to use a sharply tuned notch filter. When a dominant interference component is well removed from the signal frequency the inclusion of a notch filter leaves the signal substantially undisturbed. It should be remembered that many sources of interference such as line pick-up and breakthrough from radio transmitters are extremely stable in terms of their characteristic frequencies. In these circumstances, a tightly tuned notch filter can be accurately set and maintained for long periods without further adjustment.

4.4.2 Suppression of-harmonic responses

Subject to the limitations on system flexibility, a tuned filter can also be effective in suppressing the effect of harmonically related components of the signal which would normally be applied directly to the phase-sensitive detector.

Again, we shall suppose that the filter has a symmetrical band-pass response of the type used in the last section.

If we imagine that a filter of this type is tuned to a signal and reference frequency at angular frequency ω_R then the filter response at the Kth harmonic of ω_R is

$$H(\mathrm{j}K\omega_\mathrm{R}) = \frac{jK/Q}{(1-K^2)+jK/Q}$$

For $Q \geq 5$ we can approximate the gain magnitude by

$$|\,H(\mathrm{j}K\omega_\mathrm{R})| \cong \frac{K}{(K^2-1)Q}$$

The relative sensitivity of a switching phase-sensitive detector at the K th harmonic of the reference frequency is I/K; hence the relative sensitivity when a band-pass filter is included is

$$|\,H(\mathrm{j}K\omega_\mathrm{R})|\ /K=\frac{1}{(K^2-1)Q}$$

Thus, the relative sensitivity at the third harmonic is reduced to 1/40 (−32dB) for a Q-factor of 5.

Note that, in order to approach the 3rd harmonic sensitivity of the order of −60 dB offered by modern fundamental-only responding instruments, a Q-factor of 100 would be required, a value which would render most systems quite unusable. This assumes, of course, that a ‘standard’ second-order band-pass filter is being used. If, as an alternative, we consider the low-pass tuned filter defined in Appendix 4 we find that the roll-off beyond the centre frequency is now 12 dB/octave rather than 6 dB/octave as in the band-pass case. Repeating the calculations given above, using the appropriate frequency response function, we now obtain a relative sensitivity of approximately $1/(K^3Q)$ at the K th harmonic. In this case, a Q-factor of 40 or so would achieve the target figure of −60 dB at the 3rd harmonic, while a Q-factor of 5 would give a relative sensitivity of less than −40 dB.

This aspect of tuned-filter operation is discussed further in Chapter 8 in relation to the performance of heterodyne lock-in amplifiers. It is shown there that, for a given level of attenuation, alignment problems can be greatly reduced by using two filters in cascade, each of which has relatively low Q-factor. There is no reason in principle why this approach should not be used to improve the performance of the 'conventional' lock-in systems described so far, however, implementation is likely to be easiest when operation is confined to a small range of signal and reference frequencies.

4.5 Reference channel specifications

4.5.1 Introduction

Up to this point we have concentrated almost exclusively on the dynamic performance of lock-in amplifiers. In practice, this performance will never be realized unless the reference channel is capable of giving adequate support to the phase-sensitive detector. It turns out that the lock-in systems with the best all-round performance are those with genuine broadband reference channels giving precisely calibrated phase-shifts allied to good stability. In the following sections we shall be looking at some specifications which are commonly used in relation to reference channels with these general characteristics.

A typical broadband reference channel operates as a closed-loop control system and generates a phase shift in response to a control voltage supplied from a front-panel phase control. By ‘broadband’ we mean a system able to operate with reference frequencies covering a range of several decades without adjustment. A range of 10^5:1 to 10^6:1 is quite common in modern lock-in amplifiers, say from less than 1 Hz to 100 kHz.

At midband frequencies, good phase accuracy can be obtained with relatively short response time in the control loop, while the same control loop might not be so effective at lower reference frequencies. In view of this, most lock-in amplifiers incorporate an automatic changeover system which selects control circuitry with longer response times for reference frequencies below about 50 Hz.

If operation is required at very low frequencies, for example down to 0.1 Hz, the user can often increase the response time of the loop still further by switch-selecting a ‘slow’ or ‘low-frequency’ reference mode. The result is a considerable improvement in low-frequency phase accuracy obtained at the expense of a still more sluggish response. The provision of a switch together with an automatic changeover point ensures that the system can be made to approach its optimum performance in any frequency range of interest, but it must be expected that offset errors and noise in the control loop will always conspire to give a phase shift differing from that indicated by the phase dial. In practice, such errors might be insignificant or gross as demonstrated by the following extreme modes of operation.

The first concerns signal recovery work. Here the null-shift procedure defined in Section 3.5 can be applied to a variety of signal types and overcomes the need for a continuous phase adjustment with more than a nominal calibration. In this case a continuous phase adjustment in conjunction with a calibrated *increment* of 90° is sufficient to reach an optimum detection condition. The phase-sensitive detector is, moreover, relatively insensitive to small phase changes when adjusted for maximum output. The overall system is thus tolerant of noise in the reference control circuits and an error of a few degrees in the 90° phase increment would not seriously affect the response to a noisy signal.

The second category of measurements includes all those where the phase-shift of a signal is to be determined relative to the applied reference or where the lock-in amplifier is required to respond to small phase increments. The reference phase-shift should now be defined and calibrated in accordance with some agreed convention and the user expects to have specification limits on phase accuracy and on stability if small phase changes are to be resolved.

It was shown in Section 2.5.5 that the reference-channel phase-shift can be defined unambiguously for an applied reference voltage of any waveform. For the purpose of this section, however, we shall assume that both the signal and the external reference are of sinewave form. Besides removing any doubt which may remain about the validity of phase specification for non-sinusoidal signals, this approach also avoids problems arising from waveform distortion when non-sinusoidal signals are transmitted by a signal channel of finite bandwidth.

We shall begin by identifying the numerous components of a typical lock-in amplifier phase specification and then consider briefly the performance of a broad-band reference channel under swept-frequency conditions. The next step is to investigate how the phase accuracy of a lock-in system can be checked in practice. It turns out that the procedures involved are rigorous in the extreme and serve as useful guidelines to the more general problem of precision phase measurement with signals and systems. For this reason the principal sources of error in phase measurement are listed in a self-contained section, together with ways to minimize their effect. This last section enlarges on some of the more general properties of phase-sensitive detectors and lock-in amplifiers introduced so far, and shows how some key specifications can be brought to bear on a specific type of measurement problem.

4.5.2 Phase accuracy: points of specification

We suppose that a sinewave signal and an *in-phase* sinewave reference are applied to a lock-in amplifier. In any practical system the relative phase of the signal and reference switching waveform measured *at the phase-sensitive detector* will have several components which can be written in the form

$$\phi = \phi_R + \Delta\phi_R + \Delta\phi_s + \phi_N(t)$$

Here ϕ_R represents the phase-shift dialled on the front panel of the lock-in amplifier, $\Delta\phi_R$ is a static phase error associated with the reference control circuit and $\Delta\phi_s$ represents a residual phase-shift in the signal channel which we assume is set to its maximum bandwidth. The quantity $\phi_N(t)$ represents *phase noise* which is attributable to noise in the reference channel control circuit and has components distributed over a wide frequency range.

In principle, these individual sources of error could be evaluated separately, but this more likely to be done by the manufacturer who has access to the internal workings of the lock-in amplifier. Otherwise, it is difficult to devise a measurement which would isolate the two contributions $\Delta\phi_R$ and $\Delta\phi_s$. Fortunately, in any application where phase is an important factor, it is usually sufficient to have a measure of the total phase error given as the sum of a static error $\Delta\phi_R + \Delta\phi_s$ and a fluctuation component $\phi_N(t)$**.** We shall see how this information can be inferred from measurements carried out at the input and output terminals of the lock-in amplifier.

When comparing the phase specifications of competing equipments it is clearly necessary to check manufacturers' data sheets very carefully. In some cases the static phase errors might be specified separately; in others, the figures quoted could be for composite phase errors. In either case the specifications will be frequency-dependent and show a deterioration towards the extremes of the recommended range of signal and reference frequencies. Also, it is usual to specify only the maximum anticipated value of static phase error at any spot frequency. If more detailed information is required, for example the phase error corresponding to a particular phase setting, ϕ_R, it is up to the user to devise his own measurement procedures following the guidelines given in later sections.

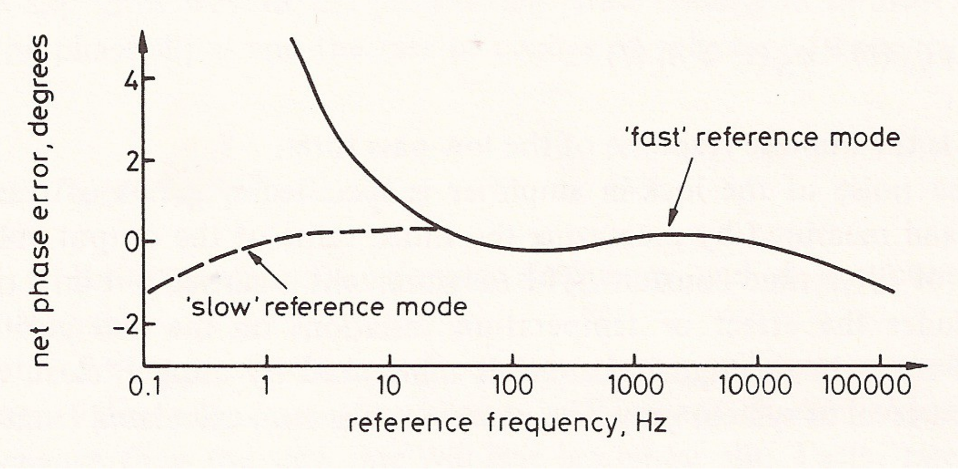


Fig.4.6 Static phase error of a typical broadband lock-in amplifier

Fig. 4.6 shows the composite, static, phase error of a typical broadband lock-in amplifier measured over the entire range of operating frequencies. This form of presentation, allied to a measure of the phase noise, is usually sufficient for all but the most demanding applications, but the figures given are strictly valid only at a quoted value of laboratory temperature. A full specification will also give a measure of the anticipated phase *drift* as a function of temperature. It turns out that the phase drift is usually far more important than the phase noise in measurements involving the detection of small phase increments in a signal.

4.5.3 Phase noise (phase jitter) and phase drift

Phase noise (or phase jitter) is specified in such a way that its effect on precision phase measurements can be easily predicted.

We suppose that the lock-in amplifier is supplied with a stable sinewave signal and reference at the reference frequency of interest. The reference phase-shifter is then adjusted to null the output from the phase-sensitive detector. The effect of reference channel phase noise is to give a residual fluctuation measured from the null point. Using the notation of the last section the net phase shift at the null point corresponds to a value

$$\phi = \phi_N(t) \pm \pi/2$$

measured at the phase-sensitive detector.

When the total error due to noise is much less than one radian the phase-sensitive output of the lock-in amplifier approximates to (Section 4.1.2)

$$v_N(t) = V_F(V_s/S_F)\ \phi_{NF}(t)$$

The variation $\phi_{NF}(t)$ represents the low-frequency components of $\phi_N(t)$ which are transmitted by the output low-pass filter. Following the approach used in Section 2.4, these are given by the convolution

$$\phi_{NF}(t) = \phi_N(t) \otimes h_L\,(t)$$

where $h_L(t)$ is the impulse response of the low-pass filter.

The phase noise of the lock-in amplifier is specified in terms of this filtered fluctuation and measured by measuring the r.m.s. value of the output voltage at a stated value of filter time constant. The measurement is carried out on a time scale which precludes the effect of temperature variations on the null position. The influence of noise in the signal channel is minimized by using a ‘clean’ sinewave signal at a low level of system gain. The phase noise is then calculated from

$$[\phi_{NF}(t)]_{\text{r.m.s.}} = (S_F/V_F V_S)\ \times\ [v_N(t)]_{\text{r.m.s.}} \times\ 180/\pi$$

Here, the factor 180/π gives a conversion from radians to degrees. This conversion brings the specification to a standard form**,** a typical specification being:

phase noise: 0.01° r.m.s. at a reference frequency of 1 kHz,
0.1 s time constant at 12 dB/octave.

In general, the r.m.s. phase noise will increase for reference frequencies at the extremes of the lock-in amplifier frequency range but its influence on the final output can always be reduced by switching to longer time constants. The effect of phase noise is to introduce noise even when, for practical purposes, the output from the signal channel is noise free. In any event, its effect, like that of phase drift, is both phase-sensitive and proportional to signal amplitude and maximized when signal and reference are brought into quadrature at the phase-sensitive detector.

Phase drift is recognized as a temperature-dependent deviation from the quadrature null point, observed under conditions where the instability of the phase-sensitive detector has negligible effect.

If any doubt persists, then the output contribution due to phase drift will be recognized by its variation with sensitivity setting while phase-sensitive detector drift can be observed under zero signal conditions, being dependent only on the output or expand gain.

The phase drift of a good-quality lock-in amplifier is of the order 0.03°/K. It is shown in Section 4.7.5 that the phase drift will usually make a far larger contribution to measurement errors than phase-sensitive detector instability when a lock-in amplifier is set up for precision phase measurement.

4.5.4 Reference channel slew rate

If a signal and reference are applied to a lock-in amplifier and the frequency is changed, the inertia of the reference channel control system will result in an instantaneous phase error known as phase 'slip'. When the frequency stops changing the phase slip gives way to the pre-existing phase condition. In most commercial systems the phase slip θ and the rate of change of reference frequency are related by:

$$\mathrm{d}f_R/\mathrm{d}t = Af_R\theta$$

where f_R is the instantaneous reference frequency and A is a constant of the reference channel.

The rate of change of reference frequency corresponding to a phase slip of 5° is a frequency-dependent quantity known as reference channel *slew rate.* Slower rates of change than the slew rate will give less phase slip. Faster rates of change give increased phase slip up to the point where phase control, and hence reference channel lock, is lost completely.

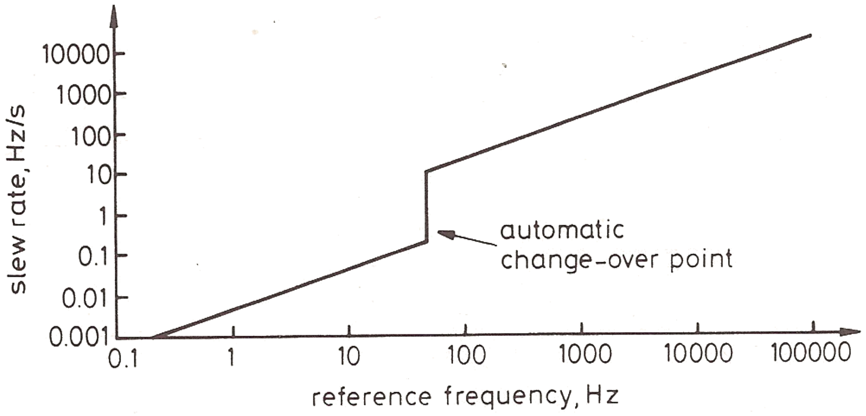


Fig. 4.7 Slew rate specification for a typical broadband lock-in amplifier

The phase slip of 5° represents the maximum phase error which could be tolerated if swept-frequency meaurements are to be carried out to a reasonable level of precision. In most applications the slew rate is considered to be the maximum usable rate of change of reference frequency.

When slew rate is specified by means of a graph, the effect of an automatic changeover point in the reference channel is to divide the graph into two distinct regions as shown in Fig. 4.7. The system is then characterized by two values of the constant A, corresponding to operation with reference frequencies above and below the changeover point. In the light of the discussion given earlier we expect the slew rate to be much lower below the changeover point where the response time of the control system is larger. In Fig. 4.7, the constants differ by a factor of 100, which is typical of' many commercial systems.

The logarithmic dependence of slew rate on reference frequency means that slew rate can be translated into 'minimum sweep time per decade'. Broadband reference channels are usually characterized by slew rates in the range 1–5 seconds/decade in the frequency region above the changeover point.

4.6 Measurement of phase accuracy

4.6.1 Introduction

In measuring phase accuracy it may be sufficient to determine whether the relative phase of the signal and reference channels, as indicated on a front panel dial, is correct to within the limits specified by the manufacturer. In other cases, where phase is of critical importance, it may be necessary to make a detailed investigation of the relative phase-shift in order to catalogue system phase errors as a function of frequency and indicated phase.

Problems arise in practice because the phase difference of the two waveforms appearing at the phase-sensitive detector is not directly accessible in a typical lock-in amplifier. This essential information must therefore be inferred from a series of auxiliary measurements made at the input and output sockets of the instrument. The usual procedure is to apply strictly in-phase sinewaves to the signal and reference inputs and then to check that the output from the phase-sensitive detector follows the $\cos\phi_R$ law as the phase-shift ϕ_R is varied by means of the front-panel phase control. Ideally we should be able to interpret deviations from the $\cos\phi_R$ law in terms of system phase errors, but this can be done only when other sources of external error have been eliminated. As we shall see, the procedures involved are rigorous to the extent that a lock-in system of the highest precision can be made to appear second-rate if they are not properly applied.

The principal sources of error affecting phase measurements are given as follows:

(i) 'Trigger' errors in the reference channel

(ii) Signal and applied reference not strictly in-phase

(iii) Errors due to oscillator distortion

(iv) Errors due to phase noise, phase drift and phase-sensitive detector instability

Ways of minimising these various sources of error are dealt with under separate headings below. It is hoped that these final sections will provide a useful cheek list for those with a special interest in phase measurements.

4.6.2 Trigger phase errors

Trigger errors are incurred at the 'front end' of the reference channel, in the input trigger circuit. The relevant waveforms are shown in Fig. 4.8.

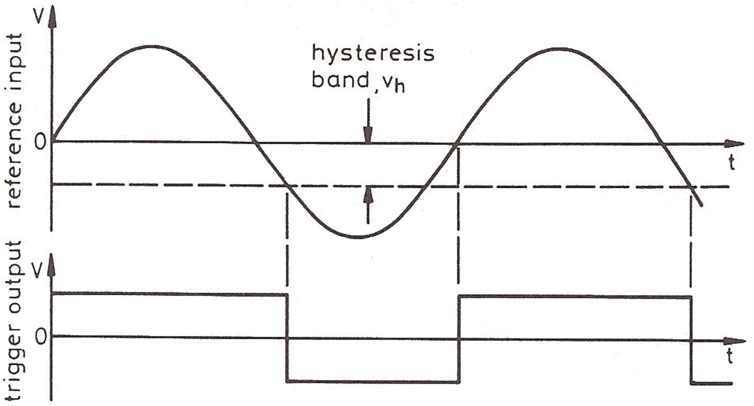


Fig. 4.8 Waveforms in a broadband trigger circuit

It is usual for the trigger circuit to be designed with a relatively large hysteresis band, v_{h} of the order of 100 mV. The input is invariably a.c. coupled and we shall assume that the phase-shift of the reference channel is to be generated with respect to the instant where the input waveform makes a positive zero crossing. Once 'fired', the output from the trigger circuit stays securely at a HIGH level and does not switch LOW until the input voltage falls below the hysteresis threshold. This is sufficiently removed from zero voltage to ensure that the trigger circuit is free from multiple triggering when the reference input is perturbed by low-level noise (amplitude up to $2v_{\mathrm{h}}$. peak-to-peak). When the reference input has relatively low amplitude, the output waveform will be highly asymmetrical. However, this is corrected in the phase-control system which follows the trigger circuit.

If phase accuracy is an important consideration we must assume that the reference waveform is free from noise. In a precisely adjusted trigger circuit the output will then accurately reflect the positive zero crossings of the input. In practice, however, we must allow for a small error, ε, in defining the zero point as illustrated in Fig. 4.9.

The phase error measured between the reference input and trigger output is simply

$$\phi_{\varepsilon} = \sin^{-1}\varepsilon/V_{\mathrm{R}}$$

$$\approx \varepsilon/V_{\mathrm{R}} \text{ radians}, \ \varepsilon << V_{\mathrm{R}}$$

Thus, for a trigger phase error of 0.1° or less we have the condition

$$V_{\mathrm{R}}/\varepsilon \geq 600$$

In a well designed circuit, ε can be held at a level of a few millivolts over the normal range of laboratory temperatures. For sinusoidal reference voltages, therefore, a peak-to-peak input of a few volts will be sufficient to ensure good trigger accuracy. Smaller inputs −of the order of 100 mV or so −will incur trigger phase errors of up to 1°, which is comparable with the expected phase accuracy of many systems. In fact, most manufacturers specify a reference level of 1 V r.m.s. for best accuracy.

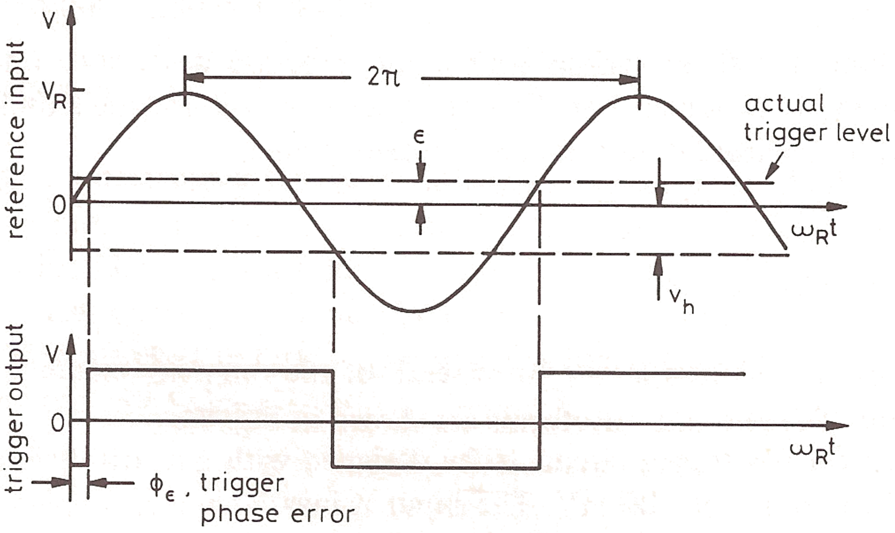


Fig.4.9 Trigger phase error

4.6.3 Defining in-phase signal and reference

The next problem is to ensure a strictly in-phase condition at the signal and reference inputs to the lock-in amplifier. Although oscillators are available giving ‘signal’ and ‘reference’ or ‘SYNC’ outputs, we cannot, in general, rely on these for the required phase precision. The only satisfactory way to achieve this is to derive both inputs from the same oscillator terminal using a resistive potential divider as shown in Fig. 4.10.

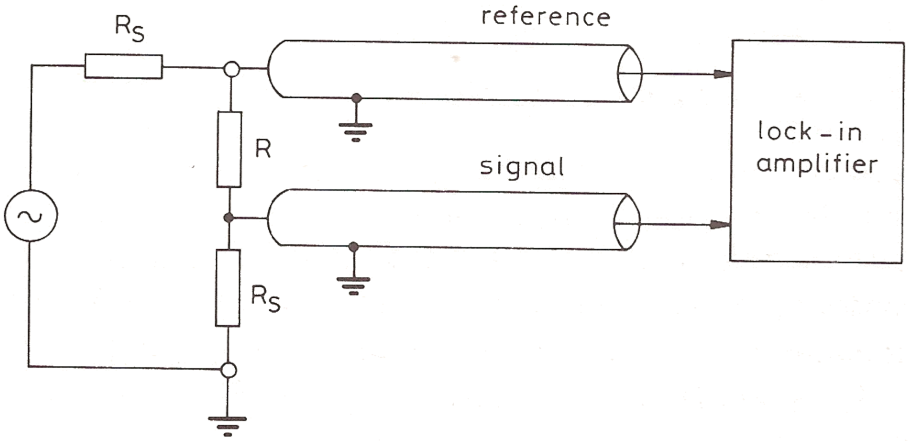


Fig. 4.10 Defining in-phase signal and reference voltages

In the arrangement shown it is assumed that the resistor R is much greater in value than the source resistance R_s and that R, in turn, is much less in value than the input resistance of both the signal and reference channels. The idea is that the screened cable connections to the lock-in amplifier are driven from sources of roughly equal resistance R_s. Provided that the screened cables are of equal length, differential phase-shifts due to cable capacitance will then be kept at a minimum .

Cable capacitance is typically of the order of 100 pF per metre run. When operating at a reference frequency f_R, the spurious phase-shift introduced by a cable of capacitance C connected to a signal source of resistance R_s is

$$\theta_e = \tan^{-1} -2\pi f_R C R_s$$

Suppose we have $C = 200$ pF with $R_s = 1$ kΩ at a frequency of 10 kHz. The cable-related phase-shift is then about 0.7°, increasing to over 7° at $f_R = 100$ kHz. Obviously an uncompensated phase-shift of this magnitude would be a serious source of error in any system set up for precision phase measurement.

4.6.4 Errors due to oscillator distortions

Let us now consider the requirements for the signal oscillator. If the accuracy of a lock-in system is to be measured in terms of its response to a sinewave as the system phase is varied, then every step must be taken to ensure that a strictly sinusoidal input is applied. We cannot expect the measurement system to follow the $\cos\phi_R$ law if the signal is grossly distorted; it follows that, for precision measurements, the distortion of the signal waveform must be correspondingly small.

The response of a phase-sensitive detector to a general periodic signal was considered in Section 3.5. From the results obtained there, it is obvious that the calculation of phase errors due to distortion components is a complicated business, suited more to numerical computation then to general analysis. Also to concentrate solely on the signal might distract attention from the reference which should similarly be free from distortion if misleading results are to be avoided.

To see why, let us turn to Fig. 4.11 which shows synchronous signal and reference voltages subjected to arbitrary harmonic distortion (greatly exaggerated for emphasis). The effect of the harmonic components is to shift the zero crossings of the reference voltage relative to those of its fundamental component. As a result, the 'zero-phase' switching waveform triggered from the positive zero crossings of the reference waveform exists in an arbitrary phase relationship with the fundamental component of either signal or reference.

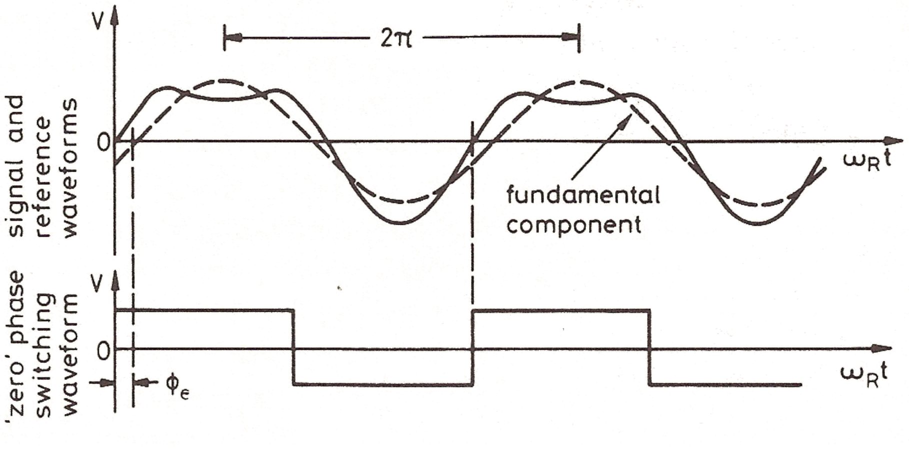

signal and reference waveforms
V
2π
0
$\omega_R t$
fundamental component
'zero' phase switching waveform
V
0
ϕ_e
$\omega_R t$

Fig. 4.11 Phase error due to waveform distortion

There is a common misconception that a 'fundamental only' responding lock-in system automatically compensates for harmonic distortion in phase measurements, but this is clearly not the case. It is true that, if the system is able to suppress harmonic responses, then the null-shift procedures given in Section 3.5 can be applied to bring the reference switching waveform directly into phase with the fundamental component of the signal. However, at this point the reference phase indicated by the phase-control dial will differ from zero degrees despite the precise synchronization of the externally applied voltages. The system will indicate an apparent phase error θ_e, which depends on the level of distortion, and the output will vary according to $\cos(\theta_R + \theta_e)$, where θ_R is the set phase of the reference channel.

When the lock-in system has harmonic responses, the null-shift procedures do not generally apply to distorted signals and the cosine dependence is no longer obtained. To arrive at a bound on the level of distortion which can be tolerated, we can take the special case of second-harmonic distortion where the positive zero crossings of a signal or reference are shifted by up to D_2/D_1 radians relative to those of the fundamental component. Here, D_1 and D_2 are the magnitudes of the fundamental and distortion components, with $D_2 < < D_1$.

Second harmonic distortion at a level of 1% can therefore contribute an apparent phase error of up to 0.01 rad or 0.5°. On this basis, and taking additional experimental evidence into account, it appears that a *total* harmonic distortion of 0.1% represents a suitable target for the signal/reference oscillator. Failure to achieve this target means that the oscillator contribution to phase error cannot be entirely ruled out: unfortunately this level of distortion is far below that of typical laboratory oscillators. Oscillators of the 'function generator' variety often have harmonic distortion in excess of 2% or 3% and are usually quite unsuited to applications where phase accuracy is a critical factor.

4.6.5 Phase noise, phase drift and phase-sensitive detector instability

When the output of a phase-sensitive detector is maximized by adjusting phase, the response is first-order independent of phase variations due to noise and drift in the reference channel. In this condition, the main sources of error in precision measurements will be offset and drift in the phase-sensitive detector, which must be minimized by using the detection system in a 'high stability' mode as defined earlier.

In most lock-in amplifiers an adjustment is provided to trim the offset of the phase-sensitive detector before measurements begin, leaving the drift component as the principal limitation on precision. In a good-quality system this can be as low as 5 to 10 p.p.m./K, equivalent to 10 μV/K or less in a 10 V output, independent of the a.c. gain provided before phase-sensitive detection.

Instability in the reference channel has its worst-case effect when measurements are carried out close to the point where the phase is adjusted for an output 'null'. The ability of the system to maintain a nulled output is then affected jointly by phase drift in the reference channel and output drift in the phase-sensitive detector.

If we suppose the system is first brought to a precise null condition, the subsequent variation of the output for a temperature change of 1 K can be expressed in the form

$$\Delta v_o = V_s(V_F/S_F)\phi_T \times \pi/180 + V_F\delta$$

where ϕ_T is the phase drift (degrees/K) and δ is the fractional output stability of the phase-sensitive detector operating with a maximum ouput voltage V_F at full-scale sensitivity S_F.

It was noted earlier that the contribution due to phase drift depends on signal magnitude and system sensitivity while the error due to phase-sensitive detector instability will be fixed for a given level of output gain. We shall assume that the latter is switched to its lowest possible value. The effect of phase-sensitive detector drift will then be negligible compared with the phase-drift contribution, provided:

$$V_s(V_F/S_F)\phi_T \times \pi/180 >> V_F\delta$$

In most phase measurements the ratio V_s/S_F is either close to unity or deliberately increased to the order of 10 or even 100 in order to enhance the phase-sensitivity of the system. Let us suppose that V_s/S_F is made equal to 10, giving a net phase sensitivity of 10 V_F volts/radian, while the phase drift is specified as 0.03°/K. If the drift of the phase-sensitive detector is not to dominate the final measurement it must satisfy the condition

$\delta << 0.0052$ (~ 5000 p.p.m./K)

We have seen that this criterion is easily satisfied with modern lock-in equipment. The conclusion is that phase drift in the reference channel is likely to make a much larger contribution to errors in phase measurement than phase-sensitive detector instability.

So far we have ignored the effect of short-ternm fluctuations due to phase noise. The time scale of these fluctuations is such that their contribution to the final output can always be reduced by increasing the output time constant. Usually the residual noise has minimal affect when setting the phase for a null output, its worst effect being to obscure the system response to small phase variations.

At reasonably large time constants the r.m.s. value of the output fluctuation due to phase noise is usually much less than the long-term deviation due to temperature related phase drift. In this case, the smallest incremental phase-shift which could be measured over a period of time would be comparable with the uncertainty due to phase-drift in the reference channel.

4.7 References

Discussion on the systems aspects of lock-in amplifiers is confined, for the most part to manufacturer's data sheets and application notes. See, for example:

1 'Specifying lock-in amplifiers'. Technical Note 116, Princeton Applied Research Corp., Princeton, NJ.

2 MUNROE, D.M. (1973): The heterodyning lock-in amplifier' Technical Bulletin, Ithaca Corp., Ithaca, NY.

Two-phase lock-in amplifiers

5.1 Introduction

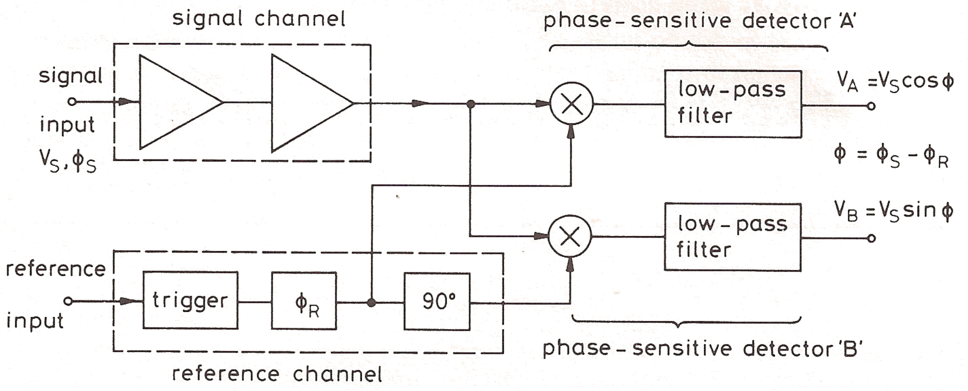


Fig. 5.1 Two-phase lock in amplifier

Two-phase lock-in amplifiers incorporate a pair of phase-sensitive detectors operated with quadrature reference waveforms as shown in Fig. 5.1. They were originally envisaged as a means of measuring the in-phase and quadrature components of a synchronous sinewave signal, a typical application being in network analysis as depicted in Fig. 5.2.

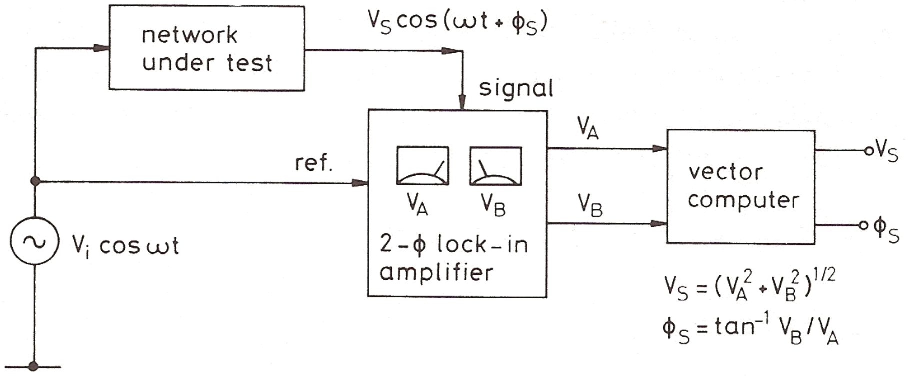

network under test
$V_S \cos(\omega t + \phi_S)$
signal
ref.
V_A
V_B
V_A
V_B
2-ϕ lock-in amplifier
vector computer
V_S
ϕ_S
$V_i \cos\omega t$
$V_S = (V_A^2 + V_B^2)^{1/2}$
$\phi_S = \tan^{-1} V_B / V_A$

Fig. 5.2 Two-phase lock-in amplifier used in network analysis. The vector computer converts the 'in-phase' and 'quadrature' outputs to polar form

For a lock-in amplifier with full-scale sensitivity S_F and output voltage swing $\pm V_F$ the output voltages V_A and V_B following the low-pass filters are:

$$V_A = V_s(V_F/S_F)\cos\phi$$

$$V_B = V_s(V_F/S_F)\sin\phi$$

where V_s is the r.m.s. value of the signal and the ratio V_F/S_F gives the scaling factor of the lock-in amplifier.

The 'vector computer' indicated in Fig. 5.2 has become a standard feature of commercial two-phase lock-in amplifiers. This is an electronic circuit which operates on the two output voltages V_A and V_B to produce voltages proportional to signal amplitude V_S and the relative phase ϕ, giving:

$$V_o = V_s(V_F/S_F) = (V_A{}^2 + V_B{}^2)^{1/2},$$

$$\phi = \tan^{-1} V_B/V_A$$

Two-phase lock-in amplifiers can thus display their outputs in either cartesian or polar form and are generally associated with measurements in two main areas. The first includes applications which make use of the quadrature components of a signal such as a.c. bridge measurements, Nyquist plotting and general impedance measurements. The second involves applications where signal magnitude is to be measured in the face of large phase variations and the polar form of the output is particularly valuable. Experiments using a swept-frequency signal and reference usually fall into this category.

The most powerful systems currently available are those which are capable of suppressing harmonic responses and so behave as if the signal is multiplied by a sinewave. These more advanced systems –to be described further in Chapters 8 and 9 –will respond only to the fundamental component of a periodic non-sinusoidal signal. In the case of two-phase systems with fundamental-only response, the amplitude of the fundamental component of a signal and its phase shift relative to the reference can be measured in a true 'vector' mode without ambiguity.

When the lock-in amplifier is of the conventional type and subject to the harmonic responses of the phase-sensitive detector, the application of a vector computer is only meaningful when the signal is of sinewave form. Also, we shall find that the vector computer falls short of the ideal when signals are very noisy, resulting in a severe limitation on dynamic range. There is, therefore, a good case for investigating two-phase techniques which extend the benefits of phase-tracking to non-sinusoidal signals while retaining the noise rejection inherent in the synchronous detection process.

It is significant that the majority of two-phase systems are catalogued as 'lock-in analysers' by their manufacturers, thereby emphasising their role in the analysis of both signals and systems. In support of this we shall be reviewing the operation of two-phase systems as wave analysers and spectrum analysers and in other applications where noise is not identified as a particular problem. This is in addition to a brief survey of some 'classic' two-phase applications. Features common to most of these applications are the need for wide dynamic range and the ability to operate with extremely high resolution in the frequency domain. We should also add 'cost- effectiveness' to this list. The relatively low cost of lock-in systems allied to their versatility and the possibility of computer-control (Chapter 10) increases their appeal as general-purpose measurement tools.

5.2 Examples of 'classic' two-phase applications

5.2.1 A.C. bridge balancing

The ability of two-phase lock-in amplifiers to simultaneously measure the in-phase and quadrature components of a low-level signal perturbed by noise makes them ideally suited to null detection on a.c. bridges. A typical experimental arrangement with a four-arm bridge is depicted in Figure 5.3. The bridge output is fed to the differential inputs of a preamplifier or directly to the lock-in amplifier if this has a differential input stage. The notes on signal connections and on the disruption of ground loops given in Appendix 6 are highly relevant to this particular application.

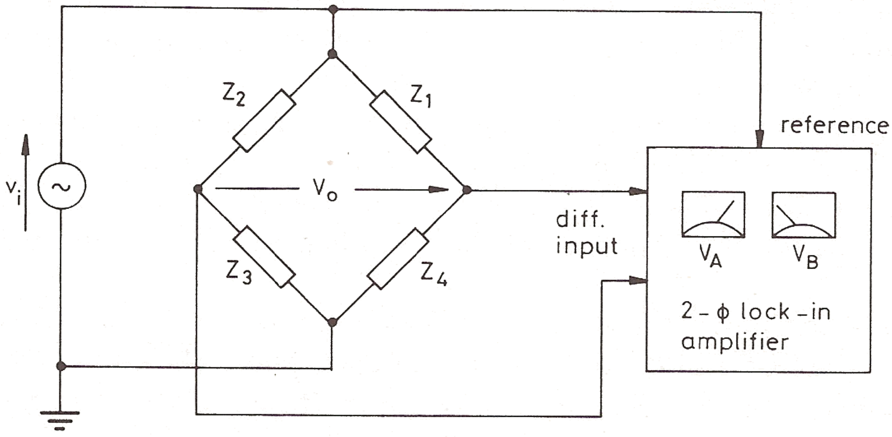


Fig. 5.3 **A.C. bridge balancing using a two-phase lock-in amplifier**

The lock-in amplifier is referenced directly from the bridge excitation source with zero phase shift in the reference channel. The in-phase and quadrature components of the bridge output voltage are measured with respect to the reference voltage and displayed directly on the twin outputs of the lock-in amplifier.

Using the notation of Fig. 5.3 the residual output voltage from the bridge is:

$$V_o = \left[\frac{Z_4}{Z_1 + Z_4} - \frac{Z_3}{Z_2 + Z_3} \right] V_i$$

Where Z_1 etc. represent the complex impedances of the bridge arms. The balance condition for a null output is now:

$$Z_1 Z_3 = Z_2 Z_4$$

And balance is only achieved when the real and imaginary parts of Z_1Z_3 and Z_2Z_4 are separately equal.

Putting the impedances in the form:

$$Z = R + \mathrm{j}X$$

we obtain the balance conditions:

$$\mathrm{Re}\ \{Z_1Z_3\} = \mathrm{Re}\ \{Z_2Z_4\}$$

$$R_1R_3 - X_1X_3 = R_2R_4 - X_2X_4$$

and

$$\mathrm{Im}\ \{Z_1Z_3\} = \mathrm{Im}\ \{Z_2Z_4\}$$

$$R_1X_3 + X_1R_3 = R_2X_4 + X_2R_4$$

We must avoid the pitfall of assuming that the in-phase and quadrature components of the bridge output can be separately and uniquely identified with each of the balance conditions. This has been responsible for abortive attempts to null one component independently of the other. Even worse is the incorrect assumption that the in-phase component can be nulled by balancing resistive elements and the quadrature component nulled by balancing reactances.* In general, we must expect that balance is obtained when the in-phase and quadrature components are both zero and that each component will only be zero when each of the null conditions is satisfied.

This rather pessimistic conclusion is tempered in practice by the relative ease of use of the two-phase lock-in amplifier in this application. No adjustments are required except to the sensitivity control which can be switched to register deeper and deeper nulls as balancing proceeds. This should be contrasted to measurement with a single-phase lock-in amplifier which would require continual switching between in-phase and quadrature components to cheek the outcome of the smallest adjustment.

Null detection makes full use of the noise-rejection properties of the lock-in system, but is subject to a source of error resulting from distortion on the bridge excitation oscillator. If the bridge is frequency selective it should be possible to achieve a null at the fundamental excitation frequency, leaving a residual voltage in the output made up from the distortion components of the oscillator. These are, of course, harmonically related to the excitation –and hence the reference – frequency. If these unwanted components are allowed to contribute to the output of the lock-in amplifier, an error will result in the determination of the null point.

Even when the excitation oscillator is of exceptionally high purity it is probable that non-linearities in the bridge elements will contribute measurable harmonic components in the bridge output. This applies particularly to bridges containing cored inductors and includes some important examples as the Maxwell, Owen and Carey-Foster bridge configurations.

The only solution is to use a lock-in amplifier with fundamental-only response, and this has become the preferred type of system in bridge balancing applications.

The extremely high sensitivity of lock-in systems, obtained through the use of external preamplifiers, means that in critical applications the level of excitation can be drastically reduced to avoid excessive dissipation in the bridge elements. The entire detection system including all connections to the bridge should be designed very carefully to avoid ground-loop problems, particularly if detection is envisaged down to a level of tens of nanovolts.

The scaling factor of lock-in amplifiers is such that a bridge offset voltage of this order could be made to provide a response of up to 10 V from the phase-sensitive detector. Such an output could then be used to provide a feedback signal for a control system; for example a temperature control system where the bridge balances about a 'set' point corresponding to a particular temperature.

5.2.2 A. C. impedance measurements

We shall use the example of semiconductor capacitance measurement which uses an experimental set-up similar to that shown in Fig. 5.4

The voltage source at the frequency of interest is arranged to have a very small output impedance by virtue of the attenuator and is coupled to the device under test by a large capacitor C_B. Application of a d.c. voltage V_{DC} causes C_B to charge and develop a ramp voltage V_{BIAS} which is applied to the device under test together with the a.c. excitation v_s. The device under test is terminated at the

* This misconception probably arises because of confusion with the method for a.c. impedance determination described in the next section.

virtual earth input of the current amplifier, and so sustains the full input voltage $v_s + V_{BIAS}$.

The signal current is now:

$$I_s = v_s\,[G_x + \mathrm{j}B_x],\quad B_x = \omega C_x$$

Which comprises components in phase and in quadrature with the excitation source.

When the phase of the reference channel is set correctly, the in-phase output of the lock-in amplifier becomes proportional to G_x and the quadrature output becomes proportional to B_x.

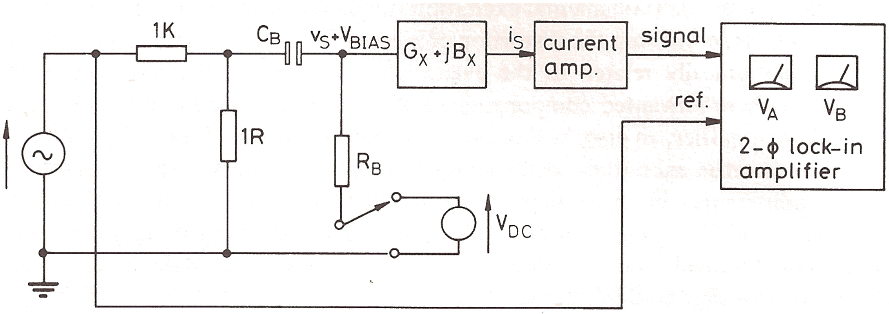


Fig. 5.4 Semiconductor capacitance measurement. Provision is made to apply a ramp bias voltage to the device under test

The system can be set up for phase and sensitivity as follows. First of all the output from the excitation source is adjusted to give as large an a.c. voltage as is allowable across the device under test. A suitable full-scale meter reading for C_x is chosen and a capacitor of this value is put in the test point. The reference phase is then adjusted to give no output from the in-phase phase-sensitive detector and the overall sensitivity is adjusted to give a full-scale reading on the quadrature output. The phasing can be recheeked by inserting a non-reactive resistor at the test point when it should be observed that the quadrature output is zero.

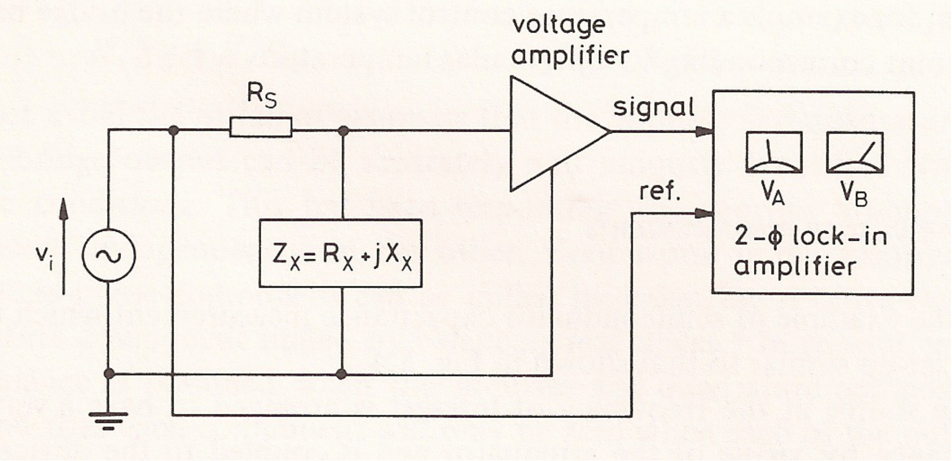


Fig. 5.5 Measurement of low impedances. The series resistor R_s is chosen so that $R_s >> | Z_x|$

An alternative approach is often used when the impedance to be measured has very low magnitude in the frequency range of interest. This is to supply the device under test from a source of relatively high impedance and to use a voltage amplifier as shown in Fig. 5.5. When the series resistor $R_s >> | Z_x|$ the device under test is supplied with substantially constant current v_i/R_s which is in-phase with the signal source. The voltage developed across Z_x is then, to a good approximation:

$$v_s = v_i\,(R_x + jX_x)/R_s$$

This method is most appropriate when accurate results are to he obtained with very low power dissipation in the device under test. The notes on amplifier noise matching given in Appendix 5 are relevant when the output voltages are particularly low since it is difficult to maintain a good noise figure when amplifying from a source of very low impedance. Fortunately, lock-in amplifiers are usually supported by a range of preamplifiers and matching transformers which enable a near-optimum noise match to be obtained under a wide range of signal and source conditions.

5.2.3 Phase measurements

We can identify three main types of application where the relative phase of a signal is to be measured, each of which involves the lock-in amplifier in a different set of measuring procedures.

The first is where the phase-shift of a sinewave signal is to be measured with respect to a reference voltage to a high order of precision. This can be achieved using the null-shift procedure introduced in Section 3.5, carried out with regard to the precautionary measures given at the end of Chapter 4. The null-shift procedure brings the signal and reference in phase at the phase-sensitive detector, whereupon the phase-shift of the signal can be read directly from the reference channel phase settings. Either a single-phase or a two-phase lock-in amplifier can be used in this type of measurement.

The second type of measurement involves the detection of small phase variations or small phase increments on a signal. The lock-in amplifier is then set up to operate as a linear phase detector by defining a ‘null’ condition with the signal and reference in quadrature at the phase-sensitive detector. The two-phase lock-in amplifier comes into its own in this type of application because the amplitude of the signal can be continuously monitored on the in-phase phase-sensitive detector while the phase variations of interest are monitored on the quadrature channel. The accuracy to which a two-phase system can maintain quadrature between the two phase- sensitive detectors is usually of a very high order, typically 0.1° at mid-band. In practice, any departure from ‘true’ quadrature can be compensated when nulling the output from the quadrature channel, leaving the system to register small phase increments to a high level of precision. Sources of error are the phase noise and phase drift of the reference channel and the output drift of the phase-sensitive detector. These must be taken into account as described in Section 4.6.5.

The general rule is to operate with the maximum possible value of time constant in order to minimize fluctuations due to phase noise and to operate at the best achievable output stability, that is with the minimum value of expand gain following the phase-sensitive detector. The overall sensitivity to phase variations can then be enhanced by increasing the a.c. gain of the system. This is a legitimate step to take in a system with large dynamic reserve; even though the output from the in-phase channel may now be greater than full-scale, the quadrature output will maintain a linear response to small phase variations provided the allowable input swing to the phase-sensitive detector is not exceeded. As demonstrated in Section 4.6.5, it is usually possible to overcome the drift of the phase-sensitive detector to give a measurement accuracy limited only by the phase drift of the reference channel. In extreme cases, entire lock-in systems have been enclosed in environmental chambers to stabilize operating conditions and minimize temperature-dependent phase errors.

The final type of measurement is the exclusive preserve of two-phase systems, where continuously changing phase angles in the range 0° to 360° are to be monitored without adjusting the controls of the lock-in amplifier. A two-phase lock-in amplifier with a vector computer is normally used in this situation; however, when the signal is particularly noisy, a vector tracking system of the type described in Section 5.4 often provides a more accurate measure of phase shift.

5.3 Noise limitations of the vector computer

The ability to measure the amplitude of signals independently of phase setting has proved to be so attractive that commercial systems have become available in which the vector magnitude provides the sole output. Aimed specifically at the spectroscopy market, these systems offer ease of use with the phase control entirely eliminated, and have become known as 'phase-insensitive' detectors or p.i.ds. There are, in addition, many research workers operating with conventional two-phase systems who habitually make use of the vector computer output in routine signal recovery applications as opposed to precision vector analysis. In view of this, it is appropriate to investigate the behaviour of the vector computer when the signal is noisy and subject to large variations in amplitude - as might be expected in a typical spectroscopy application. We, therefore, begin with an input to the lock-in amplifier of the form:

$$v(t) = \sqrt{2}\ V_s \sin(\omega_s t + \phi_s) + n(t)$$

where $n(t)$ represents wideband random noise which accompanies the signal.

The response of a synchronous detection system to random noise components is discussed in Appendix 3 and the effect of the harmonic responses of a phase-sensitive detector have been reviewed in Section 3.3. The net result is that a two-phase system produces output voltages

$$V_A = [V_s \cos\phi + n_A(t)]\ V_F/S_F$$

$$V_B = [V_s \sin\phi + n_B(t)]\ V_F/S_F$$

where $n_A(t)$ and $n_B(t)$ are uncorrelated noise voltages derived from the components of $n(t)$ which originate close to the reference frequency and its odd harmonics.

The prime advantage of synchronous detection methods over competing techniques is that the residual noise voltages which filter to the final output appear with zero average value and so make no net d.c. contribution to the overall response. As explained in Chapter 2, this is a result of the essential linearity of the synchronous detection process. Bearing this in mind, let us now consider the effects of subjecting the output voltages V_A and V_B to a stage of non-linear processing, using a vector computer. This provides an output voltage:

$$V_o = (V_A^2 + V_B^2)^{1/2}$$

The most noticeable effect of the vector computer is observed under zero or low signal conditions. The system output then becomes:

$$V_o = [n_A^2(t) + n_B^2(t)]^{1/2}\ V_F/S_F$$

The squaring operation ensures that V_o can take only positive values and is thus equivalent to a stage of rectification. The zero signal output is then a fluctuating unipolar waveform having a *non-zero* average value which we shall denote by the symbol Δ:

$$\Delta = \text{Ave. value } [n_A^2(t) + n_B^2(t)]^{1/2}\, V_F/S_F$$

It is not necessary to enter into detailed calculations to predict the general effect of this noise offset in the output. If we assume that the input noise is fixed while the signal is allowed to vary, then the input signal-to-noise ratio must be greater than some threshold value before the response to the signal begins to register in the final output. The system will be subject to gross measurement errors for signals close to the noise threshold; calculations then show that the average output voltage due to the signal must be greater than 5Δ before the vector computer is in error by less than $2^1/_2$%. The error decreases rapidly as the signal increases beyond this point, falling as the square of the signal.

When operating a spectrometer under noisy conditions the effect of a vector computer will be to introduce distortions on small features close to the baseline. An estimate of the noise offset, Δ, should always be made at the outset, therefore, to predict the range over which a linear response can be obtained. Fortunately, in many cases, this is easily achieved by temporarily shutting off the signal. If this cannot be done without losing the noise, an alternative method of estimating the noise offset is given by first of all estimating the fluctuation on either of the output voltages V_A or V_B. When operating with large time constants and correspondingly long response times this can be measured in terms of the peak-to-peak fluctuation of the output meter; otherwise it will be necessary to observe the fluctuation by connecting V_A or V_B directly to an oscilloscope. The value of the noise offset is then given to a good approximation by:

$$\Delta = V_{p\text{-}p}/4$$

where $V_{p\text{-}p}$ is the observed peak-to-peak fluctuation in volts. If the resulting value of Δ is judged to be too high compared with the smallest value of output voltage then the most obvious step is to reduce the fluctuations n_A and n_B by increasing the time constants on both phase-sensitive detectors. It will be recalled that the r.m.s. fluctuation is reduced in proportion to the square root of the time constant; hence increasing the time constant by a factor of 10 brings a reduction of $\sqrt{10}$ in Δ.

In cases where the maximum allowable time constant is established by considerations such as spectrometer scan time, it is common practice to accept a relatively poor output signal-to-noise ratio and then to introduce a stage of post-detector averaging. In most cases this consists of crude averaging ‘by eye’ of output chart records, while in others, use is made of signal averagers and waveform ‘eductors’ which store and average the results of several successive scans to a high degree of precision. Unfortunately, even these relatively sophisticated techniques cannot restore the loss of dynamic range incurred through the use of a vector computer. In general, the success of any averaging technique depends upon the noise having zero average value. It follows that the best place for such equipment is directly on the output of the phase-sensitive detector. In fact, this should present no problem in spectroscopic measurements where signal and reference usually exist in a fixed phase relationship at all times.

5.4 Vector tracking

When operating a broadband harmonically responding lock-in amplifier the use of a vector computer is restricted to sinusoidal signals. Also, if the signal is very noisy and the maximum allowable time constant is restricted by external factors (such as spectrometer scan time), the vector computer can impose a limitation on dynamic range.

We shall now investigate a mode of operation known as *vector tracking* which overcomes the noise limitations of the vector computer while retaining the ability to monitor the magnitude of non-sinusoidal signals independently of reference phase setting. The method is, in effect, an automated version of the null-shift procedures discussed in Section 3.5 and uses a two-phase system in the configuration illustrated in Fig. 5.6.

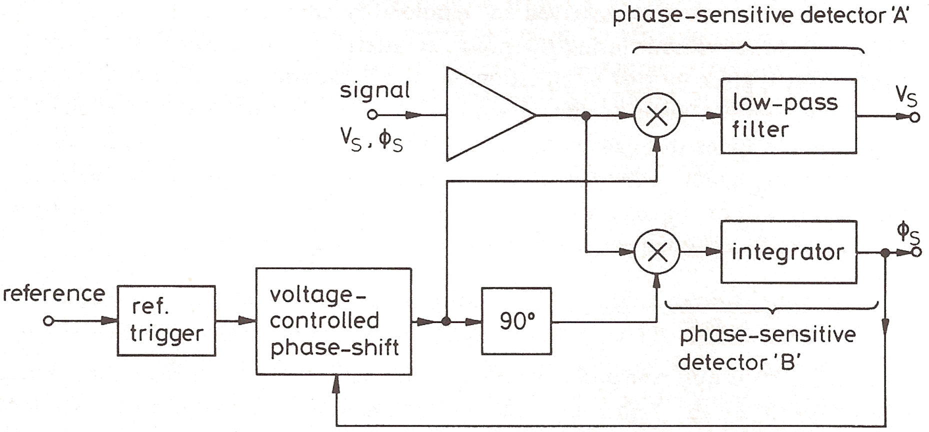

phase-sensitive detector 'A'
signal
V_S, ϕ_S
low-pass filter
V_S
integrator
ϕ_S
reference
ref. trigger
voltage-controlled phase-shift
90°
phase-sensitive detector 'B'

Fig. 5.6 Vector-track configuration

The arrangement operates as a closed-loop control system in which the output of phase-sensitive detector B is integrated and fed back to control a voltage-controlled phase-shifter in the reference channel. The control conditions are such that the net input to the integrator is zero at all times, which implies that the phase-shift is automatically adjusted to enforce a null condition at the output of phase-sensitive detector B. The reference phase at phase-sensitive detector A is displaced by 90° and will always yield a maximum response for signals which are 'symmetrical' in the sense defined in Section 3.5. Otherwise, the response will not necessarily be a maximum.* The best that we can say in the general case is that the phase is controlled to a well-defined condition which ensures that a consistent measure of signal magnitude can be obtained under adverse conditions, irrespective of the initial phase setting of the lock-in amplifier.

Vector tracking systems can accommodate variations in signal phase-shift up to a limit set by the voltage-controlled phase-shifter, usually ±100°. When the phase-shifter follows a linear law over its full range, the internal feedback voltage becomes directly proportional to the phase offset of the reference channel and can be scaled and displayed on a meter calibrated in degrees. If the signal requires a phase offset greater than the range of the phase-shifter it will be necessary to change quadrants by switching increments of 90° from the front-panel phase controls. The phase offset can then be obtained by adding the set phase of the reference channel to the phase indicated by the output meter.

The capacitive feedback of the control-loop integrator is usually switched from the time-constant control of phase-sensitive detector B which should now be treated as a means of adjusting the transient response of the overall system. The control loop contains one integrating element which is readily identifiable, but in

* Unless of course the system is fundamental-only responding. In this case, a control loop based on the null-shift procedure will always maximize the response to the fundamental component of a non-sinusoidal periodic waveform.

examining the system response we must also take into account the behaviour of the voltage-controlled phase-shifter. This cannot respond instantaneously to changes in the feedback voltage and will be characterized by a time constant which enters the control loop equations. As a result, the overall system has a response which is at least second-order and liable to exhibit oscillatory transients following abrupt changes in signal or reference-channel phase-shift. The erstwhile time-constant control on channel B is important in this respect as it can be used to control the damping of the system and adjusted to give a response free from excessive over- shoot, increased damping being equivalent to a longer time-constant selection.

The amplitude of the signal also influences the behaviour of the control loop since any change in amplitude is equivalent to adjusting the internal gain of the system. The effect is to bring about an increase in damping coupled to an increase in response time as the signal is reduced. If the transient response of the system is important in a particular application, the recommended procedure is to adjust the damping at maximum signal strength knowing that the damping can only increase* when the amplitude is reduced in the course of an experiment.

A vector tracking lock-in amplifier can thus ensure that the detection system is brought to a well-defined condition for all types of periodic signal. This is achieved with minimal effort on the part of the user and without generating spurious offsets at the output due to rectified noise components. Vector tracking can be advantageous under many circumstances, but it must be remembered that if a non-sinusoidal signal suffers large phase variations in the experiment under investigation then it is possible that the waveform will also change and that the overall calibration of the measurement system will vary accordingly. The only recourse here is to use a lock-in system which responds only to the fundamental component of a synchronous signal. This would include systems using a signal channel with a tuned filter which were discussed in Section 4.4. In this case the vector track mode serves to compensate phase variations caused by signal frequency drift relative to the centre frequency of the tuned filter.

5.5 Asynchronous operation

5.5.1 Introduction

Up to this point we have assumed that the two-phase lock-in amplifier is operating with the signal and reference precisely synchronized. In this section we shall be relaxing this restriction and looking at the possibilities of using a two-phase system fitted with a vector computer in applications where the signal and reference are asynchronous. This will provide a background to the numerous applications where two-phase lock-in amplifiers are used as wave analysers and for high-resolution spectrum analysis.

5.5.2 Operation as a wave analyser

It has been established that the response of a synchronous detection system to an asynchronous sinewave signal has the form of a ‘beat’ component at the difference frequency of the signal and reference. In practice, the only beat components which survive to perturb the final output are those which correspond to a frequency difference less than the bandwidth of the output low-pass filter.

* These observations on response should be contrasted with these given in chapter 7 in relation to phase-locked loops. We will find there that damping is *reduced* along with the signal level. The difference is due to a modification of the integrator in phase-locked operation which gives independent control of damping and response time (loop bandwidth). This feature is not usually included in standard vector systems as supplied by manufacturers.

Suppose a lock-in amplifier with output voltage swing ± V_F and a full-scale sensitivity S_F is provided with a reference at frequency ω_R and an asynchronous signal:

$$V_s(t) = \sqrt{2}\, V_s \sin(\omega_s t + \phi_s)$$

where the phase angle ϕ_s is referred to some common time origin. When $\omega_s \cong \omega_R$ the final output will be sinusoidal at frequency $\Delta\omega = |\,\omega_s - \omega_R|$ with amplitude and phase determined by the frequency-response of the low-pass filter:

$$H_L(j\omega) = A_L(\omega) \exp j\theta_L(\omega)$$

$$A_L(0) = 1$$

The final output thus takes the form:

$$V_B(t) = V_s(V_F/S_F)A_L(\Delta\omega)\sin[\Delta\omega t + \phi_R - \phi_s + \theta_L(\Delta\omega)]$$

If V_A and V_B are now applied to a vector computer capable of handling time-varying signals the system will indicate a signal magnitude:

$$V_o = [V_A^2(t) + V_B^2(t)]^{1/2}$$

$$= V_s(V_F/S_F)A_L(\Delta\omega)$$

This is a static response which can be maximized by carefully tuning the reference signal to the frequency of the asynchronous signal. If the latter represents a signal of interest, the two-phase lock-in amplifier provides a means of amplitude determination when a synchronous reference signal is not available. This approach offers an advantage over conventional, heterodyning, wave analysers in that the detection bandwidth is determined by the properties of a low-pass filter and can consequently be made very small indeed.

In deriving the response it was assumed that the quadrature phase-sensitive detectors were supplied with identical low-pass filters. Indeed, if these are not perfectly matched, the output of the vector computer appears modulated at frequencies related to the difference frequency $\Delta\omega$. Also, as we have remarked, the vector computer must be able to give a true response to time-varying signals. The usual design target is for a bandwidth in excess of 10 kHz, enabling operation over a wide range of filter time constants.

5.5.3 High-resolution spectrum analysis

In the present discussion the only distinction we wish to make between a wave analyser and a spectrum analyser is that the latter is furnished with a swept-frequency reference and is usually coupled to some form of display such as an oscilloscope or chart recorder. A typical set-up is shown in Fig. 5.7.

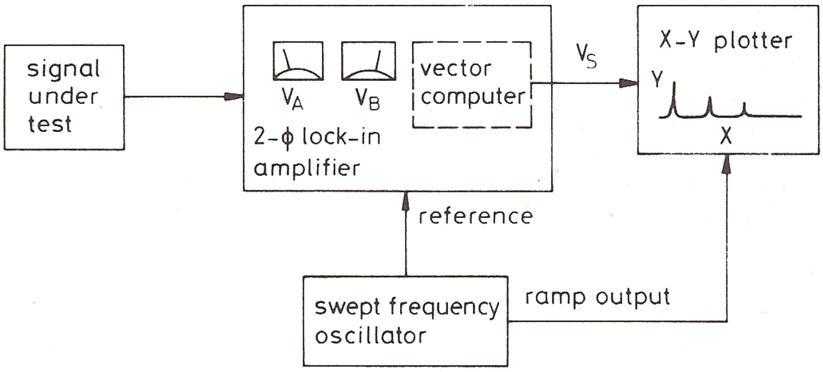


Fig. 5.7 Spectrum analysis using a two phase lock-in amplifier

When the signal of interest is characterized by a mixture of frequency components distributed over a wide frequency range the harmonic responses of a conventional broadband lock-in amplifier can have a serious affect both on the distribution and accuracy of the components in the final read-out. For this reason, it is virtually mandatory to use a system with fundamental-only response in this type of application.

In spectrum-analyser mode, the response to a single Fourier component of the signal has a 'line shape' equivalent to one of the lock-in amplifier transmission windows discussed in Section 3.3. The frequency resolution obtainable is thus equal to twice the bandwidth of the low-pass filter and is given by:

$$\Delta f = 1/(\pi T_o)$$

This result gives the −3dB resolution for a 6dB/octave filter with time constant T_O and the −6dB resolution for a 12dB/octave filter.

In deciding on the rate of frequency sweep in this and any other swept-frequency application in the slew rate of the reference channel defined in Chapter 4 must be taken into consideration. The main reason for using a lock-in amplifier is, however, that it represents a relatively inexpensive way of obtaining measurements with a very high resolution in the frequency domain. It is found, almost invariably, that this resolution −and hence the output filter bandwidth − sets the main limitation on the rate of frequency sweep; high resolution is only obtained at the expense of long observation times.

This is illustrated by Fig. 5.8. Fig. 5.8(a) demonstrates the ability of a two-phase lock-in system to resolve 10 Hz sidebands centered on a signal frequency of 10 kHz. The display was obtained at a resolution of 1 Hz and required a sweep rate of 0.2 Hz/s. The overall frequency sweep from 9975 Hz to 10025 Hz thus took approximately 4 minutes. The penalty of using too high a scan rate is illustrated in Fig. 5.8(b) for a single spectral line. As in all spectrum analyser work a rapid scan leads to a distorted display with the line peak shifted in the direction of the frequency sweep,

The oscillatory nature of the response under fast sweep conditions is characteristic of system using filters of low order, giving roll-offs of 6 dB or 12 dB/octave. The oscillatory effect is not generally observed on conventional spectrum analysers which use bandpass filters of much higher order, although the shift of the spectral peak is observed.

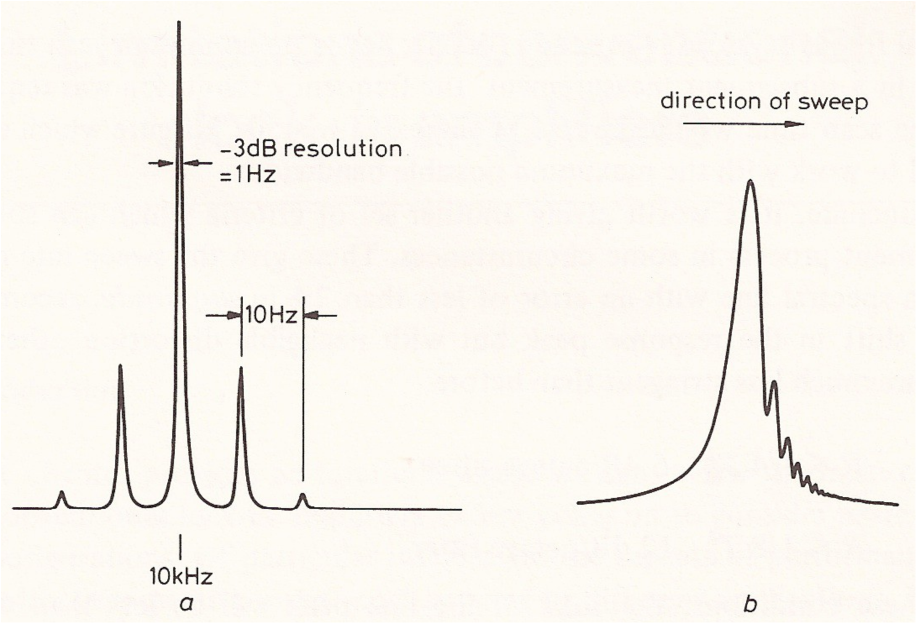


Fig. 5.8 (a) High resolution spectrum obtained with a two-phase lock-in amplifier; (b) line distortion resulting from a rapid scan

When deciding upon measurement bandwidth in a lock-in amplifier based system the maximum acceptable value should be taken, corresponding to the smallest acceptable value of output time-constant on the two phase-sensitive detectors. This is because the allowable sweep rate, R (Hz/s) has an inverse dependence on the *square* of the time constant (subject, of course, to the slew rate limitation of the reference channel not being exceeded). For a distortion-free display the sweep rate R should satisfy the criteria:

$R = 1/(50\ T_o^2)$, 6 dB/octave filter

$R = 1/(100\ T_o^2)$, 12 dB/octave filter

Note that the better frequency resolution of a 12 dB/octave filter demands a correspondingly longer sweep rate. To see the practical significance of these criteria suppose we wish to measure the frequency components of a signal appearing in the range 950Hz to 1050 Hz at a resolution of 10 Hz using a 6 dB/octave filter. The required value of time constant is:

$T_o = 1/(10\pi) \cong$ 30 ms

The recommended sweep rate for a distortion-free display is now:

$R = 1/(50\ T_o^2)$

$= 22.2$ Hz/s

The total frequency scan required is 100 Hz; hence the minimum scan time must be 4.5 s. If, in a subsequent measurement, the frequency resolution was required to be 1 Hz, the scan time would have to be increased to 450 s, a figure which emphasizes the need to work with the maximum possible bandwidth.

To conclude, it is worth giving another set of criteria which can speed up the measurement process in some circumstances. These give the sweep rate required to display a spectral line with an error of less than 2% in *amplitude*, accompanied by a small shift in the response peak but with negligible distortion otherwise. The criteria are much less stringent than before:

$R \leq 1/(4\ T_o^2)$, 6 dB/octave filter

$R \leq 1/(8\ T_o^2)$, 12 dB/octave filter

Under these conditions, the shift of the line peak will be less than the system resolution, giving an overall accuracy quite adequate for preliminary measurements.

5.6 References

References and further two-phase applications:

1 'The automatic measurement of semiconductor junction capacitance'. Application Report 5, Brookdeal Electronics, Bracknell, England

2 'Bridge balancing and the two-phase lock-in amplifier'. Application Note 104B; and 'Engineering applications of the lock-in amplifier'. Application Note 146. Princeton Applied Research Corp., Princeton, NJ

CHAPTER 6

Limitations of conventional lock-in systems

6.1 Introduction

This short chapter provides an interlude where we can review the main characteristics of conventional lock-in amplifiers before going on to consider more complex system configurations. Of particular interest are the aspects of performance which limit the effectiveness of conventional systems in different applications. If a list of these performance limitations were to be drawn up it would almost certainly include the following items:

(1) Dependence on the availability of a synchronous reference voltage

(ii) Trade-off between key specifications; for example, dynamic-range/linearity, dynamic-reserve/output-stability

(iii) Harmonic responses of the phase-sensitive detector appearing in the overall response of a lock-in amplifier

(iv) Slew-rate limitations in the reference channel

This list is not intended to be comprehensive, nor is it meant to imply that items (i) to (iv) are identified as shortcomings in every application. What we can say is that shortcomings in these areas have prevented conventional lock-in amplifiers from being adopted as general-purpose measurement tools and from performing tasks beyond their traditional role in signal recovery.

The absolute dependence of synchronous detection systems on a reference voltage would appear to be a prime candidate in this respect, yet, in practice, most research workers seem able to devise experiments where a ‘local’ reference is made available. There are, nevertheless, several examples of experiments where the signal source is remote from the detection system and a local reference must be generated by phase-locking an oscillator to the incoming signal.

An example in this category, the reception of satellite ‘beacon’ signals, is described in the next chapter. It should not be thought, however, that all phase-lock applications are of this ‘remote’ type. For example, in Fourier-transform photometry the traditional light ‘chopper’ is often replaced by a rotating optical grating. This may be so fine-ruled that the usual method of generating a reference by means of an auxiliary light source and phototransistor is impossible to apply. In this case a reference can be generated by phase-locking to the signal in the output of the experiment using the general arrangement illustrated in Fig. 6.1.

The problem of phase-locking to noisy signals merits a fairly extensive discussion in the next chapter where we shall see how phase-locked systems can be bust up using standard lock-in amplifiers and off-the-shelf modules.

Regarding the second item, it must be expected that some sort of trade-off will be encountered when any electronic instrument is operated at its performance limits. The trade-offs referred to in (ii) are inevitable when a conventional lock-in amplifier is operated with a broadband signal channel, but we have seen that the trade-offs can be improved when filters are used to eliminate unwanted components before detection. For example, the increase in signal-to-noise ratio obtained by filtering can be made equivalent to an increase in dynamic range; the noise-handling capacity of the system is improved without incurring a penalty due to increased offset and drift in the output. Alternatively, the filters could be used to reduce the total level of signal and noise at the input to the phase-sensitive detector with a possible increase in dynamic range sacrificed in favour of an improvement in linearity.

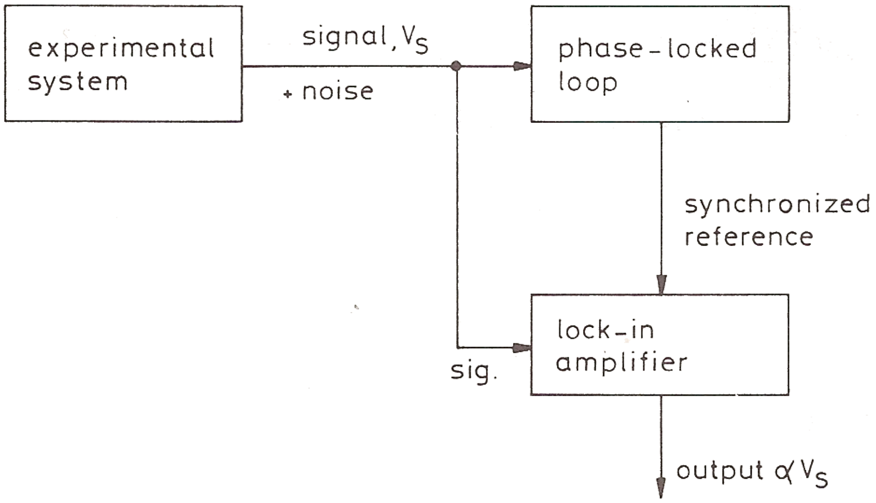


Fig. 6.1 General arrangement for generating a local reference in a lock-in recovery system. The phase-locked loop normally incorporates a second lock-in amplifier or phase-sensitive detector

In either case, the improvements attributable to the use of filters are obtained at the expense of wideband capability. Furthermore, the use of filters offers no solution to restoring the loss of dynamic range which is always observed when a phase-sensitive detector is operated towards the upper limit of its recommended frequency range. In general applications, therefore, there may come a point where the trade-offs can only be improved by moving to systems with inherently better performance. This might mean purchasing a conventional lock-in amplifier from a later generation than existing equipment. From the point of view of a designer working at the limits of available technology, it might be more appropriate to investigate techniques which enhance the dynamic range of ‘state of the art’ phase-sensitive detectors. Methods of achieving this last objective, by means of synchronous heterodyning, will be described in Chapter 8.

The remaining limitations arising from harmonic responses and slew rate can be summarized with reference to specific measurement difficulties that have come to light in earlier chapters. Some solutions will be found in Chapters 8 and 9 which deal with more complex lock-in systems operating on the heterodyne and pulse-width-modulation principles. Unlike the phase-locked systems discussed in the next chapter, heterodyning and p.w.m. lock-in amplifiers are almost invariably supplied and used as self-contained units and cannot, in general, be built up as modular systems. It will be found that some drawbacks of these more advanced systems are highlighted along with their inherent advantages. Certainly no single lock-in amplifier will address all the limitations identified at the beginning of this chapter; it is up to the user to define his own requirements and select the system most suited to his needs.

6.2 Limitations arising from harmonic responses

6.2.1 Susceptibility to interference

The response of a switching phase-sensitive detector is characterized by a set of transmission windows centered on the reference frequency and its odd harmonics. As discussed in Chapter 3, the higher-order windows have negligible effect on the final noise output when the signal is accompanied by broadband white noise, but can have the most serious effect when the signal is accompanied by narrowband noise, or by discrete interference components.

The situation is by far the worst when the lock-in amplifier is synchronized to a low-frequency reference signal at 100 Hz or less. For example, a reference frequency of 100 Hz gives rise to 499 transmission windows between 100 Hz and 100 kHz at, or near to which, the lock-in amplifier can respond to interference voltages. Moreover, at 100 kHz, where the relative magnitude of the harmonic windows is of the order 1/1000, the window frequencies are only 0.2% apart, leaving a very small margin for adjusting the reference frequency if large-scale high-frequency interference components are to be avoided.

The problem is particularly acute when operating photometric equipment incorporating infra-red detectors. Such detectors often exhibit considerable thermal inertia and so limit the usable chopping frequency to a rnaximum of 10 Hz or so. The measurement difficulties are often aggravated by the need for a matching transformer in the signal path to ensure a good noise match to the lock-in amplifier (Appendix 5). This renders the system susceptible to inductive pick-up and can result in relatively high-level interference at mains frequency. Because the chopping frequency is limited to about 10 Hz it is necessary to choose a value of this which does not have an odd harmonic within a few hertz of mains frequency.* Otherwise, measurements will be perturbed by a difference-frequency beat component in the final output. Although the beat can always be reduced by using a larger output time-constant, this may not always be acceptable in view of the increased response time of the measurement system.

The standard technique of suppressing harmonic responses by using a tuned filter in the signal channel of a lock-in amplifier was described in Section 4.4.2. From the discussion given there it is evident that a tuned filter offers only a partial solution to the present problem unless the stability of the chopping system (usually mechanical) is of a very high order. The use of a mains-frequency notch filter would usually provide a better solution to measurement difficulties in this case.

6.2.2 Ambiguity due to the harmonic responses

In a harmonically responding. system the output due to a sinusoidal signal at frequency f_R is indistinguishable from one with three times the amplitude and the correct relative phase at frequency $3f_R$. This ambiguity in the response is not so serious when the requirement is for signal recovery at a fixed frequency with synchronous signal and reference, but can give rise to misleading results elsewhere. For example, the problems with signal distortion in a.c. bridge measurements have been described in Chapter 5. Also in that chapter were examples of operation with asynchronous signals, where the lock-in amplifier is used for wideband spectrum analysis. In this case the harmonic responses. will give rise to spurious lines in the output spectrum which occur whenever an odd harmonic of the reference coincides with a frequency component of the signal

* 9.3 Hz and 11 Hz are popular choices of chopping frequency in infra-red spectroscopy for this reason.

under investigation. Clearly, the use of a tuned filter for harmonic suppression has no.relevance to this type of measurement where the frequency is changing continually.

A further example relates to the discussion on phase-locked loops in Chapter 7. If the phase detector in such a loop is provided by a phase-sensitive detector or lock-in amplifier with harmonic responses it is possible for the loop to lock securely at an odd *sub-harmonic* of the incoming signal frequency. If, in addition, the signal has squarewave rather than sinewave form, there will be a large number of additional frequencies where lock could be acquired. The effect here will be at its worst when the phase-locked loop is required to operate automatically. When the locking signal is very noisy it is recommended that the initial locking conditions are brought under manual control. This would normally be sufficient to ensure that spurious locking to harmonic components was avoided in signal-recovery applications.

6.2.3 Detection of non-sinusoidal signals

The limitation here refers to the problem of obtaining a maximum response to non-sinusoidal signals in the presence of noise. It was shown in Section 3.5 that there is no problem in principle with so-called symmetrical signals. Here, the null-shift procedures can always be applied to produce a maximum response which is first-order independent of errors in the reference-channel phase setting. Otherwise, the response will be less than maximum, resulting in a loss of output signal.to-noise ratio and an increased susceptibility to phase changes in the reference channel and in the applied signal.

The null-shift procedures can be applied with confidence to all types of periodic signal when the lock-in amplifier has fundamental-only response. The detection system is always brought to a condition where the fundamental components of the signal and reference are in phase at the phase-sensitive detector. These systems will also ensure that this condition is reached by the automatic vector tracking system described in Chapter 5 in relation to two-phase lock-in amplifiers, and when the lock-in amplifier is brought under computer control as described in Chapter 10.

In practice, the loss of sensitivity to asymmetrical signals is often marginal in conventional systems and the null-shift procedures offer the ultimate advantage in giving a phase setting which can be reproduced under the noisiest conditions. The problem of defining the phase-shift to give maximum response to an asymmetrical signal remains, however, and serves as another example of the measurement difficulties associated with harmonically responding systems.

6.3 Slew rate limitations

The specification of reference-channel slew rate was dealt with in Chapter 4. The loss of accuracy, or indeed loss of lock, that results from a rapidly changing reference frequency must be considered when setting up any experiment involving a frequency sweep at the reference input.

In most cases, when using time-constant settings of a few hundred milliseconds or greater, the primary limitation on sweep speed comes from the response of the output filter of the lock-in amplifier**.** This aspect was discussed in some detail in Chapter 5 when outlining the behaviour of a two-phase system used as a high-resolution spectrum analyser. When operating at lower resolution, however, say with resolution bandwidths of several hundred hertz, and with a relatively wide frequency sweep (covering, for example, the entire audio frequency range), the maximum allowable sweep speed will often be determined by the slew rate of the reference channel.

To take an example: suppose we operate with a 12 dB/octave output filter and a time constant setting of 1 ms. This gives a −6 dB frequency resolution of

$$\Delta f = 1/(\pi T_\mathrm{o}) \cong 300 \text{ Hz}$$

The maximum allowed sweep-rate to avoid errors greater than 2% in the output filter has been given in Chapter 5. We obtain

$$R_\mathrm{max} = 1/(8T_\mathrm{o}^2) = 125 \text{ kHz/s}$$

If it was required to sweep from 2 kHz to 20 kHz, the system would allow a sweep time

$$T_\mathrm{s} = 18.10^3/125.10^3 = 144 \text{ ms}$$

This short sweep time would allow the spectrum-analyser output to be displayed in conventional fashion on an oscilloscope. The problem is, of course, that a sweep time of 144 ms over the range 2 kHz to 20 kHz is equivalent to 72 ms/decade, a figure which is well beyond the slew rate capability of most conventional lock-in amplifiers. The reference processing circuits incorporated in pulse-width-modulated systems offer a solution to the slew rate problem as will be shown in chapter 9.

Phase-locking to noisy signals

7.1 Introduction

Phase-locked loops are used extensively in communications systems in demodulators for phase-and frequency-modulated signals[1-3] . They also form an important component in the reference channels of some conventional lock-in amplifiers and their more sophisticated counterparts to be described in Chapters 8 and 9. The requirement there is for frequency synthesis with the emphasis on precision and wideband capability. In the present context, however, we are specifically interested in the local generation of a reference waveform which is synchronized to a signal which may be obscured by noise. As explained in Chapter 6, this has relevance to all situations where it is not feasible to provide a direct reference connection.

An example of a 'remote' application is in the reception of satellite 'beacon' signals. Microwave beacons are transmitted at fixed power level while received power at an earth station is subject to variations related to atmospheric conditions on the propagation path. Detailed information about earth-space path attenuation can thus be obtained through long-term monitoring of the received signal[4].

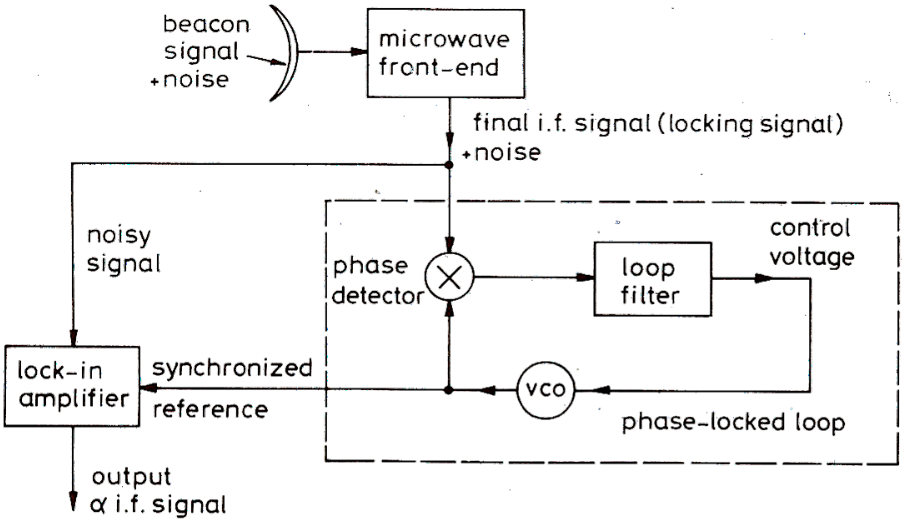


Fig. 7.1 Application of a phase-locked loop for the precision measurement of satellite beacon signals

On reception, the microwave signal is translated to successively lower frequencies and measurement of received signal strength is made at the output of the final i.f. stage. When the signal is subject to deep fading due to attenuation by rain on the propagation path, the output signal-to-noise ratio is seriously degraded and phase-sensitive detection provides the solution to precision measurement over a wide amplitude range. There is no reference directly available, so this must be obtained by phase-locking. In the most simple arrangement, shown in Fig. 7.1, a voltage-controlled oscillator is synchronized to the final i.f. signal. Usually, the beacon signal is unmodulated, which enables a very narrow bandwidth to be defined for the purpose of signal-to-noise improvement. In experimental systems the final i.f. is often chosen to be less than 1 MHz to enable the use of commercial lock-in amplifiers with wide dynamic range, both for phase detection within the loop and signal measurement.

Fig. 7.1 represents a simple form of coherent receiver in which the phase detector measures the instantaneous phase difference between the voltage-controlled oscillator and the i.f. signal, which – in this context – is referred to as the *locking signal.* The phase detector output constitutes an error signal which is filtered and fed back to the control input of the v.c.o. The noise-rejection properties of the loop are determined by the choice of loop filter which also serves to control the dynamic behaviour of the loop.

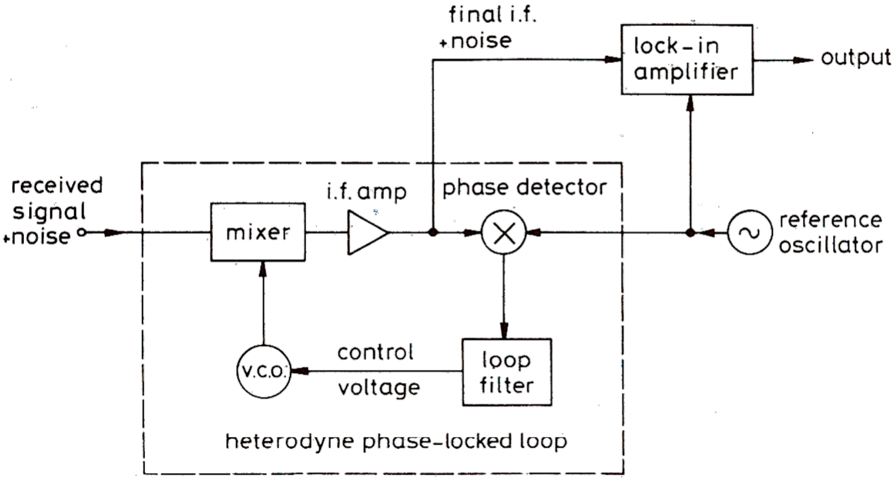


Fig. 7.2 Example of a heterodyne phase-locked loop. The loop is arranged to synchronize the output of the mixer to a stable reference oscillator

Many variations of this basic loop are used in receiver design[2,3]. Very often, the final mixer is incorporated within the loop to enable the final i.f. signal to be synchronized to a stable reference oscillator as shown in Fig. 7.2. This arrangement is widely used in satellite beacon monitoring[5]. Loops involving frequency changing are examples of 'derived' loops and broadly classified as *heterodyne* loops. A further example of a heterodyne loop used for frequency synthesis in a heterodyne lock-in amplifier is discussed in the next chapter.

The purpose of this chapter is to review the procedures whereby the basic loop of Fig. 7.1 can be made to operate successfully when the locking signal is very noisy and has variable amplitude. We shall find it possible to achieve synchronization when the locking signal is as much as 30 dB below the broadband noise level. However, to achieve such extremes of operation it is essential to understand the role of the loop filter in the locking process and to be able to select loop parameters in a systematic way. We shall therefore be looking at the mathematical background to loop operation and selecting some important results which have appeared in the literature in recent years.

In most mathematical treatments of the phase-locked loop the phase detector is modelled as an ideal multiplier. This is entirely suited to our purposes here since it enables us to extend our conclusions to systems using phase-sensitive detectors as phase detectors. In fact, multiplier-detectors are the preferred type when the locking signal is very noisy. Alternative types such as sequential or edge-triggered phase detectors are reserved for applications where the locking signal is free from noise; for example, in the reference processing systems mentioned earlier.

When describing the properties of the loop we shall maintain a clear separation between the multiplier phase detector and the loop filter, but we must recognize that this separation is not so easily achieved when both functions are incorporated in a fully integrated lock-in amplifier. In fact, many users find the problem of relating the specifications of lock-in equipment to the circuit blocks of an ideal loop model very difficult to overcome. We shall accordingly give due attention to this most important practical aspect and derive optimization procedures on the basis of ‘real’ equipment.

The literature relating to phase-locked loops is extensive but the general level of treatment is such that the non-specialist reader will probably have considerable difficulty in following the complexities of loop behaviour under conditions of extreme noise. He will also find that signal amplitude variations are often excluded from the analyses entirely. Both of these aspects are of key interest to one whose involvement in phase-locking is motivated by signal recovery, and we shall attempt to take them into account in the course of this chapter.

7.2 ‘Static’ analysis of a phase-locked loop

7.2.1 Phase detector output

We shall assume at the outset that the locking signal is noise-free and unmodulated and that the phase-locked loop has succeeded in pulling the v.c.o. frequency until it coincides exactly with that of the locking signal.

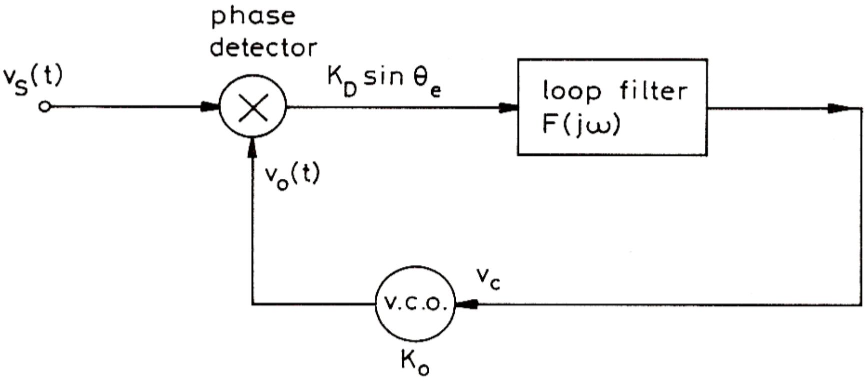


Fig. 7.3 Static analysis of a phase-locked loop
$v_s(t) = \sqrt{2}V_s\sin\omega_i t$
$v_o(t) = \sqrt{2}V_o\cos(\omega_i t + \theta_\varepsilon)$

We shall use a multiplier model for the phase detector and refer to the voltages in the loop identified in Fig. 7.3. On examining the loop in the locked condition we find an immediate consequence of using a multiplier-detector, namely that the loop must force a *quadrature* relationship between its two inputs to ensure proper operation (Section 2.4.4). We accordingly write the two phase detector inputs in the form:

$$v_S(t) = \sqrt{2}\ V_S \sin\omega_i t$$

$$v_O(t) = \sqrt{2}\ V_O \cos(\omega_i t + \theta_\varepsilon)$$

where θ_ε represents a small error measured from the ideal quadrature condition. Introducing a constant K, we obtain an output

$$v_D = KV_SV_O \sin\theta_\varepsilon$$

from the phase detector, which, in turn, supplies the feedback essential to loop control.

It is usual to express this result in the form

$$V_D = K_D \sin \theta_\varepsilon$$

Where K_D (volts/radian) measures the sensitivity of the phase detector. Since the amplitude of the v.c.o. is usually fixed we find that the phase detector sensitivity is proportional to the amplitude of the locking signal. In most of the 'standard' analysis which follows it will be assumed that K_D is fixed, but we must eventually consider the effect of variations in signal amplitude.

7.2.2 Static phase error

The notion of phase error must arise in a general description of any phase-control system. To calculate its magnitude we must first consider the operation of the v.c.o.

A linear model is to be used in which the frequency offset $\Delta\omega$ from the v.c.o. free-running frequency, ω_o, is proportional to the applied control voltage v_c. This gives

$$\Delta\omega = K_o v_c$$

where K_o is the v.c.o. gain factor with dimensions radians/V-s.

Under the essentially *static* conditions assumed here, when the amplitude and phase of the locking signal are free from variations, the v.c.o. offset frequency will be

$$\Delta\omega = \omega_i - \omega_o$$

which is maintained by a control voltage

$$v_c = K_D F(0) \sin \theta_\varepsilon$$

Here, $F(0)$ is the magnitude of the zero frequency response of the loop filter. The static frequency offset of the v.c.o. from its free-running frequency is therefore

$$\Delta\omega = K_o K_D F(0) \sin \theta_\varepsilon$$

The loop is usually designed to make the static phase error very small. In this case we can make the approximation $\sin \theta_\varepsilon \cong \theta_\varepsilon$ and so obtain

$$\theta_\varepsilon = \frac{\Delta\omega}{K_o K_D F(0)}$$

The static phase error can accordingly be reduced by increasing the d.c. gain of the loop, $K_o K_D F(0)$, and by initially setting the free-running frequency of the v.c.o. as close as possible to the incoming frequency, thus minimizing $\Delta\omega$.

7.2.3 'Hold-in' range

If we now suppose that the incoming frequency is subject to a slow drift, the loop will track and accommodate the frequency change by a change in the v.c.o. offset frequency. However, the results of the last section show that the maximum possible value of the frequency offset is

$$|\Delta\omega|_{\max} = K_0 K_D F(0)$$

corresponding to a phase error magnitude of $\pi/2$.

Beyond this point, the loop loses control and synchronization is lost. We accordingly define the *hold-in* range of the phase-locked loop:

$$\omega_H = 2K_O K_D F(0)$$

The hold-in range is proportional to the d.c. gain of the loop and in practice may be far greater than the range over which the static phase error might be considered acceptable.

7.3 Dynamic response

7.3.1 Introduction

Consideration of the static behaviour of the phase-locked loop does not involve any detailed specification of the loop filter. At this point all we can say is that the filter must be effective in suppressing unwanted components at the output of the phase detector and that a high value of gain at zero frequency makes an important contribution to reducing static phase errors and increasing the hold-in range of the loop. These must be major considerations in selecting a filter but the final choice must be consistent with an overall loop response which is stable and predictable. These considerations involve a study of the *dynamic* behaviour of the loop when locked to an incoming signal which carries phase modulation.

7.3.2 The loop equation

We begin with reference to Fig. 7.4 which shows the relevant voltages at different points in the phase-locked loop.

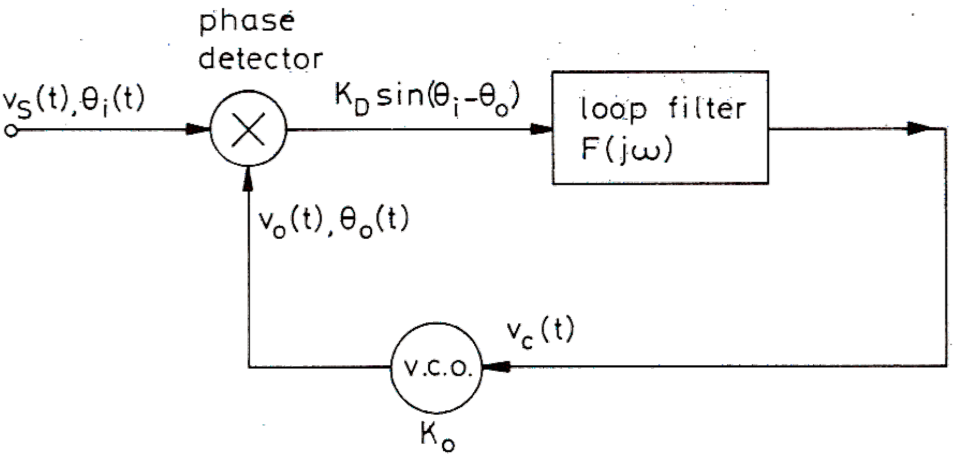


Fig. 7.4 Deriving the loop equation. The input to the loop is taken to be the phase variation of the locking signal

The loop is assumed to be in a locked condition with quadrature voltages at frequency ω_i applied to the phase detector. The v.c.o. phase is modulated by a variation $\theta_o(t)$ in response to the phase-modulation $\theta_i(t)$ carried by the locking signal. It is the nature of this response which is of interest here, in particular: how effective is the loop in 'following' the input phase variation?

The loop operates on the *difference* between $\theta_i(t)$ and $\theta_o(t)$, which produces a phase detector response

$$v_D(t) = K_D \sin[\theta_i(t) - \theta_o(t)]$$

We recall that K_D is strictly proportional to the amplitude of the incoming signal, which is assumed to be fixed throughout the following discussion.

The phase detector response is modified by the loop filter. The resulting output provides the control voltage, $v_c(t)$, to the v.c.o. and can be expressed in terms of the convolution:

$$V_c(t) = K_D \sin[\theta_i(t) - \theta_o(t)] \otimes f_L(t)$$

where $f_L(t)$ is the impulse response of the loop filter.

The resulting frequency deviation of the v.c.o. is proportional to $v_c(t)$, and for our purposes it is convenient to express this deviation in terms of the time derivative of the v.c.o. phase, giving

$$\Delta\omega(t) = \mathrm{d}\theta_o(t)/\mathrm{d}t = K_o K_D \sin[\theta_i(t) - \theta_o(t)] \otimes f_L(t)$$

This is the general non-linear equation describing the operation of the phase-locked loop. The non-linearity is inherent in the phase detector response, but can be over-come by assuming that the loop is designed to give a very small dynamic phase error. In this case, we write

$$\sin[\theta_i(t) - \theta_o(t)] \cong \theta_i(t) - \theta_o(t)$$

$$\text{for } |\theta_i(t) - \theta_o(t)| << 1 \text{ radian}$$

and so obtain the *linear* loop equation

$$\mathrm{d}\theta_o(t)/\mathrm{d}t = K_o K_D[\theta_i(t) - \theta_o(t)] \otimes f_L(t)$$

The conventional approach to this equation is to derive the transfer function relating the ‘input’ and ‘output’ phase variations. We accordingly take Laplace transforms of both sides:

$$s\theta_o(s) = K_o K_D [\theta_i(\mathrm{s}) - \theta_o(\mathrm{s})]\, F(s)$$

and rearrange to obtain the transfer function

$$H(s) = \frac{\theta_o(s)}{\theta_i(s)} = \frac{K_0 K_D F(s)}{s + K_0 K_D F(s)}$$

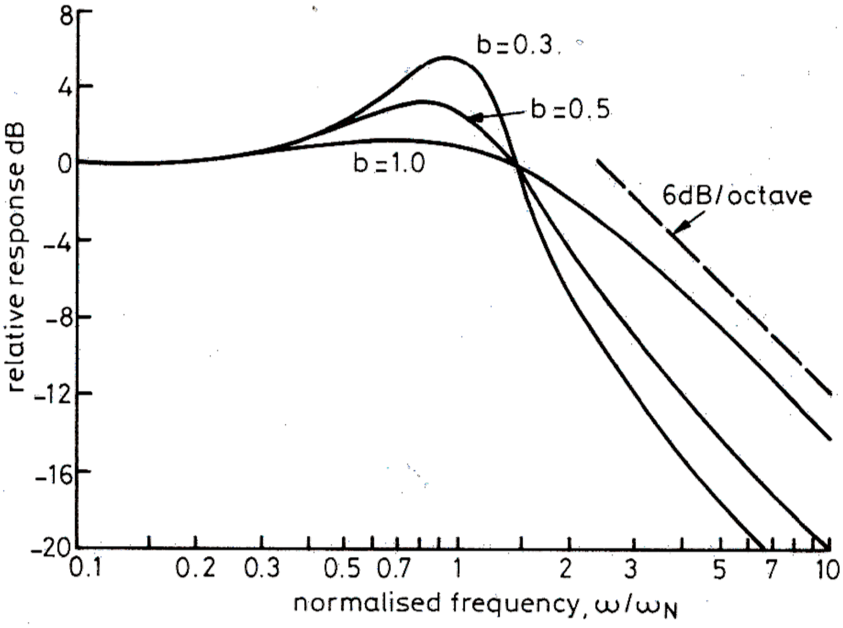


Fig. 7.5 Magnitude of a second-order closed-loop transfer function at various values of the damping ratio, *b*

Given $H(s)$ we can calculate the behaviour of the v.c.o. phase in response to variations in the phase of the locking signal. If we put $s = \mathrm{j}\omega$ we obtain the closed-loop frequency-response function $H(\mathrm{j}\omega)$ which could – in principle – be of any order, determined by the frequency-response function of the loop filter. In reality phase-locked loops are more often second-order, characterized by two parameters; natural frequency ω_N and damping ratio b. Variation of these parameters gives rise to a family of frequency responses with the magnitudes shown in Fig. 7.5. When the damping is light the loop is resonant and susceptible to oscillatory transients following abrupt changes of phase on the locking signal. As we shall see, it is important to have sufficient design variables to shape the response of the loop. The choice of loop filters in this respect is discussed below.

7.4 The second-order loop

A second-order loop can be obtained by using a loop filter of the simple *RC* low-pass type, shown as an active filter in Fig. 7.6.

The filter has a frequency-response function

$$F(\mathrm{j}\omega) = F(0)/(1 + \mathrm{j}\omega T_0)$$

$$F(0) = R_0/R_1, \quad T_0 = R_0C$$

resulting in a closed-loop frequency response function:

$$H(\mathrm{j}\omega) = \frac{K_0K_DF(0)/T_0}{K_0K_DF(0)/T_0 + \mathrm{j}\omega/T_0 - \omega^2}$$

This is the frequency response which would be obtained if a commercial phase-sensitive detector with a standard *RC* filter and 'time constant' control was used in a phase-locked loop. For the purpose of this and following sections we can put the denominator of $H(\mathrm{j}\omega)$ in the standard form

$$\omega_N{}^2 + 2\mathrm{j}b\omega\omega_N - \omega^2$$

and then identify the natural frequency and damping factor. In this case, we have

$$\omega_N = (K_0K_DF(0)/T_0)^{½}, \quad b = ½\,(K_0K_DF(0)T_0)^{½}$$

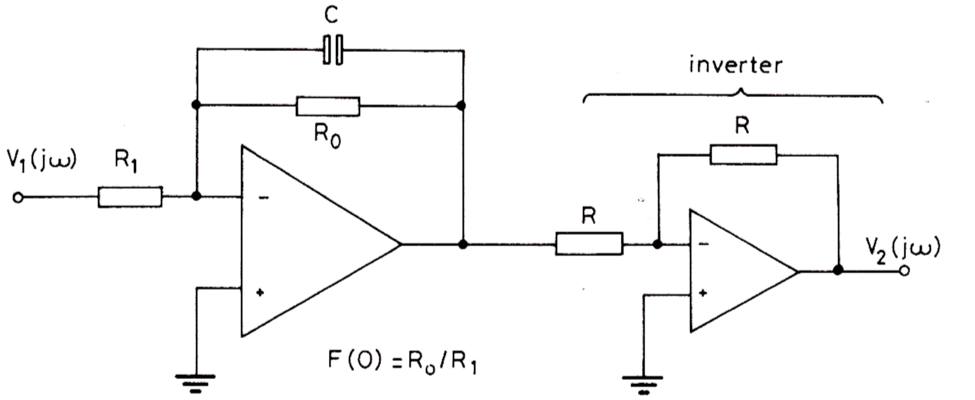


Fig. 7.6 Active RC low-pass filter. The inverter is included to correct for the phase reversal in the first amplifier. $F(\mathrm{j}\omega) = V_2(\mathrm{j}\omega)/V_1(\mathrm{j}\omega)$

The d.c. gain of the loop, $K_0K_DF(0)$ is correctly identified as the loop gain. We find that the natural frequency increases with loop gain while the damping is reduced. In fact, there are insufficient design parameters available to ensure that a particular combination of loop characteristics can be obtained. For example, at a given value of loop gain, the loop bandwidth can only be narrowed by increasing the time constant of the loop filter, resulting in a loss of damping. The transient response of the loop can thus be seriously degraded when narrow bandwidths are required.

These problems can be overcome by adopting the approach used by control engineers: to use a loop filter with a lag-lead response as in Fig. 7.7. This simple modification is effective because the additional resistor gives control of the damping of the phase-locked loop. Hence, for a fixed value of the loop gain, the natural frequency and damping can be set independently.

A variation on this filter type is shown in Fig. 7.8. We shall refer to this as an 'imperfect integrator'. Both lag-lead filters and imperfect integrators are used in commercial lock-in amplifiers adapted for phase-locking. The extremely high value of low-frequency gain obtainable with the imperfect integrator gives a phase-locked loop with wide tracking capability consistent with a low value of static phase error.

In most systems using lag-lead filters, however, the ratio R_0/R_2 is so high that there is relatively little difference in handling characteristics between the two filter types.

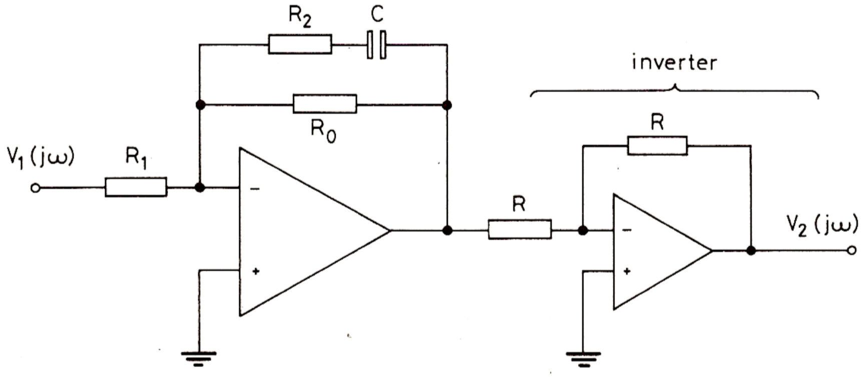


Fig. 7.7 **A lag-lead filter**
$F(\infty) = R_2/R_1,\ R_0 >> R_2$

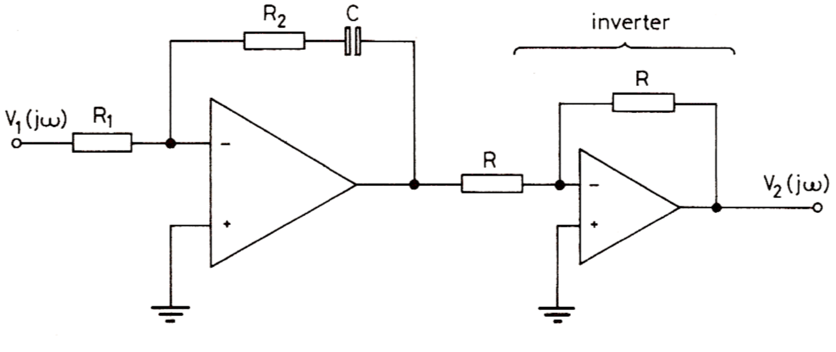


Fig. 7.8 **An imperfect integrator**
$F(\infty) = R_2/R_1$

As we shall see, the recommended procedures for optimizing the loop noise performance are the same in each case. We can adopt a uniform approach that deals correctly with both filter types by expressing the loop filter frequency response functions as follows:

$$\left.\begin{array}{l}\text{lag-lead filter } (R_0/R_2 >> 1) \\ \text{imperfect integrator}\end{array}\right\} F(\mathrm{j}\omega) = F(\infty)(1+\mathrm{j}\omega T_2)/\mathrm{j}\omega T_2,\ T_2 = R_2C$$

where in both cases, the high-frequency gain of the filter, $F(\infty)$ is a real ratio, R_2/R_1.

If the loop-frequency response function is now derived, we find, for both loop filters:

$$\omega_\mathrm{N} = (K_0K_\mathrm{D}F(\infty)/T_2)^{½}$$

and

$$b = ½\ \omega_\mathrm{N}T_2$$

The expression for b is exact for the imperfect integrator and an excellent approximation for the lag-lead filter, for all reasonable values of damping ($b^2 >> R_2/R_0$).

Gardner[1] has shown that for this type of filter the loop gain is correctly given by the quantity $K_0K_\mathrm{D}F(\infty)$.

The noise bandwidth of the loop becomes very important when we deal with the problems of phase-locking in noise. This can be expressed in terms of ω_N and b as:

$$B_\mathrm{L} = \frac{\omega_\mathrm{N}}{2}\left[b + \frac{1}{4b}\right]$$

For any value of ω_N the noise bandwidth is minimized when b = ½, and is then given by $\omega_\mathrm{N}/2$. Note that, following convention, the noise bandwidth is given in hertz, while ω_N is expressed in radians/s.

When using a lag-lead filter or imperfect integrator the natural frequency of the loop and the loop damping ratio exhibit a square-root dependence on loop gain. The practical significance of this becomes evident when the amplitude of the locking signal is allowed to vary in the course of a phase-locked experiment. We saw in section 7.2.1 that the phase-detector constant K_D is strictly proportional to V_s. If all the other loop components are held at a fixed value we obtain the following functional dependence on V_s for ω_N and b:

$$\omega_\mathrm{N} = c_1 V_\mathrm{s}^{½}$$

$$b = c_2 V_\mathrm{s}^{½}$$

where c_1 and c_2 are suitably defined constants. The variation of the noise bandwidth now takes the form:

$$B_\mathrm{L} = \frac{c_1}{8c_2}\left[1 + 4c_2^2 V_\mathrm{s}\right]$$

showing that B_L increases in direct proportion to the signal amplitude.

7.5 Noise and phase-locked loops

When the locking signal is accompanied by a random noise disturbance, noise enters the phase-locked loop via the phase detector and gives rise to a random phase error on the v.c.o. output. When the phase error is small it appears as phase noise or phase jitter on the v.c.o.[*] relative to the phase of the locking signal. In general, a small amount of jitter should not seriously affect an associated lock-in amplifier when the v.c.o. output is used as a reference voltage for measuring the amplitude of the locking signal.

If the noise input to the loop is allowed to exceed a certain level, the resulting phase error will become sufficiently large to seriously affect loop operation. In the extreme case the loop will drop out of lock completely, but before this point is reached there is a range of noise inputs where the loop is susceptible to a phenomenon known as 'cycle slipping'. In this regime the loop can be kicked temporarily out of lock by a random noise event and then restore its equilibrium an integral number of cycles away from its original condition. The effect can recur repeatedly when the noise levels are sufficiently high.

The analysis of loop operation in this regime is formidable, corresponding as it does to non-linear operation. Fortunately, our main concern is to ensure that the random phase error never reaches a point where there is a significant probability of cycle clipping. Only then can we ensure that the signal recovery system will operate for long periods of time without falling out of lock.

In principle, the random phase error due to external noise can always be reduced by designing for a loop frequency response function with small bandwidth. In this context, the noise bandwidth of the loop introduced in the last section becomes the relevant factor. Its role in evaluating the effect of noise in the input can be clarified as follows. We imaging that the phase-locked loop is a special kind of bandpass filter which accepts a noisy signal at frequency f_S. The output is a clean signal at the same frequency having a residual phase error due to noise. The filter has a frequency response function given by translating the loop frequency-response to the signal frequency f_S. This gives an effective noise bandwidth of $2B_L$ as illustrated in Fig. 7.9.

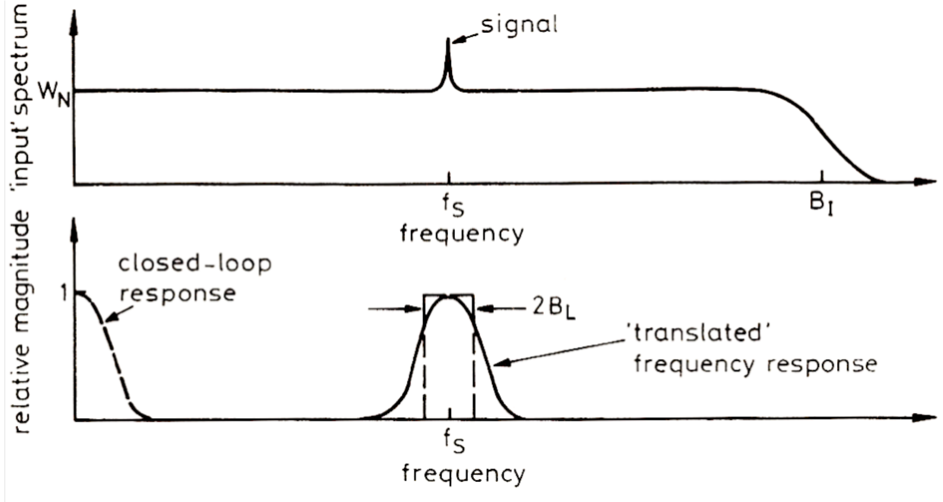


Fig. 7.9 Filter effect of a phase-locked loop
For the purpose of calculation, the loop can be regarded as a bandpass filter with noise bandwidth $2B_L$

The input noise has a bandwidth B_I with uniform spectral density W_N. The input signal-to-noise ratio is therefore

[*] The effect of phase noise *inherent* in the v.c.o. will be mentioned at a later stage. we shall generally assume that the noise in the loop, giving rise to a fluctuating phase error, is largely of external origin; that is, noise appearing with the locking signal.

$$SNR_\mathrm{I} = V_\mathrm{s}^2/(2B_\mathrm{I}W_\mathrm{N})$$

while the signal-to-noise ratio *measured within the effective bandwidth of the loop* is

$$SNR_\mathrm{L} = V_\mathrm{s}^2/(4B_\mathrm{L}W_\mathrm{N})$$

We shall not attempt to attach any physical significant to SNR_L which is often loosely called the loop signal-to-noise ratio. It can be shown nevertheless that the probability of cycle slipping will be very small provided that this noise measure is greater than about 6 dB (see, for example, the results reviewed by Gardner[1] and Blanchard[2]). We shall err on the side of safety when calculating the required value of B_L and use the criterion:

$$SNR_\mathrm{L} = V_\mathrm{s}^2/(4B_\mathrm{I}W_\mathrm{N}) \geq 10$$

In this condition the system is amenable to a small-signal analysis. The results in the literature[1] give a good estimate of the mean-square phase error[*] on the v.c.o. output, namely:

$$\overline{\theta_\mathrm{N}^2} = \frac{1}{2SNR_\mathrm{L}}$$

Let us now assume that the signal-to-noise ratio of the locking signal has been estimated and that the input bandwidth is known. We have:

$$SNR_\mathrm{L} = SNR_\mathrm{I}\, B_\mathrm{I}/(2B_\mathrm{L})$$

and, putting $SNR_\mathrm{L} \geq 10$, we obtain the condition:

$$B_\mathrm{L} \leq SNR_\mathrm{I}\, B_\mathrm{I}/20$$

to ensure that lock is maintained in the presence of noise on the locking signal. With this condition satisfied, the v.c.o. has a mean-square phase error

$$\overline{\theta_\mathrm{N}^2} \leq 0.05 \text{ radian}^2$$

The next step is to derive optimization procedures which guarantee that these conditions obtain even when the signal amplitude is allowed to vary over a wide range. This will represent a significant improvement over the ‘standard’ treatment where the signal – and hence the loop parameters – are fixed, and changes in signal-to-noise ratio are ‘arranged’ by allowing the noise intensity to vary.

It was noted earlier that we have taken no account of phase noise which is inherent in the v.c.o. itself. In designing phase-locked loops for high precision it is normally arranged that the loop bandwidth is wide enough to accommodate the bulk of these variations, and it can be shown that their effect is reduced in proportion to the loop gain of the system. This requirement is obviously incompatible with choosing a narrow loop bandwidth to combat noise of external origin. Also, it will be shown that having determined the bandwidth of the loop, consistent with a reasonably high damping ratio, the choice of loop gain is restricted. It turns out, however, that when operating in the audio-frequency range with loop components of reasonable stability, any residual jitter due to imperfections on the v.c.o. will be masked by noise on the locking signal, provided that the loop bandwidth is not made unnecessarily small. This is also an important consideration when the loop is expected to track ‘slow’ frequency variations on the locking signal as will be shown in Section 7.7.

[*] The mean square phase error is calculated on the assumption that the *static* phase error of the loop (see Section 7.2.2) has been trimmed to zero.

7.6 Optimization procedure

The optimization procedure will be derived for a classic signal recovery example, where the locking signal amplitude is varying over a wide range against a noise background with uniform spectral characteristics. It is necessary for us to have an estimate of the minimum value of the locking signal, corresponding to the worst-case signal-to-noise ratio at the input to the loop, $(SNR_{\mathrm{I}})_{\mathrm{MIN}}$.

When the loop incorporates a lag-lead filter or imperfect integrator, the noise bandwidth will be dependent upon signal level as described in Section 7.4. We accordingly design the loop to have its minimum noise bandwidth at minimum signal level. Using the result of Section 7.5, the minimum noise bandwidth is chosen to satisfy:

$$(B_{\mathrm{L}})_{\mathrm{MIN}} \leq (SNR_{\mathrm{I}})_{\mathrm{MIN}}\, B_{\mathrm{I}}/20$$

The behaviour of the loop as the signal increases from its minimum value can be predicted as follows.

First of all we note that the loop signal-to-noise ratio is given by:

$$SNR_{\mathrm{L}} = V_s^2/(4B_{\mathrm{L}}W_{\mathrm{N}})$$

Since B_{L} increases, at most, linearly with signal we find a steady improvement in SNR_{L} as the locking signal increases, causing a proportional reduction in the mean-square phase error (= $(2SNR_{\mathrm{L}})^{-1}$) due to external noise. The rise in noise bandwidth also makes the loop more effective in reducing the effect of phase noise inherent in the v.c.o. It is shown in section 7.4 that the loop damping will also rise with signal level. Fortunately, this rise is accompanied by an increase in the natural frequency ω_{N}. This joint behaviour results in a response which is not too ‘sluggish’ as would be the case if *b,* alone, were to increase.

Turning now to the question of loop damping: this would normally be chosen to have a minimum value of about ½ to ensure that the loop transient response is not marred by excessive overshoot and ‘ringing’. It also turns out that choosing this particular minimum value greatly simplifies the optimization procedures. The damping ratio falls with locking signal amplitude; we therefore arrange for a minimum damping ratio of ½ at the minimum anticipated signal level.

The general expressions for noise bandwidth and damping ratio are:

$$B_L = \frac{\omega_N}{2}\left[b + \frac{1}{4b}\right]$$

$$b = \tfrac{1}{2}\omega_{\mathrm{N}}T_2$$

Hence, if we put $b = b_{\mathrm{MIN}} = ½$ and decide on a minimum value for B_{L} we find that T_2 is given immediately:

$$T_2 = 1/(2B_{\mathrm{L}})_{\mathrm{MIN}}$$

while the minimum value of ω_{N} is given by

$$(\omega_{\mathrm{N}})_{\mathrm{MIN}} = 1/T_2 = 2(B_{\mathrm{L}})_{\mathrm{MIN}}$$

Using the value of the phase detector constant appropriate to minimum signal level, together with the required value of T_2, the loop can now be designed to give $\omega_{\mathrm{N}} = (\omega_{\mathrm{N}})_{\mathrm{MIN}}$ at $b = b_{\mathrm{MIN}}$. From the results given in Section 7.4 this implies that the minimum value of loop gain is given by

$$K_0(K_{\mathrm{D}})_{\mathrm{MIN}}\, F(\infty) = 1/T_2$$

T_2 has already been determined, so we cannot improve the minimum loop gain. We have already noted that residual phase noise on the v.c.o. will make a contribution to phase jitter under conditions of low loop bandwidth and low loop gain. The effect of these incidental variations must be checked in a trial run using a noise-free signal. This is, in any case, a useful first step in setting up a phase-locked loop of even moderate complexity.

7.7 Notes on acquisition and tracking

'Acquisition' is a general term which is used to describe the extremely complex processes by which a phase-locked loop picks up an incoming signal and moves towards lock.

When the loop is in an unlocked condition the output from the phase detector will initially be a 'beat' waveform which contains the difference frequency between the locking signal and the instantaneous frequency of the v.c.o. The maximum possible peak-to-peak value of the beat waveform is $\pm K_D$ volts which is subsequently attenuated by the effect of the loop filter. If the resulting voltage swing at the v.c.o. is sufficient to make the v.c.o. and input frequencies coincide, the system moves smoothly into lock without slipping cycles.

In a second-order system using a lag-lead filter the attenuation at high beat frequencies (corresponding to a high initial frequency difference between the locking signal and the v.c.o.) has a constant value $F(\infty)$. In this case the maximum available swing of the v.c.o. frequency is $\pm K_0 K_D F(\infty)$ which defines the *capture range* of the loop.

For initial frequency offsets within the capture range, locking is assured and fast. However, an important consequence of using an integrating loop filter is that the loop will eventually 'pull-in' to a locking signal which is at a frequency far removed from the v.c.o. frequency. The reason is that the 'beat' waveform is highly asymmetrical and contains a d.c. component. This can build up in the integrator and gradually drive the v.c.o. towards the signal. Pull-in can be a cumbersome and time-consuming process which is greatly affected by noise on the signal. It is usually overcome by manual tuning of the v.c.o. to bring the frequency difference within the capture range of the loop. Locking is then, for practical purposes, instantaneous.

It must be assumed that acquisition is always assisted when locking to very noisy signals and that the loop is finally trimmed to minimize the offset frequency of the v.c.o. As indicated in Chapter 6, manual assistance of acquisition serves also to avoid ambiguities in locking when the phase detector has responses to odd harmonics of the v.c.o. frequency.

Finally, let us look briefly at the case where the locking signal frequency is subject to a slow variation or drift. It can be shown that a loop containing an imperfect integrator is capable of tracking a locking signal with a changing frequency, but that the phase of the v.c.o. suffers a 'slip' relative to the phase of the locking signal. This is analogous to the phase slip observed in the reference channel of a lock-in amplifier in response to a changing reference frequency, as discussed in Section 4.5. In the case of the phase-locked loop the rate of change of locking frequency R is related to the phase slip θ by [1,2]:

$$2\pi R = \omega_N^2 \theta$$

where R is expressed in Hz/s. Note that the phase slip is *exclusive* of any transient phase error that may occur following the application of a frequency sweep.

When designing a loop for recovery from noise it has been suggested that the natural frequency has a minimum value

$(\omega_N)_{\text{MIN}} = 2(B_{\text{L}})_{\text{MIN}}$

If the phase slip is to have a maximum value θ_{MAX} then we obtain the condition

$R \leq 2(B_{\text{L}})^2_{\text{MIN}}\ \theta_{\text{MAX}}/\pi$

For example, a maximum phase slip of 5° gives

$\theta_{\text{MAX}} = 5 \times \pi/180$ radians

If $(B_L)_{\text{MIN}} = 100$ Hz, then

$R \leq 556$ Hz/s

The consequence of using very low noise bandwidths is a severe reduction in the allowed tracking rate. Thus if $(B_L)_{\text{MIN}}$ is only 5 Hz we obtain

$R \leq 1.39$ Hz/s

The same quadratic dependence of frequency sweep-rate on system bandwidth was noted in Section 5.5 in relation to spectrum analysis..

7.8 Using a lock-in amplifier for phase-locking

7.8.1 Introduction

So far, the treatment of phase-locked loops has been fairly general in the sense that a clear distinction has been made between the phase detector and the loop filter. As remarked in the introduction to this chapter, this separation is not so easily achieved when using a fully integrated lock-in amplifier. Let us therefore begin with Fig. 7.10, which shows the internal arrangement of a lock-in amplifier fitted with a 'phase-lock' option.

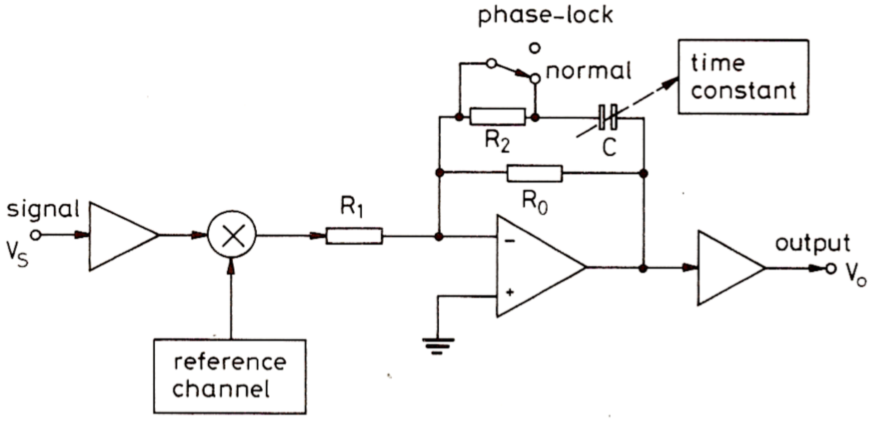


Fig. 7.10 Internal arrangement of a lock-in amplifier with 'phase-lock' option. Time constant $T_0 = CR_0$ in 'normal' mode

The phase-sensitive detector is shown schematically as a multiplier-detector while the output filter now serves a dual function. In 'normal' mode the filter capacitors are selected from the time-constant control, giving a range of values $T_0 = R_0C$. In 'phase-lock' mode the damping resistor R_2 is switched into circuit while R_0, the feedback resistor, may be removed if the loop filter is to be an imperfect integrator. One problem is that R_2 is not usually specified by manufacturers and may not be obtainable except through inquiry or by looking at the circuit diagram. It is certain, however, that the ratio R_0/R_2 will be large. R_0 is usually of the order of 10 MΩ while R_2 is commonly 10 kΩ (for example, in the Brookdeal series of phase-sensitive detectors and lock-in amplifiers).

7.8.2 Identifying the loop constants

When used as a phase detector with a time-constant filter, the observed sensitivity of a lock-in amplifier to static phase errors is, from Section 4.1:

$V_F(V_s/S_F)$ volts radian

This observed phase sensitivity is *inclusive* of the d.c. gain of the time-constant filter, R_0/R_1. In 'phase-lock' operation, the phase detector sensitivity K_D that appears in the loop equations is *exclusive* of the gain provided by the filter block. We thus obtain:

$$K_D = \frac{V_s V_F}{S_F}(R_1/R_0)$$

In practice, the ratio R_0/R_1 accounts for the bulk of the phase-sensitive detector dynamic reserve. The phase detector constant to be used in the loop equations is therefore given approximately by the overall phase sensitivity reduced by a factor equal to the dynamic reserve. For either a lag-lead filter or imperfect integrator we have $F(\infty) = R_2/R_1$. The loop gain is accordingly

$$K_0 K_D F(\infty) = K_0 \frac{V_s V_F}{S_F}\left(\frac{R_1}{R_0}\right)\frac{R_2}{R_1}$$

Putting $T_2 = R_2\mathrm{C}$, $T_0 = R_0 C$ we obtain

$$K_0 K_D F(\infty) = K_0 \frac{V_s V_F}{S_F}\left(\frac{T_2}{T_0}\right)$$

Here, T_0 is the time-constant setting on the front panel of the lock-in amplifier. In view of our earlier remarks about assigning a value to R_2, T_2 might have to be identified from a circuit diagram or by inspection of the time constant switch.

Using this value of loop gain, the natural frequency of the loop can now be put in the form

$$\omega_N = \left(\frac{K_0 V_s V_F}{S_F T_0}\right)^{1/2}$$

In the usual arrangement, T_0/T_2 appear in a fixed ratio:

$$T_0/T_2 = R_0/R_2 = r$$

The damping ratio is therefore:

$$b = \tfrac{1}{2}\omega_N T_2 = \tfrac{1}{2}\omega_N T_0/r$$

Provided that r can be identified, the loop parameters ω_N and b are now given in terms of V_s and the lock-in amplifier settings; time constant T_0 and sensitivity S_F, for an output voltage swing $\pm V_F$.

7.8.3 Optimization procedures for lock-in amplifiers

The specification for the locking signal gives us a minimum anticipated r.m.s. signal level V_{MIN}, an input noise bandwidth B_I and a worst-case signal-to-noise ratio $(SNR_I)_{MIN}$. This enables us to choose a minimum value of loop noise bandwidth $(B_L)_{MIN}$ commensurate with the input conditions:

$$(B_L)_{MIN} \le (SNR_I)_{MIN} B_I/20$$

Using the outline procedure given in Section 7.6 we design for $b_{\text{MIN}} = ½$ and immediately obtain a bound on the required value of T_0 (= rT_2):

$$T_0 \geq \frac{r}{2(B_{\text{L}})_{\text{MIN}}}$$

This leaves us to choose K_0 and S_{F} to satisfy

$$(\omega_{\text{N}})_{\text{MIN}} = r/T_0 = \left(\frac{K_0 V_{\text{MIN}} V_{\text{F}}}{S_{\text{F}} T_0}\right)^{1/2}$$

In summary:

(i) Determine minimum required value of B_{L};

(ii) Calculate the time-constant setting T_0 to ensure $(B_{\text{L}})_{\text{MIN}}$ at the smallest anticipated signal level;

(iii) Choose values of K_0 and S_F to satisfy

$$K_0 V_{\text{MIN}} / S_{\text{F}} = r^2/(V_{\text{F}} T_0).$$

The following examples will help to put the optimization procedures for lock-in amplifiers into perspective.

Example 1

A lock-in amplifier has a maximum time constant of 100 s with the ratio $r = R_0/R_2 = 1000$. What is the smallest value of noise bandwidth that can be achieved consistent with a minimum damping ratio of ½?

The smallest achievable noise bandwidth in phase-locked loop operation is

$$\frac{r}{2(T_0)_{\text{MAX}}} = 5 \text{ Hz}$$

Example 2

What is the worst possible signal-to-noise ratio on the locking signal that can be handled by a lock-in amplifier in phase-lock mode?

This is a question which is often asked but to which there is no direct answer. If we calculate the minimum achievable noise bandwidth – as in example 1 – then the best we can do is find a bound on the product $SNR_{\text{I}}B_{\text{I}}$:

$$SNR_{\text{I}}B_{\text{I}} \geq 10r/(T_0)_{\text{MAX}}$$

In example 1, the system could cope with a mean-square signal-to-noise ratio of −10 dB (1/10) in an input bandwidth of 1 kHz, or a ratio of −30 db (1/1000) in a 100 kHz bandwidth. We could also be certain that the phase-locked loop could hold lock under more adverse conditions, since a fairly conservative bound on B_{L} was taken in Section 7.5. It is likely, however, that if the design were carried out with a less stringent bound or with the minimum damping reduced much below ½, then the resulting handling characteristics and phase jitter would verge on the unacceptable.

Example 3

Using the lock-in amplifier specified in example 1, outline the optimization procedures for a signal of 50 kHz appearing at a minimum level of 1 mV with an estimated worst-case signal-to-noise ratio of −20 dB (1/10). The input noise bandwidth is set by signal conditioning filters to a value of 5 kHz and a v.c.o. is available with a sensitivity K_0 of $2\pi 10^4$ radians/V−s at the signal frequency.

First of all, the noise bandwidth of the loop: at minimum signal level this must satisfy

$$(B_L)_{MIN} \leq (SNR_I)_{MIN} B_I/20$$

$$\leq 10^{-1} \times 5 \times 10^3/20 \text{ Hz or } 25 \text{ Hz}$$

The smallest time constant consistent with this value is:

$$T_0 = r/(2 \times 25)$$

We have $r = 1000$; hence

$$T_0 = 20 \text{ s}$$

If a time constant of 20 s is not available, the next largest should be selected. The optimization procedures will ensure that the damping ratio will have a minimum of ½ as required.

Finally:

$$K_0 V_{MIN}/S_F = r^2/(V_F T_0)$$

K_0 has been given. Using $V_F = 10$ V we obtain:

$$V_{MIN}/S_F = 0.079$$

The minimum anticipated signal level is 1 mV r.m.s. The lock-in amplifier should therefore be set to a full-scale sensitivity

$$S_F = V_{MIN} / 0.079 = 12.6 \text{ millivolts}$$

In practice, a full-scale sensitivity of 10 mV will be indistinguishable from the optimum setting. Note that the locking signal may subsequently take larger values which exceed the full-scale sensitivity of the lock-in amplifier without incurring a fault condition, *provided* the total allowable swing on the signal input is not exceeded.*

If an inconveniently low or high ratio V_{MIN}/S_F is predicted, there may be scope to change the value of K_0. It should not be overlooked that many v.c.o.s found in measurement laboratories have overlapping decade ranges. By judicious choice of operating frequency it may be possible to change K_0 by a factor of 10 or even 100, depending on range selection.

When very narrow loop bandwidths are required, the ability to track signals of changing frequency (Section 7.7) must be taken into consideration. Also, a trial run using a noise-free signal should be carried out to assess the residual phase jitter in the loop arising from incidental phase- and frequency-modulation on the v.c.o. Here again, the ability to switch between overlapping ranges might prove useful and enable the v.c.o. to be operated in a region where its self-noise is lower.

7.9 The final measurement

Let us finally return to the measurement system proposed at the beginning of this chapter, where the v.c.o. output is used as a reference voltage for the detection of the locking signal in a second lock-in amplifier.

Two-phase lock-in amplifiers modified for phase-locking are ideal for this type of measurement; 'quadrature' channel B is used for phase-locking, leaving the

* This comment is in line with the procedures for increasing the phase sensitivity of lock-in amplifiers that were given in Sections 4.7.5. and 5.2.3.

signal to be measured in ‘in-phase’ channel A. When the dual phase-sensitive detectors are fitted with independently switched ‘expand’ amplifiers it is possible to operate the two channels at sensitivities differing by a factor of 10 or even 100. This is usually sufficient to ensure that the sensitivities of the two phase-sensitive detectors can be separately optimized for phase locking and for signal detection.

When the signal is noisy, the v.c.o. output inevitably exhibits phase jitter relative to the locking signal. Fortunately, as a result of the $\cos\theta$ dependence, this phase jitter has only a second-order effect on the measurement at the second phase-sensitive detector. It turns out in practice that if the phase jitter in the loop is sufficiently small to ensure long-term locked operation without slipping cycles (as we have assumed throughout), then the phase jitter will have minimal effect in the final measurement. In the worst case, at minimum signal level, its effect can be reduced by increasing the output time constant in the second lock-in channel beyond the value normally used at a given signal-to-noise ratio.

7.10 References

1 GARDNER, F.M. (1979): ‘Phase lock Techniques’ (John Wiley, New York)

2 BLANCHARD, A. (1976): 'Phase-locked loops: Application to coherent receiver design’ (John Wiley, New York)

3 LINDSEY, W.C. (1972): ‘Synchronization systems in communication and control’ (Prentice-Hall, N.J.)

4 HOGG, D.C. and CHU, T.S. (1975): ‘The role of rain in satellite communications’, Proc. *IEEE*, 63, pp. 1308-1331

5 BAYLISS, A. (1974): ‘A guide for orbital test satellite experiments’. Proc. European Conf. on Electrotechnics, Eurocon ’74, Amsterdam

CHAPTER 8

Heterodyne lock-in amplifiers

8.1 Introduction

In a heterodyne lock-in amplifier, phase-sensitive detection is carried out at a relatively high, fixed, frequency following a stage of frequency translation of the applied signal. Since the phase-sensitive detector operates at a frequency that bears no harmonic relationship to the applied signal, the harmonic responses that characterize a conventional lock-in system are suppressed. Practical lock-in amplifiers operating on the heterodyne principle thus conform very closely to ideal fundamental-only responding systems. The first of these, offering relative freedom from harmonic responses over a moderate frequency range, was the Ithaco Dynatrac lock-in amplifier. This was introduced in both single- and two-phase versions in the early 1970s.

In the early system the benefits of fundamental-only response were obtained to the detriment of performance in other areas. For example, the true frequency range was only about one decade and coverage of the audio-frequency range required a total of five sets of plug-in circuit cards. Also, the phase accuracy left much to be desired, particularly at the extremes of the individual ranges. Even under the most favourable conditions, this heterodyne system suffered by comparison with the high precision of conventional lock-in amplifiers. The system nevertheless enjoyed considerable success and served to focus attention on the limitations associated with harmonic responses which were reviewed in Chapter 6.

The block diagram of the Ithaco Dynatrac begs comparison with that of a superhet radio, and so the system attracted the "heterodyne" label from the time of its first introduction. This description is now applied more or less indiscriminately to all lock-in amplifiers that incorporate one or more stages of frequency translation. It can be argued that these systems bear only a superficial resemblance to classical heterodyne systems and that the use of expressions such as "intermediate frequency" and "i.f. filter" in relation to lock-in amplifiers is likely to cause confusion, especially with those who have a clear understanding of the conventional usage of these terms. It is therefore necessary to introduce a note of caution about the terminology employed in this chapter which reflects the usage that is now prevalent among lock-in amplifier manufacturers and appears in data sheets and publicity material.

The Ithaco Dynatrac has since been matched by alternative and improved heterodyne lock-in amplifiers from the EG&G companies, PAR and Brookdeal. These lock-in amplifiers have greatly benefited from developments in technology relating to both reference channel and phase-sensitive detector design. In this chapter we shall be taking note of these developments and aiming to highlight areas of specification which are peculiar to heterodyne lock-in amplifiers.

Of particular interest in this respect will be the problem of identifying the spurious responses which occur when the frequency of an asynchronous signal lies close to a "critical" frequency. In a conventional system these critical frequencies correspond to the odd harmonics of the reference frequency. We shall show that *in principle* heterodyne systems can be designed to be inherently free of these harmonic responses. There are nevertheless a number of additional critical frequencies, each with its related transmission window, which must be taken into account when a heterodyne model is adopted. A major objective of heterodyne system design is therefore to achieve a high degree of suppression of all unwanted responses, consistent with maintaining wideband, wide dynamic range performance.

Unfortunately, as we shall see, this last requirement cannot be achieved without sacrificing the total rejection of harmonic responses, which is inherent in an "ideal" heterodyne system. In our discussions we must therefore make a clear distinction between principles and practice and be prepared to examine the trade-offs which are necessary to produce heterodyne systems with good all-round performance.

Also included in this chapter is an appraisal of the synchronous heterodyne technique. Synchronous heterodyning offers a means of improving the dynamic range of phase-sensitive detectors and lock-in amplifiers. The technique can also be used to counteract the loss in dynamic range which occurs when an otherwise conventional phase-sensitive detector is operated towards the upper limit of its frequency range.

This last approach is used to obtain a competitive dynamic range specification in the EG&G Brookdeal heterodyne lock-in amplifiers where phase-sensitive detection is carried out at a fixed high frequency. We shall also be giving consideration to the spurious responses associated with synchronous heterodyning. The treatment given falls short of a full analysis, but it is shown how some of the major contributory factors can be overcome in practical systems.

8.2 Principles of heterodyne operation

The principles of heterodyne lock-in amplifiers can be established in terms of the idealized system shown in Fig. 8.1.

The frequency translator has inputs from the signal channel amplifier, from the applied reference signal and from an internal oscillator. The latter operates at frequency f_I which we shall call the intermediate frequency. The purpose of the frequency translator is to produce an output with magnitude proportional to the signal input but with its frequency shifted from f_s to a new and higher value $f_I + f_R - f_s$. The translated signal is then applied to a phase-sensitive detector which we assume is referenced and phase-shifted at the intermediate frequency.

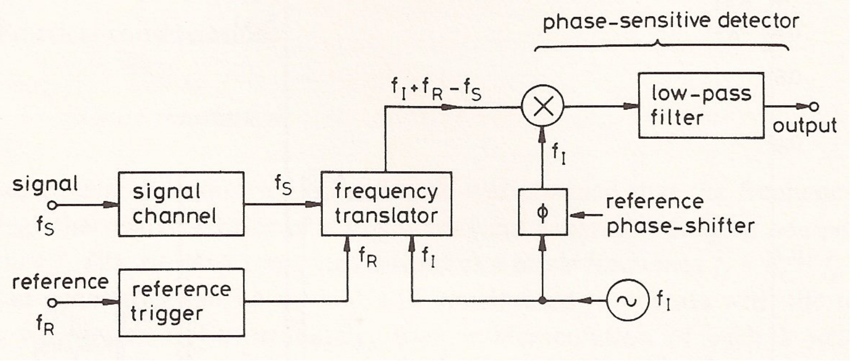


Fig. 8.1 Basic heterodyne lock-in amplifier, showing the frequencies at different points in the system

When the signal and reference are fully synchronous, $f_s = f_R$. The translated output then appears at frequency f_I and yields a classic phase-sensitive response at the phase-sensitive detector.

For signals with frequencies close to f_R the response will be an alternating output at frequency $|f_R - f_s|$, which is attenuated in the usual way by the output low-pass filter. However, in this system there is no scope for unwanted responses when the signal frequency is coincident with an odd harmonic of f_I. The latter are included on the assumption that a switching phase-sensitive detector is used with its associated harmonic transmission windows.

The critical signal frequencies are therefore those which satisfy the relationship:

$$|f_I + f_R - f_s| = Kf_I$$

where K is an odd integer. Solving for f_s, we obtain the critical frequencies:

$$f_s = |(1-K)f_I + f_R|$$

and,

$$f_s = (1+K)f_I + f_R$$

For each value of K there are therefore two frequencies where an interference component will be able to excite a harmonic response in the phase-sensitive detector. The relative sensitivity of the system to inputs at these critical frequencies is denoted by S_K and is given by

$$S_K = 1/K$$

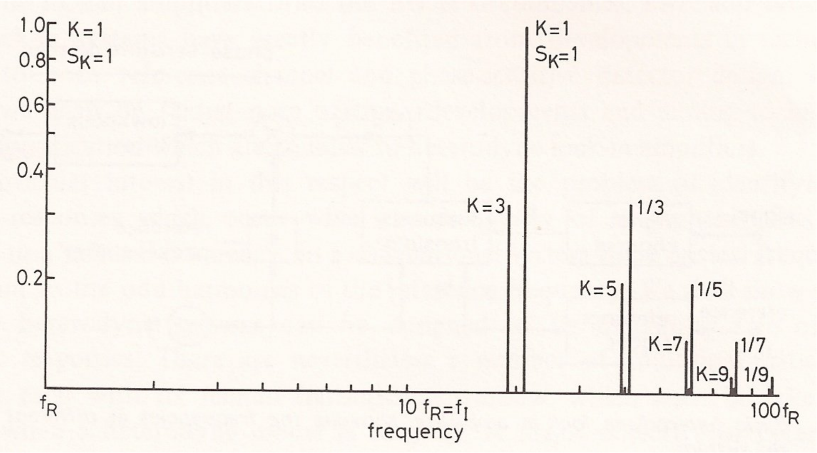


Fig. 8.2 Location and relative magnitude of transmission windows in a heterodyne system, for K = 1, 3, 5, 7 and 9. $f_I = 10f_R$

The transmission windows of the overall detection system are thus derived from the phase-sensitive detector windows and are located at frequencies determined jointly by the intermediate and reference frequencies. The transmission windows corresponding to K in the range 1 to 9 are drawn schematically in Fig 8.2 for the relationship, $f_I = 10 f_R$. The weighting factor assigned to each window is shown in each case and gives the relative sensitivity of the detection system to inputs at the appropriate critical frequency.

Fig. 8.2 reminds us that there must be two transmission windows, corresponding to $K = 1$, where the relative sensitivity is unity. These windows correspond to the "primary" response at $f_s = f_R$, and the "image" response given by signals with frequency $f_s = 2 f_I = f_R$. An image response occurs in all systems operating on the heterodyne principle. In this case, a signal for which $f_s \approx 2 f_I + f_R$ yields a translator output close to the intermediate frequency and the resulting behaviour of the detection system cannot then be distinguished from the "true" response when $f_s = f_R$. The only satisfactory way to deal with the image response is to eliminate signal components at the image frequency by filtering. The most convenient arrangement is to use a low-pass filter in the signal channel with a sharp cut-off defined at a frequency below $2 f_I + f_R$.

This filter must be introduced *before* the frequency translator and should have a cut-off frequency greater than f_{RMAX}, the maximum anticipated value of the reference frequency. An image filter is an essential component in a heterodyne lock-in amplifier, irrespective of the precise system configuration.

It turns out that in the present, ideal, case a properly designed image suppression filter would be effective in suppressing inputs at all other critical frequencies. This ideal heterodyne system is thus inherently free from responses at harmonics of the reference frequency and can be made relatively immune to the incidence of spurious responses at other, non-related, frequencies.

8.3 Practical considerations

8.3.1 Frequency translation

When discussing principles of operation it was assumed that the frequency translator had the characteristics of a single-sideband generator; a single component at frequency f_s gave rise to a translated output at a *single* frequency $f_I + f_R - f_s$. Single-sideband generators can be devised and constructed to operate with the required degree of precision. Unfortunately, the implementation of such a scheme to operate over a wide frequency range is both complex and expensive. A more cost-effective solution is to use double-sideband generation in which the required translation is achieved with a signal channel mixer and frequency synthesizer as shown in Fig. 8.3.

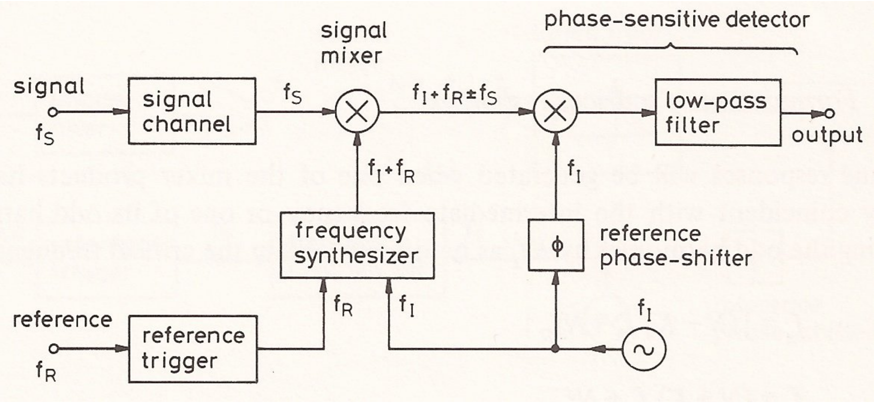


Fig. 8.3 Heterodyne system with double-sideband frequency translation

The frequency synthesizer provides an output at a precisely defined frequency $f_I + f_R$ which yields mixer products at frequencies $f_I + f_R \pm f_s$. When the signal and reference are fully synchronous, the mixer output will consist of two components, one at frequency f_I and the other at $f_I + 2 f_R$. The first of these will give rise to a phase-sensitive response as before. The second gives rise to an alternating output at frequency $2 f_R$. This is no more serious than the "ripple" component which is expected in conventional phase-sensitive detector operation. As is usually the case, the ripple component is rejected by the low-pass filter in the phase-sensitive detector output.

Turning now to asynchronous signals, we find an important consequence of using a double-sideband generator. This is a reduction in the allowable voltage swing due to asynchronous components at the input to the signal channel. Each of these gives rise to two mixer products of equal amplitude. The input voltage swing at a given sensitivity is thus limited to half the value which could be sustained when using the phase-sensitive with a single-sideband frequency translator. For a phase-sensitive detector of given characteristics the achievable dynamic reserve is consequently reduced by 6 dB.

Further complications arise in a practical implementation of the frequency translator. The signal channel mixer, like the phase-sensitive detector, is almost invariably a switching multiplier which uses a *squarewave* drive at frequency $f_{\mathrm{I}} + f_{\mathrm{R}}$. This is another example where linearity and dynamic range are obtained through the adoption of a switching operation. The result is that the output of the signal mixer consists of a large number of components at frequencies

$$N\left(f_{\mathrm{I}} + f_{\mathrm{R}}\right) \pm f_{\mathrm{s}}, \quad N \text{ odd}$$

The amplitude of each of these components is weighed by a factor $1/N$, reflecting the reduction in the harmonics of the switching waveform with increasing order.

As we shall see, the use of a switching mixer leads to a reappearance of the harmonic responses which were so successfully rejected by the ideal heterodyne scheme.

8.3.2 Formulation of spurious responses

Spurious responses will be generated when one of the mixer products has a frequency coincident with the intermediate frequency or one of its odd harmonics. Denoting the odd harmonics by Kf_{I}as before, we obtain the critical frequencies:

$$f_{\mathrm{s}} = \left|\left(N - K\right)f_{\mathrm{I}} + Nf_{\mathrm{R}}\right|$$

and

$$f_{\mathrm{s}} = \left(N + K\right)f_{\mathrm{I}} + Nf_{\mathrm{R}}$$

To calculate the relative sensitivity of the detection system to inputs at these critical frequencies, we must take the following factors into account: First of all, a factor $1/N$ resulting from the use of a squarewave drive to the mixer; secondly a factor $1/K$ to allow for the reduction of the phase-sensitive detector transmission windows with increasing harmonic order. The relative sensitivity at frequencies corresponding to given values of N and K is therefore:

$$S_{\mathrm{N,K}} = 1/NK$$

For example, when $N = K = 1$ we obtain the primary response corresponding to $f_{\mathrm{s}} = f_{\mathrm{R}}$ and the image response corresponding to $f_{\mathrm{s}} = 2f_{\mathrm{I}} + f_{\mathrm{R}}$. The relative sensitivity of the system to inputs at these frequencies is unity in both cases.

Let us now take $N = K \neq 1$. In this case, the system will be sensitive to inputs at $f_{\mathrm{s}} = Nf_{\mathrm{R}}$ and $2Nf_{\mathrm{I}} + Nf_{\mathrm{R}}$. The response to the Nth harmonic of f_{R} is thus reinstated at a relative sensitivity of $1/N^2$, compared to the figure of $1/N$ which would be obtained in a conventional system.

As noted in section 8.2 an image suppression filter would be essential in any practical system. Unfortunately this has no influence on signal components at frequencies Nf_{R}that fall within the filter bandwidth. The resulting harmonic responses can only be satisfactorily suppressed by using a second, tuned, filter in front of the phase-sensitive detector.

8.3.3 Suppression of the spurious responses

We envisage a system such as that shown in Fig. 8.4 which includes an image suppression filter in the signal channel and a tuned filter following the signal mixer.

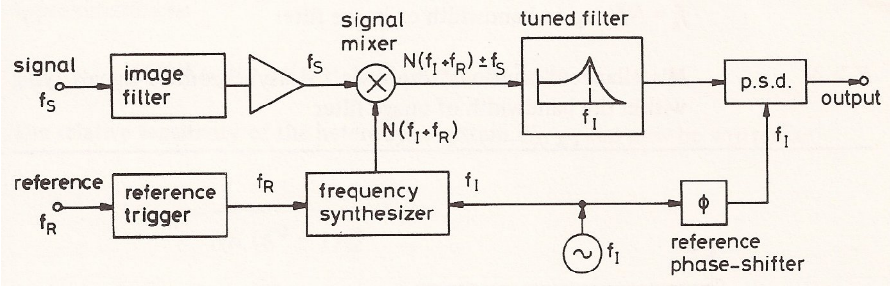


Fig. 8.4 Incorporation of tuned filter in a heterodyne system

The image suppression filter is a low-pass filter which exhibits a "flat" amplitude response for all signal frequencies up to the maximum value of the reference frequency as shown in Fig. 8.5.

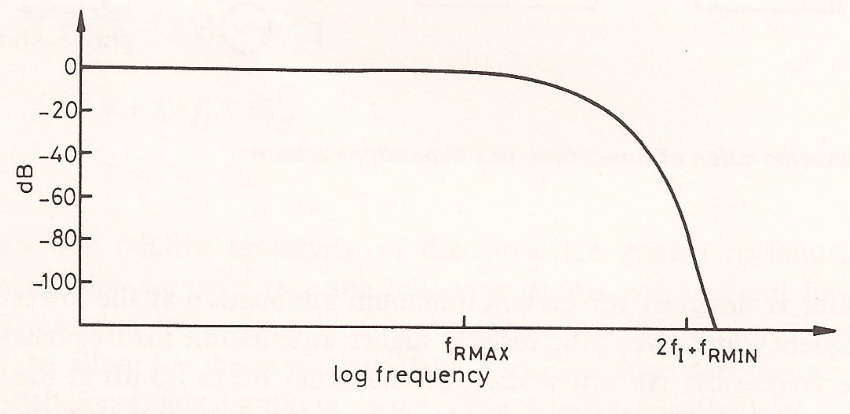


Fig. 8.5 General transmission characteristics of an image-suppression filter

The filter is designed for certain minimum attenuation at the lowest expected image frequency and gives progressively higher attenuation for frequencies beyond the image frequency. An attenuation of the order 60 to 80 dB at the image frequency would be typical of practical systems. We thus find that asynchronous signals characterized by frequencies:

$$f_s = (N+K)f_I + Nf_R; \quad N, K \geq 1$$

all lie beyond the cut-off of the image filter and are effectively eliminated before the signal mixer.

The only asynchronous signals which are likely to yield discernable responses must therefore have frequencies which satisfy the condition:

$$f_s = \left|(N-K)f_I + Nf_R\right|$$

The nature of the responses corresponding to different values of *N* and *K* are summarized in Table 8.1.

Table 8.1 Catalogue of responses for asynchronous signals with frequencies $|(N-K)f_I + Nf_R|$

N, K	Comments
$N = K = 1$	$f_s = f_R$: primary respoonse of system
$N > K$	Asynchronous signals always have frequencies $\geq$ image frequency
$N > K \neq 1$	Spurious responses whenever asynchronous signal has frequency $f_s = Nf_R$ up to bandwidth of image filter
$K > N$	Miscellaneous spurious responses for asynchronous signals lying within the bandwidth of image filter

The first of the categories listed in Table 8.1 corresponds to the "wanted" response of the system while signals in the second category are suppressed by the image filter. In the absence of a tuned filter, the system would display a relative sensitivity $S_{\mathrm{N,K}} = 1/NK$ at all critical frequencies in the final two categories. The effect of a tuned filter will be to introduce additional attenuation of these responses giving a relative sensitivity:

$$S_{\mathrm{N,K}} = \left|H\left(\mathrm{j}K\omega_{\mathrm{I}}\right)\right|/NK$$

Here, $H(\mathrm{j}\omega)$ is the frequency response of the tuned filter normalized to a magnitude of unity at the intermediate frequency.

We note that the largest spurious response will be obtained when $N = K = 3$, corresponding to an asynchronous signal with frequency $3f_R$[*]. The tuned filter should therefore be set up to reduce this response to an acceptably small value.

8.3.4 Tuned filter requirements

In all commercial systems the tuned filter is of the low-pass type described in Appendix 4. The use of a tuned filter for suppressing the harmonic responses of a phase-sensitive detector was discussed in Section 4.5.2. For a low-pass filter tuned to the intermediate frequency f_{I}, the attenuation at frequency Kf_{I} in the high-Q approximation is:

$$\left|H\left(\mathrm{j}K\omega_{\mathrm{I}}\right)\right| = \frac{1}{(K^2-1)Q}, \quad K = 3, 5, 7, \mathrm{K}$$

The relative sensitivity of the heterodyne system, $S_{\mathrm{N,K}}$, can now be written as

$$S_{\mathrm{N,K}} = \frac{1}{NK(K^2-1)Q}$$

The value of the Q-factor required for a given level of suppression can thus be calculated.

For example, suppose the system is required to have a relative sensitivity of 10^{-4} (–80 dB) to a signal with frequency close to $3f_R$. From Table 8.1 we put $N = K = 3$ and calculate the required Q-factor of the tuned filter:

[*] Note that the response due to $K = 3$, $N = 1$ corresponds to an asynchronous signal at frequency $2f_I - f_R$.This would normally be heavily attenuated by the image filter.

$$Q = 10^4 /(3 \times 3 \times 8) = 139$$

From the discussion given in Section 4.5.1 it is clear that operation with such a high value of Q-factor places severe demands on the system with regard to alignment and maintaining good phase accuracy. This is all the more troublesome in a purpose-built system where the filter is not usually accessible for routine realignment by the user.

An alternative approach which greatly eases alignment problems and which places less demands on filter performance has been used in the EG&G Brookdeal heterodyne system. This is to use two filters of relatively low Q-factor in cascade. The attenuation introduced by the filter stage then becomes:

$$\left|H\left(\mathrm{j}K\omega_{\mathrm{I}}\right)\right| = \frac{1}{(K^2-1)^2 Q^2}$$

In this case, a Q-factor of 5 is sufficient to give a relative sensitivity of –83 dB to signals with frequency $f_{\mathrm{s}} = 3 f_{\mathrm{R}}$.

Of course, these figures for harmonic rejection are strictly theoretical. There is a considerable technical hurdle to be overcome in order to realize these figures in practice and commercial systems usually specify suppression factors of around –60 dB (×1/1000). In any specific case, it is always worth checking whether the figure given refers solely to third-harmonic rejection or to the maximum level of spurious responses arising from all possible sources.

8.3.5 Phase-shifting

It has been assumed so far that the user-controlled phase-shifter which is essential for phase-sensitive detection is introduced at the intermediate frequency, on the "reference" side of the phase-sensitive detector. This has a distinct advantage in that the phase-shifter can be designed to operate at a fixed frequency rather than over a wide range of frequencies as is usually the case.

In principle, there is no reason why phase-shifting should not be carried out at the original reference frequency or, indeed, on the output of the frequency synthesizer, at frequency $f_{\mathrm{I}} + f_{\mathrm{R}}$. These are all found to be equivalent when system operation is analysed. Phase-shifting at the reference frequency requires similar broadband circuitry to that found in a conventional lock-in amplifier, whereas a phase-shifter placed at the output of the frequency synthesizer, assuming $f_{\mathrm{I}} >> f_{\mathrm{RMAX}}$, would require relatively narrowband capability.

8.4 Practical limitations

8.4.1 The frequency synthesizer

The generation of a waveform at a precisely defined frequency $f_{\mathrm{I}} + f_{\mathrm{R}}$ in response to an external signal applied at frequency f_{R} is the most demanding task facing the designers of heterodyne lock-in systems. The ultimate objective is to produce a system which:

(i) Maximises the range of reference frequencies which can be handled at a particular intermediate frequency.

(ii) Has an acceptably low level of both random and discrete phase-noise over the specified frequency range.

(iii) Minimizes phase errors between the absolute phases of the synthesized waveform and the applied reference waveform.

(iv) Has an acceptable value of reference slew rate and a small acquisition time.

A shortfall in any of these areas would be noticed by any user whose interest in lock-in amplifiers extended beyond the detection of noisy signals at a fixed frequency. The designer's difficulty is to reconcile these requirements and make the correct trade-offs to produce a system with all-round acceptable performance. The problems encountered in reaching an acceptable compromise in synthesizer performance can be highlighted by referring to the system illustrated in Fig. 8.6.

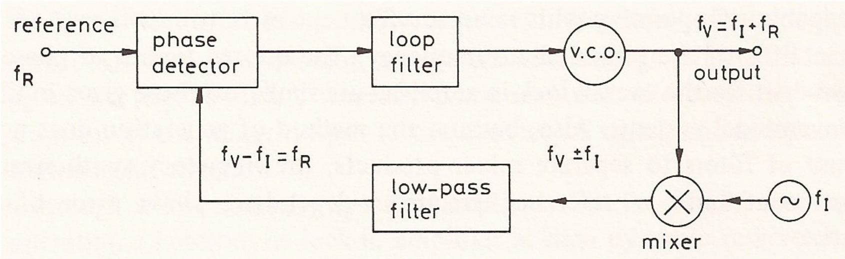


Fig. 8.6 Heterodyning phase-locked loop

This example of a "heterodyning" phase-locked loop received considerable publicity when the Ithaco Dynatrac lock-in amplifiers were first introduced[1].The phase detector compares the phase of its two inputs and its output is amplified and smoothed by the loop filter. The v.c.o. is controlled from the loop filter output and produces an output at frequency f_V, which is mixed with a signal at the intermediate frequency f_I provided from a stable sinewave oscillator. The purpose of the low-pass filter is to transmit only the low-frequency mixer product at frequency $f_V - f_I$ to the phase detector. The difference frequency $f_V - f_I$ is thus phase-locked to the incoming reference signal at frequency f_R, making f_V equal to $f_I + f_R$.

When a strictly conventional approach is taken to designing the phase-locked loop, following the treatment given in Chapter 7, the designer is faced with a difficult decision in choosing the loop characteristics. The loop bandwidth should be small enough to suppress discrete phase modulations resulting from unwanted mixer components at the output of the low-pass filter, and yet wide enough to accommodate the phase noise inherent in the v.c.o. and to give adequate slew rate performance. In addition, severe phase errors might be incurred when the low-pass filter achieves the desired separation of mixer products by virtue of a sharp cut-off. Almost inevitably, the trade-offs are such that acceptable performance can only be achieved over a relatively small range of reference frequencies. Limitations are observed in terms of poor phase tracking and excessive phase noise at the extremes of the operating frequency range.

In the loop used by EG&G Brookdeal, switching waveforms derived from the applied reference and from the intermediate-frequency oscillator are combined in a multiplexer. The output is a switching waveform in which the positive transitions occur an average rate $(f_I + f_R)/4$. This entire waveform is then used to phase-lock a v.c.o. in a phase-locked loop which incorporates a ÷4 counter. In a conventional phase-locked loop, the output from the v.c.o would be a squarewave at frequency $f_I + f_R$ subjected to a high level of discrete phase-modulation which could only be suppressed by using a loop of small bandwidth. In the EG&G Brookdeal system the unwanted modulation is well defined and can be suppressed to a high order by adding a compensation signal to the output of the phase detector. Provided the compensation signal is accurately generated and controlled, the constraints on loop bandwidth are considerably relaxed. The loop characteristics can then be optimized with respect to a low incidence of phase noise and obtaining acceptable handling characteristics over a wide range of reference frequencies. The system, in fact, proves capable of operating with reference frequencies from less than 1 Hz up to about one fifth of the intermediate frequency. The quoted figures of phase noise and phase drift for the overall lock-in amplifier are similar to those given in Chapter 4 for conventional systems. Also, because the method of generation does not rely on the use of filters to separate mixer products, the frequency synthesizer itself does not contribute a reference frequency-dependent phase error of major significance.

8.4.2 The image filter

It has been noted that the image filter should have a flat response up to signal frequencies corresponding to f_{RMAX}. The rate of cut-off beyond this point must then be extremely large in order to introduce adequate attenuation at the image frequency. In most commercial systems the ratio $f_{\mathrm{I}} / f_{\mathrm{RMAX}}$ is less than 10, so that a filter of high order, 4 to 6 pole, is required to achieve the necessary roll-off.

Like all filters used for signal conditioning, the image filter introduces phase errors into the measurement system. When the filter is of high order, the phase error can be in excess of 100° at about one half of the cut-off frequency. A fortunate consequence of using high-order filters is that the amplitude response can be made uniformly flat up to frequencies very close to cut off. In addition, the phase shift within this range can be made proportional to frequency. When the filter approximates to such a linear-phase model, the phase-shift of the signal channel can be compensated by introducing a suitable time delay in the reference channel. In the Ithaco Dynatrac, the image-filter phase characteristic was actually compensated by the characteristic of the low-pass filter in the synthesizer phase-locked loop. In other systems, such as that from EG&G Brookdeal, the synthesizer has inherently low phase error and the compensating time delay need be no more complicated than a monostable circuit operating at a fixed pulse width.

When we take matching constraints into account and add uncompensated phase errors accrued in the intermediate-frequency filter and in the synthesizer, we conclude that the overall phase precision of heterodyne systems must fall short of that obtainable in a conventional lock-in amplifier operating in the same reference frequency range.

8.4.3 The signal mixer

The most serious performance limitations associated with the signal mixer are due to non-linearity and "feedthrough". It has been stressed elsewhere that the synchronous demodulation process should be supported by linearity in all preceding stages; hence the linearity of the signal mixer should be at least of the same order as the linearity of the phase-sensitive detector. This requirement is eased in practice because the two components, being similar in concept, draw upon similar technologies.

"Feedthrough" is a phenomenon associated with voltage offsets and capacitive coupling in the signal mixer, whereby components at the switching frequency, $f_{\mathrm{I}} + f_{\mathrm{R}}$, appear at the mixer output in the absence of signal, and independently of signal channel gain selection.

Feedthrough is not normally specified explicitly but its effect becomes evident when operating a heterodyne lock-in amplifier at high dynamic reserve. It appears as an alternating component at frequency f_{R} in the output of the system and appears at its worst when operating at low reference frequencies with a relatively short time constant selected.

A level of feedthrough of 10^{-5} (–100dB) represents a reasonable target for a signal channel mixer operating with an intermediate frequency in excess of 100 kHz. When used in conjunction with a phase-sensitive detector having a dynamic reserve of $\times 10^4$ (80 dB) the output a.c. variation at frequency f_R could have a maximum value of 1/10 (–20 dB) of full scale. Unless the intermediate-frequency tuned filter has unusually high Q-factor - and correspondingly narrow bandwidth - it is unlikely to have much influence on the level of feedthrough reaching the phase-sensitive detector.

8.4.4 The phase-sensitive detector

It was noted in Chapter 3 that the dynamic range of switching phase-sensitive detectors is reduced at high reference frequencies owing to the appearance of the so-called "h.f. offset". The mechanisms at work here are very similar to those which give rise to feedthrough in switching mixers. It is claimed, quite reasonably, that in heterodyne lock-in amplifiers operating at a fixed value of intermediate frequency, the dynamic range is constant over the entire range of signal and reference frequencies. It should also be said, however, that the phase-sensitive detector is fated always to operate at a relatively high frequency where its dynamic range is less than optimum. In the Ithaco system, operation in a given frequency range involved selecting plug-in circuit cards to provide a suitably high value of intermediate frequency. Not surprisingly, the figures for dynamic range showed a marked deterioration as the system was configured to operate at higher reference frequencies and correspondingly higher values of intermediate frequency.

In the EG&G Bookdeal heterodyne lock-in amplifier the maximum intermediate frequency is 1 MHz, which allows operation with reference frequencies over the entire audio-frequency range up to 200 kHz. In this case, the dynamic range of the phase-sensitive detector is maintained at a competitively high value (>120 dB) by a technique known as *synchronous heterodyning*. It is shown in Section 8.8 that this approach can be used to improve the dynamic range of phase-sensitive detectors operating in either a conventional lock-in amplifier or in a heterodyne system. Synchronous heterodyning is not an alternative to the heterodyne mode of operation described so far; rather it is a supplementary technique applied with the objective of improving dynamic range.

8.5 Overload capability of heterodyne systems

It was the advent of the Ithaco Dynatrac that focused popular attention on the specification of overload capability in lock-in amplifiers. Clearly, a simple statement of phase-sensitive detector dynamic reserve is not sufficient when a lock-in system is so heavily supported by filters.

From the point of view of overload capability, the image suppression filter in any heterodyne lock-in amplifier can be regarded as a low-pass signal conditioning filter. The rate of cut-off of the image filter is extremely high and this leads to a dramatic improvement in overload capability for asynchronous signals having frequencies greater than about one half the intermediate frequency.

The selectivity of the tuned intermediate-frequency filter is chosen solely on the grounds of suppressing harmonic responses. Claims that the tuned filter is responsible for a significant increase in overload capability should therefore be treated with caution, particularly when the reference frequency is very low compared with the intermediate frequency. In a heterodyne system the overload characteristics are similar to those of a conventional lock-in amplifier using a tuned filter at the reference frequency with a bandwidth equal to that of the intermediate frequency filter, f_I / Q. The *effective* Q-factor is thus dependent on the reference frequency and is given by $Q_{eff} = Q \times f_R / f_I$. As a result, the filter is likely to have a significant effect on overload capability only when the reference frequency approaches its maximum value; even then the maximum reference frequency f_{RMAX} should represent a substantial fraction of f_{I}. Moreover, the *Q*-factor of the tuned filter itself should also be high; if sharp cut-off is obtained by cascading low-*Q* sections as described in Section 8.3.4. the overload characteristics will be broadly independent of frequency within the bandwidth of the image-suppression filter.

The interference rejection characteristics plotted in Fig. 8.7 serve to emphasize these points. The graphs are drawn for a system in which $f_{\mathrm{I}} = 5\,f_{\mathrm{RMAX}}$ and the intermediate-frequency tuned filter has a *Q*-factor of 20. These conditions approximate to those it the Ithaco Dynatrac lock-in amplifier. In a system such as the EG&G Brookdeal heterodyne, the reference frequency is allowed to take values up to 6 decades below the intermediate frequency. Rejection characteristics of type (a) in Fig. 8.7 are thus applicable over most of the operating frequency range.

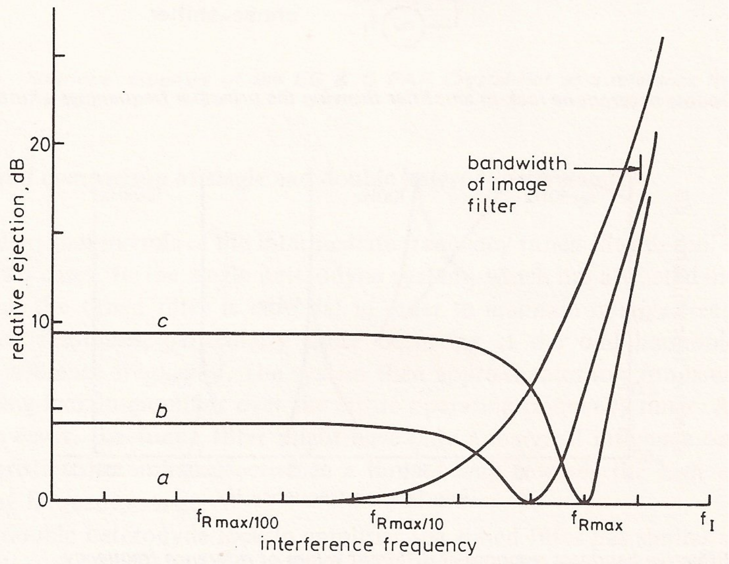


Fig. 8.7 Interference rejection in a heterodyne lock-in amplifier (a) $f_R = f_{RMAX}/10$; (b) $f_R = f_{RMAX}/2$; (c) $f_R = f_{RMAX}$

8.6 Double heterodyne lock-in amplifiers

In the heterodyne system described so far, the final detection has been carried out at the intermediate frequency. In a double heterodyne detection is carried out at the original reference frequency as shown in Fig. 8.8.

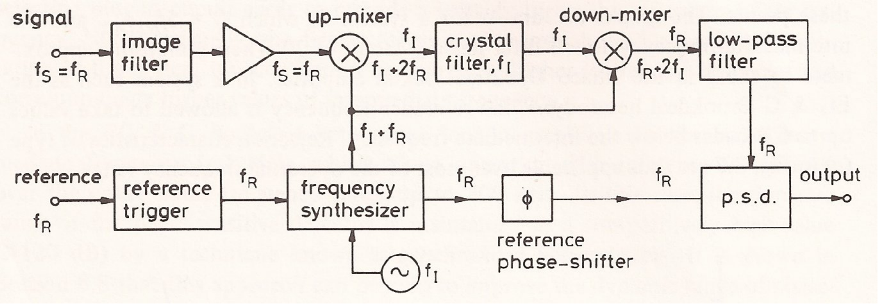


Fig. 8.8 Double-heterodyne lock-in amplifier showing the principal frequencies of interest

The intermediate frequency stage incorporates a filter of extremely high selectivity. In the case of the EG&G PAR Crystal-Het lock-in amplifier a two-section crystal filter is used to define a *Q*-factor of 50 000 at the intermediate frequency of 1MHz. The resulting 20 Hz bandwidth is significantly smaller than the reference frequency over much of the 100kHz frequency range, and thus has a profound effect in removing harmonically related components and noise before detection.

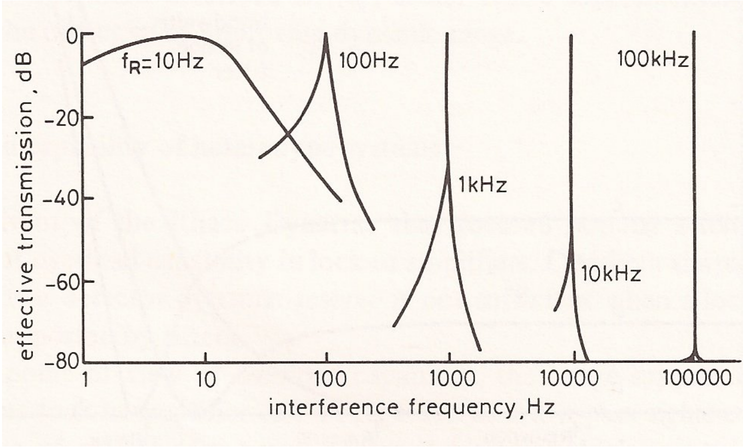


Fig 8.9 Effective bandpass response at different values of reference frequency

The system can be thought of in terms of a tracking filter with a fixed bandwidth of 20 Hz, giving an effective *Q*-factor, $Q_{\text{eff}} = 50\ 000 \times f_R / f_I$. The effectiveness of the system in rejecting interference components is illustrated in Fig. 8.9 which is drawn for different reference frequencies.

As might be expected, the phase-tracking of such a narrowband system is greatly inferior to that of a conventional lock-in amplifier. The level of spurious responses is nevertheless very low at midband. This is evident from the results given in Fig. 8.10 for operation at a reference frequency of 1kHz. The largest spurious response in this case occurs at a level of – 70dB relative to the "primary" response at 1kHz.

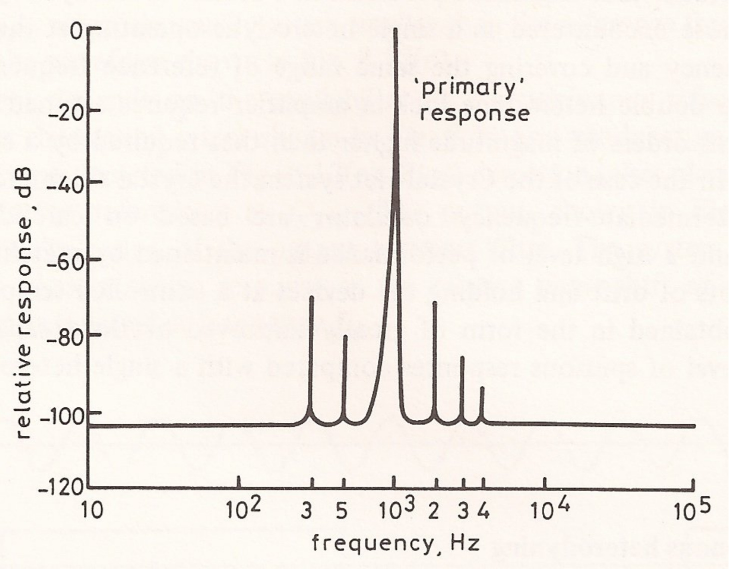


Fig. 8.10 Spurious responses of the EG&G PAR Crystal-Het at a reference frequency of 1kHz

8.7 Brief comparison of single and double heterodyne systems

It is evident that the role of the intermediate frequency tuned filter is quite different in the two cases. In the single heterodyne system, which has attracted most of our attention, the tuned filter is essential in order to maintain relative freedom from unwanted responses, particularly those occurring at the odd harmonics of the applied reference frequency. The system then approximates to a fundamental-only responding lock-in amplifier over the entire operating frequency range. As we have seen, however, the tuned filter might have only a marginal influence on overload characteristics, becoming effective in a limited way towards the high end of the operating frequency range.

In a double heterodyne lock-in amplifier the tuned signal has similar status to a tuned filter incorporated in the signal channel of a conventional lock-in amplifier. The considerations of Section 4.4 thus apply provided that allowance is made for the variation of the effective Q-factor with operating frequency. The essential difference between the two heterodyne schemes is that, in the double heterodyne system, harmonic rejection depends on the *effective* Q-factor; in a single hetrodyne, harmonic rejection depends only on the actual Q-factor of the tuned filter measured at the intermediate frequency. We can conclude from this that the double heterodyne approximates to a fundamental-only responding system only for reference frequencies greater than a specified value. In the case of the EG&G PAR Crystal-Het system referred to earlier, this minimum value of reference frequency must certainly be greater than 20 Hz, which limits the system's effectiveness in the critical low-frequency regime identified in Chapter 6.

It is also evident that alignment problems in a double heterodyne system are far worse than those encountered in a single heterodyne operating at the same intermediate frequency and covering the same range of reference frequencies. As we have seen, the double heterodyne lock-in amplifier requires a tuned filter with a Q-factor several orders of magnitude higher than that required by a single heterodyne system. In the case of the Crystal-Het system the crystal filter and the crystal-controlled intermediate-frequency oscillator are based on carefully matched components and a high level of performance is maintained by matching temperature coefficients of drift and holding the devices at a controlled temperature. The benefits are obtained in the form of greatly improved overload capability and a much lower level of spurious responses compared with a single heterodyne lock-in amplifier.

8.8 Synchronous heterodyning

8.8.1 Introduction

In the mid 1970s EG&G PAR introduced a lock-in amplifier known as the Syncro-Het which offered a considerable improvement in dynamic range over conventional lock-in amplifiers. The system enjoyed success in its own right, but attracted little in the way of direct competition. It appears that no attempt was made to refine the basic technique or to develop further commercial instruments operating on the synchronous heterodyne principle. This was the situation until recently when synchronous heterodyning was revived in the context of high-frequency lock-in systems. The objective was stated in Section 8.4.4; namely to improve the dynamic range of phase-sensitive detectors operating at a high frequency.

It was remarked in Section 3.6.5 that the d.c. response to a synchronous signal can be separated from offset voltages at the output of a phase-sensitive detector by introducing a phase reversal of 180° in the reference channel. This approach to signal detection forms the basis of the synchronous heterodyne system illustrated in Fig. 8.11. Here, the phase reversal is introduced by systematically switching the signal with a squarewave taking values +1 and –1. The switching is carried out at a frequency f_{SYN}. Final detection takes place in the phase-sensitive detector referenced at f_{SYN}. The output, shown in Fig. 8.12(f), is smoothed by the action of the output low-pass filter. The system thus yields a conventional phase-sensitive response as the relative phase-shift of the signal and reference inputs is varied at frequency f_R .

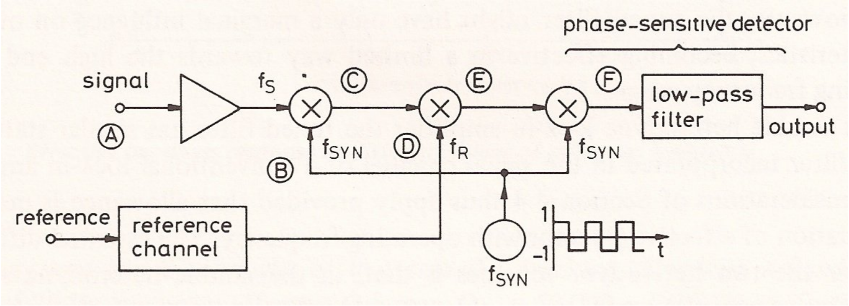


Fig. 8.11 Principles of synchronous heterodyning

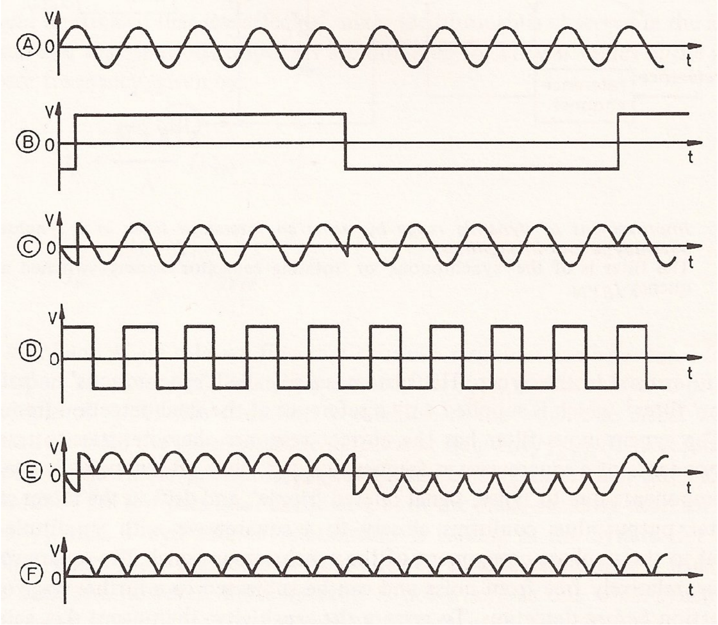


Fig. 8.12 Waveforms in the system of Fig. 8.11

In this mode of operation we find that the phase-sensitive detector can be constrained to operate in a frequency range well below the actual value of the reference frequency f_R . For example, phase-sensitive detection of a signal at 100 kHz could be achieved with a "syncrohet" frequency, f_{SYN}, of 100 Hz. The overall dynamic range would then be equivalent to that obtainable when operating the phase-sensitive detector in a conventional fashion at 100 Hz rather than the higher value of 100 kHz. It is this aspect of operation that has made the synchronous heterodyne approach so attractive in heterodyne systems where the phase-sensitive detector is referenced at a high value of intermediate frequency.

8.8.2 Dynamic range improvement

The synchronous heterodyne system described so far offers no inherent improvement in dynamic range at low reference frequencies. This was obtained in the EG&G PAR Syncro-Het system by imposing a filter between the final mixer and the phase-sensitive detector as shown in Fig. 8.13.

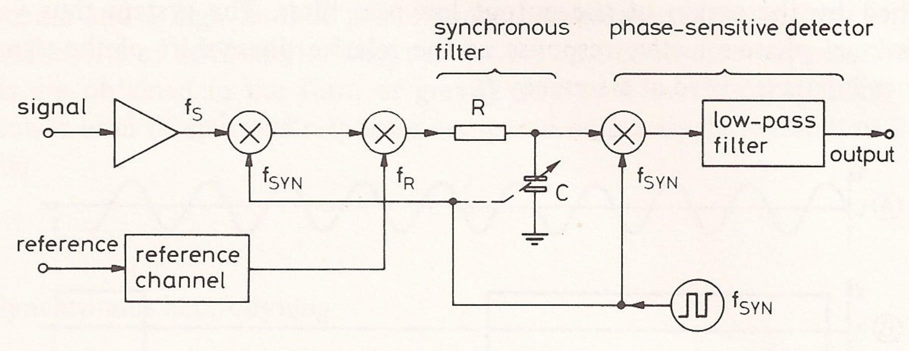


Fig 8.13 Improvement of dynamic range by using an interstage filter in a synchronous heterodyne lock-in amplifier. The filter is of the "synchronous" or "rotating capacitor" variety switched at frequency f_{SYN}

The filter used in the Syncro-Het system is a so-called "synchronous" or "rotating capacitor" filter[2] which is supplied with a reference at the final detection frequency f_{SYN}. The synchronous filter has the correct response characteristics to transmit the components of a squarewave at frequency f_{SYN}, while attenuating all asynchronous components due to noise, signal related "ripple", and drift in the mixer stages. The filter output thus conforms closely to a squarewave with a amplitude proportional to the in-phase component of the synchronous signal. The squarewave is, moreover, relatively free from noise and can be subjected to a further stage of a.c. amplification *before* detection. To restore the sensitivity, the output d.c. gain can then be reduced in proportion to the extra a.c. gain supplied in the signal path. The result is a system which offers improved output stability at a given level of dynamic reserve. In the light of the definitions given in Chapter 3, this is equivalent to a system with improved dynamic range.

The Syncro-Het lock-in amplifier operated with $f_{SYN} = 11$ Hz and proved capable of ×3000 (70 dB) dynamic reserve, consistent with an output stability of 10 p.p.m. This is equivalent to an input dynamic range of a massive 3.10^8 (170 dB). Unfortunately, the overall performance was marred by a series of spurious low-frequency outputs which occurred for critical combinations of the applied reference frequency and the internal "syncrohet" frequency f_{SYN}.

Mixer feedthrough, giving rise to a residual mixer output in the absence of signal, is a major cause of these low-frequency "beat" products. In general, the residual output from the second mixer will comprise components at combinations of f_R, f_{SYN} and their odd harmonics, characterized by frequencies:

$$|Mf_{SYN} \pm Nf_R| \;;\quad M, N \text{ odd}$$

The phase-sensitive detector is referenced at frequency f_{SYN} and is responsive to inputs at frequencies $K\,f_{SYN}$ where K is an odd integer. Spurious responses will occur whenever f_R takes values such that:

$$|Mf_{SYN} \pm Nf_R| = Kf_{SYN}$$

It should be stressed that the effect of mixer feedthrough is observed in the absence of signal and that spurious responses are obtained for critical values of the applied *reference* frequency, given by:

$$f_{\mathrm{R}} = \frac{M+K}{N} f_{\mathrm{SYN}}$$

and

$$f_{\mathrm{R}} = \frac{|M-K|}{N} f_{\mathrm{SYN}}$$

Since *M, N* and *K* are odd we find that spurious responses can be avoided (at least with respect to mixer feedthrough) provided that f_{SYN} is constrained to be an odd submultiple of f_{R}. If this condition is established, the worse-case effect will be a beat product from the phase-sensitive detector at frequency f_{SYN}. This response will not be discernible provided f_{SYN} is chosen to be greater than the maximum bandwidth of the low-pass filter in the output of the phase-sensitive detector.

Clearly, such a well-defined relationship is lacking in the Syncro-Het lock-in amplifier, where f_{SYN} is fixed and f_{R} is allowed to take values over a wide range. As we shall see, however, the constraint is not a serious one when synchronous heterodyning is applied to a heterodyne lock-in amplifier operating at a fixed intermediate frequency.

8.8.3 Application to heterodyne lock-in amplifiers

The implementation of synchronous heterodyning in the EG&G Brookdeal heterodyne lock-in amplifiers follows the block diagram shown in Fig. 8.14(a). The phase sensitive detector operates at a frequency f_{SYN} which is well below the highest intermediate frequency of 1MHz. The complexities of interstage filtering are avoided, and further reduction in complexity is achieved by using only one mixer as opposed to the double mixer stage assumed earlier.

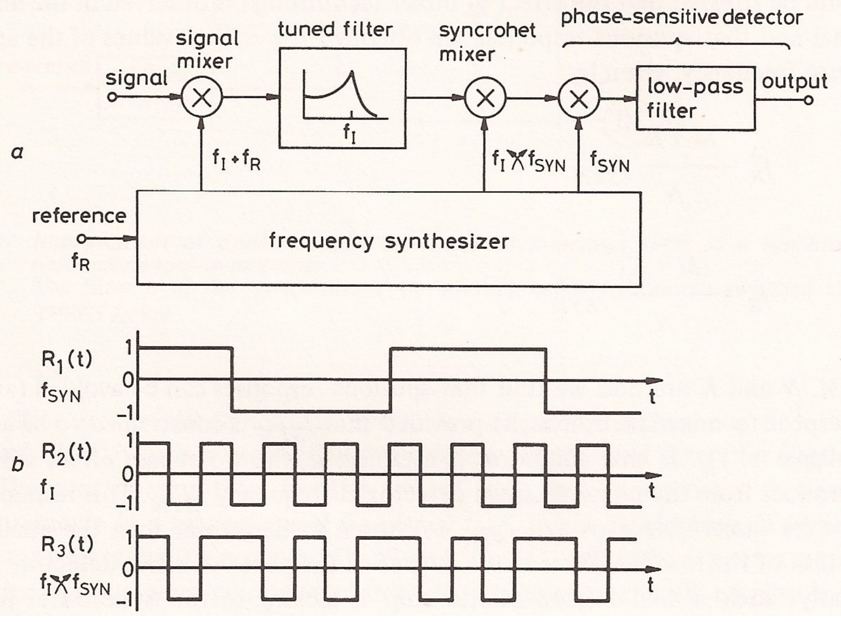


Fig. 8.14 (a) Incorporation of synchronous heterodyning in a heterodyne lock-in amplifier. Reference processing and provision for phase-shifting are omitted for clarity
(b) Generation of the switching input to the syncrohet mixer f_{SYN} is obtained by odd-integer frequency division from f_{I} to minimize spurious responses. In practice, f_{SYN} will be several orders of magnitude less than the highest value of intermediate frequency

In our original description of synchronous heterodyning, the signal was mixed successively with signals at frequencies f_{SYN} and f_R. For a signal $s(t)$, the output from the second mixer has the form:

$$s(t)\,R_1(t)\,R_2(t)$$

where $R_1(t)$ and $R_2(t)$ are squarewaves at f_{SYN} and f_R .

The alternative approach, using only one mixer, is to multiply the signal directly by a two-state waveform:

$$R_3(t) = R_1(t)\,R_2(t)$$

In practice, $R_3(t)$ can be formed by combining the individual switching waveforms $R_1(t)$ and $R_2(t)$ in an exclusive-NOR circuit. The resulting waveforms are derived in Fig. 8.14(b). In the EG&G Brookdeal heterodyne lock-in amplifier, f_{SYN} is chosen to be about 1 kHz, to lie beyond the bandwidth of the output filter, and is generated by odd-integer frequency division from the intermediate frequency oscillator. As explained in the last section this minimises the incidence of spurious outputs resulting from feedthrough in the syncrohet mixer. The method of generation avoids the coincidence of transitions in the switching waveforms $R_1(t)$ and $R_3(t)$ which is, in itself, a source of spurious outputs in systems where f_{SYN} takes arbitrary values.

8.9 Conclusions

This chapter has shown how a simple idea, that of frequency-shifting to avoid harmonic responses, has been progressively modified to allow for practical limitations in the performance of the various subsystems which comprise a heterodyne lock-in amplifier. The result is an instrument operating at a level of complexity greatly in excess of a conventional broadband lock-in amplifier and which falls short of conventional systems in several important specification areas, notably dynamic range and phase accuracy.

These drawbacks, to judge from the popularity of heterodyne systems, are more than offset by the advantages in operating with fundamental-only response allied to a low level of spurious responses. In terms of all-round performance, therefore, heterodyne systems confirm most closely with the characteristics demanded of a "general purpose" measurement tool capable of making unambiguous measurements with a variety of signal types over a wide range of operating frequencies. The availability of these instruments with a comprehensive facility for digital control adds to their appeal in a wide range of applications. Some of the characteristics of these systems are reviewed in Chapter 10, while a comparison with p.w.m. systems, which offer an alternative approach to achieving fundamental-only response, is given at the end of Chapter 9.

8.10 References

1 MUNROE, D.M. (1973): "The heterodyning lock-in amplifier". Technical Bulletin, Ithaco Corp., Ithaca, NY.

2 FRANKS, L.E. and SANDBERG, I.W. (1960): "An alternative approach to the realization of network transfer functions: The N-path filter", *Bell Syst. Tech. J.*, 39, pp. 1321-1350.

CHAPTER 9

P.W.M systems

9.1 Introduction

Single and double heterodyne systems were developed to overcome the problem of harmonic responses in lock-in detection while retaining the ability to track signals over a wide frequency range. In this chapter, we shall be looking at an alternative method of suppressing harmonic responses, by using pulse-width modulation in the reference channel.

It will become apparent that heterodyne and pulse-width modulation systems have rather different status. Whereas heterodyne lock-in amplifiers have a system configuration which is far removed from the "conventional" arrangement, the pulse-width modulation approach requires only a relatively simple modification to an otherwise basic system. As a result, the pulse-width modulation, or p.w.m., circuitry can be supplied as an option to a conventional lock-in amplifier, leaving the user to select either conventional response or fundamental-only response as required. As in the case of heterodyne systems, the suppression of harmonic responses is achieved at the expense of spurious responses at apparently arbitrary frequencies and is attended by a loss of dynamic range. We shall therefore be examining some of the trade-offs which must be made to achieve a detection system with overall acceptable characteristics.

P.W.M. lock-in amplifiers are characterized by phase accuracy of a very high order and represent a solution to suppressing harmonic responses which is both cost-effective and flexible. Other benefits, such as the potential to operate with greatly improved slew rate and the ability to operate the phase sensitive detector as an "ideal" multiplier, will also be noted.

A brief comparison of p.w.m. systems with heterodyne lock-in amplifiers is given at the end of this chapter.

9.2 Principles of operation

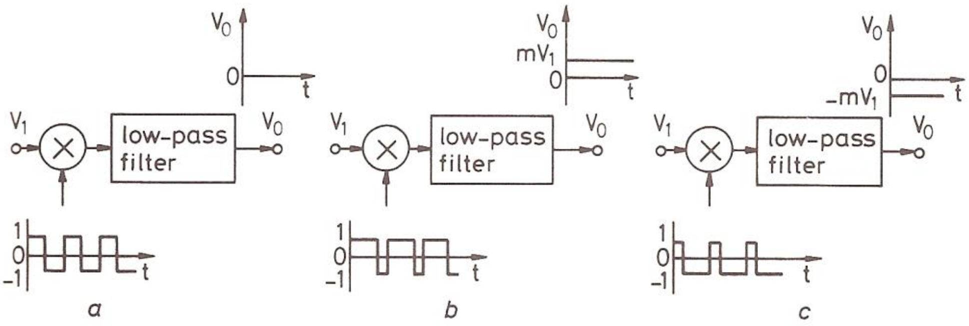


Fig. 9.1 Effect of reference waveform symmetry on the average gain of a phase-sensitive detector shown for the signal at a fixed level, V_1 volts. The average gain is (a) zero; (b) $+m$; (c) $-m$

The harmonic responses of a phase-sensitive detector result from the abrupt change of gain between +1 and –1 which occurs whenever the reference switching waveform changes polarity. If we were to measure the *average* gain of the phase-sensitive detector over many consecutive reference cycles the result would of course be zero, provided that the reference switching waveform spent an equal amount of time in its two states. As shown in Fig. 9.1(a), (b) and (c) the effect of changing the symmetry of the switching waveform is to change the average gain of the phase-sensitive detector to a value somewhere between the extremes of +1 and -1.

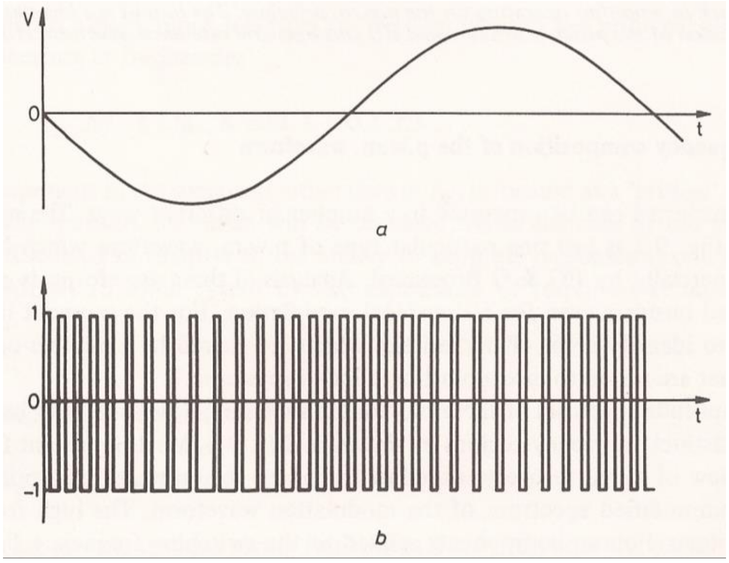


Fig. 9.2 (a) Modulation voltage; (b) pulse-width-modulated switching waveform

Suppose now that the waveform symmetry is subject to a long-term variation imposed by modulating the switching waveform as shown in Fig.9.2. The average gain of the phase-sensitive detector (Measured over a time which is long compared with a switching cycle but short compared with a modulation period) is now subject to a continuous variation which is free from discontinuities. The variation in gain, moreover, reproduces the modulation waveform exactly.

This approach, the pulse-width modulated reference channel, provides a means of achieving a sinusoidal gain variation in the signal path while retaining the dynamic range benefits of a switching phase-sensitive detector. A sinusoidal variation has been chosen because this clearly brings us a step closer to a system with fundamental-only response. To take a broader view: the system described approximates to an ideal multiplier model in which the reference input could be of any waveform, supplied in the form of a modulation voltage. Fig. 9.3 shows this configuration of a lock-in system working on this principle.

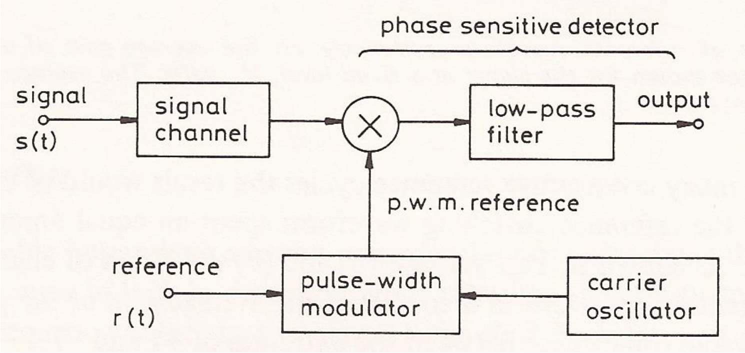


Fig. 9.3 Lock-in amplifier operating on the p.w.m. principle. The output is a low-pass filtered version of the product of the signal *s*(*t*) and a general reference waveform *r*(*t*)

9.3 Frequency composition of the p.w.m. waveform

P.W.M. waveforms can be generated in a number of different ways. The waveform shown in Fig. 9.2 is just one particular type of p.w.m. waveform which has been used commercially by EG&G Brookdeal. Analysis of these waveforms is generally a protracted business even for a sinusoidal modulation. For the moment it will be sufficient to identify some characteristics which are shared by a number of p.w.m. schemes that are suited to incorporation in lock-in systems.

The amplitude spectrum of these modulated switching waveforms can be divided into two distinct frequency regions as shown in Fig 9.4. Most important from the point of view of signal recovery is the low-frequency region which incorporates the complete unmodified spectrum of the modulation waveform. The high frequency region contains Fourier components related to the switching frequency f_0 and its harmonics, each of which carries sidebands derived from the Fourier components of the modulation.

The switching waveform thus combines the Fourier components of the modulation voltage with other, non-harmonically related, components at higher frequencies. The latter account for the switching characteristics of the waveform and each is associated with a transmission window where the detection system is susceptible to interference components.

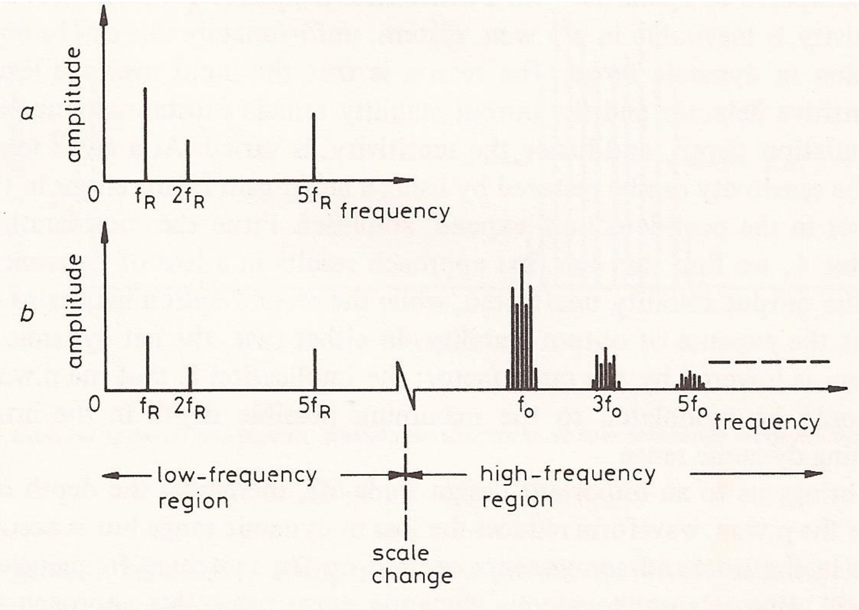


Fig 9.4 Amplitude spectra of: (a) modulation voltage; (b) a typical p.w.m. switching waveform

For fundamental-only response, the modulation will be sinusoidal at frequency f_R which is usually much less than f_0. We then identify the principal components of the spectrum at frequencies:

$$f_R,\ Kf_0 \pm Lf_R;\quad K \text{ odd},\ L = 0,\ 1,\ 2,\ 3\ ...$$

Each component in the spectrum, other than at f_R, is located at a "critical" frequency near which spurious responses will be obtained. As in the case of the heterodyne systems discussed in Chapter 8, the ability to suppress responses at odd harmonics of f_R is offset to some extent by the appearance of responses at unrelated frequencies.

9.4 Basic design considerations

9.4.1 Dynamic range

In practical p.w.m. systems giving fundamental-only response the magnitude of the component at frequency f_R is found to vary linearly with the amplitude of the modulation voltage which, in turn, determines the depth of modulation in the p.w.m. waveform. The component at f_R can take reasonably large values - of the same order of magnitude as the dominant high-frequency components - without causing over-modulation. This turns out to be very important in phase-sensitive detection where the "primary" response at $f_s = f_R$ should be as large as possible. Failure to achieve this means that the overall detection system will suffer a significant loss of sensitivity compared to operation with a conventional squarewave reference. Some loss of sensitivity is inevitable in a p.w.m. system: unfortunately this can be equated to a reduction in dynamic range. The reason is that the input overload level to the phase-sensitive detector and the output stability remain substantially unchanged as the modulation depth, and hence the sensitivity, is varied. At a given modulation depth, the sensitivity can be restored by using a larger gain factor either in the signal channel or in the post-detection "expand" amplifier. From the considerations given in Chapter 4, we find that the first approach results in a loss of dynamic reserve, leaving the output stability unaffected, while the second option maintains dynamic reserve at the expense of output stability. In either case, the net dynamic range of the system is lowered by the same factor; the implication is that the p.w.m. waveform should be modulated to the maximum possible depth in the interests of maintaining dynamic range.

This brings us to an important design trade-off. Increasing the depth of modulation on the p.w.m. waveform reduces the loss in dynamic range but is accompanied by a rise in the sideband components centred on the switching frequencies and its harmonics. Any attempt to recoup dynamic range using this approach is consequently matched by an increase in the general level of spurious responses. In addition, high-order sidebands which had negligible magnitude will now introduce transmission windows in a frequency range closer to the reference frequency, and so assume greater practical importance.

9.4.2 Spurious responses

In general, the magnitude and extent of the sideband arrays centred on f_0 and its harmonics vary non-linearly with modulation depth and can only be determined for a given modulation scheme by exact analysis. The scheme exploited by EG&G Brookdeal in their Sinetrac systems is particularly difficult to analyse because, in addition to modulating the mark/space ratio of the switching waveform, the modulating signal also causes a shift of the carrier frequency. We shall therefore restrict ourselves to a general review of system behaviour.

In the Sinetrac system, the sinewave modulation voltage at frequency f_R is specified at a standard level of 1 V r.m.s. The modulation depth corresponding to this level of modulation signal gives a loss of 10 dB in the sensitivity of the phase-detector. The sensitivity is regained by introducing an extra gain stage of 10 dB in the signal channel, with a consequent loss of 10 dB in the system dynamic reserve.

When the modulation is at a reference frequency f_R which is much lower than f_0, the sidebands centred on the carrier frequency are closely spaced at frequencies $f_0 \pm Lf_R$ but the *extent* of the sideband array is strictly limited as shown schematically[*] in Fig. 9.5. The total width, $2\Delta f$, is typically 20 kHz and substantially independent of modulation frequency when f_R takes sufficiently low values, giving a spectrum similar to that of a frequency-modulated carrier under conditions of large-index modulation.

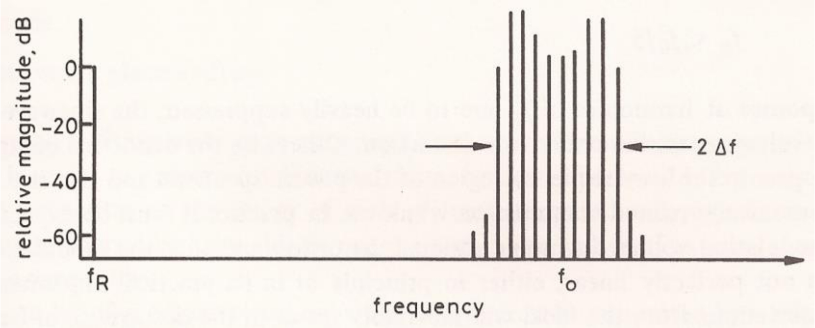


Fig. 9.5 General form of the p.w.m. waveform spectrum at low reference frequencies

This general behaviour, giving rise to a p.w.m. spectrum with sensibly fixed characteristics, corresponds to operation with f_R less than about $f_0/30$. For reference frequencies in this range, a clear separation is maintained between the low- and high-frequency regions of the spectrum. The lowest critical frequency, where a spurious response of significant magnitude could be obtained, is now in the region of $f_0 - \Delta f$,which is well removed from f_R.

When f_R is increased, keeping the depth of modulation constant, the separation of the sidebands becomes correspondingly larger. In order to predict the critical frequencies and the magnitude of their associated transmission windows, it now becomes necessary to give individual attention to each of the sidebands. In practice it will be the sidebands below the carrier frequency which prove to be the most troublesome, located at frequencies $f_0 - Lf_R$.

It turns out that for sufficiently high reference frequencies, the major contribution to spurious responses comes from the low-order sidebands for which $L \le 4$. This is illustrated by Fig. 9.6, drawn for the specific case where $f_R = f_0/10$. Here, the separation between the low- and high-frequency regions of the p.w.m. spectrum is much less well defined. Also, it is evident that, if f_R exceeds a certain value, one of the sidebands will enter the frequency range *below* f_R.

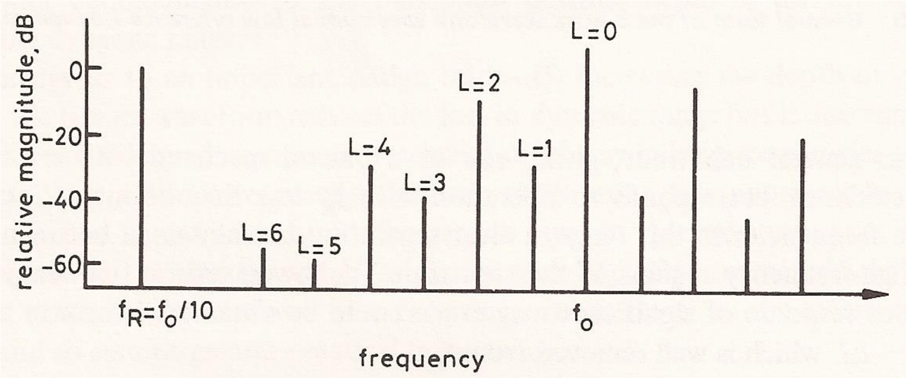


Fig. 9.6 Major sideband components for $f_R = f_0/10$, shown relative to the "primary" component at f_R

[*] Note that, in the Sinetrac system, application of the modulation voltage shifts the mean carrier frequency by about 10% from its "free-running" value. We shall ignore this effect in the following discussion.

To avoid the possibility of a transmission window appearing at an apparently arbitrary frequency below the reference, we must place an upper bound on the reference frequency. Assuming that sidebands corresponding to $L > 4$ have negligible magnitude, this gives us the condition

$$f_0 - 4f_R > f_R$$

which limits the reference frequency to the range

$$f_R < f_0/5$$

If responses at harmonics of f_R are to be heavily suppressed, the sinewave modulation voltage must have very low distortion. Otherwise the distortion components will appear in the low-frequency region of the p.w.m. spectrum and give rise to a set of harmonically-related transmission windows. In practice it must be expected that the modulation voltage has some residual distortion and that the modulation process is not perfectly linear, either in principle or in its practical implementation. Such deviations from the ideal will inevitably result in the occurrence of harmonic responses, albeit at low level.

9.4.3 Choice of switching frequency

For maximum separation between the low- and high-frequency regions of the p.w.m. spectrum, the switching frequency f_0 should be chosen to be much higher than the maximum anticipated reference frequency. The limitation on f_0 is decided ultimately, by the dynamic range of the phase-sensitive detector which deteriorates at high switching frequencies. This deterioration is compounded by the additional loss in dynamic range inherent in the p.w.m. approach. It has been observed that a p.w.m. reference channel can be configured as an option to an otherwise conventional lock-in amplifier. If this is the case, the phase-sensitive detector will be optimized over a range of frequencies rather than at a fixed high frequency. The p.w.m. switching frequency must then be chosen to be comparable with the highest frequency envisaged in conventional operation. Inevitably, the highest permitted value of f_R in p.w.m. operation must then be significantly less than this value. The EG&G Brookdeal Sinetrac lock-in amplifiers are subject to such a constraint; here the maximum recommended reference frequency in p.w.m. operation is 25 kHz in a system which can operate to frequencies above 100 kHz in conventional mode.

9.5 Reference phase-shifting

As described so far, p.w.m. systems depend on the provision of a sinewave reference voltage in order to achieve fundamental-only response. This would obviously place a severe restriction on the utility of such systems compared with, say, heterodyne lock-up amplifiers which are able to operate with a wide range of externally applied reference waveforms.

Also, to be of practical value, a p.w.m. system must be supported by a reference phase-shift network to enable the phase of the sinewave modulation voltage to be adjusted relative to a synchronous signal.

Clearly, these two drawbacks can only be overcome by adding to the complexity of the reference channel. In deciding on a suitable processing system, the following factors must be taken into account:

(i) Although we have identified an upper limit on f_R, to avoid low-frequency spurious responses, there is no fundamental limit on the lowest value of f_R which might be used. This implies that to exploit the p.w.m. technique to the full, the reference processing circuits should be capable of operating over a wide range of frequencies, amounting to several decades.

(ii) The sensitivity of the phase-sensitive detector, and hence the calibration of the overall system, depends directly on the amplitude of the sinusoidal modulation voltage, which should therefore have constant value over the entire operating range.

(iii) The sinewave applied to the pulse-width modulator should have a very low level of distortion if the detection system is to reject responses at harmonics of the reference frequency.

These requirements are fully met by the reference channel used in the EG&G Brookdeal Sinetrac series of lock-in amplifiers. The configuration is shown in Fig. 9.7. The reference input stage and the broadband phase-shift network are those of a strictly conventional lock-in amplifier, capable of operating with high precision over a frequency range in excess of five decades. The phase-shifted output of the reference channel is a closely controlled squarewave which is subsequently converted, first to a triangle and then to a sinewave. The circuits used for squarewave-to-triangle conversion have been described by Carter and Faulkner[1]; the final conversion to sinewave form is given by a piecewise linear network adjusted for a low level of distortion over the full frequency range.

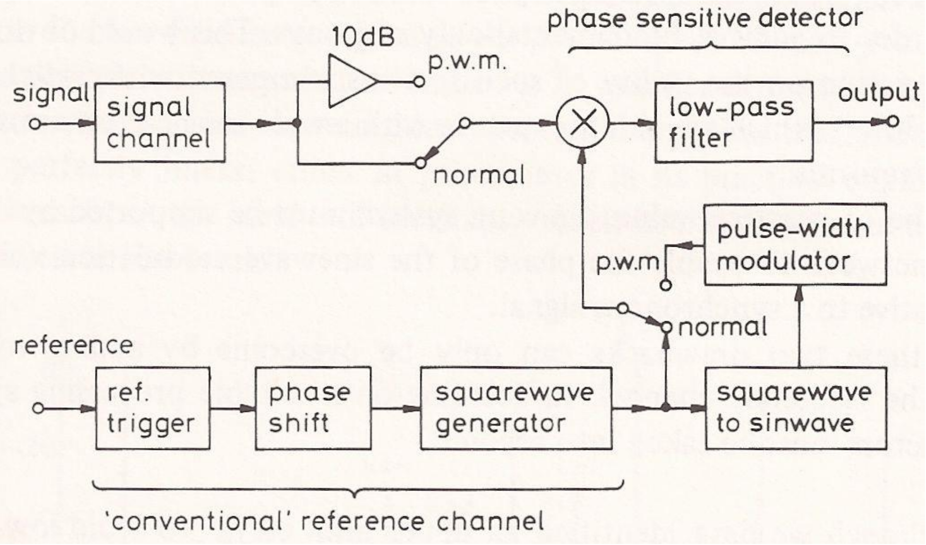


Fig. 9.7 Broadband lock-in amplifier with facility for fundamental-only operation in p.w.m. mode

Fig. 9.7 also shows the relatively simple arrangement of switches required to convert a harmonically-responding, conventional, lock-in amplifier to a system with fundamental-only response. An additional signal-channel gain stage of 10 dB is used to maintain the overall sensitivity of the system in p.w.m. mode. The diagram serves to emphasise that the p.w.m. configuration does not rely on the use of filters, and is consequently free from phase and amplitude errors due to filter misalignment. The phase accuracy and the residual phase noise are therefore comparable with those of the basic conventional system and significantly better than that of a heterodyne lock-in amplifier operating in the same reference frequency range.

9.6 Two-phase systems

The extension to two-phase systems requires an additional squarewave-to-sinewave convertor operating on the squarewave output of the quadrature reference channel. A second pulse-width modulator, operating independently of the first, is used to supply the reference input to the quadrature phase-sensitive detector.

It should be noted that p.w.m. systems do not depend on a carefully adjusted carrier frequency. In practice, the carrier frequencies of the two pulse-width modulators used in a two-phase lock-in amplifier need to be no more than nominally equal.

9.7 Analogue correlation

"Analogue correlation" is a term used in the context of p.w.m. lock-in amplifiers to cover the various modes of operation made possible when an external reference waveform is applied directly to the pulse-width modulator. This opens up numerous possibilities some of which are reviewed in the following sections.

9.7.1 Matched detection

The idea of a "matched" detector was mentioned in Section 3.5.5 in relation to the measurement of a squarewave signal using a conventional squarewave reference. If we now suppose that a periodic but non-sinusoidal waveform was used as the modulation input to a p.w.m. lock-in amplifier, the resulting detection system would be characterized by a set of transmission windows which could be exactly matched in amplitude, frequency and phase to the Fourier components of the signal. Such a system proves capable of yielding the best possible output signal-to-noise ratio for signals obscured by white noise. Unfortunately, signal recovery problems are usually associated with noise spectra far more complicated than this, so that the benefits of matched detection (which are, in many cases, marginal) are difficult to realize in practice. It usually turns out that the ability to operate in a fundamental only response mode with relative freedom from transmission windows close to the reference frequency gives a far greater advantage when measuring non-sinusoidal signals in noise. As noted in Section 3.5.5 and in Chapter 6, the null-shift procedures can be applied when the detection system has fundamental-only response, to quickly bring the reference phase-shift to an optimum setting under very noisy conditions.

9.7.2 Two-frequency lock-in analysis

In a two-phase lock-in amplifier operating on the p.w.m. principle, the pulse-width modulators associated with each phase-sensitive detector operate independently and have separate inputs. These can be supplied with external reference waveforms having different fundamental frequencies and different waveforms if so required. The ability to use independent reference inputs means that different spectral components of the signal can be separately measured on each of the phase-sensitive detectors.

Since direct connection of external references to the modulator inputs bypasses the reference phase-shifting networks, this facility is likely to be most useful in such applications as optical spectroscopy where there is minimal phase-shift between the reference and signal. An example of spectrometer operating with two chopper frequencies is shown in Fig. 9.8. The light paths from the two samples are combined at the cathode of single photomultiplier. A two-phase p.w.m. lock-in amplifier is subsequently used to measure the two chopped signals separately and simultaneously, using "analogue" reference inputs derived from the drives to the optical choppers. A ratiometer can be a useful accessory in this type of measurement; however, a digital interface of the type described in Chapter 10 provides a more flexible approach to processing the outputs from the twin channels.

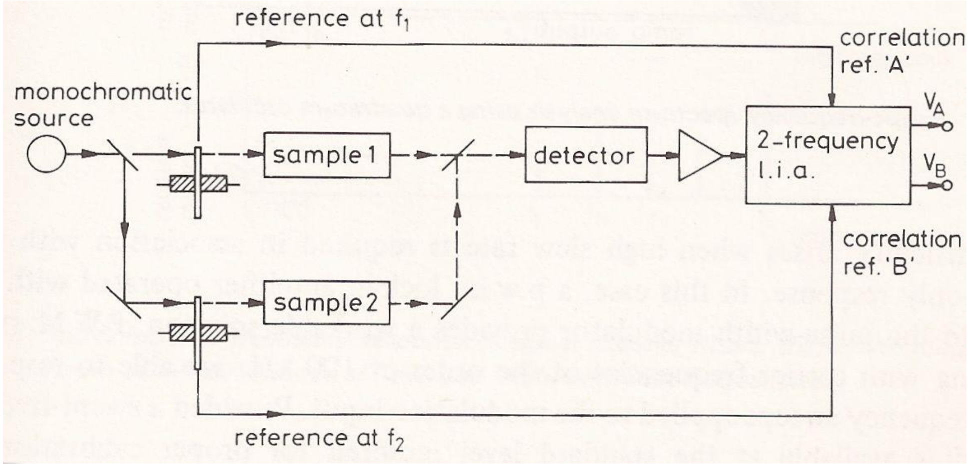


Fig. 9.8 A two-frequency lock-in amplifier application. The outputs V_A and V_B are proportional to the signals transmitted by samples 1 and 2 respectively.

9.7.3 High slew rate applications

In principle, the limited slew rate capability of conventional lock-in amplifiers could be overcome by bypassing the reference phase-shifting networks and applying a swept-frequency switching waveform directly to the phase-sensitive detector. Such a procedure is valid when the signal and reference remain sensibly in phase over the desired frequency sweep. In practice, the procedure can be applied to conventional lock-in systems only when the phase-sensitive detector is accessible in the form of a modular unit.

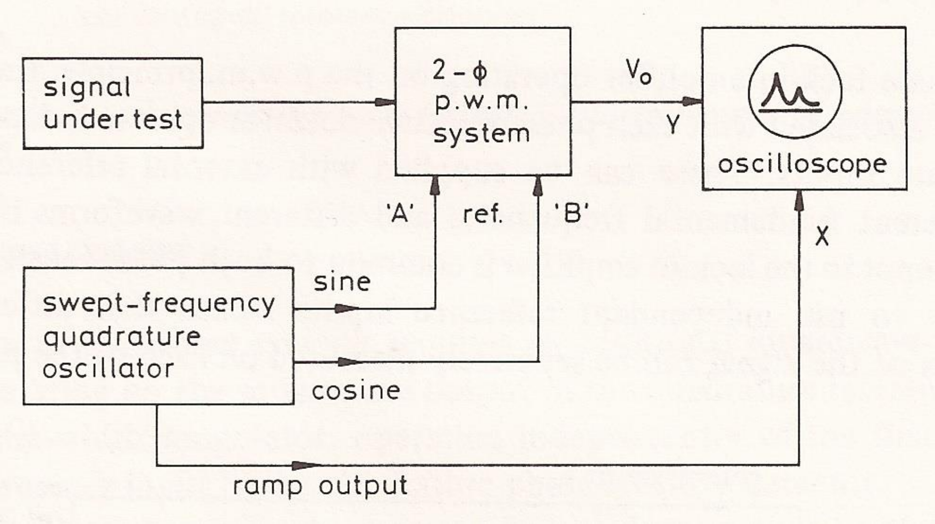


Fig. 9.9 Swept-frequency spectrum analysis using a quadrature oscillator

A difficulty arises when high slew rate is required in association with fundamental-only response. In this case, a p.w.m. lock-in amplifier operated with direct access to the pulse-width modulator provides a workable solution. P.W.M. systems operating with carrier frequencies of the order of 100 kHz are able to respond to rapid frequency sweeps applied to the modulation input. Provided a swept-frequency sinusoid is available at the standard level required for proper calibration, it is possible to obtain fundamental-only response consistent with slew rates far in excess of most practical requirements.

Wide-band swept spectrum analysis was cited in Chapter 6 as a lock-in amplifier application where fundamental-only response and high slew rate were essential joint requirements. In order to exploit the characteristics of a two-phase p.w.m.

system in this application, it is necessary that the two pulse-width modulators are operated in strict quadrature at the swept reference frequency. This highlights the need for a swept-frequency oscillator providing quadrature sinewave outputs at a standard level. Such oscillators have been made available as accessories to two-phase p.w.m. lock-in amplifiers for use in the configuration shown in Fig. 9.9. In wideband applications using high frequency resolution, the sweep-rate limitation in this type of system lies with the output filters as explained in Section 5.5.3.

9.8 Interference rejection filters

In an earlier version of the Sinetrac lock-in amplifier, the signal channel was fitted with a 2-pole low-pass filter cutting off above the maximum reference frequency of 25 kHz but well below the 100 kHz switching frequency. In addition, a notch filter was used to enhance the suppression of signal components in the region of 100 kHz. The object was to overcome the major spurious responses associated with p.w.m. operation.

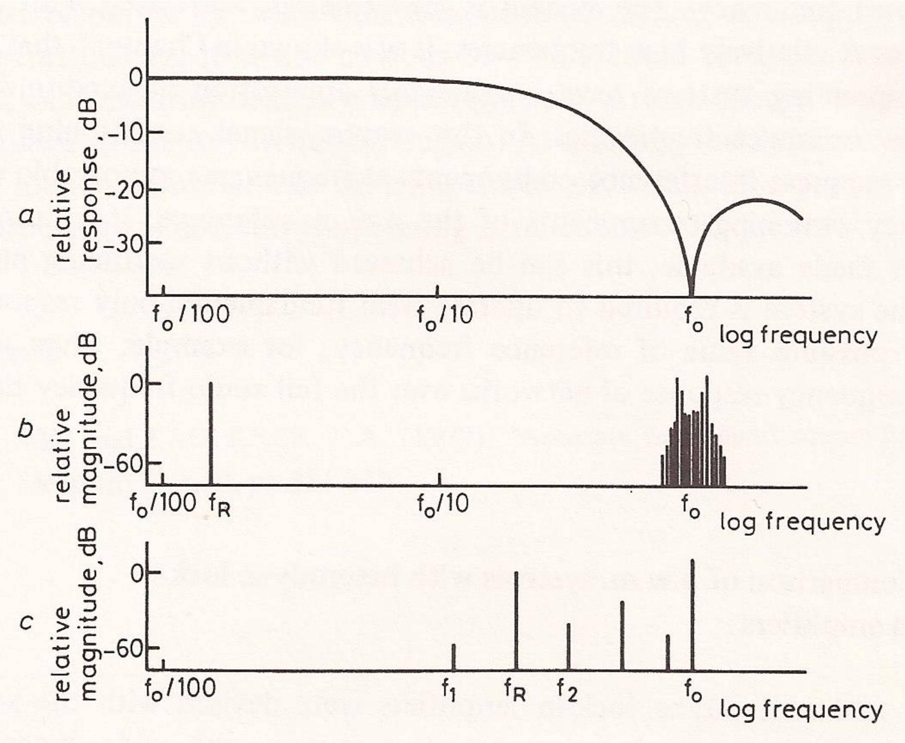


Fig 9.10 (a) Combined frequency response of signal channel filters. (b), (c) Amplitude spectrum of PWM reference at "low" and "high" reference frequencies

In practice, such a combination of filters is likely to prove desirable only at comparatively low reference frequencies for the following reasons. First of all, we have seen that at low reference frequencies the transmission windows are concentrated in the spectral regions close to the switching frequency and its odd harmonics. Spurious responses are therefore associated with high-frequency interference and the filters are most effective in suppressing these in advance of detection. Secondly, at low frequencies the phase-shift introduced by the filters will be relatively small, enabling the overall system to operate with good phase precision.

Conversely, at high reference frequencies, the signal channel filters will introduce large phase errors into the measurement system while the transmission windows move to much lower frequencies by reason of the wider sideband separation. The contrasting situations are illustrated in Fig. 9.10. We find that at higher reference frequencies there will be a number of critical frequencies, for example f_1 and f_2 which fall within the bandwidth of the signal channel filter. Interference components close to these frequencies would suffer minimal attenuation in the filter and be able to excite spurious responses. In some cases, therefore, the inclusion of the filter acts more to the detriment of the system, introducing large phase errors while having only partial success in overcoming the problem of spurious responses.

The decision to eliminate interference rejection filters with fixed characteristics in later versions of the Sinetrac system resulted in a lock-in amplifier capable of fundamental-only response, consistent with excellent phase accuracy, over many decades of frequency. The system is, nevertheless, associated with large spurious responses at relatively high frequencies. It was shown in Chapter 6 that fundamental-only responding systems have considerable application in experiments operating with low reference frequencies. In this regime, signal conditioning filters can be used to suppress interference components at frequencies comparable with the high-frequency switching components of the p.w.m. reference. If a range of optional filters is made available, this can be achieved without sacrificing phase accuracy when the system is required to operate with fundamental only response up to the highest possible value of reference frequency: for example, when measuring the swept-frequency response of networks over the full audio-frequency range.

9.9 Comparison of p.w.m. systems with heterodyne lock-in amplifiers

P.W.M. and heterodyne lock-in amplifiers were devised with the same objective in mind; to give a synchronous detection system with wide dynamic range and relative freedom from harmonic responses over a wide range of reference frequencies. As we have seen, the two approaches to this problem lead to vastly different solutions, both of which involve system designers in a number of trade-offs and compromises.

On balance, modern heterodyne lock-in amplifiers appear to offer the widest frequency range consistent with the lowest level of spurious responses, whereas commercial versions of the p.w.m. system operate up to a maximum frequency of about 25 kHz and have a number of large transmission windows accessible in the frequency range immediately beyond this value.

The overall phase accuracy of p.w.m. systems is superior to that of heterodyne lock-in amplifiers, which are susceptible to alignment errors in a number of sub-systems and generally require a far more complex configuration. The difference in complexity is reflected in system cost, since p.w.m. lock-in amplifiers usually offer a cheaper means of obtaining fundamental-only response than their heterodyne counterparts. Also p.w.m. systems can usually be converted by pushbutton selection to operate as conventional lock-in amplifiers, giving an extension of the frequency range and allowing the fundamental-only response to be traded for greater dynamic range. This flexibility in choosing the response of the system extends to choosing an arbitrary response given by the Fourier components of the applied reference waveform.

A feature that both heterodyne and p.w.m lock-in amplifiers have in common is that fundamental-only response at low frequencies is obtained by operating the phase-sensitive detector at a relatively high frequency. The result in both cases is a system with a dynamic range independent of reference frequency but less than that which might be achieved if the phase-sensitive detector was operated in conventional fashion. It was shown in Chapter 8 that the dynamic range of heterodyne systems can be recouped by using synchronous heterodyning as a supplementary technique. Unfortunately, it is not obvious how this could be applied to phase-sensitive detectors operating with a p.w.m. reference without incurring additional spurious responses in the low-frequency region. As a result, the dynamic range of a heterodyne lock-in amplifier can be comparable with that of a p.w.m. system where the phase-sensitive detector is operating at much lower frequency.

9.10 Reference

CARTER, S.F., and FAULKNER, E.A. (1977): "Accurate broadband square-to-triangle converter", Electron, Lett., 3, pp. 381-382.

CHAPTER 10

Computer-controlled lock-in amplifiers

10.1 Introduction

The advent of the microprocessor and the increasing availabity of desk-top computing power have provided a challenge to both designers and users of electronic measuring equipment. Instrument designers are faced with a demand for ‘intelligent’ instruments capable of performing programmed tasks or able to communicate with other instruments via a computer controller. As for the instrument user; he is concerned with using these new instruments to the best effect and with devising measurement procedures that take advantage of the latest developments in instrument technology.

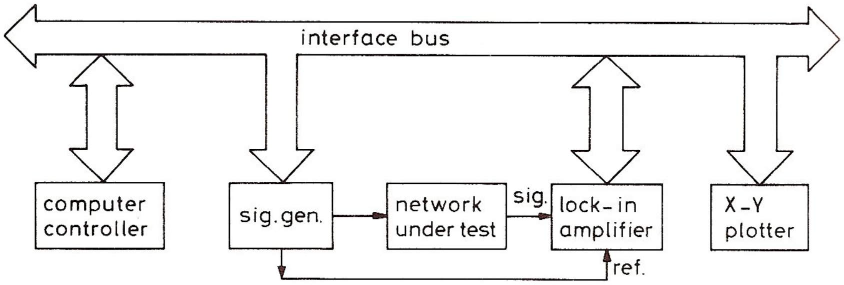


Fig. 10.1 Bus-controlled measurement system

This general situation is reflected in the specific case of instrumentation for signal recovery where increasing emphasis is being placed on computer control in its widest sense. Developments in this area have been greatly influenced by the widespread adoption of the IEEE-488 bus system for providing two-way digital communication between measuring instruments and a computer-controller. This has resulted in the availability of a large range of compatible instruments that can be, literally, plugged together to produce a computer-controlled measurement system. Many of the applications listed in Appendix 1 could benefit from such an approach; for example, Fig. 10.1 shows a bus-compatible lock-in amplifier operated in conjunction with a number of other controllable devices to provide an automatic system for frequency-response measurement.

The IEEE-488 bus protocol is rigidly defined; thus, at any time, only one device is permitted to ‘talk’, that is send data or commands over the bus, while several devices may ‘listen’ in order to receive data or commands. In the example shown, the *X-Y* plotter and signal generator would probably operate as ‘listeners’ while the lock-in amplifier would both ‘talk’ and ‘listen’, to transmit data to the computer and receive commands. Overall control comes from the computer, which is programmed to change the frequency of the signal generator in discrete steps, to manage the take-up of data from the lock-in amplifier, and to process data for presentation on the *X-Y* plotter.

In the computer-controlled lock-in amplifiers to be discussed in this chapter, the phase-sensitive detector remains intact at the heart of the system, supported by signal and reference channels having characteristics similar to those described in earlier chapters. The incorporation of digital control lines to switch the sensitivity and the internal configuration of the lock-in amplifier can be achieved without compromising key specifications such as input dynamic range and operating frequency range. In practice, therefore, the only serious limitation incurred in operating with a computer interface is with regard to output dynamic range. This is now limited by the use of an analogue-to-digital convertor on the phase-sensitive detector output. An attempt to match the 100 dB dynamic range of a typical analogue output would require a 17-bit conversion and would be difficult to justify on grounds of cost in a general-purpose measurement system. The usual provision is for a $3^1/_2$ digit conversion, giving a resolution of 10 mV in a 10 V output with 100% over-range. If this is inadequate for a particular application, the analogue output is available on its usual socket and can be separately converted to high precision if so required.[*]

The handling characteristics of the lock-in amplifier being relatively unchanged, the main problem in digital control is to create programs which reproduce the measurement routines and setting-up routines that are associated with the detection of noisy signals. In giving consideration to these routines it will be convenient to distinguish between the two main types of controllable lock-in amplifier in general use. These are ‘programmable’ lock-in amplifiers where the software control routines are resident in an external computer controller, and microprocessor-based systems – so called ‘intelligent’ lock-in amplifiers. The latter feature a number of stored software routines that can be initiated by front-panel switch selection or by a command transmitted on the interface bus.

10.2 Programmable lock-in amplifiers

In early lock-in amplifiers, the pushbutton and switch selectors controlling the overall system configuration were heavily interlocked and interlinked and required front panel assemblies that were both complex and labour-intensive in production. At a later stage, f.e.t. switches, controlled by the application of standard logic levels, became widely used for both gain selection and mode selection and there was a move to transfer hard-wired switching logic to integrated circuits mounted on the printed circuit board. This change leads to a dramatic simplification in switch design. For example, the sensitivity switch of a typical lock-in amplifier is reduced from a multi-wafer assembly to a single-pole selector as illustrated in Fig. 10.2.

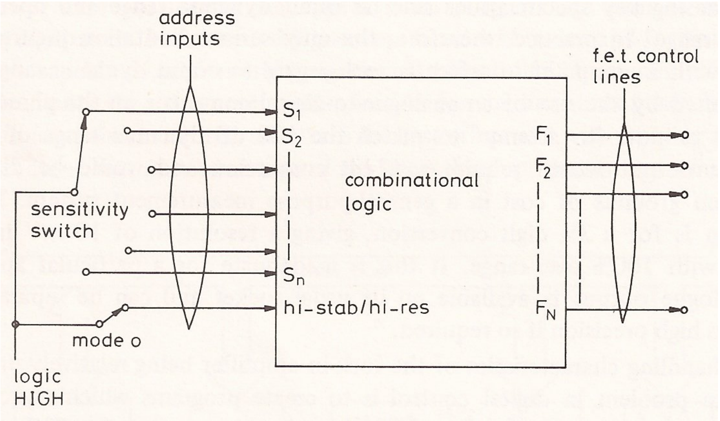


Fig. 10.2 Simplification of switching operations by using a combinational logic circuit

[*] Note that to exploit the full output dynamic range of these instruments generally places severe demands on peripheral equipment, both analogue and digital.

The sensitivity lines S_1 to S_n provide an address input to a combinational logic block which is tabled to produce the appropriate output combination on the f.e.t. control lines F_1 to F_N. These are then used to switch the gain of the amplifier stages in the lock-in amplifier signal channel. In a comprehensive detection system, giving a choice of ‘high stability’ or ‘high reserve’ operation, the address lines might be augmented by an additional input which is either HIGH or LOW depending on mode selection. The combinational logic circuit is then arranged to control the f.e.t. switches for any combination of mode and sensitivity, without adding to the complexity of the mechanical switching assembly.

The transition from a ‘standard’ instrument designed along these lines, to one where the internal switches can be controlled by the application of logic levels from an external source is relatively straightforward. In a fully ‘programmable’ lock-in amplifier, all the functions that are normally switched from the front panel, such as sensitivity, time constant, phase quadrants and ‘expand’, can be controlled from logic levels applied to a ‘digital’ input. In lock-in amplifiers having a voltage-controlled phase-shifter, the provision of a digital-to-analogue convertor enables the reference phase to be added to the list of controllable parameters. In its simplest form, the digital input might be a multi-way socket connected to a set of remotely operated switches. In modern instruments, however, the digital input is more likely to be a port having access to a standard bus system, such as the IEEE-488 bus referred to earlier. When the lock-in amplifier output is provided with an analogue-to-digital convertor, a properly defined bus system enables data to be transferred to the computer controller and to other instruments connected to, and controlled from, the bus.

In order to exploit this type of system effectively, the computer-controller must be provided with programs sufficiently powerful to undertake the management of the lock-in amplifier under a wide range of signal and noise conditions. We thus envisage a control program that defines an overall measurement procedure and contains a number of subroutines for sensitivity and phase selection. The specification of these subroutines requires a certain familiarity on the part of the user with the handling characteristics of lock-in amplifiers and would normally involve several stages of refinement before an acceptable solution was found. Some essential features of these routines are identified and discussed in Sections 10.4 and 10.5.

10.3 Microprocessor-based systems

The incorporation of a microprocessor to monitor and supervise the switching functions of a lock-in amplifier represents a significant advance in system concept and design. The result is a self-contained lock-in amplifier with the ability to undertake *sequential* switching operations controlled by software associated with the microprocessor. The control system takes additional data from

(i) front panel switch arrays

(ii) a digital interface to an external keyboard or controller

(iii) the outputs of the phase-sensitive detectors.

A lock-in system with this overall capability greatly cases demands on the user who no longer requires such detailed familiarity either with lock-in techniques in particular or signal recovery in general. For example, the resident software routines could enable the lock-in amplifier to adjust sensitivity and phase automatically to maximize the output for a given signal. As far as the user is concerned, the lock-in amplifier now operates like a special type of a.c. microvoltmeter which can read the amplitude and phase of a signal in response to a single key stroke or bus command. This reduced level of operational complexity is reflected in the amount of programming effort required to control the lock-in system when it forms a component part of a larger bus-controlled system.

There are other aspects of operation which benefit non-specialist users that apply to almost any type of microprocessor-based instrument. For example, the digital output can be scaled to reflect the input signal level, taking all factors such as sensitivity multipliers and amplifier gains into account. The system is then much less prone to operator error than a mechanically switched system fitted with a pointer scale where the danger of overlooking a scaling factor is always present. Also, since all switching operations from the front panel are supervised by the microprocessor, the system is able to inhibit or give warning of undesirable or unorthodox combinations of front-panel controls, supported by a display or print-out of the appropriate error message.

In an instrument such as a lock-in amplifier which is subject to frequency-dependent errors in the signal and reference channel circuits, there is ample scope for using the microprocessor in an automatic calibration routine. This would measure and store calibration errors over the frequency range of the instrument with the object of providing corrected results in the final measurement. In principle, the calibration routine could be extended to correct phase-sensitive detector offsets and to compensate the amplitude and phase characteristics of signal conditioning filters introduced into the signal channel.

Clearly, the incorporation of a microprocessor has progressively greater impact as the complexity of the lock-in system is increased and should, ideally, enable a greater number of facilities to be offered without sacrificing case and clarity of operation. This objective generally requires a fresh approach to front panel design. For example, the familiar phase dial of a lock-in amplifier might be replaced by a counter that can be incremented or decremented using a pushbutton switch. The reference-channel phase-shifter is then controlled from a digital-to-analogue convertor taking its input from the microprocessor data bus. The problem of monitoring the status of the instrument is overcome by displaying the phase setting on a digital panel meter which also serves to display error codes and fault conditions when the system is operated.

As an additional constraint on the system designers, experienced users would normally require that the system is able to revert to full manual control where various combinations of front-panel settings could be tried without being restricted to operate from a ‘menu’ of stored routines. This constraint would certainly apply to any microprocessor-based lock-in amplifier that was offered as a general purpose measurement tool rather than as a special-purpose instrument, rigidly programmed to perform a specific range of tasks.

The selection of routines available on commercial instruments is limited but carefully chosen to enhance the handling characteristics of the lock-in amplifier in a wide range of applications. In addition to the software routines for sensitivity and phase selection referred to earlier, there is usually the possibility to offset data by a fixed amount and to normalize data, providing an output expressed as a fraction of percentage of some predetermined level. Routines of this type are therefore applied after detection and serve as a first stage of output processing. If more complex processing is required, this would normally be carried out by a computer interfaced to the lock-in amplifier, programmed to suit the needs of a specific experiment.

Management of this interface by a microprocessor resident in the lock-in amplifier offers several advantages over a ‘hardware only’ design. Thus, transmitted data can be presented in an easily understood format, only relevant data need be transmitted, and received commands can be less complex and more meaningful. The impact of the microprocessor on the design of an IEEE-488 compatible lock-in amplifier is brought out in Table 10.1. Comparison is made with a notional design based on classical techniques, and improvements are attributed to specific characteristics of the microprocessor.

Regarding the routines for sensitivity and phase adjustment; it is essential here that the criteria for range and phase switching are clearly stated if the lock-in amplifier is to behave predictably under the worst conditions of signal and noise. The following sections give further discussion on these routines and apply equally to a detection system under software control from a microprocessor or from an external controller linked by a data bus.

Table 10.1

Instrument characteristic	Classical design techniques	Improvement in microprocessor-based design	Microprocessor characteristic leading to improvement
Digital display of output	Output in range 0 to ± 10V scaled by reference to gain setting and ×1, ×2, ×5 multipliers	Direct scaling reflecting input level	Multiplication program
Digital display of phase	0 to 99.9° Quadrant information on +90°and +180° switches	0 to 359.9° Direct indication	Addition capability
Autorange mode	Hardware design requiring physical links with sensitivity and time-constant switches	Software-only design: no additional hardware required to implement mode	Digital comparison Data manipulation Program storage
Zero offset	Manual operation	Manual-automatic operation	Program storage Data manipulation Subtraction capability
Initial Set-up	None	Automatic operation	Program storage Data manipulation Digital comparison Mathematical capability
Normalize	Manual operation	Manual/automatic operation	Program storage Data manipulation mathematical capability

Table 10.1 (continued)

Variable phase and zero-offset hardware	3-digit d.a.c. or 10-bit binary d.a.c. plus b.c.d. to binary conversion hardware	10-bit d.a.c.	B.C.D. to binary program storage
Output conversion	$\pm 3^{1}/_{2}$ digit a.d.c.	12-bit binary a.d.c.	Binary to b.c.d. program storage
IEEE 488 Transmitted data	Fixed format. Interpretation required for phase and output data	Flexible format No interpretation required No redundant data need be transmitted	Data manipulation Storage of format styles Mathematical capability Read/write storage capability
IEEE 488 Received commands	Fixed format: hardware determined	Flexible format Meaningful commands	Data manipulation Data storage Storage of format styles
Front panel control	Rotary, pushbutton and toggle switches. Hardware determined	Pushbutton switches only. Ergonomic improvements. Group of controls not constrained by internal design of instrument	Data manipulation Program storage

10.4 Automatic sensitivity selection

When a lock-in amplifier is operated manually, the response to a synchronous signal is usually adjusted by switching sensitivity to obtain an output as close as possible to a full-scale reading.

In a single-phase lock-in amplifier, the response will not necessarily be maximum and may even take negative values unless the phase of the reference channel has been correctly adjusted. When using a two-phase system, the sensitivity is usually adjusted to maximize the output of the phase-sensitive detector giving the largest response. Alternatively, when the use of a vector computer is appropriate, the sensitivity can be adjusted by observing the ‘magnitude’ output of the vector computer.

If the residual noise output of the lock-in amplifier is sufficiently large, it will be necessary to ensure that fluctuations in the output do not carry the indication due to the signal beyond full-scale. This would normally require an observation time amounting to several time constants and may result in the system being switched to lower sensitivity. Most systems offer a 1:2:5 or 1:3:10 switching sequence**,** so this final step can usually be achieved without a significant loss of output voltage.

The object of a sensitivity routine or autorange routine is to bring this sequence of operations under automatic control by comparing the *magnitude* of the lock-in amplifier output with predetermined ‘threshold’ levels. Some of the difficulties encountered when autoranging with a noisy signal are demonstrated by the following examples.

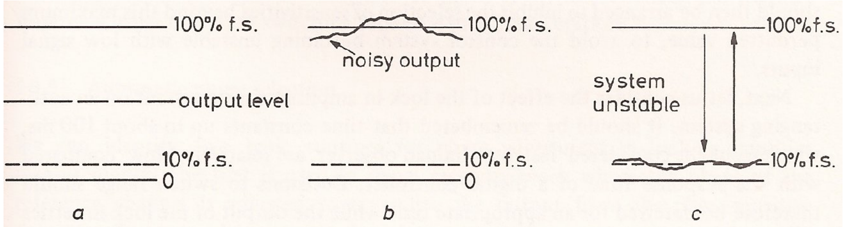


Fig. 10.3 Definition of switching thresholds in a decade autoranging system.

Suppose we have a lock-in amplifier where the sensitivity control is divided into a series of decade ranges, and that the response to a synchronous signal has been brought to a level corresponding to about half-scale output as shown in Fig. 10.3(*a*).

The lock-in amplifier controller is programmed to switch to a lower sensitivity (range-down) when the signal increases to a level corresponding to full scale output, and to switch to higher sensitivity (range-up) when the signal falls to 10% of full-scale. To be realistic we must allow for a small amount of residual noise appearing with the signal as shown in Fig. 10.3(*b*); hence the first ‘down’ transition will occur when the total output due to signal plus noise exceeds the 100% level. Fig. 10.3(*c*) shows the new situation which applies after the sensitivity has been switched. Signal and noise appear in the same relation as before, so it is only a matter of time before the total output falls below the 10% threshold level, causing the system to range-up to its original sensitivity. The situation illustrated in Fig. 10.3(*b*) is thus restored. If the signal remains at a constant value the system will attempt to switch gain alternately ‘up’ and ‘down’ with the result that a stable condition is never attained. To overcome this difficulty it is necessary to redefine the switching thresholds, for example by increasing the upper threshold to greater than 100%, or by reducing the lower threshold to less than 10%. In this way, the system can be made to tolerate residual noise on the output, at least up to a certain peak-to-peak level, and will be able to switch sensitivity to reach a well-defined condition.

Let us now look at the behaviour of the system under conditions of very low signal. If, at some point, the output signal-to-noise ratio falls drastically, or if the signal is removed, leaving only noise in the output, the sensitivity controller will attempt to switch gain to the maximum achievable value. The system will then remain in a stable condition at maximum sensitivity, *provided* the output noise peaks do not exceed the upper threshold level. If this level is exceeded, the controller will switch the sensitivity to a lower value. Unfortunately, the noise is bipolar and so repeatedly takes values close to zero voltage. The controller will thus restore the sensitivity to its maximum value at the first opportunity and subsequently make random transitions between the two most sensitive range positions.

These considerations suggest that the maximum usable sensitivity in autorange operation is where the peak output noise is just less than the full-scale output. This maximum sensitivity can be determined by experiment; the autorange program should then be arranged to inhibit the selection of sensitivities beyond this maximum permitted value, to avoid the control system becoming unstable with low signal inputs.

Next, let us consider the effect of the lock-in amplifier time constant on an auto-ranging system. It should be remembered that time constants up to about 100 ms, which would be considered ‘fast’ by a human observer, are relatively ‘slow’ compared with the response time of a digital controller. Decisions to switch range should therefore be deferred for an appropriate time while the output of the lock-in settles to a new value following a switching operation. In commercial systems supplied with an autorange facility, a settling time of four or five time constants is usually allowed between successive switching operations. In a fully integrated system, the micro-processor will be provided with the time-constant setting as a matter of course. In a bus-controlled system, the time-constant setting may have to be 'read' by the controller via the bus interface and entered into the autorange subroutine. The subroutine would usually feature a number of WAIT instructions to ensure that the program runs at an appropriate rate for the particular time-constant selection.

The final point to be taken into consideration concerns the role of 'expand' selection in determining the sensitivity in autorange operation. It is shown in Chapter 4 that, in some lock-in amplifiers, a given sensitivity can he obtained for two or more combinations of a.c. and d.c. gain in the system and that the choice of combination influences the dynamic performance of the lock-in amplifier. It follows that, if the lock-in amplifier is required to autorange in a 'high reserve' mode, the switching program should be arranged to give a gain combination that uses the maximum possible value of expand gain. Conversely, a 'high stability' switching program would be biased in favour of using high a.c. gain in order to achieve the best possible output stability for precision measurements.

In a microprocessor-based system, these factors could be taken into account automatically, depending on the mode of operation selected by the user. Other facilities that would normally be made available include a procedure for entering the maximum autorange sensitivity (in the interests of system stability as described above) and a procedure for entering the threshold switching points, usually expressed as a percentage of full-scale deflection. In some cases, the upper threshold is fixed at 110% of full-scale output while the autorange routine covers the 1:2:5 range sequence of the lock-in amplifier. A system with these characteristics would switch range until the output indication lay somewhere between 40% and 110% of full-scale with the overall switching time determined by the time-constant selected on the lock-in amplifier.

It should be acknowledged that autorange switching routines are, at best, systematic and, at worst, cumbersome. At a time constant of 1 second, a typical autorange routine would take about 1 minute to switch from minimum to maximum sensitivity in a 1:2:5 sequence. When operating with a very wide range of signal levels, using a programmable system with a choice of programs, there is a possibility to include a 'trial' routine confined to decade switching. The idea is to obtain an order-of-magnitude estimate of signal level; this estimate can then be improved using the 1:2:5 switching sequence in a final iteration.

10.5 Automatic phase selection

We can identify two basic routines for phase adjustment in lock-in amplifier measurements. The first is used in signal recovery work where the phase of the reference channel is adjusted to maximize the output from the phase-sensitive detector. The second is associated with precision phase measurement where the phase adjustment is made with the objective of nulling the output of the phase-sensitive detector.

Phase measurements might be made with either a single- or two-phase lock-in amplifier. If the latter is used, it is normally the output of the quadrature phase-sensitive detector which is to be brought to a null condition. Since phase measurements are usually made with noise-free signals, the accuracy to which the null can be set will depend on the ability of the analogue-to-digital convertor associated with the phase-sensitive detector to resolve small output changes. The resolution of the phase-shift control will be similarly limited in a digitally controlled system. In commercial systems the phase can usually be advanced in increments as small as 0.1°, which is comparable with the resolution of a conventional phase dial giving a continuous adjustment.

The phase-null routine can be initiated by subtracting increments of about 30° until the output changes sign. Smaller increments, say 5°, are then progressively added to the set phase until the output changes sign yet again. The procedure is repeated with successively smaller increments until the null is achieved to within the resolution capability of the system, or to within some specified limit.

Regarding signal recovery applications using a single-phase lock-in amplifier, an alternative approach to setting the phase is defined as follows, starting from an arbitrary initial phase condition:

(i) 'read' the in-phase value of the signal, V_A

(ii) add 90° to the set phase of the reference channel

(iii) 'read' the quadrature value of the signal, V_B.

(iv) compute $\phi = \tan^{-1} V_B/V_A$: reduce set phase by 90°

(v) add ϕ to the set phase.

This routine could be accomplished in a time equivalent to about 10 lock-in amplifier time-constants. When the lock-in amplifier has fundamental-only response, the resulting response will always be maximized and first-order independent of errors accrued in the measurement and in the computation of the signal phase. This procedure is perfectly adequate for use in general signal recovery applications and, more importantly, can be used to extend the usefulness of single-phase lock-in amplifiers in tasks which are normally reserved for two-phase systems.

Of course, if a two-phase system is available, the problem of setting phase need not arise in signal recovery work. All that is required is an autoranging control system to bring the vector magnitude to a suitable 'on-scale' value. In the case of a single-phase lock-in amplifier, an autoranging routine would normally be executed prior to setting the phase, followed by a final autorange routine to bring the maximized response within range.

APPENDIX 1

Principal applications

The following list catalogues some of the principal applications of lock-in amplifiers and phase-sensitive detectors. The compilation was made by the staff at EG & G Brookdeal and covers a wide range of disciplines in applied science and technology. It is inevitable that a number of applications appear under more than one title. To compensate for this, many readers will doubtless find a number of important applications that have been overlooked!

- Absorption spectroscopy
- A.C. bridges
- Antenna Patterns
- Astronomical spectroscopy
- Atomic absorption
- Audio amplifier frequency response
- Audiometry
- Auger spectroscopy

- Biomedical stimuli response measurements
- Bode plots

- Cochlea microphonics
- Common mode rejection measurements
- Complex impedance measurements
- Contact potential measurements
- Crosstalk in cables, amplifiers, etc.
- *C*-*V* plotting
- Cube interferometry

- De Haas Van Alphen effect
- Densitometry
- Detectivity compensation
- Displacement measurements
- Doppler measurements
- Dual-beam optical measurements

- Eddy-current flaw testing
- Edge shift in GaAs
- Electrochemistry
- Electroluminescence
- Emission spectroscopy
- E.P.R./e.s.r. spectroscopy

- Filter calibration
- Fluorescence spectroscopy
- Frequency-response measurements
- Frequency-shift measurements

- Hall effect: single frequency
- Hall effect: double frequency

- Infra-red (near and far) spectroscopy
- Interferometry

- Klystron stabilization

- Laser research
- Line ripple measurement in amplifier power supplies

- Magnetic-field measurements
- Magnetometry
- Magnetoresistance studies
- Marx gauging
- Mass spectroscopy
- Microphone calibration
- Microwave reflections, attenuation
- Microwave spectroscopy
- Moisture content measurement (*C*-*G*)
- Molecular-beam spectroscopy

- N.M.R. spectroscopy
- N.O.R. spectroscopy
- Nyquist plots

- Operational amplifier gain measurement
- Optical derivative measurements

Photometry
Plasma-physics research
Pyrometry

Radiometry
Raman spectroscopy
Ratiometric measurements
Resistance thermometry
R.F. measurements

Second sound
Seismic measurements
Semiconductor research
Source compensation
Spectrophotometry
Strain gauging

Stress-strain measurements

Temperature control
Temperature measurement
Torque measurements

Ultra-violet spectroscopy

Visible spectroscopy

Whistler signal measurements
Work function measurements

Young’s modulus

Zeeman effect

Selected topics on signals and noise

A2.1 Introduction

Fig. A2.1 shows the frequency ranges which can be assigned to a number of noise sources of practical importance. As explained in Chapter 2, we usually make a distinction between interference sources of external origin and noise which is inherent in the measurement system. The means to combat external interference sources are many and varied but inevitably involve the use of screening and attention being paid to cable runs. In many cases, susceptibility to mains-frequency pick-up can be reduced by physical re-orientation of circuits and components; factors giving rise to ground loops whereby mains-borne interference is introduced along with the signal are treated in Appendix 6.

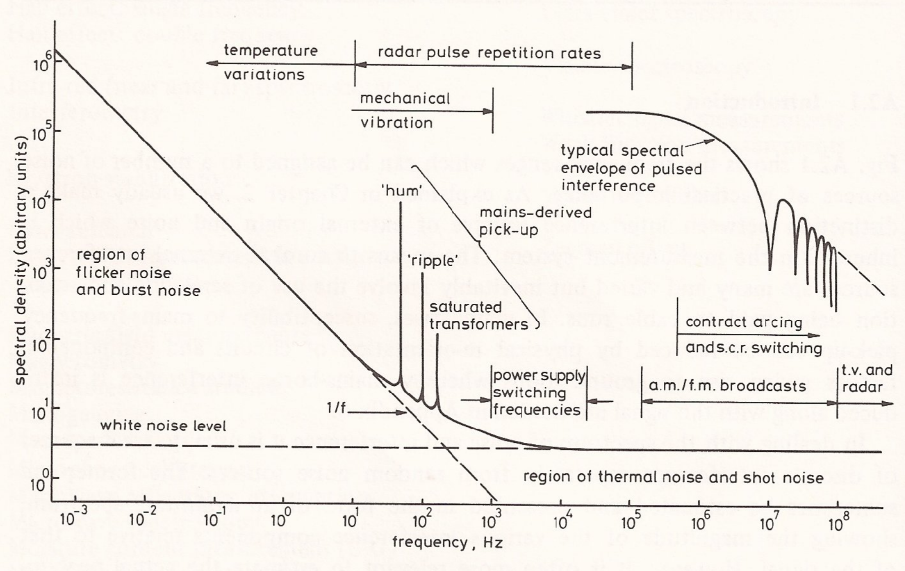


Fig. A2.1 Spectrum of noise and interference

In dealing with the spectrum of noise and interference it is usual to treat sources of discrete interference separately from random noise sources. The former can sometimes be estimated and presented in the form of an *amplitude* spectrum, showing the magnitude of the various interference components relative to that of the signal. However, it is often more relevant to estimate the actual *peak-to-peak* values of the interference components; these components often give rise to saturation in amplifiers, which is most conveniently expressed in peak-to-peak terms.

Of course, the fundamental system noise cannot be treated in terms of an amplitude spectrum. The noise manifests itself as a fluctuating voltage in the output, which is the resultant of components distributed over a wide frequency range. It is characteristic of 'well-behaved' noise sources, however, that these essentially random fluctuations deliver a consistent average power into an external load circuit. In view of this, it is appropriate to express the frequency distribution of the random noise components in terms of a *power* spectrum or, more exactly, a *power density* spectrum, $P(f)$. $P(f)$ is usually a continuous function of frequency and has dimensions of watts/Hz. By definition of a *density* spectrum the power delivered from a small frequency range Δf centered at a frequency f is simply $P(f)\Delta f$. Therefore, when $P(f)$ is specified, we can calculate the total power in any desired frequency range f_1 to f_2 from the integral

$$P_{\mathrm{TOT}} = \int_{f_1}^{f_2} P(f)\mathrm{d}f$$

This is shown graphically in Fig. A2.2 for a *white noise* spectrum with constant spectral density P_0 and for a more typical spectrum which might be encountered in practice. In both cases the total noise power in the frequency range of interest is given by the area under the spectral density plot.

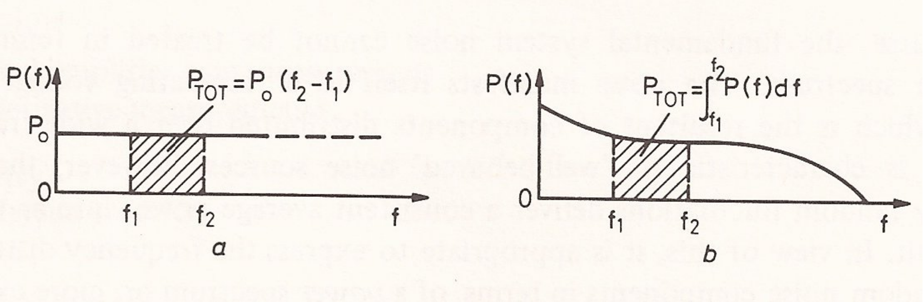


Fig. A2.2 Power density spectra for (a) white noise; (b) typical experimental noise
The total noise power in the frequency range f_1 to f_2 is given by the area of the shaded region in each case

A2.2 Voltage noise and current noise spectra

The integral of a power spectrum is given practical significance when a bandpass filter is used to reject all noise components except for those lying in a selected frequency range. The value of the integral then gives a measure of the noise power which might be measured in the filter output.

In practice, noise power measurements are usually reserved for v.h.f. and other systems operating with well-defined impedance levels. Elsewhere, it is generally more convenient to measure the filter output in terms of its *mean-square* value. Since we are concerned here with electronic systems, the output from the filter will either be a voltage fluctuation or attributable to a current fluctuation: so we express the noise intensity in a given frequency range as a total mean-square voltage or current as appropriate.

This change of emphasis leads us to define mean-square voltage and current noise spectra, $W_V(f)$ and $W_I(f)$, expressed in units V^2/Hz and A^2/Hz respectively. In a given frequency range we measure a voltage signal or a current signal in association with a total mean-square fluctuation:

$$\overline{v^2} = \int_{f_1}^{f_2} W_V(f)\mathrm{d}f$$

or

$$\overline{i^2} = \int_{f_1}^{f_2} W_I(f)\mathrm{d}f$$

A further point remains to be considered. In manufacturers' data sheets, the noise inherent in transducers and amplifiers is commonly given as an *r.m.s.* fluctuation measured in a specified bandwidth. If a noise voltage has a spectral density expressed in V^2/Hz, then the r.m.s. spectral density must have dimensions V/√Hz. Thus, doubling the measurement bandwidth for a white-noise spectrum doubles the measured intensity while the r.m.s. value increases by only √2.

A2.3 Signal spectra

It is clear that signals have a different status to noise in that a fairly precise description can often be given of their time-domain behaviour. Indeed, as has often been remarked, we must have at least an outline description for a signal before we can embark on the process of signal recovery.

In the restricted view of signal recovery which includes lock-in techniques, the signal is usually an amplitude- or phase-modulated carrier described by one of the general forms:

$$s(t) = A_0\,[1 + m(t)]\cos\omega_0 t$$

or

$$s(t) = A_0 \cos\,[\omega_0 t + m(t)]$$

In each case, the carrier frequency ω_0 is known within fairly close limits and the information or modulation signal, $m(t)$, is to be determined. Just as we have general information about the form of the signal, it is likely that the broad characteristics of $m(t)$ will also be known; the experiment itself will set limits to the maximum amplitude range of $m(t)$ and to its maximum rate of change.

In many instances, the signal is of very simple form and can be separated into sinewave components by the use of trigonometric identities, or expanded as a Fourier series. The frequency composition of the signal and its representation as an amplitude spectrum can then be inferred directly from the time domain description.

Elsewhere, a knowledge of Fourier transforms plays a role in deducing the form of the signal spectrum, but a rigorous approach is not necessarily the most beneficial. In most cases, it is usually sufficient to know: (i) the *location* of the spectrum, (ii) the *width* of the spectrum and (iii) the mean-square value of the signal. Even when the modulation $m(t)$ is only broadly specified, these three points can usually be answered. For example, in the case of amplitude-modulated carrier, we have:

$$s(t) = A_0\,[\cos\omega_0 t + m(t)\cos\omega_0 t]$$

We suppose that $m(t)$ is a slowly varying function compared with $\cos\omega_0 t$. If the Fourier transform of $m(t)$ exists, and is given by $M(\mathrm{j}\omega)$, then we can use the narrowband transformation:

$$\Im\{m(t)\cos\omega_0 t\}$$

$$= \tfrac{1}{2}M(\mathrm{j}\omega + \mathrm{j}\omega_0) + \tfrac{1}{2}M(\mathrm{j}\omega - \mathrm{j}\omega)$$

The Fourier transform of the signal is thus

$$S(\mathrm{j}\omega) = \frac{A_0}{2}\delta(\omega-\omega_0)+\frac{A_0}{2}\delta(\omega+\omega_0) + \frac{A_0}{2}M(\mathrm{j}\omega+\mathrm{j}\omega_0)+\frac{A_0}{2}M(\mathrm{j}\omega-\mathrm{j}\omega_0)$$

An example is shown in Fig. A.2.3 with the modulation strictly limited to a bandwidth B_m. A knowledge of this bandwidth is sufficient to estimate the width of the spectrum located about the carrier frequency, although the precise shape of the spectrum may not be known. In this way points (i) and (ii) raised above can be answered. Regarding the mean-square value of the signal we have:

$$\overline{s^2(t)} = \frac{A_0^2}{2}[1 + \overline{m^2(t)}]$$

If the frequency range of the signal can be estimated together with its mean-square value we can avoid the need for a formal definition of intensity spectra for periodic and other ‘deterministic’ signals. Also, we see, that provided $|m(t)| < 1$, we can estimate the signal intensity to within a factor of 2 even when $m(t)$ itself might not be precisely specified. If this seems inadequate in the light of the conventional approach to modulation systems it is no worse than the degree of approximation used in estimating the level of the background noise!

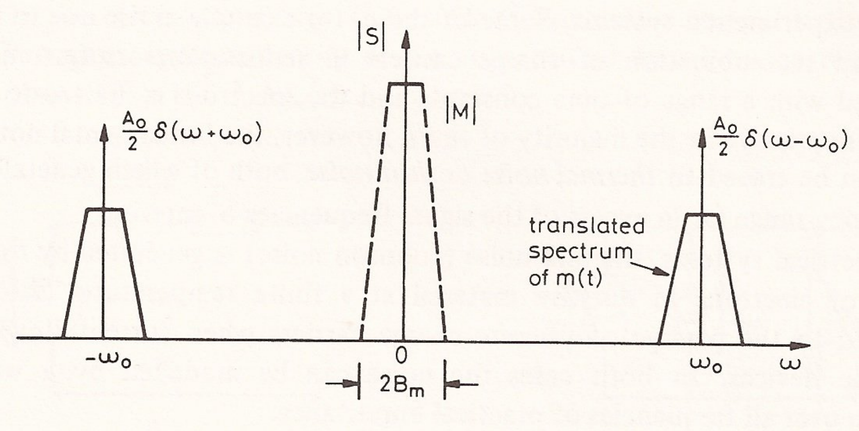

|S|
|M|
$\frac{A_0}{2}\delta(\omega+\omega_0)$
$\frac{A_0}{2}\delta(\omega-\omega_0)$
translated spectrum of m(t)
$-\omega_0$
0
ω_0
ω
$2B_m$

Fig. A2.3 Spectrum of an amplitude-modulated carrier

In the case of phase modulation: when $|m(t)| << 1$, corresponding to low-index modulation, the signal can be expressed as:

$$s(t) = A_0\,[\cos\omega_0 t\ \cos m(t) - \sin\omega_0 t\ \sin m(t)]$$

$$\cong A_0\,[\cos\omega_0 t - m(t)\sin\omega_0 t]$$

The *amplitude* spectrum and the effective bandwidth of $s(t)$ is thus the same as for amplitude modulation by the same information signal $m(t)$. Note, however, that the phase relationships are different in the two cases.

At the other extreme, when the index of modulation is large, Carson’s rule can be used to estimate the ‘spread’ of the spectrum from its centre frequency. This gives an estimate of the signal bandwidth:

$$B_\mathrm{s} = 2(|m(t)|_\mathrm{max} + 1)B_\mathrm{m}$$

where B_m is the bandwidth of the modulation or information signal.

When calculating the mean-square value of phase modulated signals, we find that the phase terms make no contribution to the final result. This is simply

$$\overline{s^2(t)} = A_0^2/2$$

In general we might expect that the signal carries *both* amplitude and phase modulation. While this is a possibility we should also note that, in the vast majority of cases, the modulations can be very slowly varying functions, often limited to a bandwidth of a few hertz. Thus, very often, the signals of interest are extremely narrowband, occupying a *relative* bandwidth of no more than a few per cent.

A2.4 Thermal noise and shot noise

There are a number of well defined mechanisms which give rise to broadband[*] noise in experimental systems. Some of these, for example, noise due to the generation and recombination of charge carriers in semiconductors (g.r. noise), are associated with a range of time constants and the spectrum is limited to an upper cut-off frequency. In the majority of cases, however, the fundamental noise mechanisms can be traced to *thermal noise* or *shot noise*, both of which generally occupy a frequency range far in excess of the signal frequencies of interest.

In electrical systems, thermal noise (Johnson noise) is generated by the random motion of electrons in resistive material at a finite temperature. Shot noise is attributed to the passage of discrete charge carriers when current flows through electronic devices. In both cases the noise can be modelled by a white-noise spectrum over all frequencies of practical importance.

First of all, regarding thermal noise: this is a fundamental source of fluctuation in all physical systems in a state of thermodynamic equilibrium. We can be sure, therefore, that it will be found in all linear, passive devices, irrespective of their form. It is often the case that such devices – and for that matter entire experimental systems – can be reduced to a simple description in terms of an equivalent *electrical analogue* circuit. The most common is either the Thévenin or the Norton form shown in Fig. A2.4.

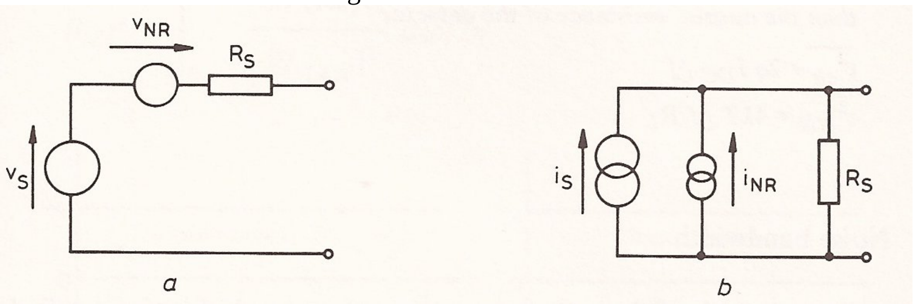


Fig. A2.4 (a) Thévenin and (b) Norton source equivalent circuits including resistance noise generators

The equivalent source resistor is defined in terms of the physical characteristics of the linear device and the thermal noise associated with the source resistance is included in the form of a voltage or current noise generator. When R_s is specified and the temperature is known we can immediately obtain the spectral density functions of these noise generators:

$W_v(f) = 4\ kTR_s$ V^2/Hz

$W_I(f) = 4\ kT/R_s$ A^2/Hz

Here, k is Boltzmann's constant (1.38×10^{-23} joules/K), R_s is the equivalent source resistance in ohms and T is the absolute temperature.

In calculations involving resistive sources at laboratory temperature it is convenient to remember that a resistor of x kilohm is associated with a random voltage generator of $4\sqrt{x}$ nanovolts/√Hz or a random current generator of $4\sqrt{x}$ picoamperes/√Hz.

[*] 'Broadband' noise is recognised by its having a spectrum that is generally free from local 'peaks' and extends to zero frequency.

Shot noise, unlike thermal noise, is always associated with current flow. The random passage of charge carriers in vacuum tubes and semiconductors gives rise to a fluctuation which depends on the average current. The spectrum is that of white noise which extends over a wide frequency range limited only by transit-time effects in the electronic device. We have

$$W_I(f) = 2qI_0 \text{ A}^2/\text{Hz}$$

where q is the electronic charge (1.6×10^{-19} coulombs) and I_0 is the average current. The r.m.s. spectral density of the noise caused by a 1 nA current flow is therefore about 1.8×10^{-14} A/√Hz.

Shot noise will be present in all semiconductor devices operating with finite bias current, and is usually the dominant source of broadband noise in optical detectors. Here, a periodic current variation due to a ‘chopped’ light beam must often be measured against a more or less steady bias current which flows in response to a much greater ‘background’ illumination due to light leakage or sample fluorescence. Many such detectors conform closely to an ideal current source, and the output can be measured by connecting the detector to an external load resistor R_L. Fig. A2.5 gives the noise equivalent circuit of this arrangement, which shows that the signal current i_s appears in competition with the shot noise of the bias current I_{DC} and the thermal noise of the load resistor. To ensure that the signal-to-noise ratio inherent in the detector is not degraded further by the thermal noise in R_L we investigate the condition:

$$2qI_{DC} \geq 4\, kT/R_L$$

which gives

$$R_L \geq 2kT/(qI_{DC})$$

The quantity $2kT/q$ is approximately equal to 50 mV at laboratory temperatures. Thus, for a bias current of 1 mA, the source will be dominated by shot noise provided that R_L is in excess of 50 Ω. In fact, the usual tendency is to choose very large values of R_L to increase the output voltage due to the signal current. In this case, the shot-noise contribution is usually the dominant one even at low bias currents. This topic is discussed further in Appendix 5 in relation to amplifier selection and the use of current amplifiers.

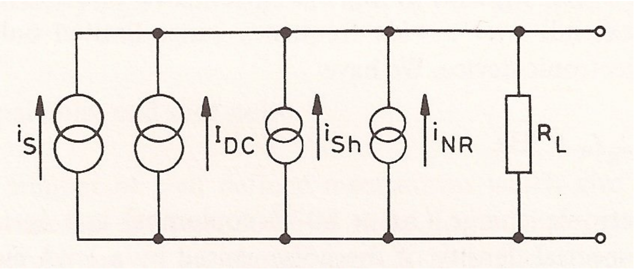


Fig. A2.5 Noise equivalent circuit for a detector terminated in a resistor R_L. R_L is much less than the output resistance of the detector

$$\overline{i_{sh}^2} = 2q\, I_{DC}\Delta f; \quad \overline{i_{NR}^2} = 4kT\Delta f\, /\, R_L$$

A2.5 Noise bandwidth

The overall bandwidth of the noise appearing with a signal of interest will always be limited to some finite value, if only because of the effect of stray reactance. More usually, however, the bandwidth is fixed at a well defined value owing to the low-pass filter effect of the transducer and output amplifier.

When the noise inherent in the experimental process is broadband in nature, the combined frequency response of the transducer/amplifier combination is often responsible for the spectral characteristics of the noise observed in the final measurement. For example, suppose the experimental noise has constant spectral density W_N over a wide frequency range, and that we can identify a frequency-response function, $H(\mathrm{j}\omega)$, with comparatively narrow bandwidth. The spectrum of the 'output' noise can then be approximated by

$$W(f) = W_N\,|H(\mathrm{j}\omega)|^2, \quad \omega = 2\pi f$$

Conversely, an observed noise spectrum can often be modelled by assuming that it originates from the passage of broadband white-noise through a filter with appropriate frequency-response characteristics. 'Broadband' spectra and 'narrowband' spectra such as those shown in Fig. A2.6 are examples where this approach is often successful.

In order to calculate the total fluctuation due to the noise we integrate its spectrum over all frequencies to obtain the mean-square value:

$$N_0 = W_N \int_0^\infty |H(\mathrm{j}\omega)|^2 \mathrm{d}\omega/2\pi$$

Since the integral depends only on the filter transfer function we can simplify all subsequent discussions by defining the *noise equivalent bandwidth* of the filter. This gives the bandwidth of the rectangular filter shown in fig. A2.6 which transmits the same fluctuation as the actual filter of interest.

The noise output of the noise-equivalent rectangular filter with bandwidth B_N is

$$N_0 = W_N\,|H_{MAX}|^2 B_N$$

hence, equating this with the integrated noise in the filter characteristic $H(j\omega)$, we obtain the filter noise-equivalent bandwidth in terms of the integral:

$$B_N = \int_0^\infty \frac{|H(\mathrm{j}\omega)|^2}{|H_{MAX}|^2}\mathrm{d}\omega/2\pi$$

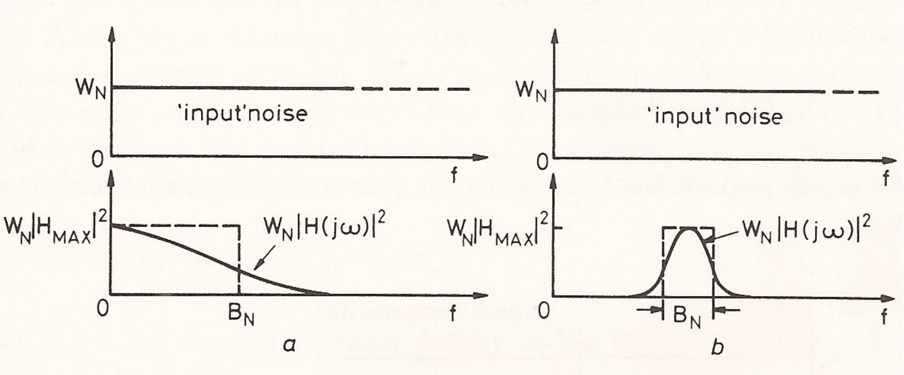


Fig. A2.6 (a) Broadband and (b) narrowband noise spectra obtained by filtering white noise. B_N denotes the noise equivalent bandwidth

The noise-equivalent bandwidth (or, simply, noise bandwidth) of a practical filter is somewhat greater than its 3 dB bandwidth, but becomes closer for filters of higher order. This reflects the sharper cut-off of high-order filters, which approximate more closely to filters with 'ideal' cut-off characteristics.

In principle, therefore, the output fluctuation of any filter in response to a white-noise input can be calculated once the maximum gain of the filter and its noise bandwidth are known. The latter is usually obtained from a catalogue of noise bandwidths such as that given in Appendix 4.

Finally, it should be noted that noise bandwidth is conventionally expressed in *hertz* and not radians/second.

A2.6 Signal-to-noise-ratio improvement by filtering

We envisage the situation shown in Fig. A2.7, where a signal, $s(t)$, appears against a background of white noise having noise bandwidth B_I and spectral density W_N.

The signal-to-noise ratio, measured on a mean-square basis, is simply

$$SNR_I = \frac{\overline{s^2(t)}}{W_N B_I}$$

A signal-conditioning filter is now used to attenuate all noise components except those lying within the frequency band occupied by the signal. The filter bandwidth is sufficiently wide to transmit the signal without distortion, but significantly smaller than the 'input' noise bandwidth B_I. In this case, the signal-to-noise ratio at the filter output is given to a good approximation by

$$SNR_0 = \frac{\overline{s^2(t)}}{W_N B_0}$$

where B_0 is the noise bandwidth of the filter. Note that the gain modulus of the filter, $|H_{MAX}|$, does not appear in the expression for signal-to-noise ratio.

Dividing the two signal-to-noise ratios we obtain the signal-to-noise ratio improvement factor

$$SNR_0/SNR_I = B_I/B_0$$

This is the classic improvement factor for the recovery of signals from white noise by filtering.

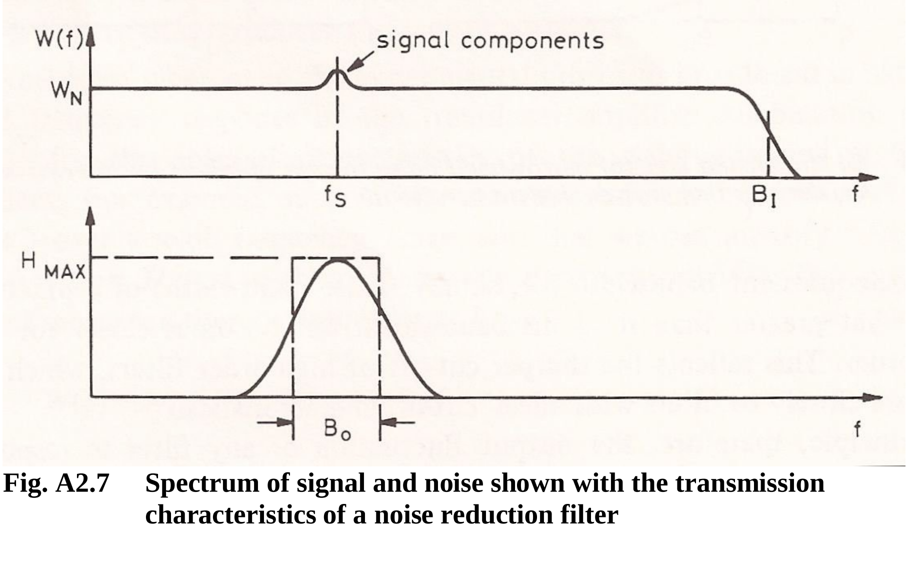


Fig. A2.7 Spectrum of signal and noise shown with the transmission characteristics of a noise reduction filter

A2.7 Low-frequency noise

‘Practical’ noise spectra almost invariably display a steady rise in spectral density as lower and lower frequencies are taken into account. This is the so-called *flicker-noise* region where the spectral density follows a law

$$W(f) = W_0/f^x$$

Here, W_0 is a constant and x takes values, typically, in the range 0.8 to 1.4. The term ‘1/*f* noise’ is also used to describe spectra of this general type.

Flicker-noise is associated with a wide range of physical processes. Although its origins are obscure, its spectral characteristics are usually well-defined for a given experimental set-up.

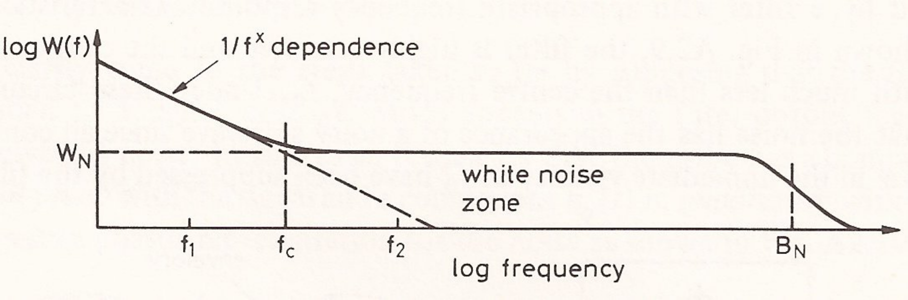


Fig. A2.8 Spectral model for broadband noise with a low-frequency noise component

Fig. A2.8 gives an example where flicker noise has added a low-frequency ‘tail’ to a broadband noise spectrum. This is typical of the noise-spectrum characteristics of a low-noise voltage amplifier where the ‘corner frequency’, f_c, marks the transition between the low-frequency and white-noise regions. We can use the corner frequency to provide the following description of the overall spectrum:

$$W(f) = W_N\,[1 + f_c/f^{\,x}], \quad f < B_N$$

where W_N is the constant spectral density in the white-noise zone and B_N is the overall noise bandwidth.

The widespread incidence of flicker-noise accounts for the equally widespread use of a.c. excitation in experimental work, the object being to bring the signal of interest into the spectral region above the corner frequency. If a clear separation is not achieved, then it may sometimes be necessary to calculate the total fluctuation from the frequency interval f_1 to f_2 shown in Fig. A2.8 which includes the corner frequency. For the purpose of calculation it is usual to assume that $x = 1$. To do otherwise implies that the spectrum of low-frequency noise has been characterised very carefully.

Our spectral model gives a total mean-square fluctuation:

$$\int_{f_1}^{f_2} W_N\,[1 + f_c/f]\,\mathrm{d}f = W_N\,(f_1 - f_2) + W_N\,f_c\;1\mathrm{n}\,(f_2/f_1)$$

We thus find that the white-noise component gives rise to a mean-square fluctuation proportional to the measurement bandwidth, $(f_2 - f_1)$, while the flicker-noise contribution depends on the frequency *ratio.* We conclude from this that the flicker-noise fluctuation measured per octave or per decade is constant over all frequencies.

A2.8 More about narrowband noise

The term ‘narrowband noise’ is generally used to describe noise that has zero spectral density in the vicinity of $f = 0$. It was remarked in Section A2.5 that narrowband noise can often be modelled by supposing that white noise has been transmitted by a filter with appropriate frequency-response characteristics. In the example shown in Fig. A2.9, the filter is highly selective and the output noise has a bandwidth much less than the centre frequency, f_0. Under these circumstances we find that the noise has the appearance of a noisy sinewave since all components except those in the immediate vicinity of f_0 have been suppressed by the filter.

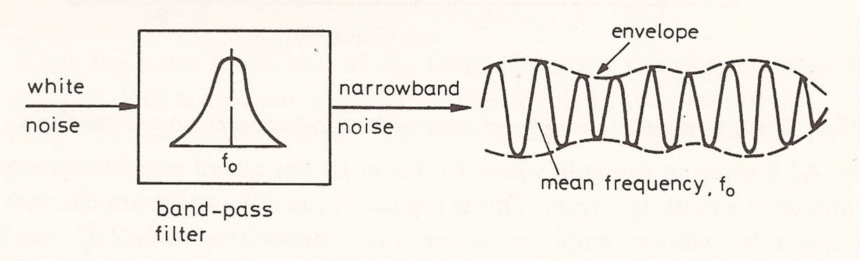


Fig. A2.9 Generation of narrowband noise

We shall find that the structure of narrowband noise lends itself to a time-domain description which proves to be very useful when considering the response of synchronous detectors to noise inputs. To provide a time-domain model we suppose that we start with a ‘clean’ sinewave at frequency f_0 and then impose random variations on its instantaneous amplitude and phase. The result is a voltage:

$$n(t) = R(t) \cos [\omega_0 t + \phi(t)]$$

$R(t)$ and $\phi(t)$ are random modulations that vary very slowly in comparison with $\cos\omega_0 t$. $\phi(t)$ is a simple phase modulation while Fig. A2.9 shows that we can interpret $R(t)$ as the *envelope*[*] of the noise. Because $R(t)$ is a relatively slow variation, we find that there is no dramatic change in the envelope over several cycles at frequency f_0.

We now expand $n(t)$ into its constituent components to obtain:

$$n(t) = R(t) \cos \phi(t) \cos\omega_0 t - R(\mathrm{t}) \sin \phi(t) \sin\omega_0 t$$

and then define:

$$n_\mathrm{i}(t) = R(t) \cos \phi(t)$$

$$n_\mathrm{q}(t) = R(t) \sin \phi(t)$$

Thus:

$$n(t) = n_\mathrm{i}(t) \cos\omega_{¥} t - n_\mathrm{q}(t) \sin\omega_0 t$$

where

$$[n_\mathrm{i}^2 (t) + n_\mathrm{q}^{2}(t)]^{1/2} = R(t)$$

and

$$n_\mathrm{q}(t)/n_\mathrm{i}(t) = \tan\phi(t)$$

[*] An ideal rectifier, incorporating an output low-pass filter, would deliver an output voltage which varied in response to the envelope function, $R(t)$.

We can clarify some of the steps taken so far by supposing that the filter is precisely tuned to a signal, cos$\omega_0 t$, which appears in the filter output together with the narrowband noise. In this case, it appears that the noise has a component $n_i(t)$ that lies *in-phase* with the signal and a component $n_q(t)$ in *quadrature* with the signal. This suggest a phasor representation for the noise as shown in Fig. A2.10.

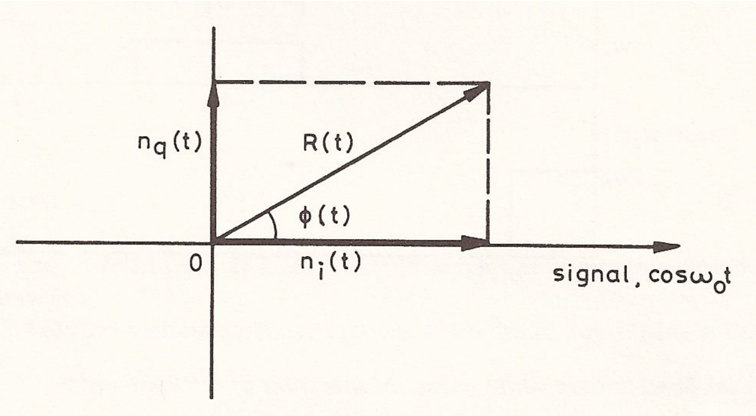


Fig. A2.10 Phasor representation of narrowband noise

Relationships between $n(t)$, $n_i(t)$ and $n_q(t)$ and their spectra are derived in most books on communication theory; for example, Taub and Schilling[1], and Haykin[2], but – for our purposes – it will be sufficient to note the following properties:

(i) *Mean value:* We are, dealing with noise at the output of a bandpass filter which implies that $n(t)$, and hence $n_i(t)$ and $n_q(t)$, have zero mean value, that is:

$$\overline{n(t)} = \overline{n_i(t)} = \overline{n_q(t)} = 0$$

(ii) *Mean-square value:* The phasor diagram shown in Fig. A2.10 is a 'snapshot' taken at a particular instant. The fact that n_i and n_q have zero mean values implies that the noise phasor spends an equal amount of time – on average – in all four quadrants. This symmetry suggests that n_i and n_q have equal mean square values and it can be shown that this is indeed the case:

$$\overline{n_i^2(t)} = \overline{n_q^2(t)}$$

In general, the noise processes n_i(t) and n_q(t) are uncorrelated. Hence, the total mean-square fluctuation of the narrowband noise, $n(t)$, is:

$$\overline{n^2(t)} = \overline{n_i^2(t)\cos^2\omega_0 t} + \overline{n_q^2(t)\sin^2\omega_0 t} = \tfrac{1}{2}\overline{n_i^2(t)} + \tfrac{1}{2}\overline{n_q^2(t)}$$

We thus obtain:

$$\overline{n^2(t)} = \overline{n_i^2(t)} = \overline{n_q^2(t)}$$

(iii) *Spectral density of* $n_i(t)$ and $n_q(t)$: If we denote the spectrum of the narrow-band noise by $W_n(f)$, then the spectra of $n_i(t)$ and $n_q(t)$ are identical, obtained through the transformation[1,2]:

$$W_i(f) = W_q(f) = W_n(f - f_0) + W_n(f + f_0)$$

where f_0 is the 'centre frequency' of the narrowband noise process.

$W_i(f)$ and $W_q(f)$ generally have the form of *low-pass* spectra. In the special case where $W_n(f)$ is symmetrical about $f = f_0$, we obtain:

$$W_i(f) = W_q(f) = 2W_n\,(f + f_0)$$

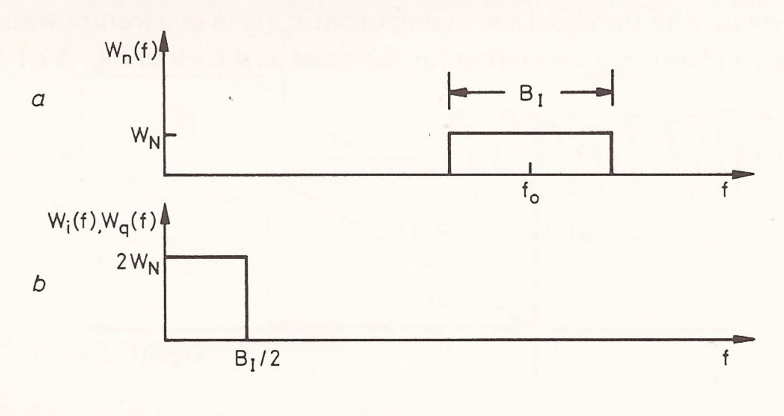


Fig. A2.11 (a) Band-limited white noise; (b) spectrum of n_I(t) and n_q(t)

For example, suppose the narrowband noise has the form of band-limited white noise as shown in Fig. A2.11. We have:

$$W_n(f) = \begin{cases} = W_N, & f_0 - B_I/2 \le f \le f_0 + B_I/2 \\ = 0, & \text{elsewhere} \end{cases}$$

In this case, $W_i(f)$ and $W_q(f)$ take the form:

$$W_i(f) = W_q(f) \begin{cases} = 2W_N, & f \le B_I/2 \\ = 0, & \text{elsewhere} \end{cases}$$

Integrating $W_i(f)$, $W_q(f)$ and $W_n(f)$ over all frequencies we confirm the results given earlier, namely:

$$\overline{n_i^2(t)} = \overline{n_q^2(t)} = \overline{n^2(t)}$$

where, in this case:

$$\overline{n^2(t)} = W_N B_I$$

A2.9 References

1 TAUB, H., and SCHILLING, D.L. (1971): Principles of communication systems’ (New York, McGraw Hill)

2 HAYKIN, S. (1978): ‘Communications systems’ (New York, John Wiley & Sons)

Synchronous detection and noise

A3.1 Signal-to-noise-ratio improvement

Let us consider the response of a synchronous detector to an amplitude-modulated signal perturbed by random noise, giving an input of the form:

$$v_{\text{in}}(t) = m(t)\cos\omega_0 t + n(t)$$

We have seen that the operation of synchronous detectors does not necessarily depend on the elimination of unwanted noise components by filtering in advance of detection. However, it was shown in Appendix 2 that the structure of narrow-band noise is particularly convenient when it comes to performing calculations in the time domain. We shall therefore assume that the input noise is band-limited as shown in Fig. A3.1 with a bandwidth B_{I} much greater than the signal bandwidth $2B_{\text{M}}$.

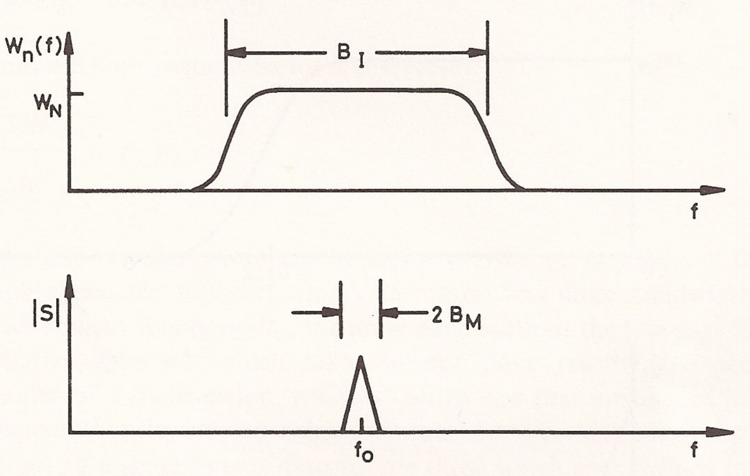


Fig. A3.1 Spectra of: (a) input noise and (b) amplitude-modulated signal

If we further assume that the noise spectrum is centered on the signal frequency, we can use the results of Appendix A2.8 and write:

$$n(t) = n_{\text{i}}(t)\cos\omega_0 t - n_{\text{q}}(t)\sin\omega_0 t$$

The input voltage to the synchronous detector now has the form:

$$v_{\text{in}}(t) = \left[m(t) + n_{\text{i}}(t)\right]\cos\omega_0 t - n_{\text{q}}(t)\sin\omega_0 t$$

Following the arguments developed in Appendix A2.8, we identify $n_{\text{i}}(t)$ and $n_{\text{q}}(t)$ as the components of the noise lying, respectively, in phase and in quadrature with the signal.

A reference voltage, synchronous with the signal of interest, is now introduced at the synchronous detector and the phase is adjusted to bring signal and reference in phase. The reference voltage is:

$$v_{\mathrm{R}}(t) = \sqrt{2} V_{\mathrm{R}} \cos \omega_0 t$$

giving multiplication products:

$$v_{\mathrm{in}}(t) v_{\mathrm{R}}(t) = \frac{V_{\mathrm{R}}}{\sqrt{2}} \left[m(t) + n_{\mathrm{i}}(t)\right]\left[1 + \cos 2\omega_0 t\right] - \frac{V_{\mathrm{R}}}{\sqrt{2}} n_{\mathrm{q}}(t) \sin 2\omega_0 t$$

At this stage we shall assume that the output low-pass filter is used only to eliminate components centered on frequency $2\omega_0$ without modifying the low-frequency output of the multiplier. We obtain an output voltage:

$$v_0(t) = \frac{V_{\mathrm{R}}}{\sqrt{2}} \left[m(t) + n_{\mathrm{i}}(t)\right]$$

giving an output signal-to-noise ratio:

$$SNR_0 = \overline{m^2(t)} / \overline{n_{\mathrm{i}}^{\,2}(t)}$$

This important result shows that the output signal-to-noise ratio is given in terms of the noise components that lie *in-phase* with reference voltage. We thus conclude that the *quadrature* noise components $n_{\mathrm{q}}(t)$ are rejected at the point of detection and so make no contribution to the low-frequency output.

The input signal-to-noise ratio is:

$$SNR_{\mathrm{I}} = \overline{\left[m^2(t) \cos^2 \omega_0 t\right]} / \overline{n^2(t)} = \tfrac{1}{2} \overline{m^2(t)} / \overline{n^2(t)}$$

From Section A2.8 we have:

$$\overline{n_{\mathrm{i}}^{\,2}(t)} = \overline{n^2(t)} = W_{\mathrm{N}} B_{\mathrm{I}}$$

Hence:

$$\frac{SNR_0}{SNR_{\mathrm{I}}} = 2$$

A signal-to-noise improvement factor of 2 is thus inherent in the operation of the synchronous detector.

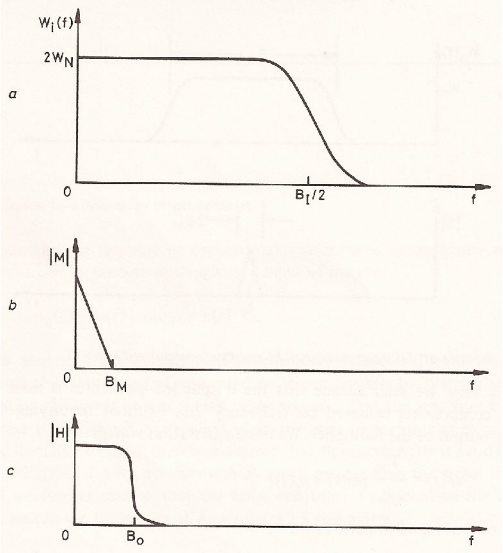


Fig. A3.2 (a), (b) Spectra of output noise and recovered modulation signal in a synchronous detector; (c) frequency response of noise-reduction filter in final output

The spectra of the recovered modulation signal and the output noise are shown in Fig. A3.2 for the case where $V_{\mathrm{R}} = \sqrt{2}$. We now suppose that the bandwidth of the low-pass filter is greatly reduced in order to eliminate noise components from the final output. The output noise bandwidth is accordingly set to a value B_0 as indicated in Fig. A3.2 where B_0 is sufficiently wide to transmit the recovered modulation signal without distortion.

When $B_0 << B_{\mathrm{I}}$ we can approximate the mean-square value of the noise following the low-pass filter by:

$$N_0 = 2W_{\mathrm{N}}B_0$$

and so obtain the output signal-to-noise ratio:

$$SNR_0 = \frac{\overline{m^2(t)}}{2W_{\mathrm{N}}B_0}$$

compared with the input signal-to-noise ratio:

$$SNR_{\mathrm{I}} = \frac{\overline{m^2(t)}}{2W_{\mathrm{N}}B_{\mathrm{I}}}$$

The signal-to-noise improvement factor is therefore:

$$\frac{SNR_0}{SNR_{\mathrm{I}}} = B_{\mathrm{I}}/B_0$$

The "classic" signal-to-noise improvement factor derived in Appendix 2 for linear filters is thus applicable to synchronous detectors. The noise bandwidth of the detector is determined simply by B_0, the noise bandwidth of the low-pass filter.

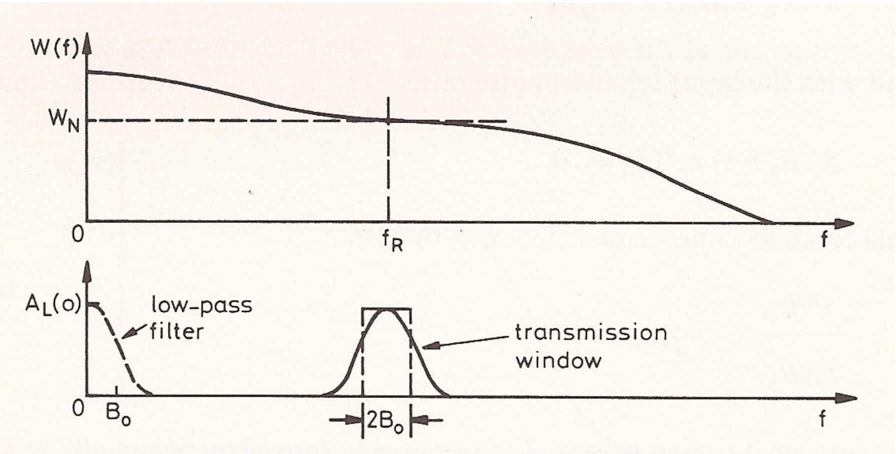


Fig. A3.3 Synchronous detector with arbitrary input noise spectrum

An alternative approach which helps to put these results into perspective involves the idea of a transmission "window" which was first introduced in Section 2.4. We have seen that the only asynchronous components which survive to perturb the final output of a synchronous detector are those which are confined to a transmission window centred on the reference frequency, having a noise bandwidth equal to *twice* the noise bandwidth of the low-pass filter. Using the spectral model shown in Fig. A3.3 for an arbitrary noise spectrum, the mean-square fluctuation associated with the components within the transmission windows is, approximately:

$$N_0 \approx 2B_0W_{\mathrm{N}}$$

However, from the results given above, it is evident that the synchronous detector responds only to the components of the noise that lie *in-phase* with the reference. Of the total mean-square fluctuation appearing within the transmission window we can ascribe one half to noise components lying in phase with the reference and one half to noise components in quadrature with the reference. If we now suppose that a synchronous detector has a full-scale sensitivity S_F and a full-scale output V_F, the noise appearing in the final output will have a mean-square value:

$$\overline{v_N^2} = \tfrac{1}{2}(V_F / S_F)^2 N_0 = (V_F / S_F)^2 B_0 W_N$$

The factor $^1/_2$ accounts for the loss of the quadrature noise components at the point of detection. Although the synchronous detector transmission window has a noise bandwidth $2B_0$, the rejection of the quadrature noise components results in an *effective* noise bandwidth of B_0.

Note that if a two-phase lock-in amplifier is used with a noisy input, the residual noise outputs from the two phase-sensitive detectors will originate respectively from the in-phase and quadrature components of the noise. The fluctuations observed at the two outputs will, in general, be uncorrelated but otherwise have similar statistical properties.

A3.2 Noise measurements

Lock-in amplifiers are often used in a noise-measurement mode where the reference frequency and low-pass filter are selected to define a narrow measurement bandwidth centered on a spectral region of interest. A "noise measurement" unit is then used to measure the r.m.s. noise output from the low-pass filter. The result is a measure of the "spot" spectral density in the immediate vicinity of the reference frequency. This is clearly an application for a fundamental-only responding lock-in system: otherwise noise "leakage" from harmonic transmission windows could seriously affect the outcome of a measurement.

The noise bandwidth of a two-section *RC* low-pass filter is:

$$B_0 = 1/(8T_0)$$

where T_0 is the selected time constant. If the noise has an r.m.s. spectral density V_N at the reference frequency, the noise appearing in the final output will have an r.m.s.value:

$$V_{\mathrm{r.m.s}} = \frac{V_N V_F}{\sqrt{2} S_F}\left(\frac{1}{4T_0}\right)^{1/2}$$

If $V_{\mathrm{r.m.s}}$ is measured with a noise measurement unit, the noise voltage spectral density referred to input is:

$$V_N = 2A_N V_{\mathrm{r.m.s}} S_F (2T_0)^{1/2} / V_F$$

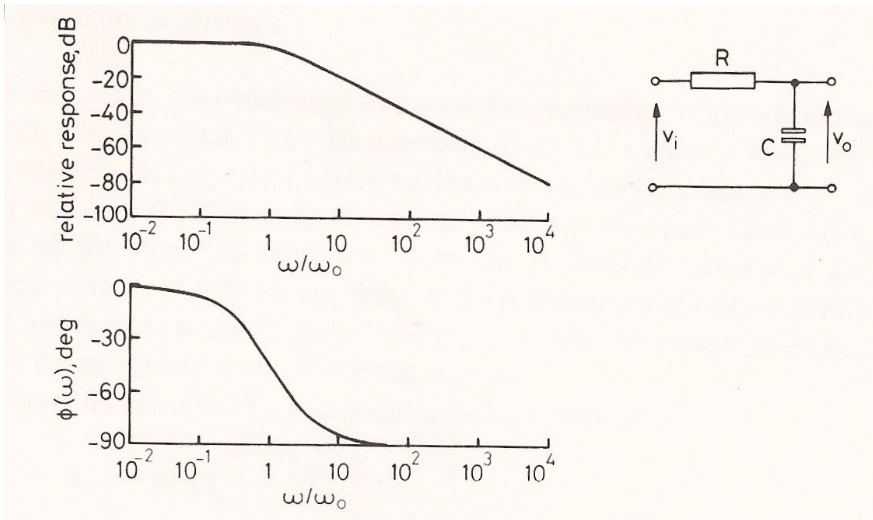
where A_N is a scaling factor specified for the noise measurement unit.

Lock-in amplifiers give a unique mode of measurement whereby the noise can be measured in the presence of a synchronous signal without errors due to intermodulation. In practice, this means that noise-measurement units are almost invariably a.c. coupled to ensure that d.c. components due to detected signal do not affect the measurement of the r.m.s. value of the noise output.

Signal conditioning filters

A4.1 Low-pass Filters

A4.1.1First-order



$$\omega_0 = 1/RC = 1/T_0$$

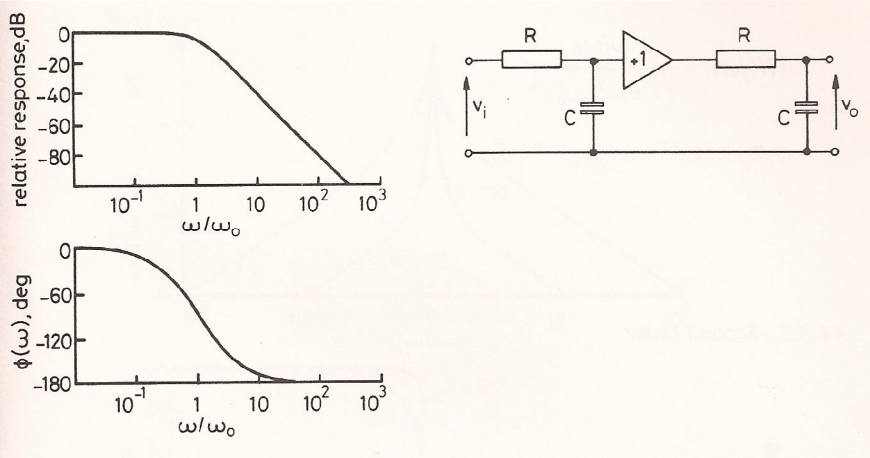
$$\left|H(j\omega)\right| = 1/\left(1+\omega^2/\omega_0^2\right)^{1/2}$$

$$\phi(\omega) = -\tan^{-1}\omega/\omega_0$$

Noise bandwidth:

$$B_{\mathrm{N}} = \omega_0/4 = 1/(4T_0), \quad (\mathrm{Hz})$$

A4.1.2Second order



$$\omega_0 = 1/RC = 1/T_0$$

$$|H(j\omega)| = 1/\left(1+\omega^2/\omega_0^2\right)$$

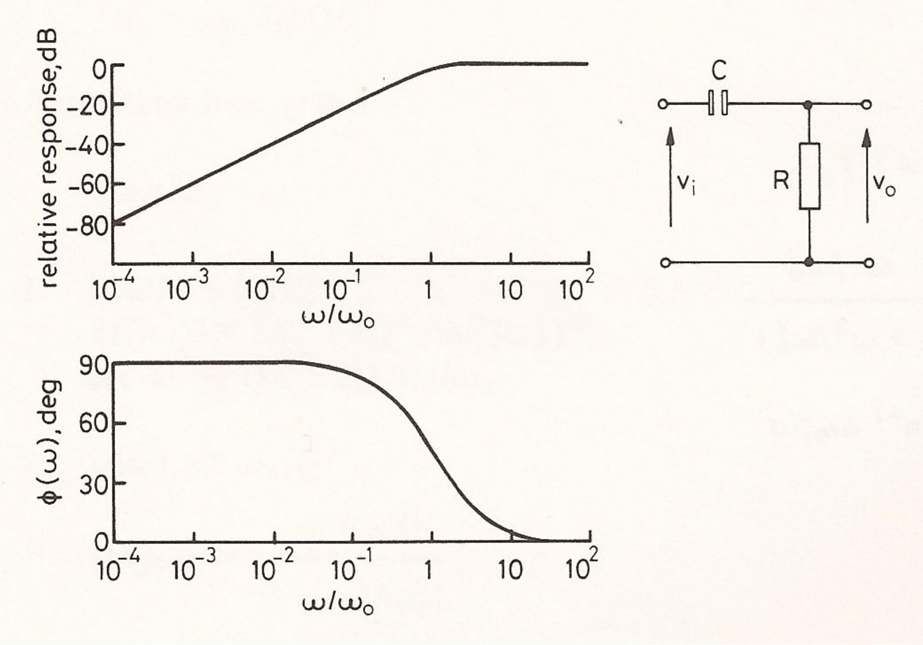
$$\phi(\omega) = -2\tan^{-1}\omega/\omega_0$$

Noise bandwidth:

$$B_\mathrm{N} = \omega_0/8 = 1/(8T_0), \quad (\mathrm{Hz})$$

A4.2 High-pass Filters

A4.2.1First order

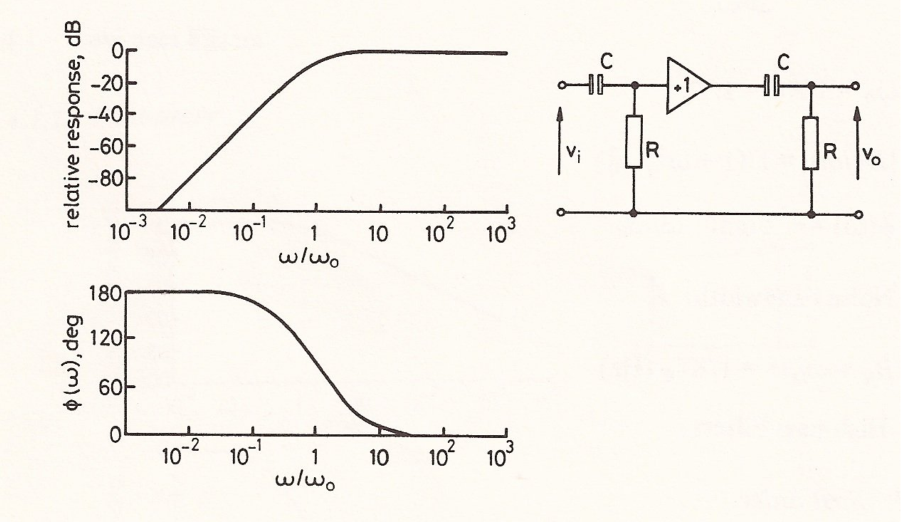


$$\omega_0 = 1/RC = 1/T_0$$

$$|H(j\omega)| = \frac{\omega/\omega_0}{\left(1+\omega^2/\omega_0^2\right)^{1/2}}$$

$$\phi(\omega) = -\tan^{-1}\omega_0/\omega$$

A4.2.2Second order



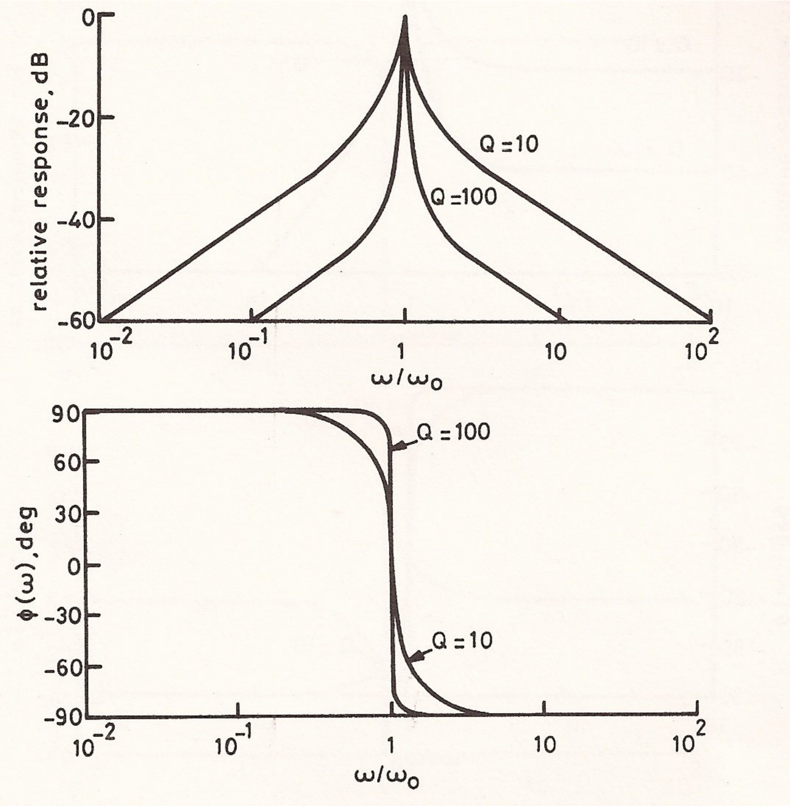
$$\omega_0 = 1/RC = 1/T_0$$

$$\left|H(j\omega)\right| = \frac{\omega/\omega_0^2}{\left(1+\omega^2/\omega_0^2\right)}$$

$$\phi(\omega) = -2\tan^{-1}\omega_0/\omega$$

A4.3 Active tuned filters

A4.3.1Band pass



$$H(\mathrm{j}\omega) = \frac{\mathrm{j}\omega\omega_0/Q}{\omega_0^2 + \mathrm{j}\omega\omega_0/Q - \omega^2}$$

–3dB bandwidth: ω_0/Q

Noise bandwidth (for $Q > \frac{1}{2}$):

$$B_{\mathrm{N}} = \omega_0/(4Q), \quad (\mathrm{Hz})$$

Approximations, $Q \geq 5, \Delta\omega = \omega_0 - \omega$:

1. $|\Delta\omega| << \omega_0/Q$:

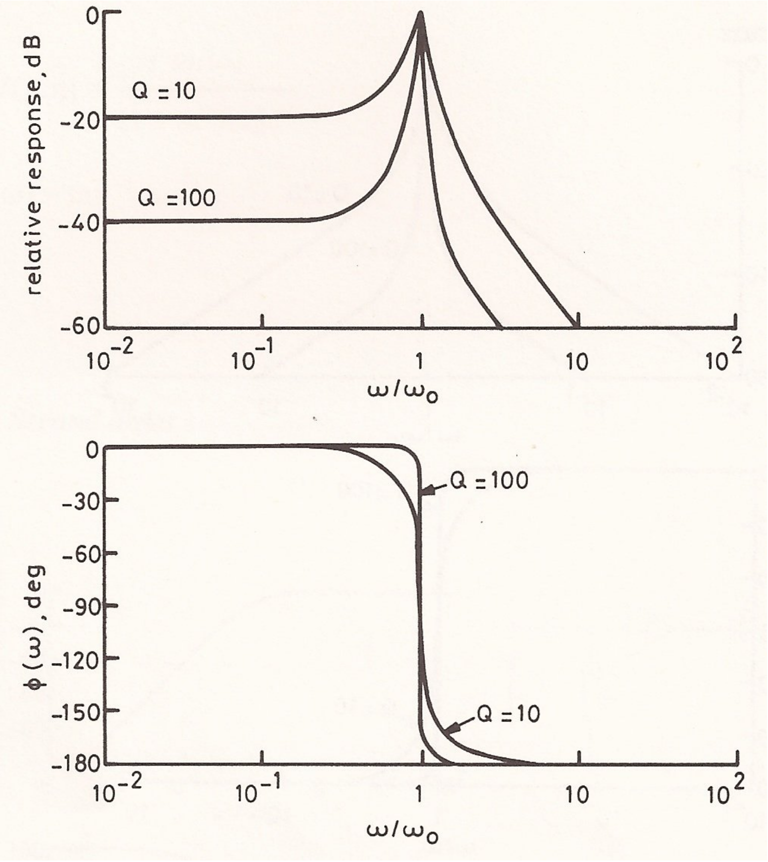
$$\left|H(\mathrm{j}\omega)\right| \approx 1/\left(1+4Q^2\Delta\omega^2/\omega_0^2\right)^{1/2}$$

$$\phi(\omega) \approx \tan^{-1} 2Q\Delta\omega/\omega_0$$

2. $|\Delta\omega| >> \omega_0/Q$:

$$\left|H(\mathrm{j}\omega)\right| \approx \frac{\omega/\omega_0}{Q\left|1-\omega^2/\omega_0^2\right|}$$

A4.3.2Low-pass



$$\left|H(\mathrm{j}\omega)\right| \approx \frac{\omega_0^2/Q}{\omega_0^2 + \mathrm{j}\omega\omega_0/Q - \omega^2}$$

–3dB bandwidth: ω_0/Q

Noise bandwidth (for $Q > \frac{1}{2}$):

$$B_\mathrm{N} = \omega_0/(4Q), \quad (\mathrm{Hz})$$

Approximations, $Q \geq 5, \Delta\omega = \omega_0 - \omega$:

1. $|\Delta\omega| << \omega_0/Q$

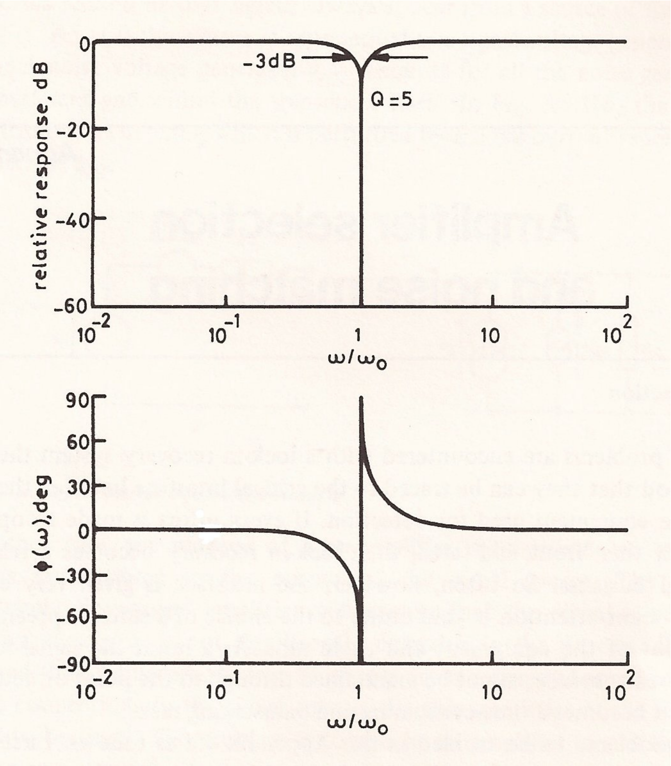
$$\left|H(\mathrm{j}\omega)\right| \approx 1/\left(1 + 4Q^2\Delta\omega^2/\omega_0^2\right)^{1/2}$$

$$\phi(\omega) \approx \tan^{-1} 2Q\Delta\omega/\omega_0 - \pi/2$$

2. $|\Delta\omega| \;>> \omega_0/Q$

$$\left|H(\mathrm{j}\omega)\right| \approx \frac{1}{Q\left|1 - \omega^2/\omega_0^2\right|}$$

A4.4 Active notch filter



$$\left|H(\mathrm{j}\omega)\right| \approx \frac{\omega_0^2/\omega^2}{\omega_0^2 + \mathrm{j}\omega\omega_0/Q - \omega^2}$$

Notch width at –3dB points:

$$\approx \omega_0/Q$$

Attenuation at $\omega = \omega_0$:

> –70 dB (typical)

APPENDIX 5

Amplifier selection and noise matching

A5.1 Introduction

If operational problems are encountered with a lock-in recovery system there is a strong likelihood that they can be traced to the critical interface between the signal source and the equipment used for detection. If every effort is made to optimize performance in the "front end" area, then lock-in recovery becomes a relatively straightforward business. So often, however, the interface is given very cursory treatment and scant attention is paid either to the choice of a suitable preamplifier or to the layout of the equipment and cable runs. As a result the signal-to-noise ratio encountered *at source* cannot be maintained through the point of detection, so measurement becomes a time-consuming and painstaking task.

The basic problems to be tacked in this Appendix are as follows. First of all, how to decide on the *type* of amplifier to be used in a given application so as to ensure that the signal is handled in the most effective and predictable way. Secondly, how to ensure that the input signal-to-noise ratio is not unduly degraded in the process of amplification, recognizing that even a "low-noise" amplifier can generate a significant amount of noise in some circumstances.

It will be assumed throughout that the principal noise limitations arise from thermal noise and shot noise. For an extension to more complicated noise models, reference should be made to a paper by Faulkner[1]. Note that design aspects of low-noise amplifiers is a topic excluded from the present treatment.

A5.2 What kind of amplifier?

We have assumed throughout our earlier discussions that the signal of interest appears in the form of an electrical signal, usually at the output of an electrical transducer. From now on we must regard this as a signal source and we shall find it convenient to represent it in the form of a circuit *model.* This is a necessary step if we are to consider the effect of making external connections.

The exact model may be more or less complex, but in many practical situations it is sufficient to use the Thevenin or Norton forms shown in Fig. A5.1 These basic forms remind us that signals always appear from a source of finite inpedance Z_s. In Fig. A5.1(a) the source is represented as a signal voltage generator v_s and the additional noise voltage generator v_{Ns} accounts for all the noise generated within the experiment and within the transducer itself. In Fig. A5.1(b) the signal appears in the form of a current i_s which is perturbed by a noise current represented here as the generator i_{Ns}.

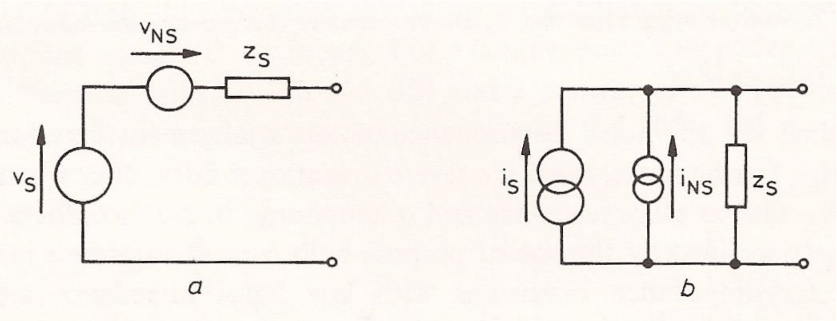


Fig. A5.1 Signal source equivalent circuits

In either case, the purpose of a preamplifier is to provide an output voltage which is proportional to the output of the signal source. The choice of an appropriately "low-noise" amplifier will ensure that the signal-to-noise ratio encountered at source is not significantly degraded in the process. It must be remembered that, while we are free to choose from a range of amplifiers or to make external connections to the signal source, the characteristics of the source must be assumed to be fixed. The problem is to choose an amplifier to suit the source, not the other way around!

In general, it is possible to define the characteristics of an amplifier to ensure that the final output depends on v_s or i_s alone, rather than on some combination of v_s or i_s with Z_s. For example, with a voltage source such as Fig. A5.1(a) it is sufficient to use an amplifier in which the input inpedance has magnitude $\left|Z_{in}\right| >> \left|Z_s\right|$. This will normally be specified as a voltage amplifier having a well-defined voltage gain over the frequency range of interest.

In the case of the current source of Fig. A5.1(b) we require an amplifier with a very *low* input inpedance compared with Z_s in order to measure the output current independently of Z_s. Such an amplifier would normally deliver an output voltage which is proportional to i_s and so is often referred to as a *transimpedance* amplifier. A common way to achieve a transimpedance amplifier is to use a voltage amplifier with a low impedance Z_L shunted across its input terminals as shown in Fig. A5.2. Those familiar with operating photomultiplier tubes will recognize this arrangement where Z_L is replaced by R_L, the anode load resistor. One reason for the popularity of this configuration is that R_L converts a hitherto "unseen" current into a more readily observable voltage variation, i_sR_L, which is subsequently amplified to give an output $A_vi_sR_L$. Dividing output by input we obtain a transimpedance A_vR_L although the overall operation is very rarely thought of in this light. Thus, increasing the load resistor is almost invariably looked upon as a means of increasing the signal voltage rather than as a means of obtaining a larger transimpedance.

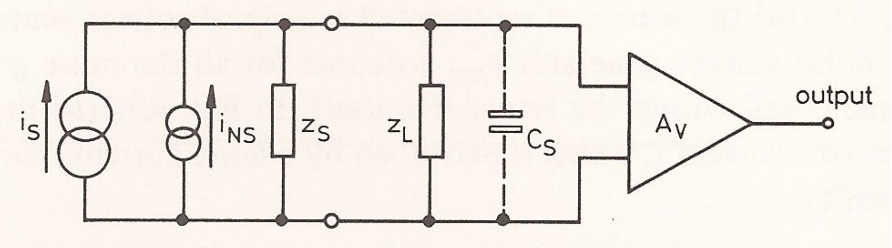


Fig. A5.2 Transimpedance amplifier. C_s represents stray reactance in the input circuit. $|Z_{in}| >> |Z_s|$

As we shall see, the noise performance of this arrangement deteriorates at low values of R_L. Furthermore, there are severe operational difficulties when using high values of R_L due to stray reactance and microphony. In practice, these difficulties can be largely avoided by the use of purpose-built current amplifiers in which high values of transimpedance consistent with low input inpedance are obtained through the use of parallel feedback.

For a given noise performance with high-impedance sources such as photomultipliers, current amplifiers can give much improved handling characteristics in terms of gain stability and relative freedom from "cable" effects. These characteristics are reviewed in Section A5.4, with special reference to photometric measurements.

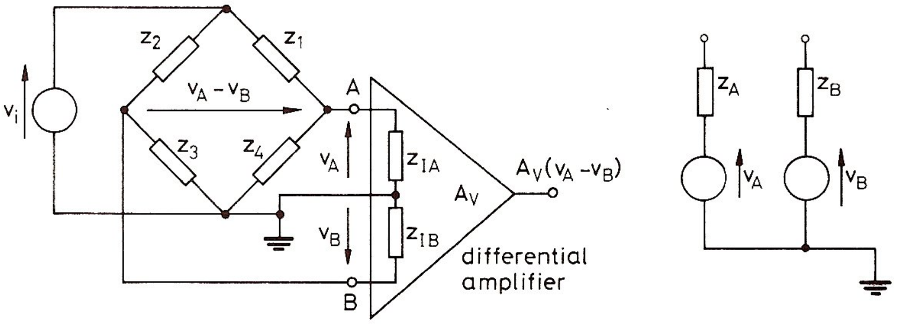


Fig. A5.3 (a) Application for a differential amplifier; (b) equivalent circuit of source. $Z_A = Z_2 // Z_3$; $Z_B = Z_1 // Z_4$

Finally, let us look at another type of voltage source exemplified by the bridge circuit shown in Fig. A5.3. Here, the signal of interest appears as the difference in potential between two points in the bridge where neither point is at ground potential. This gives us a typical application for a *differential* voltage amplifier connected as in Fig. A5.3(a). The usual arrangement is that the impedance of each amplifier input is much larger than the source inpedance presented by the bridge. This can be found by applying Thevenin's theorem to obtain the source equivalent circuit shown in Fig. A5.3(b). When the output inpedances Z_A and Z_B are identical the bridge is said to behave as a *balanced* source (not to be confused with a balanced bridge). In an unbalanced source there may be a large difference between the impedances of the two arms; however, the output inpedances are usually affected only slightly by the small adjustments which are made to the bridge at its null point.

A differential amplifier has three input terminals: A, B and ground. A voltage applied between terminals A and B is called a series or differential mode voltage. The mean voltage of A and B with respect to ground is called the common-mode voltage. A prime specification of a differential amplifier is its common-mode rejection ratio (c.m.r.r.) which gives the ratio of the series-mode gain to the common-mode gain. C.M.R.R. thus measures the ability of a differential amplifier to reject a voltage applied equally to its inputs. For example, for an amplifier with a series mode or differential gain of 100 (40 dB) and a c.m.r.r. of 10^5 (100 dB), a 1V common-mode voltage would produce an output of $1\ \mathrm{V} \times 100 \div 10^5 = 1\ \mathrm{mV}$.

C.M.R.R. is usually specified at a midband frequency, say 1 kHz, and will be in excess of 120 dB for a good-quality amplifier. Rejection falls with increasing frequency and a front-panel adjustment is often provided to maximize rejection at a frequency of interest.

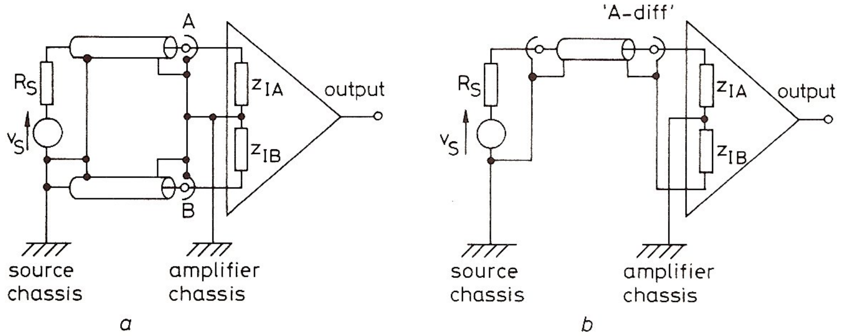


Fig. A5.4 (a) Differential with single-ended source; (b) alternative configuration using a single cable and an amplifier switched to "A-diff" mode

Differential amplifiers have an important role to play in suppressing ground loops in measurement systems. For this reason they are often specified for use with single-ended voltage sources as shown in Fig. A5.4(a). This arrangement is discussed further in Appendix 6. To facilitate connections with BNC terminations, some amplifiers incorporate a switch position labelled "A-diff" whereby the connections illustrated in Fig. A5.4(b) are made automatically. If the differential amplifier has identical input impedances on inputs A and B the inputs are said to be *balanced*. Other amplifiers offer a *pseudo-differential* or *unbalanced* input where a single input connector is wired as shown in Fig. A5.4(b) but the two input impedances are not the same. In this case, the common-mode rejection ratio is not so spectacular (usually around 80 dB), but it is possible to achieve moderate differential performance for ground-loop suppression consistent with extremely low noise. "True" differential amplifiers are generally 3 dB more noisy than their single-ended counterparts and are normally used where ultra-low-noise performance is not required. In practice, the increased immunity of a true differential stage to common-mode inputs may outweigh its additional noise contribution, and this is usually the preferred configuration for "general purpose" amplifiers.

A5.3 Noise in voltage amplifiers

A5.3.1Introduction

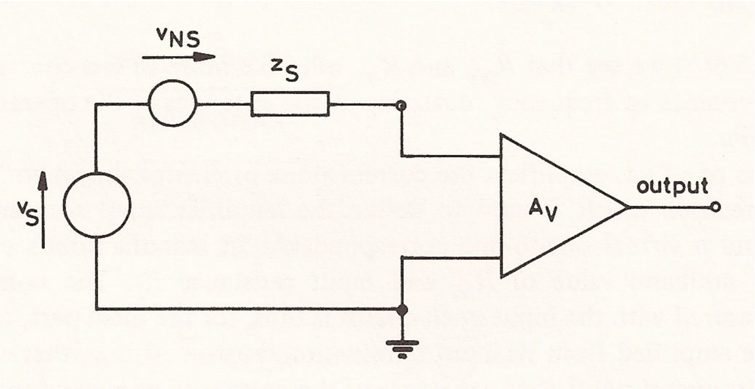


Fig. A5.5 Voltage source with amplifier

We begin with Fig. A5.5, which shows a voltage amplifier connected to a signal source. The entire source noise is accounted for by the random voltage generator

v_{Ns} which has spectral density $W(f_0)$ in the vicinity of the signal frequency f_0. The source signal-to-noise ratio measured in a small bandwidth Δf centered on f_0 is therefore

$$SNR_I = \frac{\overline{v_s^2}}{W(f_0)\Delta f}$$

We shall now assume that the signal originates in a linear passive device and that the source inpedance is resistive with value R_s over the frequency range of interest. In this important case the limitation at source is due to thermal noise so that the best possible value of source signal-to-noise ratio is

$$SNR_I = \overline{v_s^2}/(4kTR_s\Delta f)$$

where k is Boltzmann's constant and T the equilibrium temperature.

The output of the amplifier would, ideally, be an amplified version of the total input from the signal source. In practice, we must allow for noise generated within the amplifier which gives rise to an additional output fluctuation with mean-square value $\overline{v_A^2}$. The output signal-to-noise ratio is therefore less than SNR_I, and is given by:

$$SNR_0 = \frac{A_v^2\overline{v_s^2}}{\overline{v_A^2} + A_v^2 4kTR_s\Delta f}$$

where A_v is the gain of the amplifier at the signal frequency.

We now define the noise figure of the amplifying system

$$F = \frac{\text{best possible } SNR_0}{\text{actual } SNR_0}$$

This ratio will always be greater than unity for any real combination of voltage source and amplifier. In the present example, the noise figure takes the form:

$$F = 1 + \frac{\overline{v_A^2}}{A_v^2 4kTR_s\Delta f}$$

Let us now turn to the noise model shown in Fig. A5.6(a) which will enable us to predict the behaviour of the amplifier under a wide range of operating conditions.

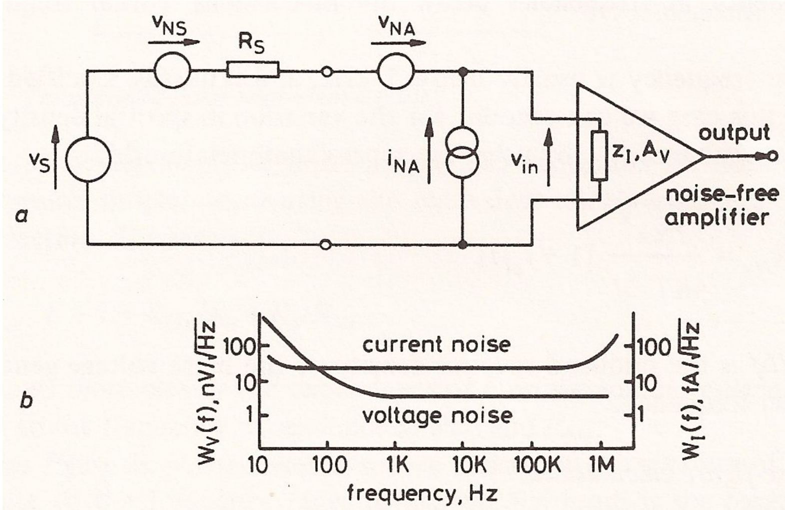


Fig. A5.6 (a) Noise model for a voltage amplifier (b) Typical spectral densities of the noise generators v_{NA} and i_{NA}

Here, the total noise of the amplifier is attributed to a pair of random-noise generators connected at its input. As is usual in such models, the amplifier itself, including the input impedance Z_I, is assumed to be noise-free. In this representation, it is clear that the signal is in competition with the amplifier noise generators v_{NA} and i_{NA}. in addition to the noise associated with the source. A fully documented voltage amplifier will have v_{NA} and i_{NA} specified in terms of their r.m.s. spectral densities over the entire frequency range of the amplifier. Fig A.5.6(b) gives an example which is typical of modern amplifiers using a j.f.e.t. input stage.

An alternative presentation which proves to be extremely useful in practicc is derived as follows. We note that the mean-square fluctuations $\overline{v_{NA}^2}$ and $\overline{i_{NA}^2}$ appearing in a bandwidth Δf centred on any frequency of interest can always be associated with *equivalent noise resistances* R_{Nv} and R_{Ni} defined by

$$R_{Nv} = \frac{\overline{v_{NA}^2}}{4kT\Delta f}$$

$$R_{Ni} = 4kT\Delta f / \overline{i_{NA}^2}$$

From Fig. A5.6(b) we see that R_{Nv} and R_{Ni} will take more or less constant values over several decades of frequency, deviating at the extremes of the operating range of the amplifier.

In the case of j.f.e.t. amplifiers the current noise originates almost entirely with the passive resistor which is used to define the amplifier input resistance. As a result, we find a virtual one-to-one correspondence in manufacturer's catalogues between the midband value of R_{Ni} and input resistance R_I. The noise of the amplifier measured with the input open-circuit is thus, for the most part, due to the thermal noise amplified from its input termination resistor. At the other end of the scale, the short-circuit noise gives a measure of the noise-voltage generator V_{NA} but here the results depend very much on the selection of input transistors and on the circuit configuration. A common feature is that the voltage noise exhibits a flicker-noise dependence at frequencies below the flicker-noise corner frequency (see Appendix 2).

The corner frequency is usually below 1kHz, and is usually specified by manufacturers. In this case we can account for the variation in spectral density and find the appropriate value of R_{Nv} by using the approximate relationship:

$$R_{Nv} = \frac{\overline{v_{NA}^2}}{4kT\Delta f}\left[1 + f_c / f\right]$$

Where $\overline{v_{NA}^2} / \Delta f$ is the midband spectral density of the noise voltage generator and f_c is the corner frequency.

A5.3.2 Noise-figure calculations

We shall use our amplifier noise model to calculate the output signal-to-noise ratio when the source is limited by the thermal noise of the source resistor.

To do this it is helpful to transform the input circuit of Fig. A5.6(a) to obtain the modified noise equivalent circuit of Fig.A5.7. This shows clearly how the contribution of i_{NA} depends on the source resistance R_S. We shall assume that v_{NA} and i_{NA} are independent noise sources (that is they exhibit no correlation) and that they give rise to mean-square fluctuations $\overline{v_{NA}^2}$ and $\overline{i_{NA}^2}$ in a frequency interval Δf centred on the signal frequency. The signal then appears in association with a total fluctuation

$$v_{\mathrm{T}}^{2} = 4kTR_{\mathrm{s}}\Delta f + \overline{v_{\mathrm{NA}}^{2}} + R_{\mathrm{s}}^{2}\overline{i_{\mathrm{NA}}^{2}}$$

and the signal-to-noise ratio measured at the amplifier output becomes

$$SNR_{0} = \overline{v_{\mathrm{s}}^{2}}/[4kTR_{\mathrm{s}}\Delta f + \overline{v_{\mathrm{NA}}^{2}} + R_{\mathrm{s}}^{2}\overline{i_{\mathrm{NA}}^{2}}]$$

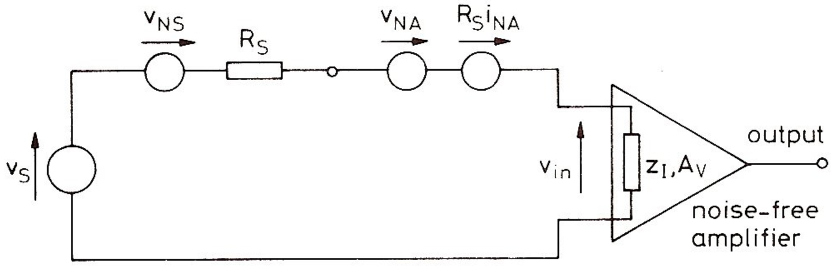


Fig. A5.7 Transformed noise-equivalent circuit

An ideal amplifier would give an output signal-to-noise ratio equal to the value measured at source, $\overline{v_{\mathrm{s}}^{2}}/4kTR_{\mathrm{s}}\Delta f$. The actual value is therefore worse by a factor

$$F = \frac{\text{best possible } SNR_{0}}{\text{actual } SNR_{0}}$$

$$= 1 + \frac{\overline{v_{\mathrm{NA}}^{2}} + R_{\mathrm{s}}^{2}\overline{i_{\mathrm{NA}}^{2}}}{4kTR_{\mathrm{s}}\Delta f}$$

With our particular amplifier noise model, the expression for noise figure can be put into much simpler form using the equivalent noise resistances defined in the previous section. The result is

$$F = 1 + R_{\mathrm{Nv}}/R_{\mathrm{s}} + R_{\mathrm{s}}/R_{\mathrm{Ni}}$$

which shows more clearly the dependence of F on the source resistance R_{s}. Also, F is subject to the frequency dependence of R_{Nv} and R_{Ni}.

If noise figure is plotted versus source resistance, using values of R_{Nv} and R_{Ni} appropriate to the frequency range of interest, the result is the parabolic curve of Fig. A5.8 which exhibits a minimum value for a value of source resistance given by $R_{\mathrm{s}} = \sqrt{R_{\mathrm{Nv}}R_{\mathrm{Ni}}}$.

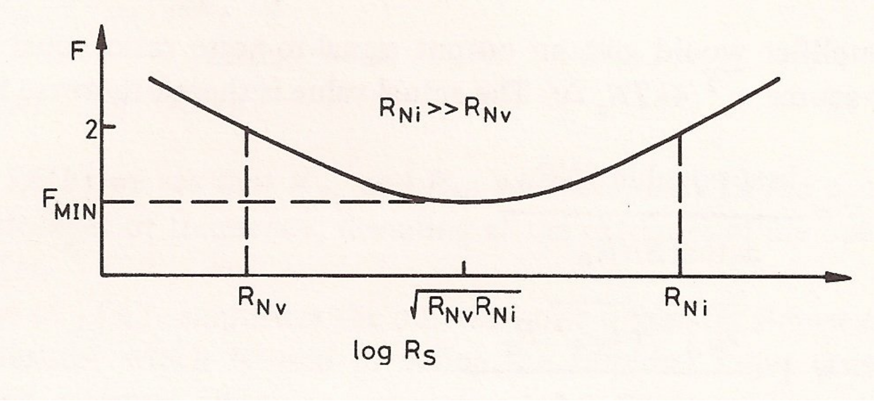


Fig. A5.8 Dependence of noise figure on source resistance.

The rise in noise figure for values of R_s less than R_{Nv} is a reminder that we cannot expect an amplifier to generate less noise than the thermal noise of an arbitrarily small source resistance. The noise figure similarly increases as R_s exceeds R_{Ni}: from the remarks made in the previous section this would normally correspond to operating a voltage amplifier from a source resistance which is greater than the amplifier input resistance.

It is evident from Fig. A5.8 that the noise-figure graph has a very broad minimum when $R_{Ni} >> R_{Nv}$. Under this condition we obtain a noise figure of 3 dB (F = 2) when $R_s = R_{Nv}$ or when $R_s = R_{Ni,}$ and a value of $F \cong 1$ for $R_s = \sqrt{R_{Nv} R_{Ni}}$.

Although it is instructive to observe how the noise figure of a given amplifier varies with source resistance, in practice we are usually faced with a fixed value of source resistance. We must then choose an amplifier from a wide range of competing devices which gives an acceptable noise figure. Any amplifier which achieves this end with a noise figure of less than 3 dB can be said to be "low noise" within the context of a given experiment, and an amplifier which is capable of giving this performance over a wide range of source resistance is said to have a *high figure of merit*, *M*, defined by

$$M = \sqrt{R_{Ni} / R_{Nv}}$$

Modern j.f.e.t. amplifiers have figures of merit in the range 100 to 1000 and they can be roughly classified according to the lowest value of source resistance which can be handled with an acceptable noise figure. Thus, an amplifier catalogued as "low noise" would normally be useful for source resistances as low as 1 kΩ (voltage noise of $4\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$ at midband) while "ultra low-noise" units extend the useful range to 40 Ω or 50 Ω (voltage noise of $800\ \mathrm{pV}/\sqrt{\mathrm{Hz}}$ at midband).

A5.3.3 Minimum noise figure and optimum source resistance

The value of source resistance which minimizes the noise figure is known as the *optimum source resistance*:

$$R_{opt} = \sqrt{R_{Nv} R_{Ni}}$$

The minimum noise figure obtained with $R_s = R_{opt}$ is:

$$F_{MIN} = 1 + 2\sqrt{R_{Nv} / R_{Ni}} = 1 + 2/M$$

For practical purposes we can say that a noise figure of 1 dB (F = 1.26) is indistinguishable from the best possible figure of 0 dB. Since figures of merit of 100 and greater are obtainable, the minimum noise figure is easily achieved when R_s falls in the correct range. More to the point, however, we see that noise figures indistinguishable from the ideal can be obtained from a wide range of source resistances - even when R_s differs from R_{opt} by an order of magnitude - as shown in the first example below.

Example 1

A low level signal is to be measured from a source of 10 kΩ resistance at a frequency of 5 kHz. An amplifier is available with the following specification:

r.m.s. noise voltage density ($f \geq 1$ kHz):

$$4\ \mathrm{nV}/\sqrt{\mathrm{Hz}}, \quad R_{Nv} = 1\ \mathrm{k\Omega}$$

r.m.s. noise current density

$$14\ \mathrm{fA}/\sqrt{\mathrm{Hz}}, \quad R_{Ni} = 100\ \mathrm{M\Omega}$$

What is the noise figure that can be achieved with this combination?

The noise figure is given by:

$$F = 1 + R_{\mathrm{Nv}}/R_{\mathrm{s}} + R_{\mathrm{s}}/R_{\mathrm{Ni}}$$

$$= 1 + 0.1 + 10^{-5}$$

$$= 1.1 \text{ (0.4 dB)}$$

Let us now calculate the minimum noise figure which can be obtained using this amplifier in the same frequency range but with the optimum value of source resistance. This is

$$F_{\mathrm{MIN}} = 1 + 2\sqrt{R_{\mathrm{Nv}}/R_{\mathrm{Ni}}}$$

$$= 1.006 \text{ (0.27 dB)}$$

which is obtained at a value of source resistance given by

$$R_{\mathrm{opt}} = \sqrt{R_{\mathrm{Nv}} R_{\mathrm{Ni}}}$$

$$= 316 \text{ k}\Omega$$

If the signal is sufficiently strong that a degradation of 3 dB in signal-to-noise ratio can be tolerated, this same amplifier will be suitable for sources with resistances in the range 1 kΩ to 100 MΩ.

Example 2

A signal of 100 nV r.m.s. is to be measured from a source of resistance 100Ω in a bandwidth of 1 kHz using the same amplifier as in example 1.

In this case the system will have a noise figure

$$F = 1 + 10 + 10^{-6}$$

$$= 11 \text{ (10.4 dB)}$$

The r.m.s. noise voltage associated with a source resistance of x kΩ at laboratory temperature is (Appendix 2):

$$4\sqrt{x} \text{ nV}/\sqrt{\text{Hz}}$$

Hence the input signal-to-noise ratio is

$$SNR_{\mathrm{I}} = \frac{(100 \times 10^{-9})^2}{16.10^{-18} \times 0.1 \times \Delta f}, \quad \Delta f = 1 \text{ kHz}$$

$$= 6.25 \text{ (8 dB)}$$

which will be reduced to 6.25/11 when the signal is amplified.

There is a clear case for seeking an amplifier with better noise performance

Suppose now, however, that the signal appears at a level of 1 µV r.m.s. The input signal-to-noise ratio is now increased to

$$SNR_{\mathrm{I}} = 625 \text{ (28 dB)}$$

In this case an amplifier noise figure of 10 dB or so would reduce the output signal-to-noise ratio to about 18 dB, which might be considered quite adequate if the signal is to be measured in a recovery system. A decision to select an "ultra low-noise" amplifier with $R_{\mathrm{Nv}} = 100\ \Omega$ or less may then be uneconomical provided that 1 µV r.m.s. represents the minimum value of the signal for all time. If, at a later stage, the signal amplitude is likely to be reduced beyond this value, then the question of amplifier noise will undoubtedly be raised again.

Finally it should be noted that these examples are for operation at a fixed frequency, using the values of R_{Nv} and R_{Ni} appropriate to that frequency. If operation over a wider frequency range is envisaged, then more information is required. Fortunately, this is usually available, as discussed in the following section.

A5.3.4Noise-figure contours

Since noise figure depends upon frequency in a fairly complicated way, most manufacturers elect to present their data graphically in the form of *noise-figure contours.*

Figure A5.9 shows a single contour, the 3 dB contour, drawn against axes labelled with source resistance and frequency on logarithmic scales. The shape of the contour is derived as follows: first of all, the regions (i) and (ii). These define the lower and upper limits of R_s required to give a noise figure of 3 dB at midband. They thus coincide with the midband values of R_{Nv} and R_{Ni}. The rise in the lower contour in region (iii) results from the rise in R_{Nv} at low frequencies due to flicker-noise effects in the amplifier, and shows that the value of R_s required to maintain a 3 dB noise figure becomes progressively larger as the operating frequency is reduced.

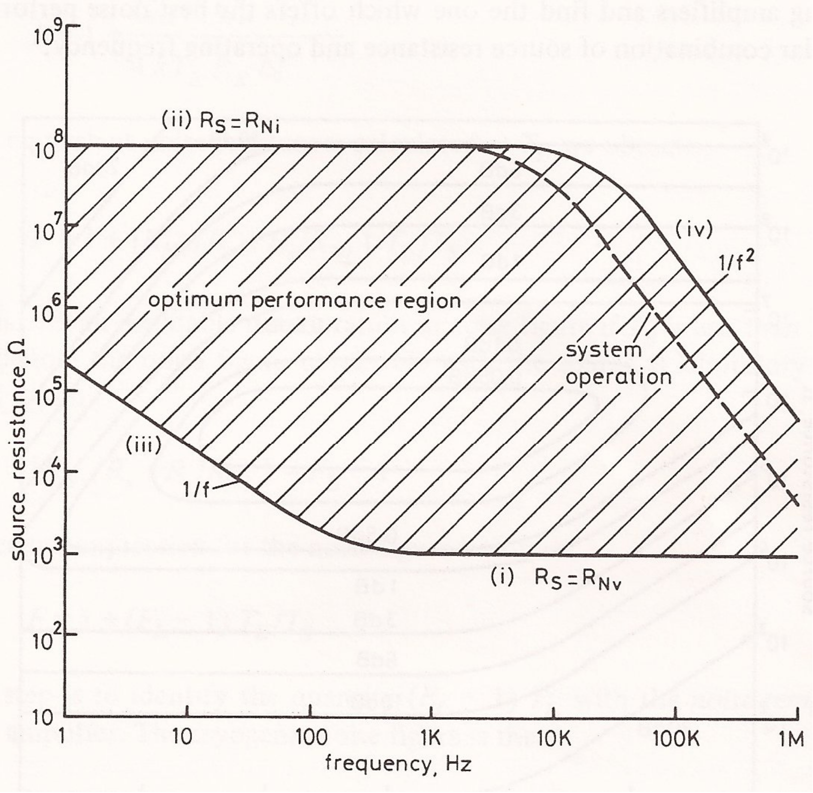


Fig. A5.9 The 3 dB noise-figure contour for a low-noise voltage amplifier

The sloping characteristic in region (iv) indicates that R_{Ni} is, in fact, *reduced* at high frequencies and that R_s must be reduced in proportion if the noise figure is to be held at the midband value of 3 dB. This is one area where the contours are particularly useful since the high-frequency cut-off depends on the "noise capacitance" of the amplifier which may not otherwise be specified. We can be sure, however, that if the amplifier input is heavily loaded with cable capacitance, the turn-over in the upper contour will shift to lower frequencies as indicated by the broken line appropriate to "system" operation.

The area enclosed by the 3 dB contour is often called the "optimum performance" region. Its vertical extent is maximized by selecting an amplifier with a high figure of merit, but this should be consistent with a low value of R_{Nv} if optimum performance is required from low source resistances. In many voltage-amplifier applications where the source resistance is less than 1 MΩ, the loss of performance in region (iv) is not normally significant. However, the turn-over frequency can be as low as 1 kHz for amplifiers with extremely high input resistance, of the order of 1 GΩ.

A complete set of contours for a general-purpose low-noise voltage amplifier is shown in Fig. A5.10 exactly in the form which might be encountered in a manufacturer's catalogue. The contours provide the means for a researcher to compare competing amplifiers and find the one which offers the best noise performance for a particular combination of source resistance and operating frequency.

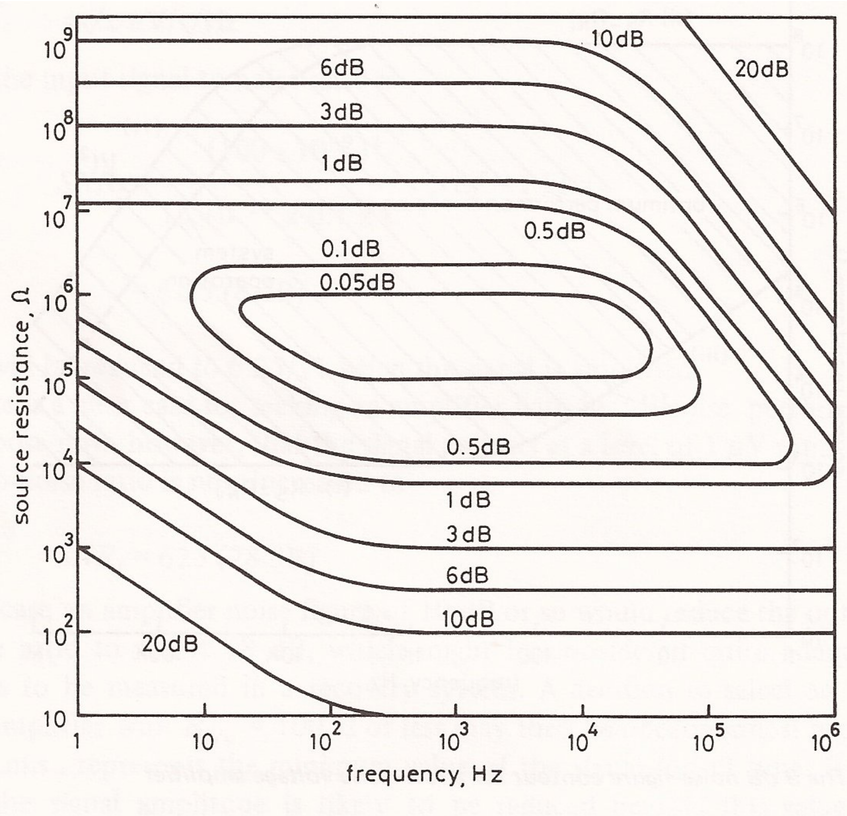


Fig. A5.10 Typical noise-figure contours

Alternatively, if an amplifier is available and there is sufficient latitude in, say, the choice of experimental frequency, it may be possible to arrange to operate within the optimum performance region of the amplifier. For example, if an optical detector provides a signal from a source resistance of 10 kΩ and the amplifier contours are those given by Fig. A5.10, we find that changing the optical chopping frequency from 10 Hz to 100 Hz brings an improvement in noise figure from 3 dB to less than 1 dB even though both frequencies lie below the flicker-noise corner frequency of the amplifier.

A5.3.5Cryogenic sources

The noise figure data supplied by manufacturers almost invariably refer to the source and amplifier at normal laboratory temperatures. Let us now return to our original definition of noise figure and rework the results for the more general case where the source is at a different temperature to the amplifier. We shall assume that v_{NA} and i_{NA} are specified at laboratory temperature T_{L}, say 290 K, and represent the source temperature by T_{s}. The noise figure is now:

$$F = 1 + \frac{\overline{v_{\mathrm{NA}}^2} + \overline{i_{\mathrm{NA}}^2} R_{\mathrm{s}}{}^2}{4kT_{\mathrm{s}}R_{\mathrm{s}}\Delta f}$$

Using the equivalent noise resistances calculated at T_{L} we obtain:

$$F = 1 + \left[R_{\mathrm{Nv}} / R_{\mathrm{s}} + R_{\mathrm{s}} / R_{\mathrm{Ni}}\right] T_{\mathrm{L}} / T_{\mathrm{S}}$$

which indicates an inevitable deterioration in noise figure if T_{s} is less than T_{L}.

If we denote the noise figure obtainable with the source at laboratory temperature by F_{L}, then

$$\left[R_{\mathrm{Nv}} / R_{\mathrm{s}} + R_{\mathrm{s}} / R_{\mathrm{Ni}}\right] = F_{\mathrm{L}} - 1$$

and the general expression for the noise figure becomes:

$$F = 1 + \left(F_{\mathrm{L}} - 1\right) T_{\mathrm{L}} / T_{\mathrm{S}}$$

The next step is to identify the quantity $\left(F_{\mathrm{L}} - 1\right) T_{\mathrm{L}}$ with the *noise temperature*, T_{e}, of the amplifier. The cryogenic noise figure is then

$$F = 1 + T_{\mathrm{e}} / T_{\mathrm{s}}$$

Thus, we find that to give a noise figure of 2 (3 dB) with a cryogenic source, the noise temperature of the amplifier must equal the temperature of the source.

Unfortunately, low-frequency voltage amplifiers are rarely specified directly in terms of noise temperature. The following examples show the sort of calculation which must usually be undertaken.

Example 1

An amplifier has a noise figure of 2 dB when operated with an optical detector at room temperature (F_{L} = 1.58). What is the achievable noise figure when operating with a detector of the same resistance at a temperature of 77 K?

We have

$$F_{\mathrm{L}} = 1.58$$

so the noise figure obtainable at a source temperature of 77 K is:

$$F = 1 + (1.58 - 1) \times 290 / 77$$
$$= 3.18 \text{ (5 dB)}$$

Example 2

For a given value of source resistance, what noise figure must be achieved with a source at laboratory temperature to ensure a cryogenic noise figure of 3 dB for sources at (a) 77 K and (b) 4 K?

The amplifier should be capable of operating at a noise temperature of 77 K in the first case and 4 K in the second. The maximum acceptable noise figure with the source at laboratory temperature is then

$$F_{\mathrm{L}} = 1 + T_{\mathrm{e}} / 290$$

This gives (a) F_{L} = 1.26 (1 dB) and (b) F_{L} = 1.014 (0.06 dB).

In this situation it is clear that the "optimum performance" region - which is bounded by the 3dB noise-figure contour for operation with sources at laboratory temperature - is now considerably reduced in area and effectively replaced by a smaller region bounded by the 1 dB or even the 0.05 dB contour. This places restrictions on the choice of operating frequency and defines a much tighter bound on the value of source resistance required to maintain an acceptable noise figure. We can conclude that extremely low noise figures have rather more than academic interest when cryogenic sources are involved. If the source resistance differs widely from the optimum value appropriate to a given amplifier, it will be necessary to introduce a stage of noise "matching" using a signal transformer.

A5.3.6Transformer noise matching

The voltage noise resistance R_{Nv} of even an "ultra low-noise" amplifier may, in many cases, be too high to give an acceptable noise figure from sources either of low resistance or at low temperature. In such cases it may be necessary to resort to noise "matching" whereby the source resistance is transformed to a new value which is much closer to the optimum sources resistance of a given amplifier. It should be noted that *noise* matching is achieved in the interest of maximizing the signal-to-noise ratio at the output of an amplifier, and is quite distinct from any attempt made to maximize either the signal voltage or the signal power through *impedance* matching.

As a first step we can disregard any attempt to "transform" the source resistance by the addition of resistors between the source and the amplifier. Series resistors merely add to R_{Nv} when performing noise calculations, and parallel resistors cause a reduction in R_{Ni} and will always degrade the signal-to-noise ratio. Far from "reducing signal and noise equally" an input attenuator will always introduce noise at the expense of the signal.

The usual approach is to introduce a transformer of turns ratio n_{T} as shown in Fig. A5.11. We shall assume for the moment that the transformer is ideal with no loss, wide bandwidth and infinite self-inductance. The transformer reflects a voltage $n_{\mathrm{T}}v_{\mathrm{s}}$ into its secondary circuit and a resistance $n_{\mathrm{T}}^2 R_{\mathrm{s}}$. The signal-to-noise ratio at the transformer output is thus unchanged and remains at its "best possible" value while the amplifier "sees" a source of resistance $n_{\mathrm{T}}^2 R_{\mathrm{s}}$. By suitable choice of n_{T}, therefore, we can arrange for the noise matching condition:

$$n_{\mathrm{T}}^2 R_{\mathrm{s}} = R_{\mathrm{opt}}$$

and so ensure that the overall system operates at its minimum noise figure.

Practical transformers can bring about a significant improvement in system performance, but, nevertheless, fall short of the ideal in almost every respect. Among the factors which must be taken into account are a reduced response when the transformer is operated outside its recommended frequency range and the effect of noise generated within the transformer itself. The latter includes the effects of vibration and the susceptibility of the transformer to pick-up, particularly at line-related frequencies.

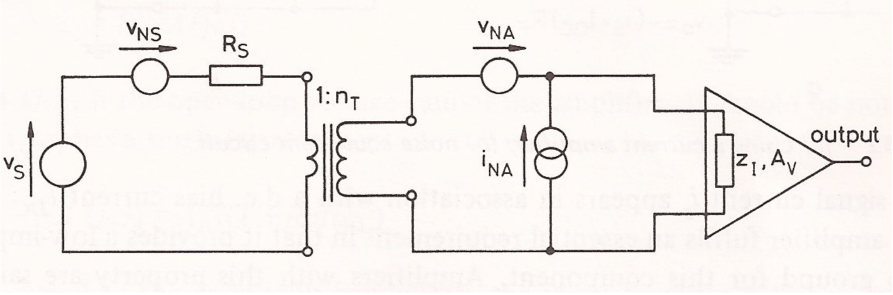


Fig. A5.11 Noise matching using a transformer

The useful frequency range of a transformer depends jointly on the source resistance and, in those with multiple tappings, on the selected turns ratio. The usual behaviour is a restriction on bandwidth when either R_s or n_T is increased; the information is most usefully presented in graphical form.

In commercial transformers the pick-up problem is reduced by packaging the transformer in a heavily screened box, while the effects of vibration and microphony are suppressed by the use of shock-absorbing mounting materials. This leaves the resistance of the windings as the main source of internally generated noise since the thermal noise of the primary coil plus the noise reflected from the secondary effectively add to the applied signal.

The noise resistance of a transformer is given in terms of the primary and secondary coil resistances R_1 and R_2 by:

$$R_T = R_1 + R_2 / n_T^2$$

It is thus possible to define a noise figure for a transformer and - as in the case of amplifiers - to present noise figure as a function of source resistance and operating frequency. Most useful of all, however, are the noise-figure contours plotted directly for a given combination of transformer and amplifier that are made available by some manufacturers.

A5.4 Noise in current amplifiers

Our interest in this section is with current amplifiers obtained by applying parallel feedback to a high-gain, low-noise voltage amplifier. A typical arrangement is shown in Fig. A5.12(a) in which the current amplifier* is connected to a high-impedance transducer modelled as an ideal current source.

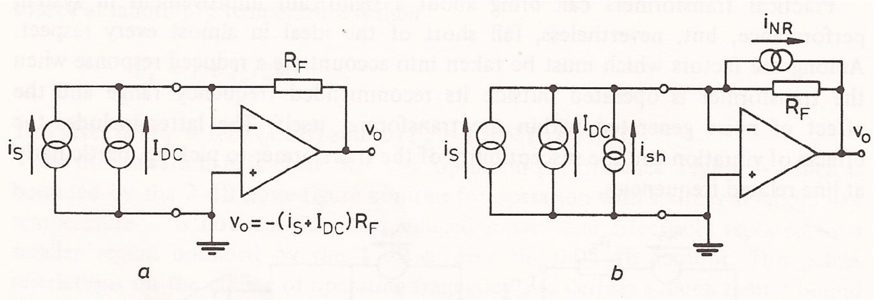


Fig. A5.12 (a) Using a current amplifier; (b) noise equivalent circuit

The current signal i_s appears in association with a d.c. bias current I_{DC} and the current amplifier fulfils an essential requirement in that it provides a low-impedance path to ground for this component. Amplifiers with this property are said to be able to "sink" a d.c. current (which may be many orders of magnitude greater than the signal current).

The dominant sources of noise are due to the shot noise of the current source and thermal noise in the feedback resistor R_F, which are included in the noise equivalent circuit of Fig. A5.12(b). If the amplifier is not to degrade the signal-to-noise ratio encountered at source, then the shot-noise contribution must exceed the thermal-noise contribution. Following the arguments developed in Section A2.4, we obtain the condition:

* Although strictly a transimpedance amplifier, the arrangement in Fig. A5.12(a) is usually catalogued as a current amplifier when the transimpedance is substantially real and constant over the operating frequency range. The transimpedance is given in this case directly by the feedback resistor, R_F.

$$I_{\mathrm{DC}} \geq \frac{2kT}{qR_{\mathrm{F}}}, \quad (2kT/q \approx 50\ \mathrm{mV\ at}\ T = 290\ \mathrm{K})$$

This defines a lower limit on I_{DC} to maintain an acceptable noise performance. At large values of I_{DC} we find a further limitation; this is given by the maximum value of bias current which can be sunk without driving the amplifier into saturation. We obtain:

$$I_{\mathrm{DC}} \leq (V_{\mathrm{OUT}})_{\mathrm{MAX}} / R_{\mathrm{F}}$$

and so arrive at the allowed range of I_{DC}:

$$50\ \mathrm{mV} / R_{\mathrm{F}} \leq I_{\mathrm{DC}} \leq (V_{\mathrm{OUT}})_{\mathrm{MAX}} / R_{\mathrm{F}}$$

Even if $(V_{\mathrm{OUT}})_{\mathrm{MAX}}$ were little more than 50 mV, there would always be a value of R_{F} which gave the required current sinking and which contributed noise not greater than the input current shot noise. In practice $(V_{\mathrm{OUT}})_{\mathrm{MAX}}$ is usually of the order ± 10 V. Thus, for any given value of R_{F}, the smallest and largest currents which can be accepted by the amplifier without, on the one hand, suffering significant noise degradation, and, on the other hand, exceeding the amplifier's current sinking capability, may be in the ratio 50 mV:10 V or 1:200.

One reason which is often cited for choosing "true" current amplifiers is their relative immunity to "cable" effects. This refers to the effects of pick-up, capacitive loading and microphony which can result when a high-impedance source is terminated by a large load resistor and connections are made via long cable lengths. When a current amplifier is used, the additional impedance introduced by such a cable is shunted by the relatively low input impedance of the amplifier given by:

$$Z_{\mathrm{I}} = R_{\mathrm{F}} / A(\mathrm{j}\omega)$$

where $A(\mathrm{j}\omega)$ is the open-loop voltage gain of the amplifier. It should be noted that when $A(\mathrm{j}\omega)$ has a single-lag response:

$$A(\mathrm{j}\omega) = A_0 / (1 + j\omega / \omega_1)$$

with ω_1 equal to the open-loop bandwidth, the input impedance has an inductive component and takes the general form:

$$Z_{\mathrm{I}} = R + \mathrm{j}\omega L$$

where

$$R = R_{\mathrm{F}} / A_0$$

and

$$L = R_{\mathrm{F}} / \omega_1 A_0$$

Although current amplifiers are particularly immune to the effect of stray capacitance on the input cable, this capacitance can have an adverse affect on noise performance at sufficiently high frequencies. To see this, we must take into account the input voltage noise generator of the amplifier v_{NA} which gives rise to an input noise current:

$$i_{\mathrm{Nv}} = v_{\mathrm{NA}} 2\pi f C_{\mathrm{s}}$$

C_{s} represents the total capacitance of the source and the amplifier input. Its effect is illustrated graphically in Fig. A5.13 for the case where v_{NA} has a midband value of $4\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$. The graph shows the value of C_{s} that causes an increase of

3 dB in the amplifier noise contribution as a function of feedback resistor R_F and frequency.

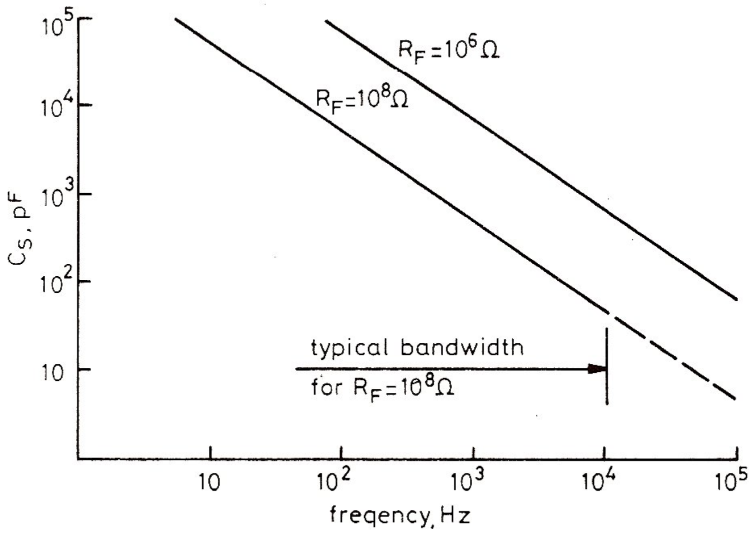


Fig. A5.13 Maximum allowed value of C_s for optimum noise performance of a current amplifier

A5.5 References and further reading

1 FAULKNER, E.A. (1975): "The principles of impedance optimisation and noise matching", *J. Phys. E: Sci. Instrum.*, 8, pp. 533-540.

2 Technical Note 101 (1077): "The use of current preamplifiers", (EG&G Brookdeal, Bracknell, England).

3 Technical |Note 243 (1976): "Noise in amplifiers", (EG&G Princeton Applied Research Corp., Princeton, NJ).

4 "Noise figure contours" (1969), (EG&G Princeton Applied Research Corp., Princeton, NJ).

Interference and ground loop suppression

A6.1 Introduction

Although consideration has been given to sources of fundamental noise in experiments and to the noise contribution of amplifiers, it is noise injected by pick-up from external sources that is usually the most troublesome in practice. In this Appendix we shall therefore be dealing with some of the problems that are met when small signals are handled in a typical laboratory environment.

A useful starting point is to list the principal mechanisms by which interference couples to experiments:

(i) capacitive pick-up;

(ii) inductive pick-up;

(iii) electromagnetic interference (e.m.i.);

(iv) high-frequency interference superimposed on mains supplies;

(v) ground loops.

Problems associated with ground loops will be left until a later stage. Otherwise, means of overcoming the first four sources of interference are generally well known and have been thoroughly documented[1,2]. These can be summarized as follows:

(a) Use screened cables to reduce capacitive "hum" pick-up between signal and power lines to suppress crosstalk between adjacent signal cables. Reinforce this approach by ensuring that low-level signal cables are routed separately from mains cords and digital highways. Reduce point-to-point capacitive pick-up within an experiment by the use of metal enclosures or fine mesh screens.

(b) Arrange for a large separation of signal lines from sources of power-frequency magnetic fields such as transformers and electric motors. Reduce the susceptibility of circuits to stray magnetic fields by eliminating large circuit loops.

Transmit signals via screened twisted pairs where spurious voltages induced in successive small loops tend to cancel. These would normally be essential first steps before resorting to expensive solutions involving high-permeability screening. Note that lock-in amplifier construction, toroidal transformers having low external fields are almost always used in order to reduce "hum" pick-up within the instrument case.

(c) When laying out experiments it should be remembered that loops of wire act as antennas at radio frequencies and that the nature and quality of a signal ground is considerably obscured when the length of the ground path becomes comparable with a wavelength at the interference frequency[3].

The effect of electromagnetic interference in a long cable run can be suppressed by techniques that raise the r.f. impedance of the cable. For example, it is worth investigating the effects of coiling the cable to form an r.f. choke. Other transmission-line techniques involve the use of transmission-line transformers and *coaxializers*. The most simple form consists of a few turns of screened cable wound on a ferrite toroid as illustrated in Fig. A6.1. This arrangement has the practical advantage of maintaining d.c. continuity throughout the length of the cable and is sometimes known as a *longitudinal choke*[1]. The tight coupling introduced by the winding results in a considerable attentuation of r.f. common-mode voltage but presents a low impedance to signal current. This technique can also be applied successfully when a differential signal is transmitted over a screened twisted pair[4].

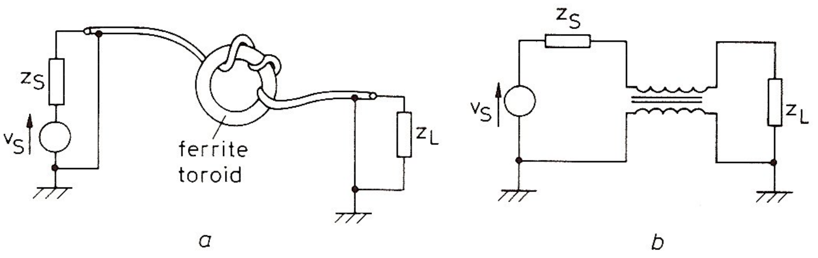


Fig. A6.1 (a) A longitudinal choke; (b) equivalent circuit

In general, the only satisfactory approach to suppressing electromagnetic interference is to systematically trace the interference source and to identify the path by which it couples to the experiment. In severe environments it may be necessary to erect a mesh screen forming a Faraday cage either around the experiment or around the offending source when this is most convenient. The screen should be earthed at one point only, using the most direct route possible.

(d) In many cases, the source of high-frequency interference can be traced to the coupling of large transients to the powerlines from pulsed lasers, thyristor controllers, laboratory ovens and other ancillary equipment used in experimental work. The solution here may be to use plug-in r.f. filters on the mains inputs of sensitive instruments to prevent transient interference appearing on instrument power supplies.

Let us now turn to the last item in our original list: ground loops. The fact is that even when detailed attention has been paid to screening and laying out cable runs, experiments can still be plagued with interference. The reason is that screened connections have finite inpedance, and so are able to support spurious voltage drops. These can give rise to severe measurement difficulties unless careful attention is paid to experiment design. Lock-in amplifier-based experiments are particularly prone to earthing problems. The result can be a large component of synchronous voltage appearing at the signal input that could be much larger than the "true" synchronous signal of interest. This aspect is discussed further in Section A6.4.

A6.2 Ground loops: single-ended amplifiers

Fig. A6.2 shows the connection of a transducer voltage source to a single-ended amplifier via a screened cable. The cable screen is securely referenced to the amplifier earthing point and is thus effective in shielding the sensitive inner conductor against capacitive pick-up from external sources.

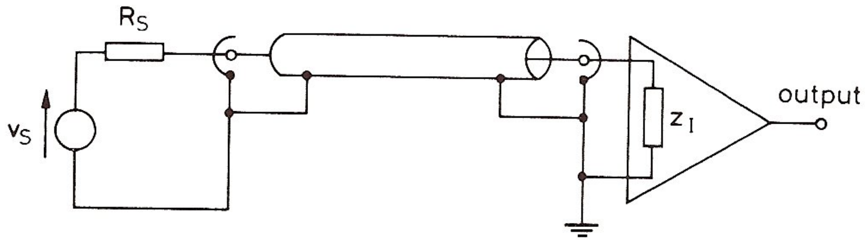


Fig. A6.2 Connection of a "floating" source and grounded amplifier

Unfortunately, it is not always possible to maintain the screen at the same potential along its entire length. When this cannot be achieved, the effectiveness of the screen is reduced because variations in screen potential become capacitively coupled to the inner conductor. Also, in some circumstances the voltage developed across the cable screen is able to add directly to the signal voltage.

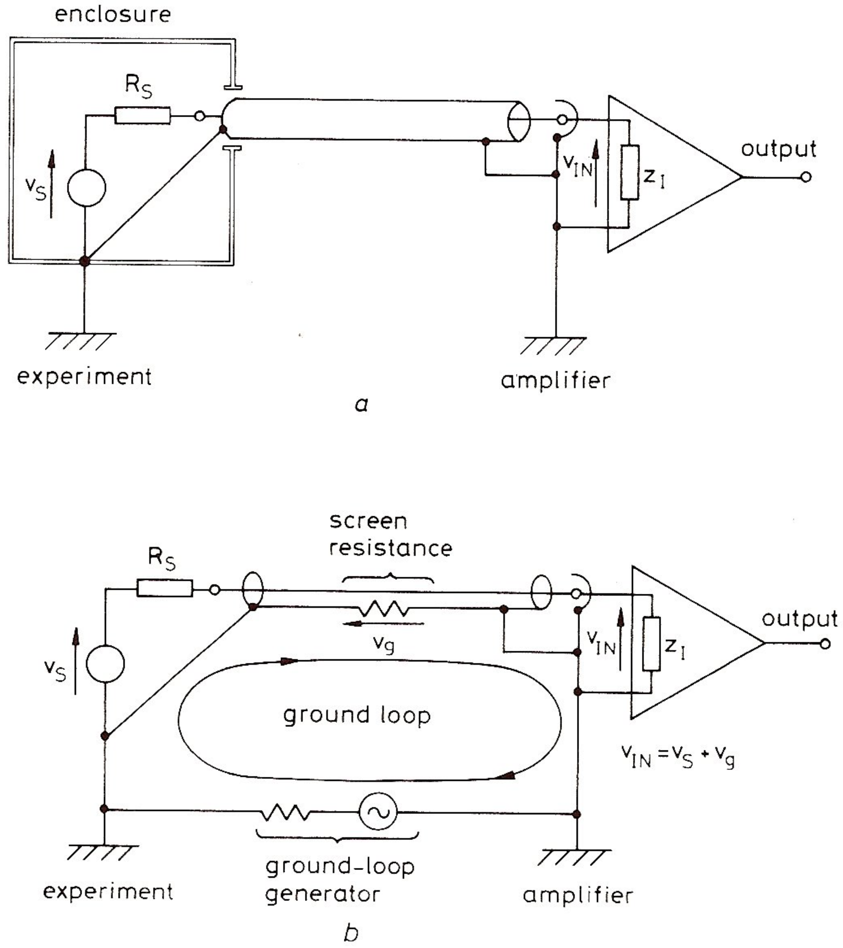


Fig. A6.3 Ground loop established when items of "grounded" equipment are interconnected

Consider, for example, the connections shown in Fig. A6.3(a). Here, the signal source is located inside a screening enclosure. Following "good practice" for optimum screening, the signal-source common and the cable screen have been connected to a single earth point to ensure that no signal currents or earth currents flow through the enclosure. The main problem in this, or any similar, arrangement, is to define a true "wet earth" for screen connections. The use of chassis symbols with separate labels for "experiment" and "amplifier" reminds us that in an extensive laboratory installation it is not unusual to find a.c. potential differences of several hundred millivolts between adjacent chasses. Any attempt to connect these via a cable screen of finite impedance then results in a *ground loop* as indicated in Fig. A6.3(b). this loop is sensitive to any potential difference between the two chasses and is additionally susceptible to inductive coupling to stray magnetic fields. Both these effects are accounted for by the inclusion of a ground loop generator in Fig. A6.3(b) which develops a voltage drop v_g across the cable screen. Because the amplifier senses the potential difference between the screen and the inner conductor of the connecting cable, the so-called *common mode voltage*, v_g is effectively added to the signal.

This spurious input may dominate the measurement of the signal unless appropriate steps are taken. In principle, the common-mode signal can be eliminated by bringing the source and amplifier chasses to the same potential, but attempts to achieve this are rarely successful in practice. Even when units are brought into close proximity and bolted to a metal plate it is not unusual to find large potential differences between "earth" points only a few inches apart. A far better approach is therefore to investigate ways of "floating" either the source or the amplifier with a view to breaking the ground loop completely.

In the case of this source this might be achieved by using insulating bolts and washers to prevent direct contact between the transducer case and the experimental chassis. Where this is not feasible, a battery-powered preamplifier provides a reliable (and safe) way of isolating the amplifier input. An alternative and usually more convenient approach is to use a *semi-floating* amplifier in which a "float" resistor, typically in the range 50 Ω to 1 Ω, is used to provide a degree of isolation which approaches actual breaking of the ground loop.
Fig. A6.4 shows how the float resistor is inserted between the cable screen and the amplifier chassis; as a result, most of the ground loop voltage is dropped across the float resistor, leaving a relatively small fraction across the cable screen. In effect, the common-mode signal v_g is reduced by the ratio of the screen resistance (typically a few tens of milliohms) to the float resistance. Note that the reduction of the spurious voltage across the screen also results in a corresponding reduction in interference capacitively coupled from the screen to the inner conductor of the connecting cable.

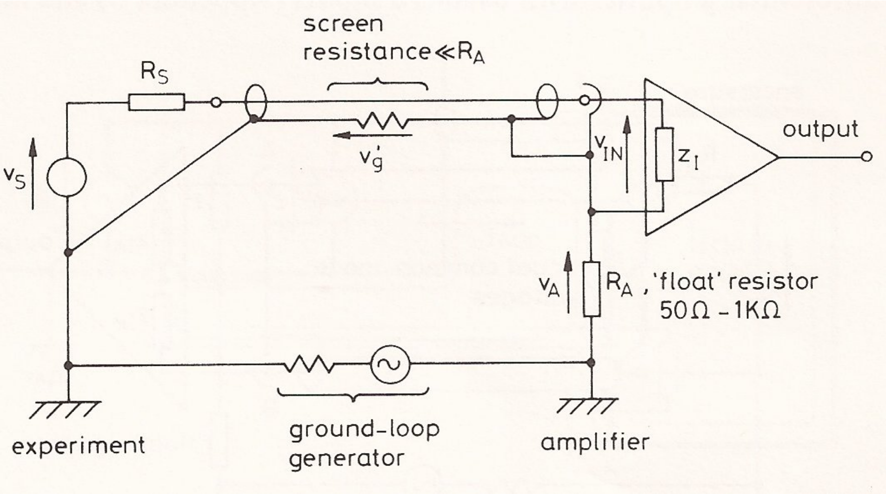


Fig. A6.4 Introduction of an amplifier "float" resistor, resulting in a much reduced common-mode votlage v_g

A6.3 Ground loops: differential amplifiers

The ground-loop suppression afforded by a semi-floating single-ended amplifier is impressive, but yet may be insufficient in some circumstances. When, for reasons of safety or practicability, it is not possible to use a floating source, residual connection problems can usually be overcome by using a differential amplifier as shown in Fig. A6.5. As before, an amplifier float resistor is used to bring about a large reduction in the spurious cable screen voltages which now appear as equal common-mode voltages across the amplifier inputs. As explained in Appendix 5, the extremely high common-mode rejection of the amplifier will then ensure suppression of ground-loop interference at power frequencies.

In order to maximize suppression it is essential to provide identical routes between source and amplifier for both cables to ensure that there is no differential pick-up in the two screens. This problem can be overcome by arranging to transmit differential signals over twisted pairs in a common screen, an approach favoured by instrumentation engineers. The symmetry of a twisted-pair connection also tends to equalize capacitive pick-up between the screen and the two conductors, and is additionally effective in reducing inductive pick-up.

The common-mode rejection ratio of a differential amplifier falls at high frequencies. Common-mode pick-up at radio frequencies should therefore be reduced as far as possible using screening and the coaxilizer techniques referred to earlier. Even at moderate interference frequencies, spurious phase shifts caused by the distributed cable capacitances acting with unequal resistances in the two paths may cause incomplete cancellation of common-mode voltages. An improvement will usually be obtained when a fully balanced source such as an a.c. bridge is used with a "true" differential amplifier with balanced inputs (Appendix 5) and matched cable lengths.

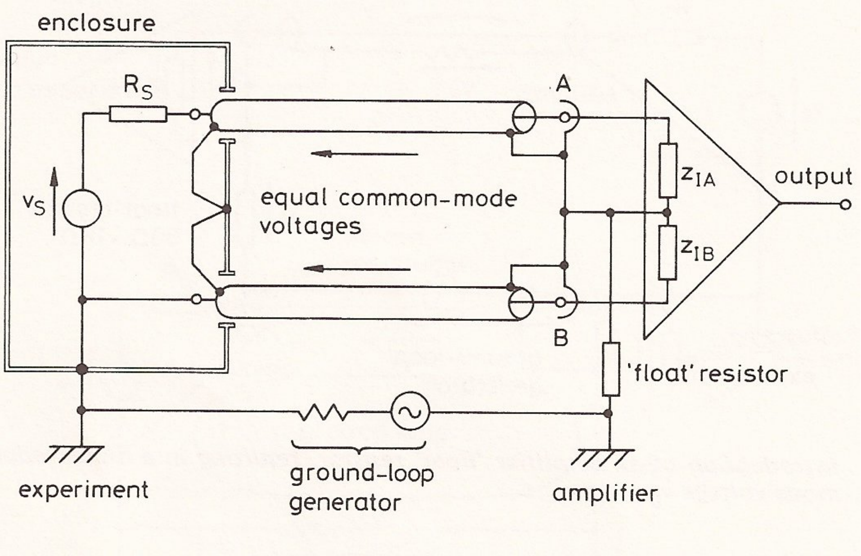


Fig. A6.5 Using a differential amplifier to overcome grounding problems

A6.4 Ground loops and lock-in recovery

Careless connection of signal and reference cables to a lock-in amplifier can lead to ground-loop problems over and above those described so far.

In most signal-recovery experiments, an oscillator is used as both an excitation and reference source, while the signal is taken from a "single-ended" transducer output. Fig. A6.6 gives an example where the oscillator and transducer output are strapped to the chassis of the experiment with direct connections made to the reference and signal inputs of a lock-in amplifier. The impedance Z_E represents the load provided by the excitation circuit to the experiment. To take a specific example, the oscillator could provide the drive to a vibration or "shaker" table, while the signal output is derived from a vibration transducer mounted on the device under test. In this case, the drive current could be several orders of magnitude greater than the current flowing in either the reference or signal circuits.

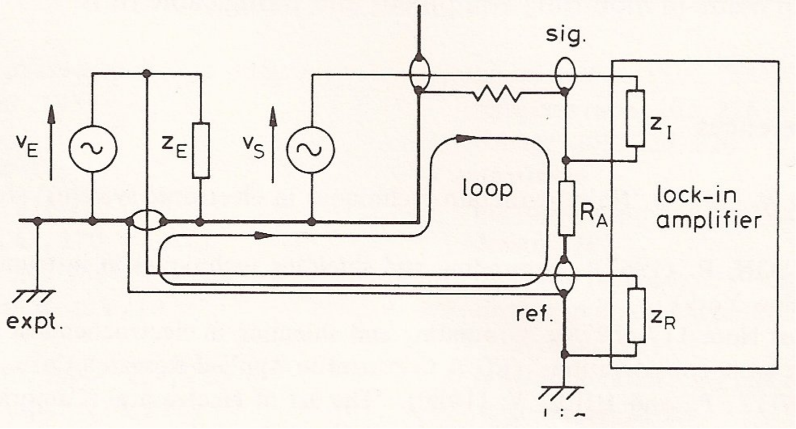


Fig. A6.6 Indiscriminate connections in a lock-in amplifier-based experiment causing a synchronous loop

Problems will arise in practice with the arrangement shown because the return path from Z_E to the excitation oscillator is shunted by a second path formed by the screen connections on the reference and signal connection cables. A fraction of the drive current is thus able to circulate in this path and generate a *synchronous* voltage drop across the signal screen in series with the signal voltage of interest. Although the amplifier "float" resistance R_A will attenuate the voltage considerably, the spurious voltage could well be greater than the signal voltage. The fact that spurious voltage is also synchronous with the applied reference would severely restrict the range over which measurements could usefully be performed.

Note that if the drive load were omitted completely, the signal screen would continue to provide a return path for a fraction of the current flowing in the reference circuit. Fortunately, the reference current is limited by the input

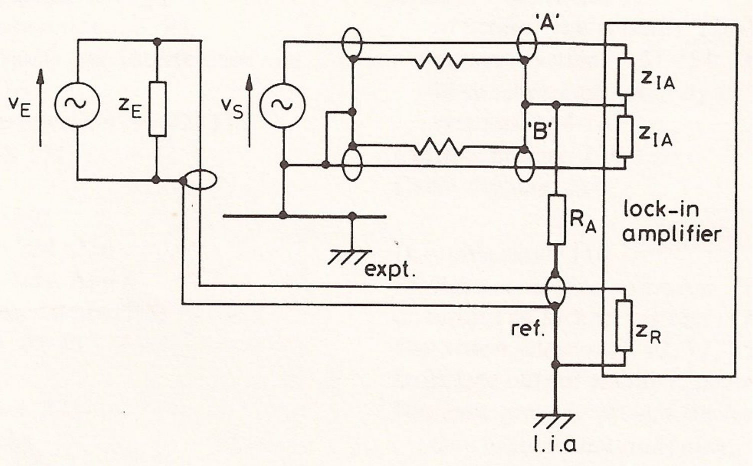


Fig A6.7 Improved experimental layout

resistance of the reference circuit. However, the spurious, synchronous, voltage set up in the signal input is often troublesome in low-level applications. A floating differential input can provide the solution in most cases. This leads us to the improved arrangement illustrated in Fig. A6.7, which also shows how to overcome the effect of the drive current. Here, the drive circuit is completely isolated from the sensitive signal circuit and contact is made to the experimental chassis at a single point. It is certainly possible that this, or some equivalent, arrangement could be arrived at by trial and error. It would be far better, however, to give some preliminary thought to ground loop problems; the alternative is almost invariable an unco-ordinated attempt to achieve "on-line" solutions when a large investment has already been made in mounting equipment and fixing cable runs.

A6.5 References

1. OTT, H.W (1976): "Noise reduction techniques in electronic systems" (John Wiley, New York).
2. MORRISON, R. (1967): "Grounding and shielding techniques in instrumentation" (John Wiley, New York).
3. Technical Note 117 (1978): "Grounding and shielding in electrochemical instrumentation - Some basic considerations" (EG&G Princeton Applied Research Corp., Princeton, NJ).
4. HOROWITZ, P., and HILL, W. (1980): "The art of electronics" (Cambridge Univ. Press, Cambridge).