A Signal Constellation and Carrier Recovery Technique for Voice-Band Modems

Iain Andrew Blair Lindsay

Ph.D.

University of Edinburgh

1986

"Je ne regret rien....."

Contents

Declaration of Originality

I hereby declare that this thesis and the work reported herein was composed and originated entirely by myself, in The Department of Electrical Engineering at The University of Edinburgh, between October 1981 and August 1986.

I.A.B. Lindsay

Abstract

This thesis may be divided into two parts; it covers the design and analysis of a signal constellation intended for the transmission of digital data over a voiceband channel at 9600 bps and a phase-locked loop structure and implementation intended for carrier regeneration and symbol timing estimation from a multiple phase shift keyed telegraph signal.

An introductory chapter on modems and channel characteristics is followed by a chapter on signal design, focusing on the major factors influencing the design and selection of a signal constellation for a specific application. Several candidate designs for transmitting 9600 bps in a voice band channel are described and compared. A new design is proposed (C1−5−10N) which offers good error probability performance and ease of decoding, particularly when employed in conjunction with the phase-locked loop technique described in subsequent chapters.

A detailed analysis of C1−5−10N is given, that describes the optimum position of the signal points and the error susceptibility when decoded with minimum distance boundaries and with independent phase/magnitude decision boundaries. A numerical technique for calculating probability of error is described. The distribution of error events over signal states is calculated and used to build an impure grey code. The zero-crossing characteristics of the signal are analysed, based on a representative test sequence of randomly selected symbols.

An introduction to phase-locked loop techniques for carrier regeneration includes descriptions of the non-linearity, Costas loop and remodulation loop methods. A new sampling phase-locked loop structure, the Signal Driven Phase-Locked Loop, is presented. This loop uses information derived from the zero-crossing instants of a signal, which also provide the loop clock signal. A key feature of the loop is its potential for simple implementation. The underlying mode of action of the loop is explained and critical properties of the individual elements of the loop are reviewed. In particular, there is a comparison of alternative techniques for implementing a numerically controlled clock (oscillator).

A detailed description of a prototype digital implementation of the new phase-locked loop is complemented with a Z-plane model. Properties of the noise on the regenerated carrier signal are analysed. Using an unmodulated carrier in both moderate and severe noise, the transient and steady-state behaviour of the prototype is measured and compared with the predictions of the Z-plane model.

The closing chapter summarises the current state of work on C1−5−10N and the Signal Driven Phase-Locked Loop. The utility of both is discussed, in the context of present day trends in modem design. Suggestions for future development are proposed.

List of Abbreviations

AGC	automatic gain control
APK	amplitude and phase shift keying
ARM	accumulator rate multiplier
ARQ	automatic request repeat
ASK	amplitude shift keying
AWGN	additive white gaussian noise
BPSK	binary phase shift keying
BRM	binary rate multiplier
CDF	cumulative frequency distribution
CNR	carrier to noise ratio
DCE	data circuit-terminating equipment
DPSK	differential phase shift keying
DTE	data terminal equipment
ECC	error control coding
FDM	frequency division multiplex
FDX	full duplex
FEC	forward error correction
FSK	frequency shift keying
HDX	half duplex
$I\Phi - M$	independent phase-magnitude
ISI	intersymbol interference
MSK	minimum shift keying
MSNR	mean signal to noise ratio
NCC	numerically controlled clock
NRZ	non-return-to-zero
OOK	on off keying

PC	phase counter
PDF	probability density function
PRBS	pseudo-random binary sequence
PSK	phase shift keying
PSNR	peak signal to noise ratio
PSTN	public switched telephone network
$P(\epsilon)$	probability of error
$P(\epsilon_b)$	probability of bit error
$P(\epsilon_s)$	probability of symbol error
QAM	quadrature amplitude modulation
QPSK	quaternary phase shift keying
RC	ramp counter
ROM	read-only memory
SDPLL	signal driven phase-locked loop
SNR	signal to noise ratio
SX	simplex
TDM	time division multiplex
TWT	travelling wave tube

Acknowledgements

Philosophers, in olden days,
Were wont to proffer glowing praise

To those, their patrons, who made offers
Of remuneration from their coffers.

And also to those learned men,
Whose knowledge raised their lowly ken

That they might also someday be
Possessors of a Ph.D.

So now, in turn, I do the same,
And mention herein some, by name.

To Jimmy Dripps I owe my chance
For academic dalliance,

Funded by The Treasury,
Through the hands of SRC.†

And Racal too, I here recall,
For funding and three months in thrall.‡

Chris Ash, with Terry, Brian and Tim,
Made working life seem not so grim.

In Edinburgh, so many folk
Have made the academic yoke

Sit lighter on me, that it might,
To list each one here, take all night.

I trust they will not be offended
By thanks which are so open-ended.

And finally I mention here
The people that I hold most dear,

My Mum and Dad, for giving me
The chance to be.

† Well, it was "SRC" when I started and "SERC" wreaks havoc with the scansion.
‡ It wasn't that bad, but I waive accuracy under poetic licence

1. Introduction

The initial aim of the project on which this thesis is based, was the development of a hardware prototype of the Signal Driven Phase-Locked Loop, in order to determine if the structure was viable in a practical environment and to investigate its potential in the demodulation of phase-shift keyed signals. A further specification was that the implementation be entirely digital and hence suitable for microfabrication.

This developed into a plan for a complete modem design based on the signal driven phase-locked loop, which led in turn to the study and design of a signalling format for use by that modem. A voice-band channel was chosen as the application area for two reasons. First, the original difficulty of estimating the upper frequency limit of a fully digital implementation and second, the range of applications of a voice-band modem. A high speed (9600 bits per second) voice-band modem, implemented as a single integrated circuit, has a wide market potential for inclusion in both portable and fixed data capture and processing equipment. The simplicity of implementation initially envisaged for such a device, when based on the signal driven phase-locked loop, promised a significant saving in equipment cost, size and power consumption.

1.1. Organisation of Chapters

The main body of this thesis may be divided into two parts, covering two aspects of modem design: The design and analysis of a signal constellation intended for the transmission of digital data over a voiceband channel at 9600 bits per second (chapters two and three) and a phase-locked loop structure and implementation intended for carrier regeneration and symbol timing estimation from a multiple phase shift keyed telegraph signal (chapters four and five). The introduction and conclusion (chapters one and six) place these two parts within the context of an overall modem design.

The introductory chapter provides a common base for the two streams within the main body, by providing an overview of the principal aspects of modem design. After a brief

description of the theoretical limitations to information transmission, the limitations and impairments commonly encountered on practical channels are presented. The different modulation techniques which may be employed in a modem are mentioned, with the constellation diagram being introduced as a method of describing signals in a manner which emphasizes their data transmission properties, rather than their methods of implementation. The chapter concludes with a summary of the main subsystems which may be found in a modem and some aspects of the way in which modems may be used in conjunction with the terminal equipment.

Chapter two presents the theoretical and practical aspects of the general signal design problem. The constellation diagram is used throughout as the means of identifying the signals discussed. Having defined the two major divisions of signal design as being the noise- and bandwidth-limited cases, the discussion concentrates on the bandwidth-limited area and presents and contrasts the properties of four specific constellations. In chapter three, a new signal design is proposed (C1−5−10N) which offers good error probability performance and ease of decoding, particularly when employed in conjunction with the phase-locked loop technique described in chapters four and five. A specification of the design requirements is followed by a survey of previously published designs, leading to the identification of a suitable structure. Methods of calculating the probability of error of a constellation in additive noise are discussed and a numerical approach is described in detail. This numerical procedure is used to refine the structure of the constellation selected, which is compared with those designs presented in chapter two. Detailed properties of C1−5−10N are presented.

Chapter four provides an introduction to phase-locked loop techniques for carrier regeneration. Following an introduction to the general carrier regeneration problem, descriptions of the non-linearity, Costas loop and remodulation loop methods are given. A description of the key features of phase-locked loop structures is followed by the presentation of a new sampling phase-locked loop structure, the Signal Driven Phase-Locked Loop. The underlying mode of operation of the loop is explained and critical properties of the individual elements of the loop are reviewed. In chapter five is a detailed report of investigations of a prototype digital (hardware) implementation of the new phase-locked loop, complemented with an initial attempt to produce a corresponding Z-plane model. Properties of the noise on the regenerated carrier signal are analysed.

Using an unmodulated carrier in both moderate and severe noise, the transient and steady-state behaviour of the prototype is measured and compared with the predictions of the Z-plane model.

The closing chapter summarises the current state of work on C1−5−10N and the Signal Driven Phase-Locked Loop. The utility of both is discussed, in the context of present day trends in modem design. Suggestions for future development are proposed.

1.2. Context

The rapid proliferation of computers over the past decade has brought with it an increased need to transfer digitally encoded information. While the physical movement of for example, magnetic tapes, is one way to achieve this transfer, it has many disadvantages and is inappropriate for interactive and other time critical applications. Electronic transmission of data can be fast and reliable, and is equally suitable for short or long messages. The principal requirement is a suitable communication link, which may take a variety of physical forms. The following definitions define major partitions of such a link:

i) MEDIUM; the physical structure which conveys the signal, or its method of use; e.g. coaxial cable, high-frequency radio.

ii) CHANNEL; the combination of the medium and any transducers or other devices needed to provide a suitable analogue electrical interface to it.

iii) LINE; as in *line signal*, that part of the link which carries the signal in analogue form.

iv) LINK; the channel and ancillary equipment, such as modems or line drivers, which may be needed to provide a digital data path. The ancillary equipment, referred to as Data Circuit-Terminating Equipment (DCE), provides a standardised digital interface to the Data Terminal Equipment (DTE) which uses the link.

For transmission over short distances, usually less than one or two miles, dedicated cables often form a suitable medium on which to base a link. By operating at baseband, using moderately simple hardware, communication at rates of several tens of megabits per second can be achieved. For longer distances, this baseband transmission approach becomes less suitable, partly because of the way in which the baseband signals become

distorted, leading to errors in transmission. Furthermore, it may be impractical to install dedicated cables, in which case the public switched telephone network (PSTN), leased lines, or some other medium such as radio, must be employed. Whatever medium is chosen, for moderate to long distances its usable passband will rarely extend down to zero frequency, and is unlikely to be well matched to the spectral distribution of energy in a baseband digital signal. A device which translates and modifies the frequency spectrum of a digital signal is known as a data modem (often simply 'a modem', being a contraction of *mo*dulator-*dem*odulator).

1.3. Fundamental Modem Limitations

As a DCE element in a data link, the modem's function is to convert an information bearing electrical signal between a discrete digital, and a continuous analogue format. The conversion must preserve the information content of the signal, and provide immunity to the distortions and interference encountered by the analogue signal, during its passage through the channel. At the same time, the modem should be efficient in its use of the channel, providing a high rate of data transfer, commensurate with the limitations imposed by the channel's characteristics. An example of a data link configuration is illustrated in figure 1.1.

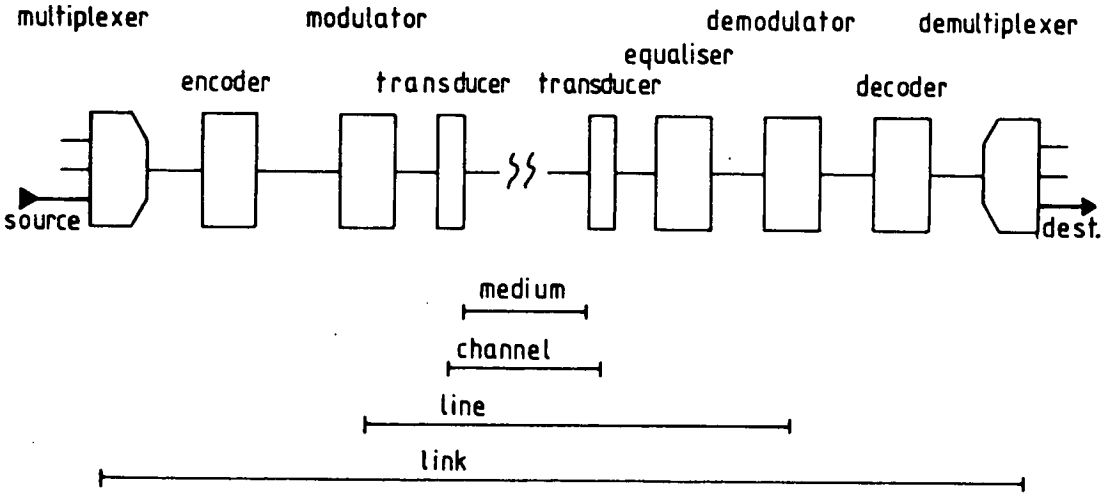


Figure 1.1 Example Data Link (simplex)

1.3.1. Quantifying Information

In communications engineering, the word *information* has a well defined meaning. The concept of information was formalised by Hartley [1], who introduced a logarithmic measure of the information content of a data source. This measure, which he called the *entropy*, is a function of the unpredictability of the data generated by a source. For example, if a source is only capable of transmitting two messages, and it continuously transmits these two messages, each after the other, then that source has minimal entropy because it is possible to predict its output with certainty. In this case, receiving the signal from the source does not yield any information, the message sequence could be reconstructed perfectly without it. Conversely, if the source transmits each of its two messages at the same average rate, but the choice of which message is transmitted in any given epoch does not depend on the messages transmitted in any other epoch, then the source has maximal entropy. The message sequence cannot be reconstructed without information from the source. Any attempt at reconstructing the sequence without this information will, on average, only succeed in putting the correct messages in a certain fraction of the places (in this case, fifty percent). Notice that creating a sequence in which the wrong message is always chosen for a given epoch, requires information. For the binary example just given, this is the same amount as is required to choose always the correct message.

The entropy of most data sources is significantly less than the maximum which could be obtained using the same set of messages. Conceptually, the output of such sources can be split into two components, information and redundancy. By reducing the redundant content of the raw data stream, the information may be transmitted more efficiently.

1.3.2. Manipulating Redundancy

Source encoding, the process whereby redundancy in a signal may be reduced, is normally considered to be outwith the scope of modem design. This is partly because there is no one technique of sufficiently wide applicability that it can be built into a general purpose modem. Where such data rate compression is feasible and desirable, it is achieved by a pre-processor, separate from the modem. As a result, except for the aspect described under 'scrambling' below, a data modem treats its input as if it were pure information with no redundancy.

Increasing the redundancy of a source does enter into modem design. Adding redundancy is a means of protecting the transmitted data against corruption, by repeating the information in a manner which enables the receiving modem to cross check its interpretation of the line signals effectively. This use of error control coding (ECC) to increase the redundancy of data is widespread, and not restricted purely to communication systems. The ubiquitous parity check bit appended to ASCII coded data words is a common, although not especially efficient, example.

There are two forms of ECC. Automatic request repeat (ARQ) relies on coding to detect the presence of erroneous bits in a block of received data. The receiver then uses a *backward channel* to request the transmitter to repeat the corrupt block. This is a closed loop technique and has the advantage of being able to adapt to varying levels of interference. The disadvantages are the requirement of a second channel (of relatively low bandwidth) and a catastrophic collapse in performance as the probability of receiving an uncorrupt block falls. Consequently, ARQ is only practicable on channels with moderate or low levels of interference. The presence of a long delay between transmission and reception increases the complexity of an efficient implementation, by requiring the transmitter to store blocks for possible re-transmission.

Forward error correction (FEC) involves adding sufficient redundant data to the transmission that the receiver can deduce both the presence and position of erroneous bits. This is an open-loop technique; there is no feedback to the transmitter. FEC must be able to deal with the worst expected case, which results in a poor average throughput if the interference level is low or very variable. It has the advantage of not requiring a backward channel and being applicable to any level of interference.

Originally, ECC for modem applications was developed as an independent element of the data link, dealing with the purely digital representation of information (algebraic encoding), the assignment of analogue channel symbols to digital words (symbol encoding) being a separate problem. However, by merging the algebraic and symbol encoding, performance gains in the range 2dB to 6dB. may be achieved [2,3]. Several recent designs for high performance modems [4,5,6,7] have incorporated ECC, carefully matched with the symbol encoding and international standards [8] are being defined, in an attempt to ensure compatibility between modems using this approach.

1.3.3. The Nyquist Limit

A direct effect of ECC is an increase in the number of data bits which must be transmitted in order to signal the information. If the information rate is to remain the same after coding, then the data rate through the channel must increase proportionally. The effects of increasing the rate of transmission through a channel were studied by Nyquist [9] in connexion with the behaviour of orthogonal (non-overlapping) baseband pulses on telegraph lines. He established that in order for such pulses to remain orthogonal, their maximum rate of transmission was restricted to be equal to the bandwidth of the channel. He then went on to note the inefficiency of the "carrier telegraph", which required twice the bandwidth of the "dc-telegraph", suggesting that one remedy for this inefficiency was the transmission of two independent telegraph signals, one modulating a sinusoid and the other, a cosinusoid, both superimposed within the same bandwidth, but separable at the receiver by *coherent* detection.

This use of quadrature carriers increases the *dimensionality* of the transmission, enabling two orthogonal pulses to coexist within the same timeslot, thereby doubling the pulse-rate through the channel. Hence the limiting relationship between pulse rate and bandwidth for orthogonal signalling:

$$R \leq nB \tag{1.1}$$

where **n** is the dimensionality and **B** is the single-sided bandwidth of the channel. For the common case of coherent demodulation, **n** = 2, but this value may be raised if there are other parameters of the carrier medium which may be independently modulated, such as polarisation. Manipulating the dimensionality of a transmission is discussed further in chapter two.

Because neither perfectly time limited, nor perfectly band limited signals are realistic entities in a practical situation [10], a theory for designing easily separable signals is required. By choosing suitable spectral shape factors, having vestigial symmetry, for the signals used, Nyquist determined that it is possible to generate pulses which he referred to as *non-distorting* [9 p621]. These pulses, although spread over several timeslots, do not interfere with one another, provided that the value of each pulse is taken at its mid-point. By using signals in this class, orthogonal signalling approaching the Nyquist limit may be achieved. Orthogonality is a desirable property of a signal set, because it results

in a line signal with good immunity to additive noise. Sacrificing the requirement that symbols (pulses) be orthogonal leads to the bandwidth efficient, or *dense* signal sets, with which it is possible to increase the data rate through the channel, but at the cost of an increased probability of error ($P(\epsilon)$).

1.3.4. The Shannon Limit

The principles of Hartley and Nyquist were combined by Shannon [11,12], who included the effects of noise, to produce a comprehensive theory of information transmission. He showed that for a certain type of channel, there was a well defined upper limit to the rate at which information could be transmitted. The equation formulated by Shannon, for the capacity of a continuous, mean power limited, white gaussian noise channel is:

$$C = W \log_2 \left(\frac{P + N}{N} \right) \tag{1.2}$$

Where the capacity **C** is in bits per second, the bandwidth **W** is in Hertz and **P** and **N** are the signal and noise power respectively. He also showed that this was the limiting form of the capacity equation for peak power limited and coloured noise channels. From equation (1.2), it can be seen that bandwidth and signal power may be traded off against each other, to provide the required capacity for transmission. There is no threshold level for either parameter†. As long as both are non-zero, the theory predicts a certain finite mean rate at which information may be transmitted through the channel. It is theoretically possible to transmit information with zero $P(\epsilon)$, at rates up to the limiting rate. Above this limit, there will always be a finite $P(\epsilon)$, so that although the number of data bits per second may exceed the capacity, the number of information bits per second will not. The one practical deficiency of Shannon's prediction is that it is an average over an infinite timespan, so there is no bound on how long one might have to wait in order to receive an error-free message. Transmitting at a rate less than the limit does not evade this problem. If a message must be passed within a finite time, a finite $P(\epsilon)$ is inevitable.

Although few real channels are as well behaved as Shannon's model, the general implications of his work are widely applicable. Thus, the object of modem design is to

† There is a lower limit of -1.6dB to the ratio of *symbol energy* to noise spectral density, corresponding to infinite bandwidth.

minimise the $P(\epsilon)$, while approaching as closely as possible to a theoretical estimate of the channel's capacity. Shannon likened the processing of a modem to a "statistical impedance matching transformer", whose function is to maximise information flow through a channel. It varies the physical structure of a signal, while maintaining the logical content.

1.4. The Channel

The channel and its disturbances may have many physical forms and origins, each of which may be used to classify the expected behaviour. The modem deals with an electrical interface to the channel, so it is natural to express the properties of the channel in terms of their effect on the electrical signals at this interface.

Shannon's capacity theorem indicates that the principal division of channel limitations is between bandwidth and interference. These may be further sub-divided as follows.

1.4.1. Bandwidth

The definition of bandwidth is a somewhat arbitrary procedure [10] and depends on the intended use of the parameter. Because of this, equation (1.1) is as much a definition of bandwidth as of rate of transmission, a fact emphasised by the variability of the 'constant' factor of two, used in (1.1) [13]. In particular, the *usable* bandwidth of a channel is limited by the linear distortions which it impresses on a signal. Linear distortion covers those aspects of the channel transfer characteristic which produce reversible effects on the signal. Thus, by suitable processing and with sufficient information, these defects can be corrected. The extent to which such compensation may usefully be extended is ultimately limited by the omnipresent law of diminishing returns. Thus, compensating for large values of attenuation in a magnitude transfer characteristic will most probably result in the addition of more noise than signal.

1.4.1.1. Frequency Distortion

This is a misleading name for a channel amplitude transfer characteristic which is a non-constant function of frequency, within the band of interest. Two typical examples are a constant slope of increasing attenuation with frequency, encountered on short haul (i.e. baseband cable) telephone circuits [14] and notches caused by multipath propagation over radio channels [15]. Such impairments reduce the effective bandwidth of the channel

such that signals, which are usually designed to have a comparatively flat in-band spectrum, develop intersymbol interference (ISI) when transmitted through such channels. Additionally, the noise in the channel is likely to be coloured and this affects the statistical properties of the resulting error events. A comparative study over three different signalling techniques, of the effects of frequency distortion and parabolic phase distortion, is given in [16].

1.4.1.2. Phase Distortion

A channel is said to introduce phase distortion when its phase transfer characteristic is not a linear function of frequency and hence its group delay is not constant, over the signal bandwidth. As with frequency distortion, phase distortion may also result in ISI, in this case caused by dispersion of the energy in each symbol as a function of frequency. Phase distortion differs from frequency distortion in that energy is not selectively removed from the signal, rather, it is displaced in time; it is rare to encounter either phase or frequency distortion without the other. Although a commonly desired characteristic, phase linearity (constant group delay) is not always an essential condition for optimum pulse transmission [17,18,19], being dependent on the approach taken in designing/adjusting the channel response.

1.4.1.3. Echoes

Echoes are a dispersive phenomenon, caused by abrupt anisotropy in the medium and also by multiple path propagation where the relative delay between paths is much greater than the correlation time of the channel. As a crude analogy, echoes may be considered a time limitation, in a similar sense to bandwidth being a frequency limitation. A transmitter may send for a period limited by the arrival delay of the first echo at the receiver. It then refrains from sending until the echo has decayed, whereupon it commences a further period of transmission. In practice, such a procedure is unnecessary since it is possible to cancel the echo at the receiver, allowing continuous transmission with unimpaired reception. The characteristics of the discrete echo [20] are basically similar to those of the primary signal, except that, due to the different propagation route, they will have been affected differently by the channel. Echoes from one transmitter can affect the receivers at both ends of a channel; this unique property is an important consideration in full duplex operation.

1.4.2. Non-Linear Distortion

The most obvious non-linearity in any system is an inability to handle unlimited peak power levels. While conservative design may avoid this in most instances, the amplitude transfer characteristic over a limited range will still not be totally linear. This may be due to the physical properties of the medium, or because of imperfect transducers or other line equipment. The ionospheric channel is non-linear because of the properties of the reflecting medium, telephone local loops incorporate iron cored transformers and any line signal which undergoes frequency translation is subjected to an essentially non-linear process. Whereas the distortions mentioned in the previous section are (theoretically) reversible and affect each signal in the channel independently, the effects of non-linear distortion are potentially more serious. A method of characterising bandpass non-linearities, which expands the usefulness of the common two-tone intermodulation test, is given by Maseng [21].

1.4.2.1. Gain Compression

Gain compression is a term describing the first-order effect of limited power handling. High power signals appear to suffer more attenuation than those of lower power, as their peak magnitude approaches the limit. This effect can seriously reduce the efficiency of microwave channels which employ travelling wave tube (TWT) amplifiers, for example. Pre-distortion of the signal may be used to compensate partially for the compression, but will exacerbate other effects. Particularly for the current generation of satellite channels, gain compression may be a significant deciding factor in the choice of signal design for a modem.

1.4.2.2. Spectral Shift

Spectral Shift is encountered on links where, at some intermediate stage, the line signal has been translated in frequency with imperfect syntonisation of the corresponding heterodyne oscillators. It may also occur as doppler shift caused by relative motion of the two ends of the link (mobile communications), or of an intermediate propagator (non-geostationary satellite or ionospheric layer) and is commonly encountered on older trunk telephone channels (e.g. intercontinental cables [20]). The receiving modem may be capable of tracking the nominal centre frequency of a signal, in which case this effect may be compensated for over a limited range.

1.4.2.3. Phase Noise

Phase noise, also known as jitter, has the same origins as spectral shift, but results from instability. It reduces the coherence of a signal and, in extreme cases, the resultant signal may have the same envelope as the original, but with random phase. The severity with which phase noise affects a signal is independent of signal level, as it is the result of a modulating process. Low frequency phase noise, i.e. below the data rate of the signal, may be cancelled by suitable phase tracking.

1.4.2.4. Intermodulation

Intermodulation is the generation of spectral components within the passband of a channel, originating from the action of third and higher-order non-linearities on the primary signals in the channel. The generation of intermodulation products requires that the primary channel signal have a minimum of two spectral components, whose harmonics can be combined to produce discrete interferers. The structure for a more complex signal rapidly approaches a noise-like form, with the cure for intermodulation distortion being to eliminate the non-linearity.

Harmonic distortion products are integer multiples of spectral components within the passband. Since such multiples usually lie outwith the channel passband, they are not of direct concern, but they may cause interference to other channels.

1.4.3. Interference

The energy arriving at a receiver is an aggregate of energy from the transmitter, plus energy from a variety of other external sources. Of the energy which derives from the transmitter, some may be ascribed to the action of the distortions mentioned above. These distortion products, plus the contributions from other external sources (additive noise), constitute interference which is responsible for errors in the output data. Since the distortion products have already been dealt with, this section will cover the characteristics of the additive noise, that which exists independently of the desired signal. The extent to which this interference may be eliminated depends on its predictability, or more precisely, on its dissimilarity to the wanted signal. When the characteristics of the wanted signal are the same as those of the interference in the channel, then no amount of further processing at the receiver will help. For the purpose of calculating error probabilities, noise encountered in communication systems is often modelled as being

white (constant power per unit bandwidth) over the band of interest and as having a normal or gaussian amplitude probability density function (PDF). This additive white gaussian noise (AWGN) model is representative of the unstructured or smooth noise present in all channels, but does not adequately represent the impulsive or structured noise which is also present on many real channels [22,23,24]. These two classifications of noise will be dealt with separately.

1.4.3.1. Smooth (Unstructured) Noise

The principal source of smooth noise is the thermal agitation of charge carriers. The ensemble average of the very large number of these present in any real channel, results in a PDF which closely approaches gaussian. This accords with the central limit theorem, the overall PDF of a group of sources is the convolution of the individual PDFs and, as the number of independent sources increases, the composite PDF tends to a gaussian curve, irrespective of the PDF of individual contributions. The spectral distribution of thermal noise is white up to a limit determined by the noise temperature of the channel. For a noise temperature of 300K, this limit is in the sub-millimetre wavelength region. The shape of the gaussian curve is symmetrical about the mean and contains most of its area within two and a half standard deviations of the mean. The residue (less than two percent) is distributed in the 'tails', which extend towards infinity. A wide variety of noise environments have a PDF which is gaussian within one or two standard deviations [25], but this range is not usually important in determining the performance of a data transmission system. For most real systems, operating at error probabilities of less than 0.1%, the errors are a result of the noise represented by the tails. The correspondence between the theoretical curve and actual shape of these tails determines the usefulness of the gaussian noise model. There are few classes of channel, extra-terrestrial communications being one exception, where errors are almost entirely due to gaussian noise. While individual cases of purely gaussian noise have been observed [23 p3250] the majority of channels have a significant level of impulsive noise. The popularity of the gaussian model in theoretical work may be partly due to its mathematical tractability, its undoubted presence as a component in all systems and the previous lack of any general description of real noise environments. Attempts at alternative models are noted in the next sub-section.

All real noise is effectively bandlimited [10] and therefore not strictly white. However, if the noise process has a spectrum which is flat over a bandwidth greater than that of the receiver, its deficiency outwith the range of the receiver will go undetected. Thus the noise may effectively be considered to be white. Even so, amplitude distortion, present in most channels, means that the noise emerging from a channel will be coloured to some extent, even if the original source was white. The difficulties of treating coloured gaussian noise theoretically are outlined in [26].

Despite the practical shortcomings of the AWGN model, it has the advantage of being the basis of a standard test of modem performance. Theoretical and controlled experimental results using AWGN are widely available and can be directly compared. Also, the model meets the reductionist principle of minimising the number of parameters qualifying a result.

1.4.3.2. Impulsive (Structured) Noise

The distinguishing feature of impulsive noise is its high probability of large interference levels. It encompasses wideband signals such as short pulses (time domain impulsive) and narrowband signals such as single tones (frequency domain impulsive). These signals may originate from natural sources, e.g. lightning, or may be man-made, e.g. cross-talk. Attempts to model impulsive noise have often relied on measuring the noise in a specific environment, then using this information, in conjunction with knowledge of the underlying physical processes, to create a model for that particular environment [22,23,24,27]. Such models are of limited applicability and lack the widespread acceptance of the AWGN model. A model which does not relate to specific types of interference has been proposed by Middleton [25]. He classifies all non-gaussian noise as impulsive and then divides this into two principal classes (class A and class B). A third class (class C) is the combination of the two principal classes. The model for class A noise, which relates to sources having a bandwidth less than the receiver, requires three parameters. The class B model for wideband sources requires six parameters. Procedures for determining these parameters for a particular environment, given measured data, are given in [28], and examples using this model to describe a wide variety of noise environments[25,28,29] show good correspondence.

1.5. Modulation Techniques

The principal action of a modem is translation of the frequency spectrum of a signal between baseband and a passband based on some nominal carrier frequency. The value of the carrier frequency chosen is determined by characteristics of the medium or the interface to the medium and does not in itself affect the parameters of the data transmission. Only the bandwidth of the translated signal is important and this is governed by the rate and type of modulation employed. There are three primary classes of digital modulation, based on the classification of their analogue counterparts: Amplitude Shift Keying (ASK), Frequency Shift Keying (FSK) and Phase Shift Keying (PSK). There is a fourth class, Amplitude and Phase shift Keying (APK), which is a combination of ASK and PSK. This modulation format was noted very early on in the history of digital signalling [9], although its analogue counterpart has never been widely employed. Within these very general categories, there exists a plethora of acronyms for various subdivisions, e.g. binary phase shift keying (BPSK), minimum (frequency) shift keying (MSK), on off keying (OOK), quaternary phase shift keying (QPSK) &c.. In some instances there are several different acronyms, not all of them in the same one of the above four divisions, which apply to the same modulation format, e.g. the synonyms QPSK, 4PSK, 4QAM (four point quadrature amplitude modulation). There are further complications to this naming scheme which will not be discussed here. Furthermore, the ease with which arbitrary waveforms may now be synthesised renders obsolete a naming scheme based on the explicit modulation of a carrier. An alternative naming scheme might be based on the characteristics of the constellation diagram corresponding to the given symbol set. The constellation diagram provides a graphical description of the line signal of a modem, without reference to any particular modulation technique. It reveals similarities between apparently disparate formats and allows rapid appraisal of their basic features.

A wide-ranging comparison of modulation techniques for digital radio applications has been published by Oetting [30], which contains an extensive bibliography.

1.6. The Constellation Diagram

The constellation diagram represents the output of a modem as a set of points, each point corresponding to one of the channel symbols used by the modem to convey a particular

bit pattern. These points are distributed over N dimensions, where N is the minimum number of orthonormal basis functions required to synthesise all of the symbols in the set. Because the constellation diagram does not require any specific group of orthonormal functions to be used, a waveform may be analysed by whatever technique is most suitable, although orthogonal sinusoids are usually assumed unless there is an explicit statement to the contrary. The resulting constellation diagram may be directly compared against any other, regardless of the fact that the axes of the two diagrams may not represent identical functions. The essence of the constellation diagram is that it illustrates the distance properties of a signal set and to this end, all groups of orthonormal functions are equivalent. Distance, in this context, is a measure of how distinguishable two signals are from each other. Points separated by large euclidean distance represent signals which are very distinct and may be considerably perturbed without compromising their identity (causing errors in reception). Real signals are perturbed by passage through a channel, so that knowledge of the effect on the constellation diagram of a particular disturbance allows error prone regions to be identified. Chapter two deals with this subject in more detail.

1.7. Modem Subsystems

Figure 1.2 is a block diagram showing the principal elements which may be found in a modem for use with telephone channels. Not every modem will incorporate all the elements shown in figure 1.2 and some may have a different configuration of elements, but it is generally illustrative of medium to high speed synchronous modems which use PSK or APK signalling.

1.7.1. Scrambling

The function of a scrambler in a modem is to ensure that the symbol stream seen by the receiver does not have certain undesirable characteristics. These scramblers have nothing to do with encryption for confidentiality. Many data sources have a high propensity for generating long strings of identical symbols, or other patterns which may make it difficult for the receiving modem to maintain correct synchronisation of its subsystems and which foil the efforts of tracking equalisers. The scrambler reduces the probability of occurrence of these troublesome patterns (but does not eliminate them). It is usually implemented as an m-sequence polynomial code generator, whose output is convolved

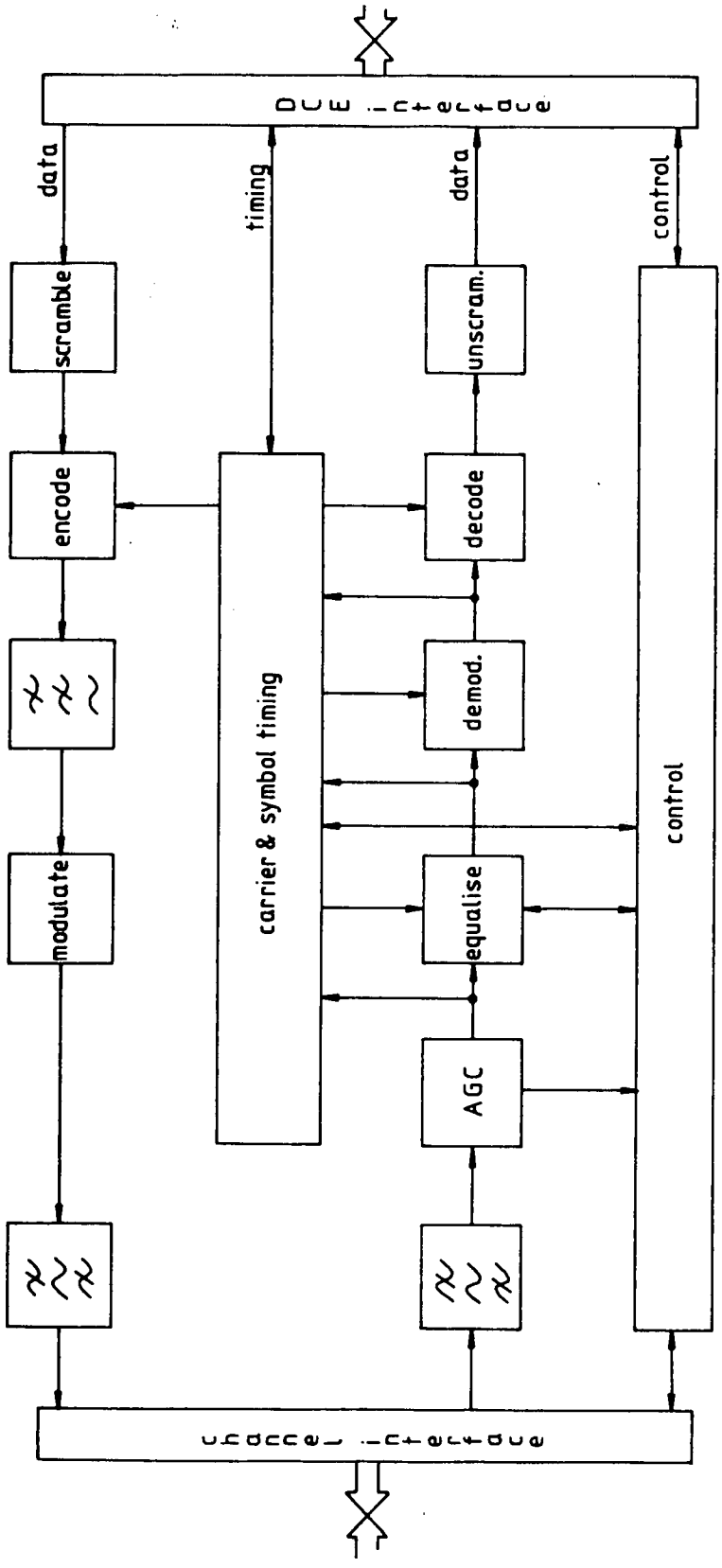


Figure 1.2 Elements of a Modem

with the data to be transmitted. The effect of this is to disperse spikes in the PDF of data symbols, whitening the frequency spectrum and ensuring a random occurrence of transitions in the data stream from the demodulator. By suitable design, these scramblers may be made to be self synchronising, so the receiver does not require additional circuitry to synchronise with the transmitter scrambler.

1.7.2. Symbol Encoding

The symbol encoder collects binary digits from the input stream and uses them to select channel symbols for transmission. This may either be done on a block basis, with each bit affecting only a single channel symbol, or convolutionally, when a single bit will affect the selection of several channel symbols. When the number of symbols available is a power of two, block encoding is straightforward and commonly used. Otherwise, a hybrid or pure convolutional encoding may be employed. Where phase modulation is used to generate the channel symbols, it is common to encode bit patterns as transitions between symbols, rather than as particular symbols. This differential technique is used to avoid the requirement for an absolute phase reference and results in a small performance penalty which is acceptable in most cases. Differential keying may be avoided by using a preamble to establish an absolute reference phase. The optimum allocation of bit patterns to symbols (or symbol transitions) depends on the the inclusion of ECC and on the demodulation technique. In the common situation, where hard decisions are taken as to which symbol has been received, bit patterns are allocated so that erroneous reception affects as few bits as possible. This is termed *grey coding* and is important when symbol errors are predominantly to nearest neighbours, as is the case when the interference is mainly smooth (non-impulsive), low level noise. An alternative strategy, termed *set partitioning* [2], may be used when the modem incorporates ECC and the demodulator makes soft decisions on the received symbols. The use of ECC with hard decision demodulators requires a grey coding scheme [3]. The expected nature of the channel disturbances is important when deciding on details of the ECC implementation. Principal parameters of interest are the expected length of error bursts and the nature of any correlation between successive error events. The block length of a block code and the constraint length of a convolutional code, limit the extent to which clustered error bits can be detected or corrected. Very generally, for example, a code which can deal with three errors in fifteen will be more effective than one which can

handle one error in five. Both have the same naive correcting power, but only the first one could deal with three consecutive error bits. Similarly, when error events are correlated, because of the nature of the noise source, the presence of ISI, or the use of differential encoding for example, the number of bits over which a coding scheme can exert its influence will govern its effectiveness. Interleaving is a simple way to combat correlated errors, by grouping individual bits selected at an interval greater than the correlation distance. Alternatively, codes may be specifically designed to deal with structured error patterns [31]. From an implementation viewpoint, block codes generally introduce less delay and are more straightforward to decode than convolutional codes, but offer less potential correcting power for a given implementation complexity. An interesting comparison study is given in [4].

1.7.3. Filtering and Modulation / Demodulation

The modulator is the point at which digital bit patterns are converted to continuous analogue waveforms, suitable for transmission through the channel. The exact structure of the modulator depends on the signalling format used and on the technology available to implement it. The precise and repeatable control of the waveform afforded by the use of digital signal processing, allows the implementation of signalling formats which might otherwise be prevented by the tolerances required on analogue components. Where the modulation rate is too high for digital waveform synthesis, regularity of the signal constellation is an important consideration. If the constellation allows the modulator to be broken down into a number of canonical sub-sections, the requirements for balance and matching of components are eased. An example of this is the 16QAM signal set, which has proven popular for microwave digital radio channels [32]. The same comments apply to the design of the demodulator, which must be able to maintain accurately placed threshold levels.

Filters may be used both before and after a modulator, to control pulse shape and line signal spectrum. The bandwidth of the transmitted signal is determined mainly by the modulation rate and type of modulation employed. The exception to this occurs in systems which use partial response signalling. In this case, filters are used to limit the bandwidth of the transmitted signal to a range significantly less than the nyquist limit. The resulting controlled intersymbol interference generates a constellation with an increased number of transmission levels, and is one way of generating dense signal sets.

The filtering is then an integral part of the modulation process.

If the modulation process is linear, filtering pulses at baseband immediately prior to modulation, may be used to ensure symmetry in the spectrum of the line signal about the carrier frequency. This may be difficult to achieve with post-modulator filtering, if the geometric mean of the passband signal is not sufficiently close to the carrier frequency. Similarly a baseband filter, driven by impulses, may be used to generate the pulses sent to the modulator, with a complementary filter used in the receiver to provide matched filter detection. If the modulation process is non-linear, e.g. FSK, pre- and post-modulator filtering are not equivalent and so cannot be interchanged.

As well as exercising control on the shape of the main lobe of the transmitted spectrum, filters are commonly used to reduce the sidelobe level, in order to prevent interference to signals on adjacent frequencies. Corresponding filters are required prior to the demodulator, to remove interference and out of band noise. To minimise group delay distortion, the phase transfer characteristic of modem filters is usually tightly specified. This restricts the class of filter designs which may be used and leads to filters of quite high order. The specification may sometimes be relaxed for receiver filters preceding an equaliser, in which case elliptic and similar classes having a steep cutoff characteristic, may be used.

1.7.4. Equalisation

Equalisation may be used to compensate for some of the linear distortion of a signal, principally ISI, introduced by the channel. Although ISI is not usually severe enough to cause errors by itself, it moves signals closer to decision thresholds, reducing the margin against noise. Simple modems may not incorporate equalisers, particularly when they operate well below the limitations of the channel. If the characteristics of the channel are sufficiently reliable, a pre-set network may be incorporated, adjusted to equalise the nominal characteristics of the expected distortion. For operation close to the ultimate channel capacity, an adaptive equaliser is essential. In its simplest form, this is based on a transversal filter usually having between sixteen and thirty two taps, the weightings on which may be controlled to provide a filter response which is the inverse of the channel. The algorithms used to control the tap weights often aim to minimise the mean square error between the equaliser output and some desired signal [33]. A training signal is

often used to speed up the initial adaptation process, after which the weights are either frozen, or the equaliser uses decision feedback from the demodulator to track changes in the channel. This relies on the scrambled data transmission to provide an adequate sounding of the channel and a suitably low error rate in the decision feedback path.

In addition to compensation of the frequency and phase characteristic of the channel, it may be necessary to stabilise the received signal level, where this is not already done by other line equipment. Demodulators which have decision thresholds at zero are generally immune to variable signal levels. If the modulation format produces a signal which has an essentially constant envelope and the signal to noise ratio is acceptable, a limiter can be used to ensure a constant signal level. Other cases require an automatic gain control (AGC) circuit, to ensure that the demodulated signal corresponds to the pre-set threshold levels.

1.7.5. Carrier Recovery

Unmodulated energy at the carrier frequency conveys no information. Therefore, efficient modulation techniques generate a line signal without a coherent carrier frequency component (i.e. suppressed carrier formats). Coherent demodulation, which requires a locally synchronised carrier, offers FSK and ASK a desirable signal to noise advantage and is essential for PSK and APK. There are four basic approaches to providing the receiver with a local synchronous carrier. Differential phase shift keying (DPSK) uses each line symbol as the reference for its successor. This is adequate when the SNR is good, but degrades rapidly when conditions worsen, since the channel noise affects the receiver via two paths. Narrow-band and tracking filters (phase-locked loops) are widely employed to extract, either a low level pilot carrier deliberately injected at the transmitter, or a carrier regenerated by various non-linear operations on the received signal sidebands. As a fourth alternative for some wideband applications, the modem may operate in burst mode and use a preamble to establish a reference phase. A primary frequency standard provides sufficient phase and frequency stability for the duration of the burst, to obviate the need to derive reference information from the line signal. Chapter four discusses carrier recovery techniques in more detail.

1.7.6. Timing Recovery

A data symbol or bit timing clock is required by a modem to ensure that decisions on the output of the demodulator are taken at the optimum instant and to provide a timing signal on the data output of synchronous modems, when appropriate. Since the data format is normally non-return-to-zero (NRZ), there is no guarantee of a stable spectral component at the clock rate, so that techniques equivalent to those used for carrier recovery must be employed. Unlike the carrier recovery problem, which often requires frequency as well as phase tracking due to the possibility of frequency offset on the received signal, symbol clock synchronisation is essentially a phase tracking procedure. The data timing is not affected by carrier frequency shift and since any data regenerating stage along the link transfers data isochronously, only the (moderately tight) tolerance on the timing sources at the local and remote modems allows any frequency difference. For voice-band circuit operation, a timing frequency offset of the order of one part in ten thousand may be allowed, whereas carrier offsets may be fifty times greater than this [34]. Data timing may be derived from either the line signal, or the demodulated baseband data stream. The line signal approach gives faster acquisition, particularly when the remote modem sends a startup preamble, although the performance of data derived timing is more reliable during normal operation. Incorporating both techniques [35] gives the advantages of each, at the cost of increased complexity.

1.8. Control

As well as the business of ensuring accurate data transmission, the modem must recognise certain line disciplines and be able to control interface signals to the DTE. Line discipline is the low level protocol used to establish communication between modems. Dealing with such diverse things as data rate and format, modulation format and line access procedures, it is a combination of static parameters established when the link is installed or initially configured and dynamic parameters which may be changed during operation, either by the DCE or DTE. Some of the more important of these parameters are mentioned here.

1.8.1. Synchronous and Asynchronous Data Formatting

A modem can be designed to work with either a synchronous or an asynchronous data stream (not to be confused with the demodulation process, which may be coherent or

non-coherent). The simpler of the two types is the asynchronous modem. Used mainly for communication with interactive terminals, it accepts short bursts of data, usually between five to nine bits long, bracketed by start and stop pulses. The start and stop pulses are used to synchronise the timing in the remote DTE. A burst may be transmitted at any time, the modem does not maintain a clock at the data rate and usually requires no knowledge of the data rate, except insofar as it affects such parameters as the bandwidth of filters in the modem. In between bursts of data, the output of the modem is steady. Because of the presence of the start and stop pulses, the maximum line utilisation of an asynchronous modem is less than eighty percent. Consequently, where maximum efficiency is required, a synchronous modem is used. Synchronous modems maintain an internal clock locked to the timing of the data stream and transmit data continuously. If the DTE has no valid data to send, it generates an idle pattern which is then transmitted by the modem. This may result in some loss of transparency, certain data codes being reserved for data framing and idle filling, but this is a problem for the DTE and does not directly concern the modem. Because there are no start and stop pulses, line utilisation may approach one hundred percent. Balanced against this is the extra complexity of the modem, which must include additional control logic and maintain a synchronised bit clock derived from the channel signal.

1.8.2. Network Configuration

Depending on the medium and application, communication between modems may be simplex (SX), half duplex (HDX), or full duplex (FDX). For simplex operation, the modem is required to either transmit or receive only. For duplex communication, both transmission and reception are required; simultaneously in the case of full duplex, alternately for half duplex.

The startup time of a modem may be crucial to efficient line utilisation, depending on the way in which a communications link is operated. When a simplex or full duplex link is established and then maintained for a long period, the time required for the various elements of the modem to establish synchronism and adapt to the channel is usually unimportant. The turn-around time when the link is half-duplex, or the set-up time when the link is a segment of a polled network, may become limiting factors when these configurations are used to transmit short messages.

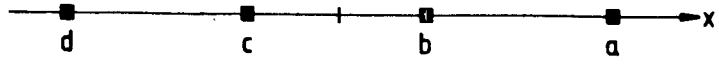
2. Signal Constellation Design

Chapter one having covered the limitations and disturbances of typical channels and the ways in which these affect the suitability of a particular channel for digital transmission, this chapter will consider the design of sets of signal waveforms to be used for transmitting data through the channel.

Certain types of signals, having low cross-correlations, make good signalling elements and are referred to as *naturally distant* signals. A description of these signals is followed by a consideration of the bandwidth limitations of channels and the utility of dense signal constellations for low noise channels. This leads on to the synthesis of artificially distant signals, by the application of coding and suitable demodulation techniques. The sphere hardening argument is presented as an explanation of Shannon's capacity theorem and of the effect of increasing the time-bandwidth product of a signal set. Basic characteristics of the constellation diagram are described, followed by a summary of approaches to the design of signal sets. The chapter concludes with descriptions of four sixteen point, two-dimensional constellations, which provide background for the study of the constellation presented in chapter three.

2.1. Naturally Distant Signals

Assuming that the linear distortions of the channel have been minimised by the use of appropriate equalisation techniques, the raw error rate due to additive noise is determined by the distance between symbols in the signal constellation. For additive interference, uncorrelated with the signal, this distance is the simple euclidean distance between the symbols, as represented in signal space on the constellation diagram. Naturally distant signals fall into three classes: orthogonal, bi-orthogonal and simplex. To illustrate them, figures 2.1[b-d] show the constellation diagrams for a four symbol signal set of each class. Figure 2.1a does not represent a naturally distant set, but is included for reference. It corresponds to a coherent ASK format, in which each channel

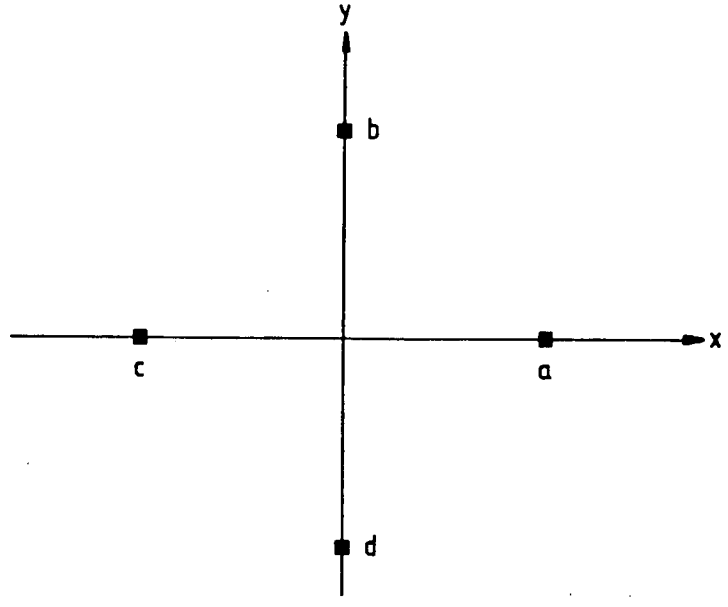


Mean Energy = 1

Minimum Distance = $2/\sqrt{5}$

■	x
a	$3/\sqrt{5}$
b	$1/\sqrt{5}$
c	$-1/\sqrt{5}$
d	$-3/\sqrt{5}$

Figure 2.1a Four Symbols in 1 Dimension (*not naturally distant*)

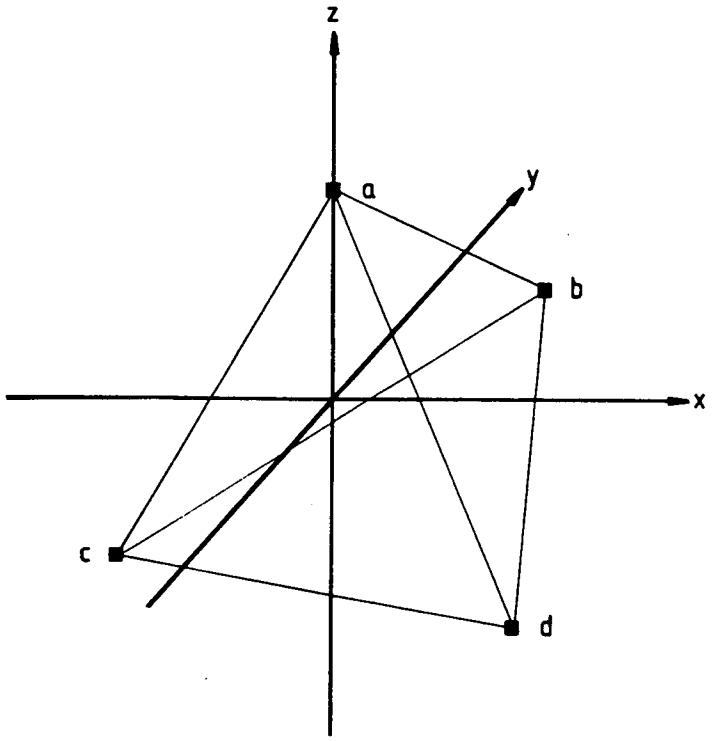


Mean Energy = 1

Minimum Distance = $\sqrt{2}$

■	x	y
a	1,	0
b	0,	1
c	-1,	0
d	0,	-1

Figure 2.1b Four Symbols in 2 Dimensions (Bi-orthogonal)

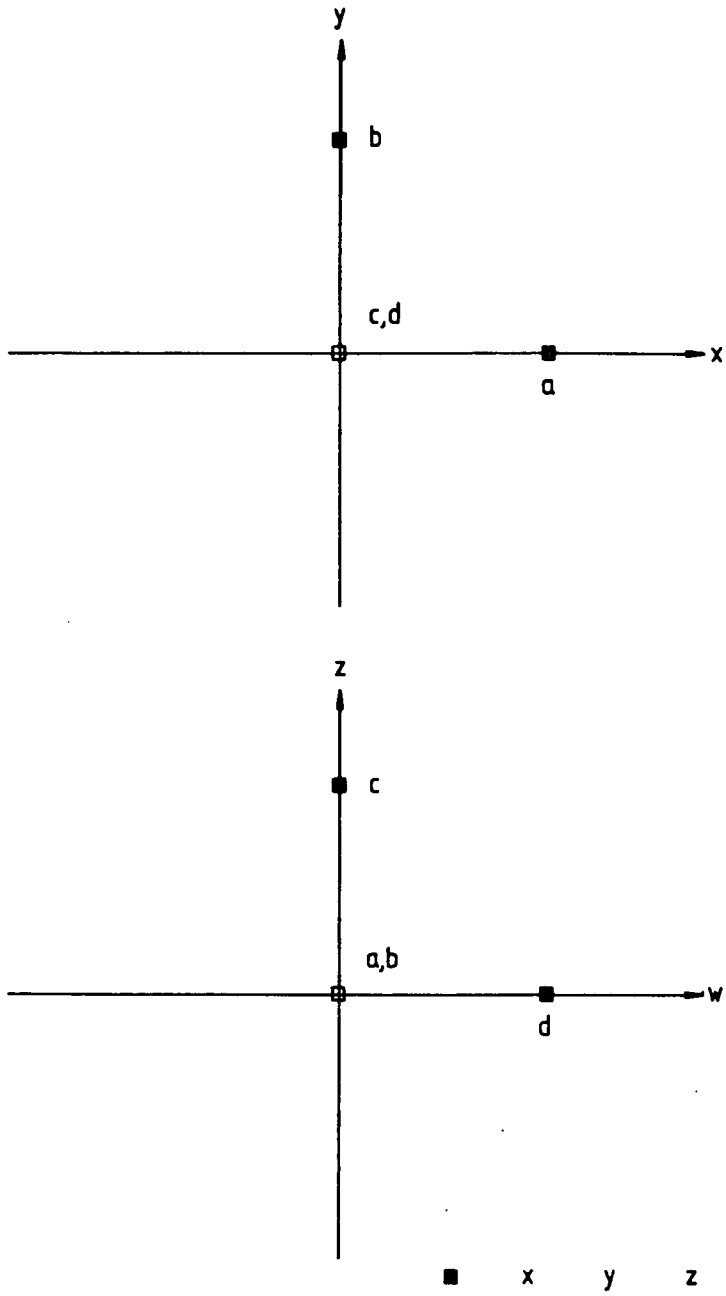


Mean Energy = 1

Minimum Distance = $\sqrt{24}/3$

	x	y	z
a =	0,	0,	1
b =	0,	$\sqrt{8}/3$,	$-1/3$
c =	$-\sqrt{6}/3$,	$-\sqrt{2}/3$,	$-1/3$
d =	$\sqrt{6}/3$,	$-\sqrt{2}/3$,	$-1/3$

Figure 2.1c Four Symbols in 3 Dimensions (Simplex)



Mean Energy = 1

Minimum Distance = $\sqrt{2}$

■	x	y	z	w
a =	1,	0,	0,	0
b =	0,	1,	0,	0
c =	0,	0,	1,	0
d =	0,	0,	0,	1

Figure 2.1d Four Symbols in 4 Dimensions (Orthogonal)

symbol is represented by a different amplitude of the carrier wave, unlike a naturally distant set where all symbols have equal energy.

Orthogonal signals are characterised by having zero cross-correlation between symbols. Therefore, an orthogonal signal set has one symbol per dimension of signal space, as illustrated in figure 2.1b. An FSK signal, where the tones are separated by a frequency of one half of the signalling rate, forms an orthogonal set when demodulated coherently. When non-coherent demodulation is employed, the tones must be spaced with a frequency difference equal to the signalling rate, to maintain orthogonality. This places each tone at a null in the spectrum of the energy resulting from every other tone.

Bi-orthogonal (or trans-orthogonal) signals are formed from orthogonal sets by the addition of the negative of each signal. Each signal in the set has a correlation of zero with every other signal except one, with which it has a unit negative correlation. Because they make use of negative cross-correlations to distinguish signals, bi-orthogonal sets must be demodulated coherently. Possessing two symbols per dimension of signal space, bi-orthogonal sets have the highest density of the three classes and are used where bandwidth is at a premium. An example of this class is QPSK, the constellation diagram for which is shown in figure 2.1c.

The third class of naturally distant signals form simplex (or antipodal) sets. They also require coherent demodulation and are characterised by each signal having the same (maximum) negative correlation with every other signal. This gives simplex sets the largest minimum distance of any constellation and makes them the optimum set for an AWGN channel, when bandwidth is unconstrained. Simplex signals form a linearly dependent set, such that each signal may be constructed as the negative of a summation of the others. Consequently, a set of M simplex signals requires $(M - 1)$ dimensions in signal space. Compared to orthogonal signals, simplex signals require less energy by a factor of $(1 - 1/M)$ for the same minimum distance, corresponding to an SNR advantage of 3 dB when $M = 2$, which falls to 1.25 dB for $M = 4$. The noise performance of all three classes of naturally distant signals converges, as the number of symbols in the set increases. The most commonly encountered simplex set is BPSK; figure 2.1d illustrates the tetrahedral constellation diagram of a four point simplex set.

2.2. Bandwidth Considerations

Real channels are usually either bandwidth limited, or noise limited. That is to say, if the channel is unable to support orthogonal binary signalling at the limiting rate imposed by its theoretical capacity, then it is bandwidth limited; otherwise, it is noise limited. The relation between these two cases is illustrated in figure 2.2, which is a plot of normalised SNR (SNR per bit) versus normalised bit rate (rate per unit bandwidth) and which has the location of some common signal sets marked.

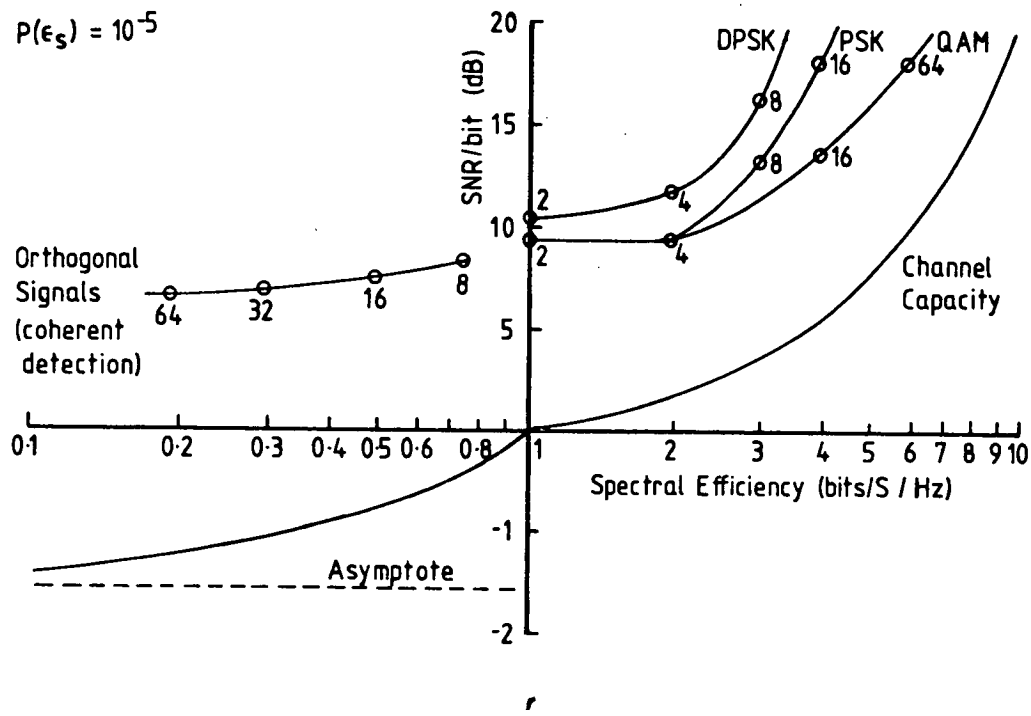


Figure 2.2 Bandwidth/SNR Tradeoff Diagram (After [36 p192]).

It is rare to encounter a real channel which is neither bandwidth nor noise limited, i.e. matched to the signal. Consequently, the initial aim in designing a signal set is either to compress or to expand the spectrum, relative to the orthogonal binary set. In many cases, bandwidth restrictions are artificially imposed, in order to implement channel sharing by the use of frequency division multiplex (FDM). This is not always the most efficient technique, alternatives being time division multiplex (TDM) as used on some multi-access satellite channels and spread-spectrum approaches such as code division multiplex. These alternatives and the power-bandwidth tradeoff have been reviewed by Tou and Roy [37].

2.2.1. Compression for Spectral Efficiency

Spectral efficiency, measured in bits/second/Hz, is a commonly employed measure of the effectiveness of a signal set in reducing the bandwidth required to transmit a digital signal. Bandwidth compression is achieved by the use of dense, non-orthogonal signal sets, which use each symbol to represent several data bits, exploiting the power-bandwidth tradeoff [38]. Since the bandwidth of a signal is governed primarily by the signalling rate, making each symbol represent more than one bit per dimension increases the data rate, without increasing the bandwidth requirement. In bandwidth limited cases, noise is not a major limiting factor in the channel, therefore, the attendant reduction in distance between signals has little practical effect on the $P(\epsilon)$.

The telephone channel is an example of the bandwidth limited case. Having been engineered to provide an SNR of 20 dB to 30 dB over a bandwidth of 3 kHz, it has an excess SNR of about 13 dB over that required for satisfactory orthogonal signalling and an estimated capacity of approximately 25,000 bits per second. Many terrestrial channels tend to be bandwidth limited, either by the need to share a common medium amongst several users by using frequency channelisation, or because the implementation technology makes it the most practicable proposition. Noise limited channels arise chiefly where one end of a link is mobile and consequently has a very limited power budget.

2.2.2. Expansion for Noise Immunity

Channels which are noise limited have an excess of available bandwidth. In order to utilise this bandwidth, the data may be algebraically coded to increase the redundancy of the signal. The modem must then transmit more data (but not *information*) in a given time, in order to maintain the decoded data rate at the output of the receiving modem. If the modem maintains the use of a naturally distant signalling format, this may only be achieved by increasing the signalling rate and hence the bandwidth of the line signal. Consequently, the dimensionality of the signal space used to transmit each information bit is increased and if the algebraic coding has been designed optimally, this produces a greater minimum distance between adjacent codewords, for the same mean energy per information bit. As an illustration of how increasing the dimensionality of a signal set can increase its minimum distance, without requiring an increase in mean symbol energy,

consider the examples of figure 2.1, where the minimum distance increases from 0.89 to 1.63 (equivalent to a 5.2dB change in SNR) as the dimensionality per bit is increased from 0.5 to 1.5.

The four dimensional orthogonal constellation in figure 2.1d has the same minimum distance (1.41) as the bi-orthogonal arrangement (figure 2.1b), which is worse than the three dimensional simplex (1.63) (figure 2.1c). This is not surprising, as the orthogonal arrangement does not make best use of the four dimensions and improved performance can be obtained by use of a four dimensional simplex. On a cautionary note, while the minimum distance of a constellation is important in determining the $P(\epsilon)$, particularly at high SNR, it is not the only factor and two constellations with the same minimum distance do not necessarily have the same $P(\epsilon)$ when all distances are taken into account.

As an alternative to naturally occurring distant signals, such as BPSK or QPSK, others may be synthesised from signals which do not, themselves, necessarily possess good distance properties. An example of this approach is the N-orthogonal block coding scheme [39], in which each bit of the input data generates several consecutive channel bits, or *chips*. These chips are used to construct a limited set of signals, all of which are orthogonal, even though the individual chips may not be orthogonal. To gain the ultimate performance from this type of signal set, the receiver must make a decision about which symbol was received by examining the line signal over the duration of the entire symbol.

In practice, many receivers for this type of signal set take decisions on the identity of each chip and then send this (binary quantised) information on to a decoder, which subsequently identifies the symbol. This type of *hard-decision* detection, where each chip is uniquely identified by the demodulator, results in a performance loss approaching 2dB at low SNR, because information is lost in the process of identifying each chip independently of the others. In essence, this type of signal design which assumes hard decisions, is equivalent to straightforward algebraic encoding of the digital data and highlights the similarity between ECC and channel symbol coding.

A better procedure is to make *soft-decisions* on each chip. In this case, the output from the demodulator is a weighted indication of the likely identity of each chip. Rather than a binary quantised indication that a particular chip was, or was not received, the demodulator generates signals, typically quantised to four or eight levels, indicating the *a*

posteriori estimate of the probable identity of each chip. Thus, an uncorrupted chip will be definitely identified by the demodulator, while a disturbed chip will result in a set of more and less likely candidates. The final decision is taken, implicitly, by the decoder, when it selects the symbol represented by the ensemble of chips.

This time-domain approach of constructing symbols from consecutive chips expands the bandwidth of the line signal by increasing the basic signalling rate, thus requiring shorter pulses to be transmitted. An alternative, frequency domain, approach uses multiple carriers, separated in frequency so that they are orthogonal. Each carrier is modulated by a chip stream, derived by algebraic encoding of the data stream, demodulation being subject to the same restrictions as for the time-domain case above. This *parallel tone* implementation of a constellation may also be used directly, without coding the bit stream into chips and is commonly encountered in applications where the channel is highly unstable, such as one involving ionospheric propagation [40].

An example of a modem which combines both time and frequency domain approaches is given by Ralphs [41]. In this modem, seven bits of input data are combined with one bit of supervisory data, to give one digital symbol to be transmitted. This digital symbol is then represented by a channel symbol comprised of two chips, each chip being signaled by a pulse from one of twelve carriers. There are restrictions on the possible combination of chips, so that not all sequences are possible.

2.3. Signal Dimensionality

Shannon's capacity theorem is based on the assumption of a signal set of infinite dimensionality, or time-bandwidth product. Practical signalling techniques, which use a finite alphabet to transmit messages from a potentially infinite set, have their dimensionality limited by the maximum permissible delay for short messages and by the capacity of the transmitter and receiver to store sections of long messages. Consequently, the statistics of the noise as seen by a receiver, being based on a finite sample set, vary from message block to message block. This variability condemns any finite coding scheme to be too powerful for some message blocks, resulting in wasted capacity and too weak for others, resulting in errors. Although this means that messages will always have a non-zero $P(\epsilon)$, it brings the consolation that there is also a non-minimal probability of correct reception for all cases where some transmitted energy

reaches the receiver, albeit with a very low value which is unlikely to be useful in most applications. Specifically, when the received energy per bit to noise spectral density ratio (E_b/N_o) falls below -1.6dB, a practical transmission system will perform better than the Shannon limit, which predicts the worst possible bit error probability of 0.5 below this value. For BPSK signalling in AWGN, the bit error probability at -1.6dB E_b/N_o is 0.12.

The behaviour of error probability statistics as the dimensionality of a signal set is increased in the presence of additive noise and particularly Shannon's limiting case of perfect transmission, can be explained by the *sphere hardening* argument. If the noise in a channel is additive and uncorrelated with the signal, then it may be modelled as a set of vectors, having a variable uniformly distributed phase and a variable magnitude, centred on the points of the constellation. The magnitude distribution of these vectors as a function of time, depends on the amplitude distribution of the additive noise, being of poisson form for gaussian noise. The strength of the noise may be indicated on the constellation diagram by a circle or sphere of appropriate dimension, marking the mean radius of the noise vector. The normalised length of this radius is a function of the noise density and is independent of the number of dimensions of the constellation diagram. However, the variance of the square of the mean radius is a linearly decreasing function of the dimensionality, so that as the dimensionality increases, the initially fuzzy shell described by the tip of the noise vector, condenses to an increasingly well defined surface. This behaviour leads to the use of the graphic term sphere hardening [42].

Sphere hardening is a geometric interpretation of Shannon's capacity theorem. For signal sets of low dimensionality, the circle or sphere is a poor approximation to the extent of influence of the noise, but as the number of dimensions is increased, corresponding to the use of longer messages and more bandwidth, the variability of the radius approaches zero. Thus, in the limit, determination of the optimum signal set is the classic sphere packing problem†. Voids between the noise spheres represent under-utilisation of the channel, whereas intersecting spheres mean that errors are certain and the $P(\epsilon)$ rises rapidly as they merge. The sphere hardening phenomenon can be shown to be a property of a very general group of noise types.

---

† The sphere packing problem has been intensively studied by mathematicians. The best results known to date are tabulated by Sloane [43].

When the noise is correlated with the signal, or not purely additive, the boundary described by the mean radius of the noise vector is not spherical, but may still permit of a reasonable description in some cases. Foschini et al. [44] have derived a contour of constant distance for a combination of phase and additive noise in two dimensions, the result being somewhat banana shaped.

The performance of a constellation is ultimately bounded by the capacity of the channel, so that continued extension of the signal set dimensionality produces ever diminishing returns [45]. A general guide for the dimensionality of a signal set, is to allow two signals per dimension. This corresponds to the use of bi-orthogonal signals, which may be natural or synthesised, according as the bandwidth of the channel permits. Some of the best coded signalling schemes approach to within 6 dB of the theoretical capacity of the target channel and perform about 5 dB better than the best uncoded ones [46].

2.4. Characteristics of the Constellation Diagram

That points on a constellation diagram correspond to a (pulse) waveform representing a digital word, has already been explained in the previous chapter. Because it is usual to shape a symbol pulse prior to transmission and because real channels do not have unlimited bandwidth, these points really only represent the symbols at the sampling instant in the receiver. Between sampling instants, the transition from one symbol to another describes a locus through signal space. Such loci may be seen in some of the oscillographs presented in [47].

If the modulation process is linear, these loci are straight lines. If, however, the modulation process is non-linear, as for FSK, the loci are curved and the expansion of effective distance between signal points which this curvature represents, corresponds to a companding effect. This produces anomalous behaviour in noise, expressed in the form of noise suppression at high SNR and a threshold effect at low SNR, where the noise resistance of the signal set collapses suddenly. More detail on these *twisted modulation* schemes may be found in [42]; this chapter will only be concerned with the properties of constellations implemented via linear modulation processes.

2.4.1. Radial Point Distribution

The optimum point distribution depends on the nature of the power limitation imposed by the channel. If the channel is mean power limited and AWGN is the primary

interference, the distribution of symbol magnitudes should be gaussian. For a peak power limitation, the distribution should be a linearly increasing function, up to the limiting power level. These are rough criteria for the evaluation of a constellation and are based on the principle that the statistical properties of the signal should be similar to those of the channel. Clearly, the difference will not be pronounced for small constellations of reasonable design, but for those with more than sixteen symbols, the mean to peak power ratio of competing designs may be several decibels. Decisions based on $P(\epsilon)$ comparisons are then strongly affected by the choice of either peak or mean signal power as a parameter.

2.4.2. Decision Regions

Just as the transmitted symbols in a signal set may be represented by points on the constellation diagram, so the regions in signal space which the demodulator associates with each received symbol, may be delimited on the same diagram by lines or surfaces, the *decision boundaries*. The optimum hard-decision demodulator chooses that symbol in the signalling alphabet which is closest, in signal space, to the received waveform. Therefore, its decision boundaries perpendicularly bisect the set of lines joining all nearest neighbours in the constellation, forming closed polytopes around the central region of points and open ones around peripheral points. Practical limitations of the channel will usually provide some bounding on the open decision regions.

If the signal set has a canonical structure, optimum decision boundaries may be relatively simple to implement, otherwise it may be worthwhile to consider a sub-optimum set. An example of such a decision region structure is the use of independent decisions on phase and magnitude to identify symbols. Some early investigations of APK signals [47,48] restricted the points of the constellation to lie on concentric circles and a limited set of radii, partly to facilitate signal generation and detection.

The decision regions of soft decision decoders may likewise be shown on a constellation diagram of the channel *symbols*. The representation of the decision regions of the *chips* on diagram(s) of the sub-constellation(s)† is more difficult to visualise, as it involves a larger number of polytopes, each carrying a representation of the chip weightings to be

† There is no theoretical reason to restrict all chips in a symbol to use the same sub-constellation [49].

inferred from it.

2.4.3. Manipulating the Constellation Diagram

The variation of some characteristics of a signal can be related directly to quite simple transformations of the constellation diagram. These may be used to compare similar constellations and to optimise tentative designs.

2.4.3.1. Translation

Translating a constellation in signal space leaves the relative distance of symbols unchanged and thus does not affect the performance of the constellation in additive noise. The main effect of translation is to move the centroid of the constellation and hence change the mean power of the line signal. When the centroid is located at the origin and all symbols are equiprobable, the power of the line signal is minimised and the resulting suppressed carrier signal has maximum power efficiency. Since the carrier conveys no information, its loss does not affect the accuracy of transmission. The only effect of its elimination is to make the task of the receiver more complicated, in those cases where a coherent reference is required for demodulation. The subject of carrier recovery is covered in later chapters; at the moment it is sufficient to note that the carrier can be reconstructed from information in the sidebands. In some implementations of a constellation, a low level pilot carrier may be transmitted along with the sidebands. The position of the centroid then defines the amplitude and phase of this pilot.

Translation of a constellation will affect its performance in the face of multiplicative noise and channel non-linearity. Although this point should be borne in mind, it is not of great practical significance, since translation invariance is principally employed as a tool when iteratively optimising a constellation (§2.5); it allows the position of a single point to be changed, followed by a translation of the whole constellation to restore carrier suppression.

2.4.3.2. Rotation

Assuming that the noise has identical properties in all dimensions of the signal space, rotation of the constellation about the origin has no effect other than to change the arbitrary reference phase of the signal set. An anisotropic noise distribution is unlikely to be encountered with a two dimensional common carrier frequency signal, unless

certain types of coherent interference are a major influence in the channel. Even in that case, if the constellation has a high degree of rotational symmetry, the variation in performance in-between congruences may be expected to be small. For constellations of high dimensionality, the possibility of an anisotropic noise distribution may be worth considering.

2.4.3.3. Radial Expansion

Radial expansion, or explosion, of the constellation corresponds to increasing the power of the line signal, without affecting the relative energy of points in the constellation. Obviously, this affects the performance of the constellation in additive noise, as all points are made relatively more distant. This property and its inverse, contraction, are used to renormalise a constellation during iterative optimisation, which might otherwise lead to an endlessly expanding solution. The renormalisation may be used to maintain either a peak or a mean power limit, as appropriate for the application.

2.4.3.4. Symmetries

Many practical constellations have a high degree of symmetry. That is to say, if each point is given a unique label, the constellation diagram may be rotated and reflected in many ways, to produce a similar constellation, differing from the original only in the position of labels attached to the points. Symmetry in a constellation brings both advantages and disadvantages. Of the disadvantages, decoding ambiguity is the most frequently encountered. Since the receiver will not usually have an absolute phase reference, it is not possible to decide which symbol, of a set of isomorphically corresponding symbols, has been received. Differential encoding is a common solution to this ambiguity problem, another alternative being the use of a known initial sequence (preamble) to establish an absolute reference at the receiver. Of the advantages brought by symmetry, simplification of the transmitter and receiver structures is very useful in a practical sense. Symmetry is useful in prediction of signal set performance, where the statistics of the transmitted symbol stream are uncertain. Most analyses assume an equiprobable occurrence of symbols, in order to compute a value of average probability of symbol error ($P(\epsilon_s)$) over the entire set. If a constellation is perfectly symmetric, i.e. all symbols have equal $P(\epsilon_s)$, then its performance is unaffected by the statistics of the symbol stream. On the other hand, the true performance bound for other

constellations is the $P(\epsilon_s)$ of the worst case symbol, with the actual performance being somewhat pattern sensitive.

2.5. Design Approaches to the Optimum Signal Set

As well as producing designs for specific applications, there has been much attention paid to the problem of finding globally optimum sets, both for the AWGN channel and others. The structure of an optimum signal constellation depends on many influences and there is no single structure which is always optimum. However, it can be claimed that a regular simplex structure is, or forms a subset of, the optimum structure for a large number of cases [26]. An obvious exception is the design of signals for non-coherent channels, where the lack of coherence, which precludes the use of negative cross-correlations, may be due to the properties of the channel, or to the design of the modem. In considering the design problem for such channels, Scholtz and Weber [50] have shown that the orthogonal structure is locally optimum at all SNR, when all signals have equal energy and there is no bandwidth restriction. Another exception occurs when it is infeasible to generate a constellation with sufficient dimensionality to contain a simplex or orthogonal structure. For these applications, the optimum design is strongly influenced by the SNR, as discussed in the comparison of two-dimensional constellations (*v.i.*).

The references just cited [26,50], are examples of one approach to signal design, namely the search by analytic modelling. This approach is usually directed toward producing generalised results for channels with few restrictions, which may then be used to identify critical parameters in a more specialised design. The results often include mathematical expressions for $P(\epsilon)$ which may be exact, asymptotically exact, or bounds. As well as being useful in their own right, such expressions are often used in computationally aided search approaches, but in all cases, if the expression used is asymptotic or a bound, it is important to ensure that it is valid over the range of parameters under investigation.

An approach which bridges the gap between purely theoretical considerations and the problems of implementation is exemplified by the work of Zetterberg and Brändström [49]. They performed an analysis of the properties of certain crystallographic structures, in a search for good dense signal sets in four dimensions. From their occurrence as naturally selected packing forms, these structures may be expected to have acceptable

distance properties. In addition, their regular nature suggests the possibility of straightforward implementation.

A third approach is that of the computationally aided search. There are many variations in technique possible here and the capability to include many factors which are difficult to express analytically. This method is probably best suited to the search for specific designs, rather than the production of generalised results, although it may be a fruitful source of material for generalised inductive reasoning. The basic method involves selecting an initial signal point configuration and then perturbing this under the influence of given constraints, while using a suitable technique to monitor the progress of the design, which is iterative and converges (hopefully) to a local optimum. All three elements of the procedure are open to choice. The initial configuration may be randomly selected, or may be based on a design whose performance is known or anticipated. The perturbation may be directed or random, with the constraints involving geometry, power limitations or other factors. Finally, the performance measure may be focussed on a specific type of noise or distortion and may take the form of the evaluation of an analytic expression or bound, or be the result of a numerical finite element analysis.

2.6. Dense Signal Sets in Two Dimensions

This chapter concludes with some examples of two dimensional, sixteen symbol signal sets, which illustrate many of the facets of signal design. The designs are compared principally on the basis of $P(\epsilon_s)$ in AWGN and implementation complexity. Note particularly that in this section *symbol* and not bit error probability curves are plotted, since the bit error performance will vary with the symbol coding and demodulation technique employed. Some pertinent comments on this distinction and other measures such as channel and bandwidth efficiency are made by Viterbi [51], who notes that if the data transmitted by a modem have an implicit symbol structure, e.g. alphanumeric character codes from a terminal, it is the $P(\epsilon)$ of these units of transmission which is of prime relevance, rather than the bit-error probability ($P(\epsilon_b)$). On the other hand, if FEC is to be applied independently of the modem, then the $P(\epsilon_b)$ is the important parameter.

The first set is optimised for minimum $P(\epsilon_s)$ in an AWGN channel at high SNR and represents a bound on the performance of two dimensional constellations. The second is

a PSK design, all symbols having equal energy and being equally spaced at constant magnitude. It has not been used widely, but is included as a representative of simple one parameter modulation techniques and illustrates the waning effectiveness of this class of constellation as the alphabet size increases. The third and fourth examples, each of which has been optimised for different channel and implementation considerations, have been widely employed and show some of the compromises made in signal design. The four constellations are drawn in figures 2.3 to 2.6 and curves displaying their $P(\epsilon_s)$ performance in AWGN are presented in figure 2.7.

2.6.1. The Gaussian Optimum Constellation

The location of points which minimises the $P(\epsilon_s)$ over an AWGN channel, with a mean power limitation, has been calculated in [52]. The method used to determine this constellation employed an asymptotic (high SNR) expression for $P(\epsilon_s)$, which was applied to an initially random distribution of points. Under the control of an average power constrained gradient descent algorithm, this point distribution was perturbed until a local minimum was found. The search was executed using several initial configurations and the resulting local minima (some of which were equivalent) were compared. The authors of [52] suggest that the best local minimum so found is very likely to be equivalent to the (two dimensional) global minimum, but this statement should be considered in the light of the low SNR effects referred to in §2.6.2. The constellation and its optimum decision boundaries are illustrated in figure 2.3. The locations of these points lie close to an equilateral triangular lattice, concordant with the statement in §2.3, that a sphere packing represents the optimum distribution for large signal sets *of unrestricted dimensionality*, when channel disturbances are uncorrelated and additive. A constellation based on an exactly triangular lattice has been studied by Simon and Smith [53] and Thomas et. al. [54] and trials of a prototype modem using a constellation of twenty-four points on a sparse triangular lattice are reported by Yanagidaira [55].

An obvious drawback of the optimum gaussian constellation is its lack of symmetry, which makes implementation awkward. While it would be feasible, using digital processing, to construct a modem based on this constellation, it is doubtful that the reduced $P(\epsilon)$, when compared to competing designs such as 16QAM, would justify the complexity. However, it is a useful marker of the performance which other two dimensional constellations might be expected to approach. The decision boundaries, used

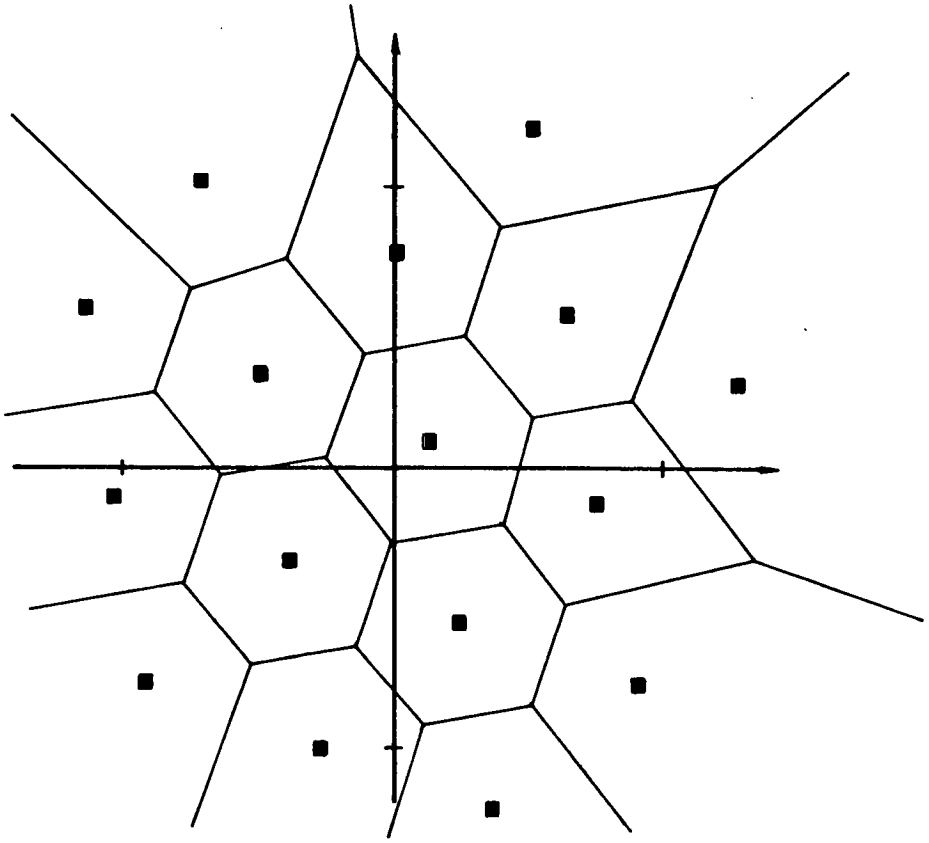


Figure 2.3 Gaussian Optimum Constellation

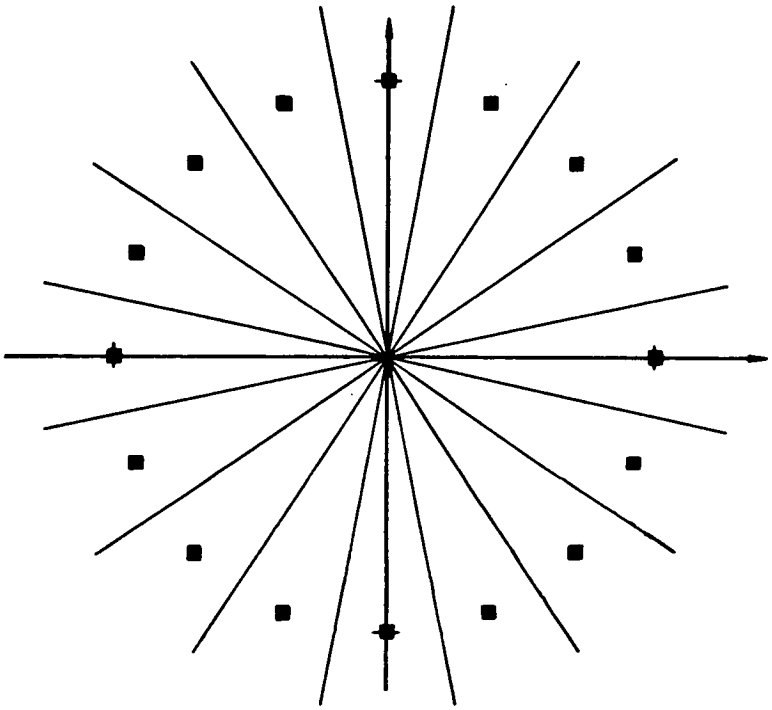


Figure 2.4 Sixteen Symbol PSK

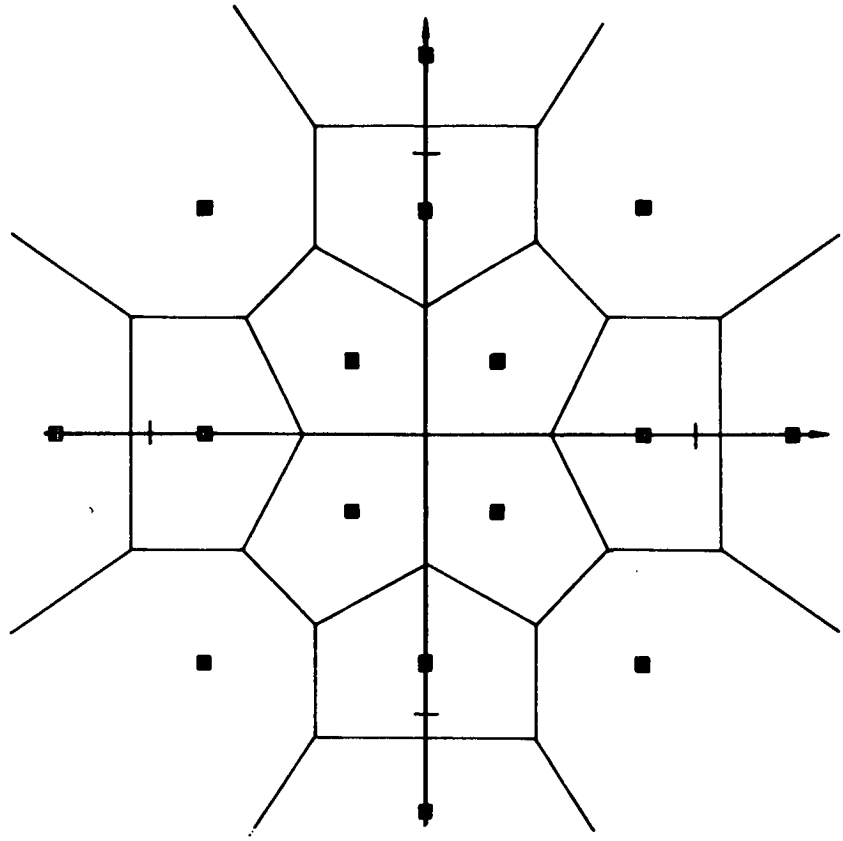


Figure 2.5 CCITT V.29 Recommendation

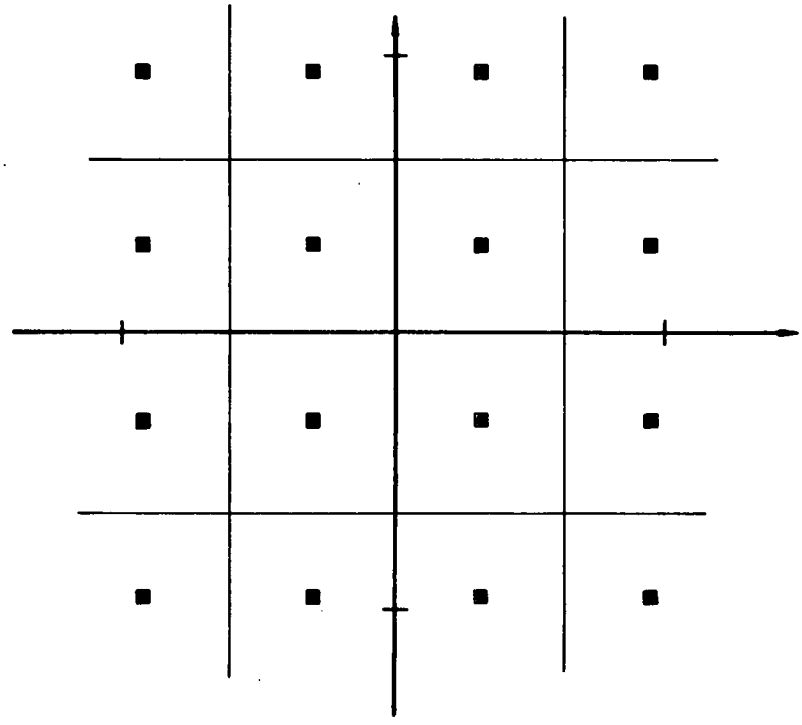


Figure 2.6 Sixteen Symbol QAM (V.29*bis*)

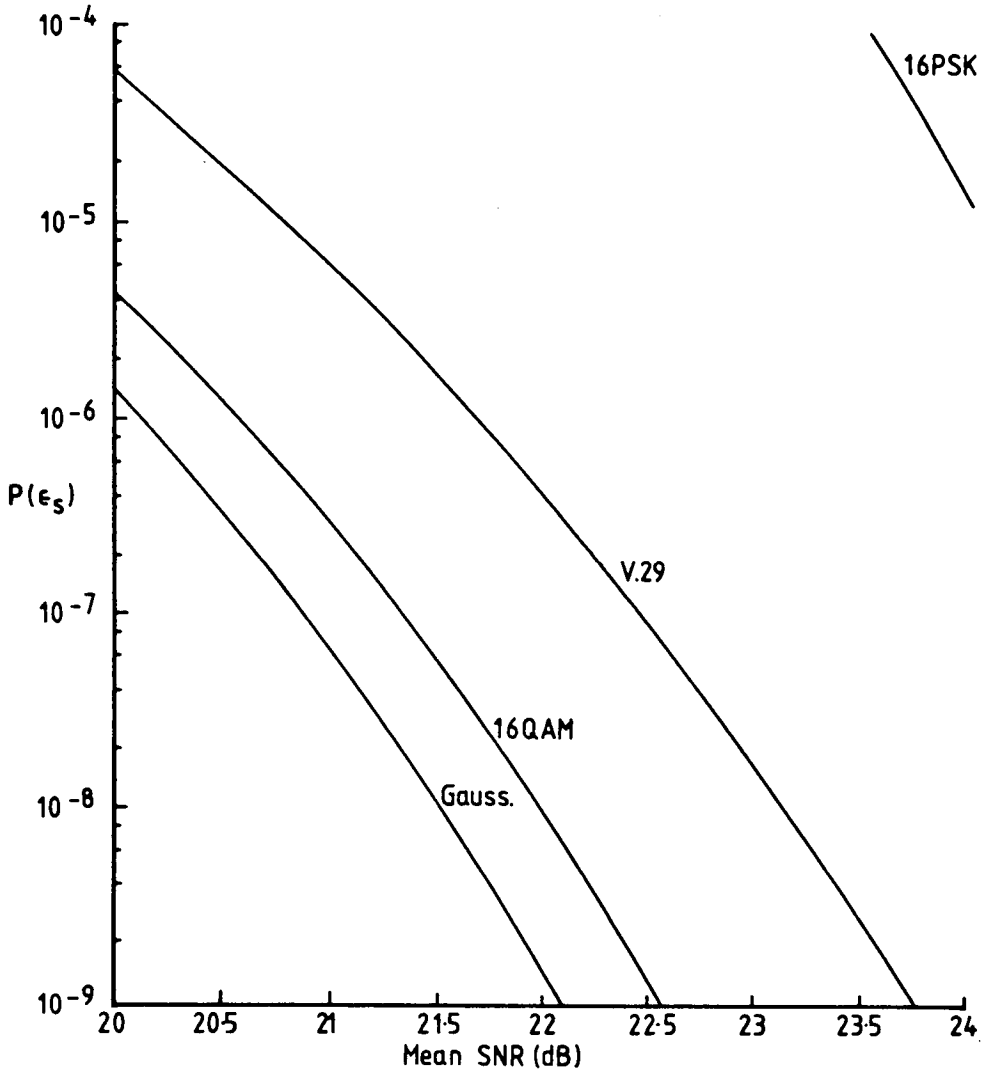


Figure 2.7 $P(\epsilon_s)$ Comparison of figures 2.3-2.6 in AWGN

with hard decision decoding, are based on hexagons, thus it is not possible to implement a pure grey code. The error penalty this imposes at moderate to low error probabilities may eliminate the advantage over 16QAM, which does have a pure grey code. It is reported in [53] that the grey code penalty factor is 1.33, a value which reduces the SNR margin over 16QAM from 0.33 dB to 0.22 dB at a $P(\epsilon_s)$ of 10^{-5}. For certain numbers of symbols, in particular seven, nineteen and thirty-seven [52,53], the gaussian optimum constellation has perfect rotational symmetry, making implementation more tractable.

2.6.2. The 16PSK Constellation

The 16PSK constellation, shown in figure 2.4, is a natural progression from the 4PSK and 8PSK designs commonly used in microwave digital radio. Unlike the other designs considered in this chapter, it is a one parameter modulation scheme. Implementation is very simple, both for the modulator/demodulator and carrier recovery subsystems and a pure grey code is easily arranged as each symbol has only two nearest neighbours. Being a constant envelope format, 16PSK is little affected by amplifier non-linearity and this gives it an advantage of several decibels in some applications [32]. Its major disadvantage is a rather poor $P(\epsilon)$ performance at high SNR, when compared to other designs (figure. 2.7). It is particularly susceptible to the effects of phase noise and ISI, because of the unfavourable distribution of symbols in signal space. The spacing allows great latitude for amplitude disturbances, but is very sensitive to rotational disturbance of the signal points. 16PSK has not found much practical application, usually being rejected in favour of 16QAM, or constellations generated by partial response techniques, which also offer a competitive alternative to 8PSK.

The potential performance of 16PSK at low SNR appears interesting. As was pointed out in the analysis by Lucky and Hancock [56], designs which result in some bounded decision regions, as is the case with the three other designs considered here, may be expected to suffer large $P(\epsilon_s)$ at low SNR, as the $P(\epsilon)$ for the symbols within such regions approaches unity with falling SNR. Decision regions with unbounded magnitudes are able to maintain an asymptotically constant proportion of perturbed signal points at low SNR, although as the distribution of erroneous symbols also approaches a constant, the ultimate endpoint is simply the worst possible $P(\epsilon_s)$ for N symbols, i.e. $1 - 1/N$. Viterbi and Stiffler [57] have studied the use of high order PSK signals in N-orthogonal codes [39] with a sub-optimal receiver and their results suggest that the density of the PSK set used should increase as the SNR is reduced.

2.6.3. The CCITT V.29 Constellation

In contrast to 16PSK, the V.29 [34] constellation (figure 2.5) is very resilient in the face of angular disturbances. This resistance to phase noise is gained at the expense of its performance in additive noise. In the light of its specific application to telephone channels, which often have an SNR of more than 25 dB, but which may introduce large

amounts of ISI and phase noise, such a trade-off is understandable. However, the widespread acceptance of the 16QAM constellation as an alternative, in the form of recommendation V.29*bis* [58], suggests that the significance of these disturbances was over estimated. The V.29 constellation is a clear example of optimisation in favour of hardware simplification. The CCITT recommendation specifies that a conforming modem shall be capable of operating at reduced data rates, using constellations which are composed of a subset of the points in the main constellation. Using the specified bit to symbol mappings, this is readily achieved by padding the bit stream with zeros or modulo-2 sums of other bit pairs. Furthermore, the placement of all signal points on a regularly spaced set of eight radii simplifies the carrier recovery arrangements, particularly when changes of transmission rate are contemplated.

2.6.4. The 16QAM Constellation

The constellation illustrated in figure 2.6 and referred to here as 16QAM, is widely implemented in modems for microwave digital radio systems [32,59] and also telephone channels, where it is recognised as the V.29*bis* standard [58]. 16QAM has good immunity to additive noise, being about 1dB better than V.29 under both peak and mean power limited constraints and maintains an advantage in the face of jitter of up to 2° rms. The regular lattice structure of 16QAM enables straightforward implementation of optimum (minimum distance) decision boundaries. To illustrate the canonical structure of the constellation, it may be analysed as a superposition of two 4PSK signals with 6dB power ratio, with each 4PSK signal viewed as the superposition of two quadrature BPSK signals. Alternatively, as the mnemonic suggests, 16QAM may be constructed from two quadrature 4ASK (suppressed carrier) signals, each of which could be generated as the superposition of two 2ASK signals with a 6dB ratio. There are two corresponding demodulation processes, but it should be noted that the use of one process for modulation and the other for demodulation does not permit the implementation of a pure grey code [59], which is only possible when corresponding processes are used. Because the phases of the signal vectors lack a simple common denominator, carrier recovery is more complex than for 16PSK or V.29. There are several solutions to this problem, some of which are discussed in later chapters.

3. The C1−5−10N Constellation

This chapter reports work done on the design and analysis of a two dimensional, sixteen symbol signal set, for the transmission of binary data at 9600 bps over a voice-band (telephone) circuit.

3.1. Design Specification

The design of this constellation was motivated primarily by the need for a signal set with which to test the operation of a carrier-recovery technique, the signal driven phase-locked loop (SDPLL) reported in chapter five. This required a constellation with regular phase values, both for the operation of the carrier recovery system and also to allow the possibility of producing a scheme for independent phase and magnitude ($I\Phi-M$) detection. Although a sub-optimal technique, $I\Phi-M$ detection was seen as a route to the production of a simple modem implementation. The constellation was required to be competitive with 16QAM and V.29, having a bandwidth efficiency of at least three bits per second per hertz and a comparable $P(\epsilon)$ performance. The $P(\epsilon)$ figures used are based on peak signal to noise ratio (PSNR) as the author is of the opinion that in the area of application considered here, limitation of the peak signal energy is more stringent than mean energy, this limit being primarily due to nonlinearity (gain compression) in the channel. This constraint places constellations with a near unity peak/mean power ratio at an advantage. Specifically, the ratio for the design to be presented (C1−5−10N) is 1.53 dB, which may be compared with the figure of 2.55 dB for 16QAM. This difference should be borne in mind when considering purely mean power limited applications.

The design chosen is referred to here as C1−5−10N and is similar to that referred to in the literature as "(1,5,10)", herein referred to as C1−5−10S. In the nomenclature chosen, the prefix C indicates a constellation based on concentric circles, the hyphenated number sequence gives the number of signal points per circle, from the centre outwards, while the suffices S and N indicate a staggered and non-staggered alignment of points in

adjacent rings, respectively. Reference to figures 3.1 and 3.2, which show the two constellations and their minimum distance and $I\Phi-M$ decision boundaries, will clarify the difference between them.

3.2. Existing Published Work

In the course of this investigation, which concluded with the adoption of the C1−5−10N constellation, believed to be previously unreported, the author found four published references to investigations of the C1−5−10S constellation.

The most comprehensive of these is by Foschini et al. [44], who made a detailed comparison of C1−5−10S with the V.29 (referred to as "(4,90°)") and 16QAM designs. The study was done for both mean and peak power constraints† and considered the combined effects of both AWGN and phase jitter. The phase jitter was modelled by substituting an approximate non-euclidean distance measure into a commonly used asymptotic (high SNR) approximation for $P(\epsilon_s)$, which relates $P(\epsilon)$ for a given symbol to the minimum distance from that symbol to the decision boundary. They conclude that C1−5−10S is capable of good performance under both peak and average power constraints and in the presence of moderate (less than 1.5°) phase jitter. A second contemporaneous paper by Foschini et al. [52] concentrates on the technique of computational search and optimisation used in their design and analysis of constellations.

Simon and Smith [53] concentrated primarily on constellations with hexagonal decision boundaries, comparing them with 16QAM, C1−5−10S and C8−8S. Rather than using asymptotic expressions or bounds to estimate $P(\epsilon)$, they derive an expression for the error probability across a line segment of the decision boundary and then solve this numerically, for each segment of each boundary. A normalised average is then taken, to give the mean $P(\epsilon_s)$, from which the $P(\epsilon_b)$ is derived by multiplication with the *grey code penalty*, calculated as the average Hamming distance between adjacent symbols. While this gives an asymptotically exact value of $P(\epsilon_b)$ for those constellations possessing a pure grey code, or those where there is uniform partitioning of error events to all adjacent symbols, it is inaccurate when there is no pure grey code and the partitioning of error probabilities to adjacent symbols is not uniform. In such cases, the

†The values for the mean and peak power of 16QAM in figure 3 of [44] are erroneous, actually being *magnitude* values. However, the correct ratio has been used in deriving the mean and peak $P(\epsilon)$ values elsewhere in the paper.

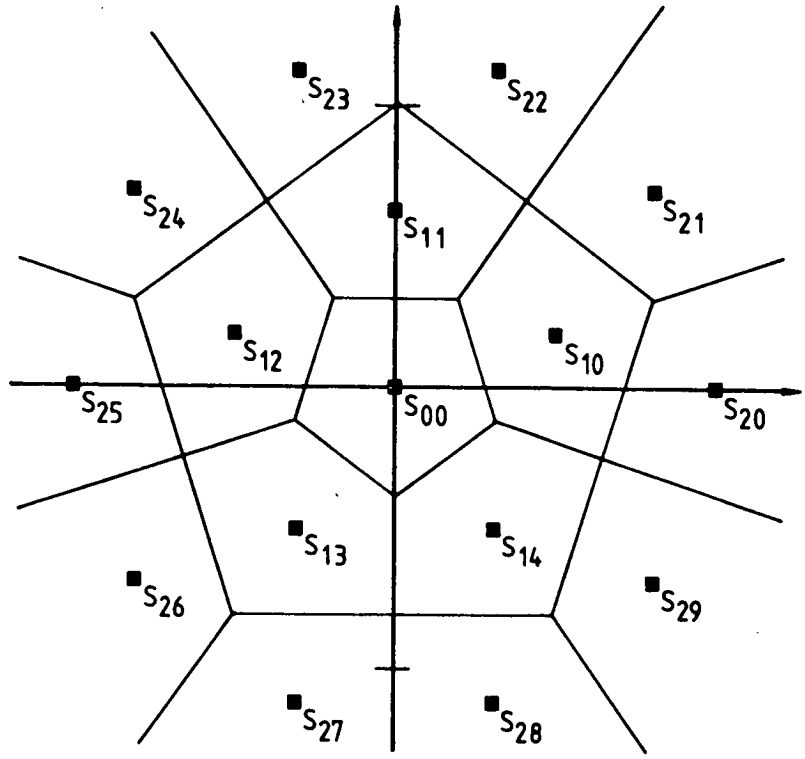


Figure 3.1a C1−5−10S, Minimum Distance Decision Boundaries

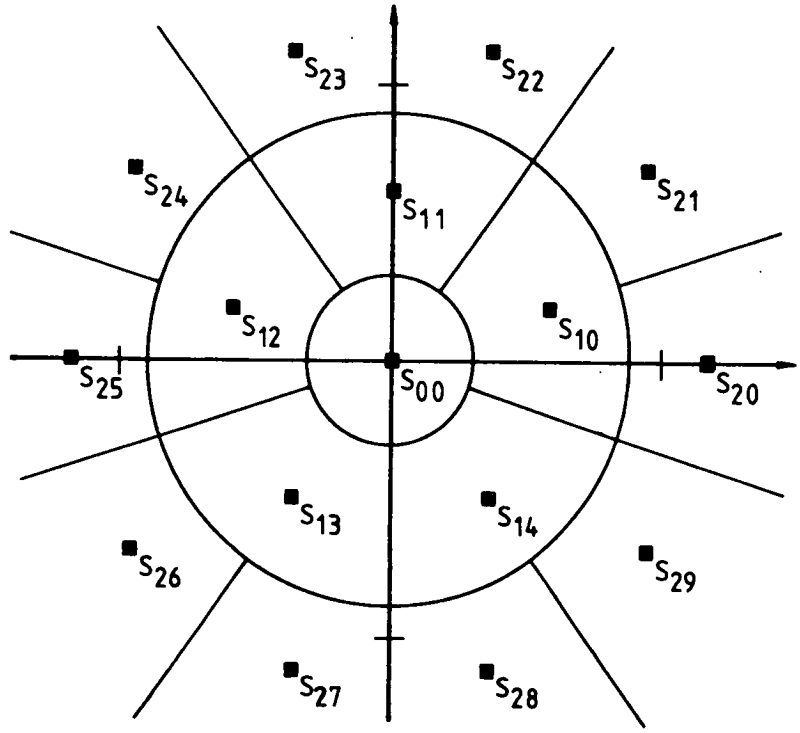


Figure 3.1b C1−5−10S, Independent Phase-Magnitude Decision Boundaries

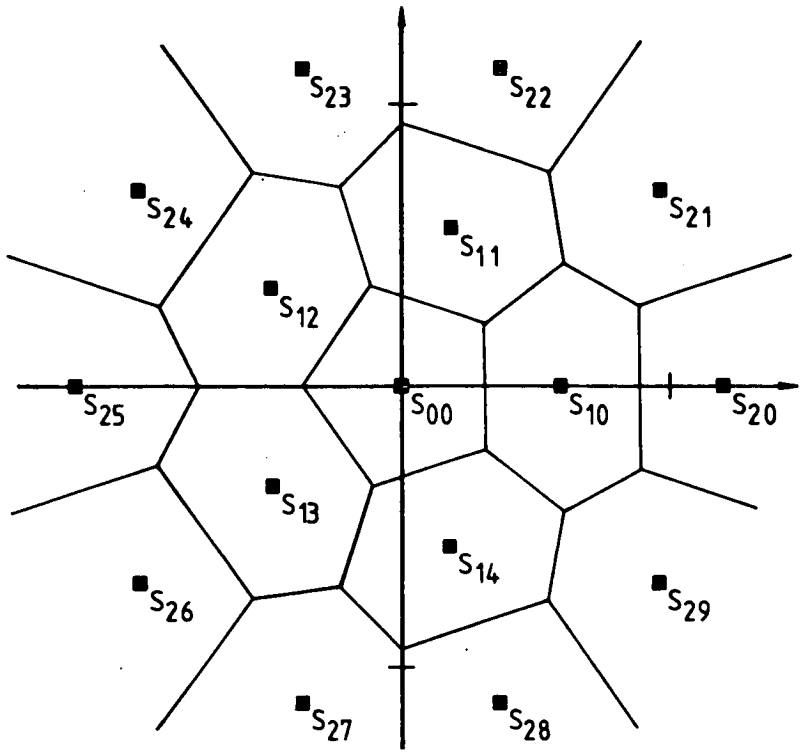


Figure 3.2a C1−5−10N, Minimum Distance Decision Boundaries

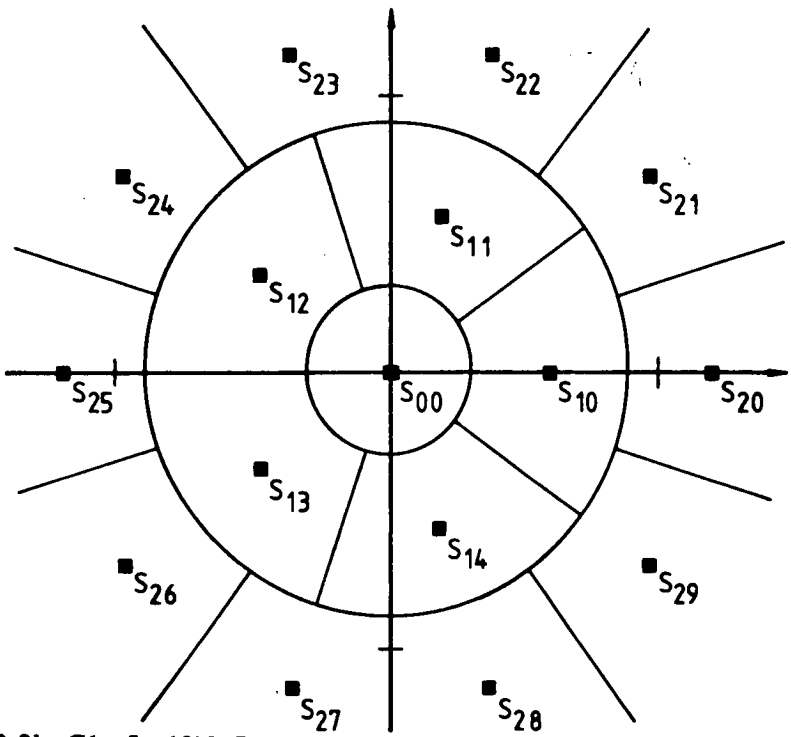


Figure 3.2b C1−5−10N, Independent Phase-Magnitude Decision Boundaries

Hamming distance to an adjacent symbol should be weighted by the proportion of errors resulting in that symbol, when calculating the average grey code penalty.

Simon and Smith conclude in favour of 16QAM, principally on the basis of an assumed relative simplicity of implementation and a $P(\epsilon)$ performance comparison in AWGN. They suppose that the only practical implementation of a C1−5−10S receiver is a maximum-likelihood detector, which they dismiss as excessively complex.

Thomas et al. [54] present twenty-eight constellations with symbol populations covering a range from four to one hundred and twenty eight, of which seven have a population of sixteen symbols. As well as making a $P(\epsilon_s)$ comparison on the basis of peak and mean SNR in AWGN, they devote a large part of the paper to implementation considerations, in particular, the non-linearity effects of TWT amplifiers. The $P(\epsilon_s)$ behaviour is calculated using the same asymptotic formula† as used by Foschini et al., but without including the jitter function.

The results presented in this chapter show that C1−5−10N, while not as good as C1−5−10S in $P(\epsilon)$ performance, is not seriously inferior and may be an alternative to other 16 symbol constellations. C1−5−10N has the advantage that it is amenable to simple demodulation, using the SDPLL and $I\Phi-M$ detection. The reduction in $P(\epsilon)$ performance is not great, being an increase by a factor of three going from the optimum (minimum distance) detection of the S constellation, to the sub-optimum ($I\Phi-M$) detection of the N. Minimum distance detection of the N constellation gives a comparison which is even more favourable, making it attractive for interworking between simple and complex modems, where network topology makes such cost−performance tradeoffs possible.

3.3. The Selection of C1−5−10

Based on existing published data, there were nine candidates for consideration as the constellation to be used for testing the SDPLL. These were: C16 (16PSK), C8−8, C5−11, C4−12, C4−4−4−4 (V.29), C1−5−10, square lattice (16QAM, V.29*bis*), sparse triangular lattice and hexagonal lattice.

---

†There is a typographical error in equation (8) of [54], however, the $P(\epsilon_s)$ values have been calculated with the correct formula.

C16 was rejected immediately, principally for its poor $P(\epsilon_s)$ (figure 2.7). The three constellations based on polygonal lattices were also rejected, because of the irregularity of the phase spacing of their symbols.

The square lattice (16QAM, V.29*bis*) was examined very carefully, before being rejected, as it has several points in its favour. It has good $P(\epsilon_s)$ (figure 2.7) and is a well established standard [58], bringing the considerable advantage of compatibility with existing systems. Work on 16QAM has been reported [60] in which only certain symbols, having regular phase separation, are used for the carrier recovery. This approach was not favoured, as the added complication of symbol gating would detract from and obscure the performance analysis of the SDPLL.

Of the five remaining annular constellations, C4−4−4−4 has the advantage of being an international standard [34]. However, its position as a widely implemented standard is poor, due to its substantial $P(\epsilon_s)$ inferiority to V.29*bis* and also because it is protected by patent.

C8−8 was rejected, as being of distinctly inferior $P(\epsilon_s)$ performance in comparison with the other three, which were all very similar in this respect.

Of the remaining three, the one with best $P(\epsilon_s)$ performance, C5−11, was rejected because there was no suitable alignment between phase positions of the inner and outer ring symbols. C1−5−10 was chosen in preference to C4−12 on the basis of its marginally better $P(\epsilon_s)$ and because its symbols had a common denominator of twenty, as opposed to twenty-four, regularly spaced phase angles. The wisdom of this decision is questioned later on in this chapter.

Having settled on C1−5−10 as a preliminary candidate, the next step was to check the parameters of the constellation and compare its performance with the established constellations, C4−4−4−4 (V.29) and 16QAM (V.29*bis*).

3.4. $P(\epsilon)$ Calculation Procedure

Many previous investigators of constellation performance have used various analytic approximations in deriving an expression for the $P(\epsilon)$, usually commenting that the resulting expression was expected to be accurate at high SNR, but not giving any expected range of applicability or error. Furthermore, the derivation of these expressions often involves the assumption of decision boundaries which are composed of

line segments, which is not the case for the magnitude thresholds of an $I\Phi - M$ detector. In order to gauge the expected accuracy of the commonly employed asymptotic expression:

$$P(\epsilon_s) = \frac{1}{M}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \sum_{\substack{i=1 \\ j\neq i}}^{M}\sum_{j=1}^{M} \frac{\exp\left(-|s_i - s_j|^2 + 4\sigma^2\right)}{|s_i - s_j| + \sigma\sqrt{2}} \tag{3.1}$$

where M is the number of symbols in the constellation, s_i and s_j are sets of vectors to the symbols and σ is the standard deviation of the noise amplitude; it was used to calculate values on the $P(\epsilon_s)$ curve for 16QAM, the results of which were then compared with those from an exact expression [36 p187] which exists for that constellation ($M = 16$):

$$P(\epsilon_s) = 1 - (1 - P_m)^2 \tag{3.2}$$

Where:

$$P_m = \left(1 - \frac{1}{\sqrt{M}}\right) \times erfc\left(\frac{3}{2(M-1)}\gamma_a\right)^{\frac{1}{2}}$$

$$\gamma_a = mean\ SNR$$

A plot of the error of this asymptotic expression as a function of PSNR is shown in figure 3.3, from which it may be seen that (3.1) predicts a $P(\epsilon_s)$ of twice the correct value at 11.3dB PSNR, with a ten percent excess at 19.0dB PSNR. The expected $P(\epsilon_s)$ for a two-dimensional sixteen point constellation at 19.0dB PSNR is of order 10^{-3} to 10^{-2}, below which most applications not using highly redundant FEC coding would not accept the channel as usable. The value of 10% at 19dB PSNR was taken as a reference point for the performance of the computational analysis technique used in this investigation, which was designed to have a total inaccuracy not greater than ten percent over the range 19dB to 23dB PSNR. While it can be argued that calculation to an accuracy of better than a factor of two in the predicted $P(\epsilon)$ of a constellation is of little practical significance, such pedantry is relevant when making comparative judgements of competing candidates. The 23dB limit is based on the expectation of a $P(\epsilon_s)$ of order 10^{-6} in AWGN, with the observation that beyond this, the gaussian nature of the noise

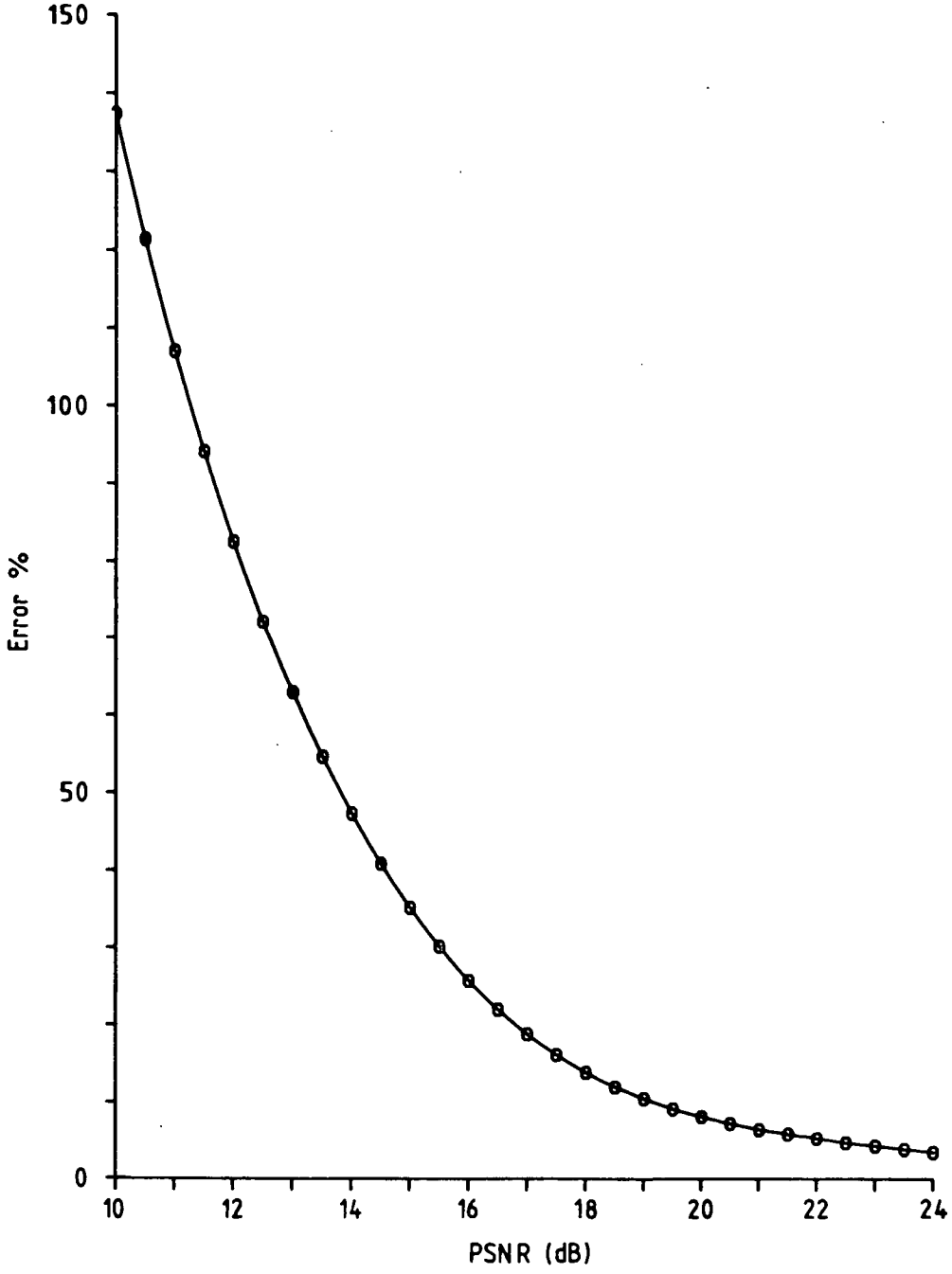


Figure 3.3 Error of the $P(\epsilon_s)$ Expression (3.1) for 16QAM

in any real channel is highly suspect. Furthermore, the reliable (and efficient) attainment of $P(\epsilon_s)$ beyond 10^{-6} usually requires the incorporation of some form of ECC. This results in performance which is principally a function of the coding technique employed, a subject which is beyond the scope of this study.

Because the use of ECC in combination with suitable channel encoding can make efficient use of channels with a high raw $P(\epsilon)$, the relative performance of constellations below a $P(\epsilon_s)$ of 10^{-3} is also of interest. As is illustrated by figure 3.3, the accuracy of (3.1) falls off dramatically as PSNR is reduced. The computational technique to be described may be designed to yield accurate results with no such catastrophic degradation, but limited primarily by the quantisation of the calculated noise integral. This advantage was exploited to determine good grey codes for C1−5−10N, with attention being drawn to the variation in performance of such mapping, particularly at low PSNR. While the use of soft decoding, in conjunction with e.g. Viterbi's algorithm, may be expected to yield better results at low PSNR, the behaviour of hard detection with grey coding is still of interest, particularly where the channel may be subject to quasi-static changes in PSNR; i.e. being generally of good quality, but exhibiting periods of high noise level.

3.4.1. The Numerical Analysis Approach

For any stationary noise distribution, the effect on a signal constellation may be determined by locating the origin of the noise PDF at the position of one symbol of the constellation, then scaling the axes of the PDF appropriately for the ratio of signal to noise power. The proportion of the normalised PDF volume not included in the decision region for the symbol gives the $P(\epsilon_s)$ for that symbol. This process may be repeated for all unique† symbols, a weighted average of which gives the $P(\epsilon_s)$ performance for the entire constellation. In general, it is not possible to derive an analytic expression for the integral of the PDF delimited by a decision boundary. Therefore, suitable algebraic approximations must be used, or the integral must be evaluated numerically.

The numerical approach chosen, hereafter referred to as ERPART, splits the integral up into a large number of small, finite elements, calculating an approximate volume for each of these elements and adding all of these results to give an approximation to the value of

† If, due to symmetry, a group of symbols within a constellation have isomorphic decision regions, the calculation need be done for only one of them.

the integral. By making each finite element sufficiently small, linear approximations to functions may be employed, while still maintaining acceptable accuracy.

The calculation may be further sub-divided into two stages: the evaluation of the PDF as a set of finite elements, followed by the summation of the two groups of elements as delimited by a suitably scaled decision boundary function. The first stage need be done once only for each noise distribution investigated, the results being stored for reference each time the second stage is executed.

An advantage of this approach is that it is easily adaptable to any stationary noise distribution, particularly those obtained as a result of channel measurements, for which no adequate algebraic expression may exist (most of the work reported in this chapter was done before the papers by Middleton [25,28,29] came to the author's attention). Furthermore, it does not rely on any particular property of the decision region boundaries. Finally, ERPART has the advantage of being a straightforward procedure, simple to implement. The computational resources available to the author meant that it was not necessary to program complex algorithms in order to perform the calculations within an acceptable time limit.

3.4.2. Parameters of the Numerical Technique (ERPART)

There are several parameters to be considered in the numerical approach chosen. They are: the computing resources required for the calculation, the radial extent of the noise PDF before truncation, the accuracy of calculation of each element and of the summation of elements and the fineness of quantisation.

3.4.2.1. Machine Resources

All the analysis programs were developed and run on the Edinburgh University central mainframe facility [61]. There was no real financial constraint on machine time, consequently turn-around became the main time criterion and was generally less than two hours for jobs requiring less than three hundred seconds of processor time. There was a maximum file size limit of 2Mbyte, which set a ceiling on the size of the file used for the noise PDF table. This did not turn out to be a real constraint. The programs were written in IMP80, an Algol style language, with the main programming emphasis on reducing development time and not running-time. Therefore, the algorithms employed were deliberately simple, rather than highly efficient. This was valid strategy given the

resources available and the limited short term use of the programs. Continued use of the programs would justify re-writing them for improved efficiency.

3.4.2.2. PDF Truncation

A gaussian PDF, which was the only type used in the study reported here, extends indefinitely and so must be truncated in the analysis. The radius of truncation is determined by the maximum SNR and the required accuracy of the $P(\epsilon)$ at that SNR; the inaccuracy is absolute and its significance diminishes rapidly with falling SNR. There is a secondary consideration, that the number of elements in the resulting finite element table increases as the square of the truncation radius. Preliminary estimates showed that the quantisation error was likely to be the most expensive to minimise and was, therefore, allocated the largest share of the overall 10% target. Since the truncation error is a static offset, it can be compensated for in most situations, so that its principal effect is to limit the minimum resolvable $P(\epsilon_s)$.

The range of PSNR of interest was 13 dB to 23 dB, which gives a minimum $P(\epsilon_s)$ of the order of 10^{-7}. Allowing a worst-case truncation error of one percent and using the equation [62 p383]:

$$\int_b^\infty \frac{r}{\sigma^2} exp\left(\frac{-r^2}{2\sigma^2}\right) dr = exp\left(\frac{-b^2}{2\sigma^2}\right) \qquad (r \geq 0) \qquad (3.3)$$

for the volume under a two-dimensional gaussian bell excluded by truncation at **b** standard deviations (σ), the required radial extent of the table is 6.44σ. Although it is unlikely that the noise in any real channel would follow a gaussian PDF beyond 4σ, this model was adhered to for the reasons given in chapter two. The noise table was based on cartesian co-ordinates, so the radial PDF limit set the length of the x and y axes. The use of polar coordinates would have allowed simplification of the table to a one-dimensional array, requiring very little storage space. However, this would have increased the amount of calculation required to perform the relocation and scaling of the table for a given symbol and SNR.

3.4.2.3. Machine Accuracy

The calculations were performed on an ICL 2976 processor, using double word floating point arithmetic. Each double word is 64 bits long, of which 56 bits are used for the

mantissa, giving over 16 decimal places of precision. The accumulator is 128 bits long and extends the mantissa for intermediate results to 120 bits. The IMP80 language provides access to a set of library procedures based on the ICL supplied library, but modified for IMP80 calling conventions. The accuracy of these library procedures and their methods of implementation are described in [63]. Three procedures were reviewed as potentially useful for computing the values of the finite table elements; they were the complementary error function ICL9CM2CERF, the error function ICL9CM2ERF and the exponential function ICL9CM2EXP. The complementary error and error functions were rejected, due to problems with accuracy over certain ranges of interest and increased complication in programming. By using the exponential function, which had sufficient accuracy for all values of the argument of interest, the program could be readily adapted to the generation of any other PDF which could be expressed analytically.

3.4.2.4. Quantisation Error

Quantisation step size is responsible for most of the inaccuracy in the implementation described here. The strong curvature of the gaussian PDF makes it a candidate for the use of a non-linear quantisation scheme, but this would complicate the use of the resulting table and probably cause a substantial increase in run-time for the analyses, or require the use of an auxiliary table. Preliminary estimates indicated that, with the method of use of the table, a variation in step length by a factor of two only was required for use over the $P(\epsilon_s)$ range 10^{-2} to 10^{-7}. This was considered too small to be worthwhile and so a fixed increment scheme was adopted, which produces maximal quantisation effects at both high and low SNR.

At high SNR, consider the situation illustrated in figure 3.4, which shows a circular decision boundary lying in an annular quantisation step. Given sufficiently fine tessellation for the rectangular grid of noise elements, this is a good approximation to the actual situation with a circular decision boundary and moreover, is the worst case scenario. The assumption of a circular decision boundary is also worst case. Then the quantisation error is maximal when the midline of the annulus **r**, falls just inside the decision boundary **b**, resulting in an erroneously low value for the $P(\epsilon)$.

This presupposes that the mid-point of each square based parallelepiped is used to represent the centroid of that element, which is only correct for special cases. The error

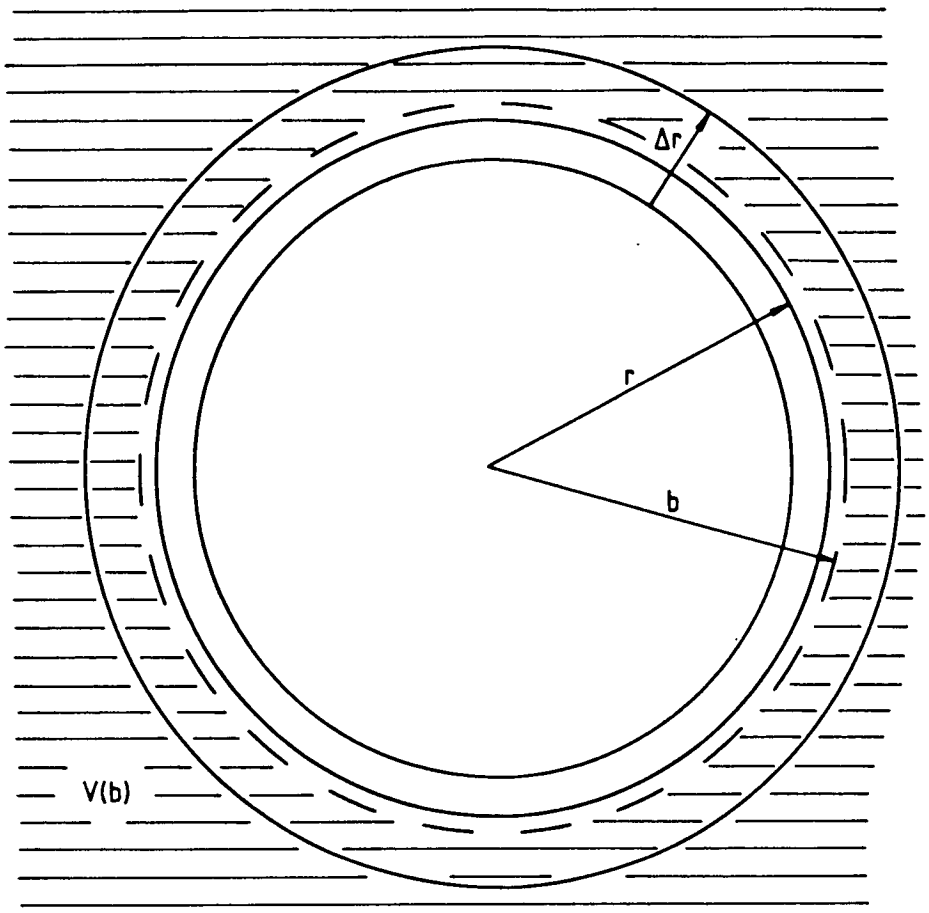


Figure 3.4 Annular Quantisation Error Regions

from this supposition is worst where the noise PDF is most strongly curved, i.e. at high SNR, but is mitigated by the increased number of elements traversed by the decision boundary per decibel increment at high SNR. An initial estimate indicated that this error would be negligible and this is confirmed by the fine structure analysis of the resulting $P(\epsilon)$ curves given later in this section.

Referring to figure 3.4, the volume under the tails of the gaussian bell, excluded by a boundary drawn at radius **b** from the centre of the distribution, is given by [62 p383]:

$$V(b) = exp\left(\frac{-b^2}{2\sigma^2} \right) \tag{3.4}$$

Taking radial quantisation steps of Δr, then at radius r, the worst case quantisation error

(absolute) around boundary **b** is:

$$Q(b,\Delta r) = V(b) - V(r + \Delta r/2) \tag{3.5}$$

$$Q(b,\Delta r) = exp\left(\frac{-b^2}{2\sigma^2} \right) - exp\left(\frac{-(b + \Delta r/2)^2}{2\sigma^2} \right) \qquad (r = b_-) \tag{3.6}$$

$$Q(b,\Delta r) = \left[1 - exp\left(\frac{\Delta^2 r/4 + b\Delta r}{2\sigma^2} \right) \right] \times exp\left(\frac{-(b + \Delta r/2)^2}{2\sigma^2} \right) \tag{3.7}$$

Therefore, the relative error is given by:

$$E(b,\Delta r) = \frac{Q(b,\Delta r)}{exp(-b^2/2\sigma^2)} \tag{3.8}$$

$$E(b,\Delta r) = 1 - exp\left(\frac{b\Delta r + \Delta^2 r/4}{2\sigma^2} \right) \tag{3.9}$$

Whence the required step size for a given relative error is derived as the positive root of the equation:

$$\Delta^2 r + 4b\Delta r - 8\sigma^2 ln(1 - E) = 0 \tag{3.10}$$

$$\Delta r = -2b + 2\left(b^2 + 2\sigma^2 ln(1 - E)\right)^{\frac{1}{2}} \qquad (E < 0) \tag{3.11}$$

The maximum positive quantisation error is derived in like fashion; the corresponding step size equation is:

$$\Delta r = 2b - 2\left(b^2 - 2\sigma^2 ln(E + 1)\right)^{\frac{1}{2}} \qquad (E > 0) \tag{3.12}$$

A decision boundary at a radius of 5.6777σ gives rise to a $P(\epsilon_s)$ of 10^{-7}, corresponding

to a PSNR of approximately 23 dB for the C1−5−10 constellation. Table 3.1 gives a selection of values of E and Δr for this circumstance, along with values for the size and number of steps per half axis of the table (all integer values rounded up).

At r = 5.6777σ, $log[P(\epsilon)] = -7$			
% Error E	Step Δr	6.44σ/Δr	Kbytes
−50	0.14194	46	17
−20	0.06404	101	80
−10	0.03352	193	292
+10	0.03362	192	288
−5	0.01717	375	1099
−2	0.00697	924	6671
−1	0.00350	1838	26393

Table 3.1 Quantisation Error as a Function of Step Length

Clearly, reducing the quantisation error to the same proportion as the truncation error, element error (*v.i.*) and machine error is expensive in table storage. Furthermore, the run time of the program is proportional to the table size; so most of the initial target error allowance was used here, with a table size of 200 elements in 6.44σ being used, to give an expected worst case quantisation error of 9.59% at $P(\epsilon_s) = 10^{-7}$. This worst case would be approached by a rectangular decision boundary which made an unfortunate alignment with the quantisation steps of the table. Nonlinear and non-axially aligned boundary segments result in an average quantisation error approaching zero (there is a static error of $< -0.03\%$ at $P(\epsilon_s) = 10^{-7}$). Because of the quadrantial symmetry of the table, it is only necessary to store 40,000 of the 160,000 elements in the table which, therefore, occupied only 313Kbytes (table 3.1 reflects this).

As an exact result was available for 16QAM (equation 3.2), a constellation which provides a nearly worst-case test of the quantisation error, corresponding sets of results were compared. The comparison indicates a peak error of approximately ±8.5% at 23 db PSNR and ±5% at 19 dB, which is reasonable and compares favourably with the error of the asymptotic expression (3.1) which is +4% at 23 dB and +10.4% at 19 dB PSNR. As

a further test of the influence of quantisation, a fine-grained examination was made of results for the C1−5−10N decision regions around 23 dB PSNR, where the radial position of a boundary moves through one quantisation step in 0.05 dB. Plots of $log[P(\epsilon_s)]$ vs. PSNR for the nine different decision regions of C1−5−10N, over approximately two quantisation steps, are shown in figure 3.5. The most severe effect was seen for S_{20} (figure 3.5e), where the decision boundary most closely approaches the aligned rectangle case. With a linear approximation of the S_{21} result used as a reference, the curve oscillates with a peak deviation of -2.9% +1.6%. Since this worst-case inaccuracy applies only to one symbol, its significance is reduced when calculating average symbol error probabilities.

At very low SNR, the effective quantisation becomes coarse, as few table elements are included within (as opposed to outwith) the boundary. Since this does not take effect until the SNR is well below the range of interest here, it will not be considered further.

3.4.3. Noise Table Generation

3.4.3.1. Calculation of Individual Elements

The value of each table element, which is the volume of a vertical parallelepiped under the gaussian bell, was calculated by sub-dividing each element and calculating the volume of the sub-element as the average height of the vertical edges times the base area, normalised so that the total integral under the PDF was unity. The number of sub-elements required was estimated by examining the volume of an element at the periphery of the distribution, where the curvature of the PDF is strongest, estimating the asymptote of this volume calculation as the number of sub-elements used was increased. A measurement of the time required to calculate this single element was also used, to estimate the time needed for the whole table. It was found that using 4 sub-elements gave a result to within 0.02% for this worst case example, which is essentially perfect. Consequently, this approximation made negligible contribution to the overall error in the $P(\epsilon)$ calculations, for little computational effort. The table generation required 8.5 seconds of processing time.

3.4.3.2. Trimming and Accuracy check

Because the table file is a square array and the PDF, of which it represents one quadrant,

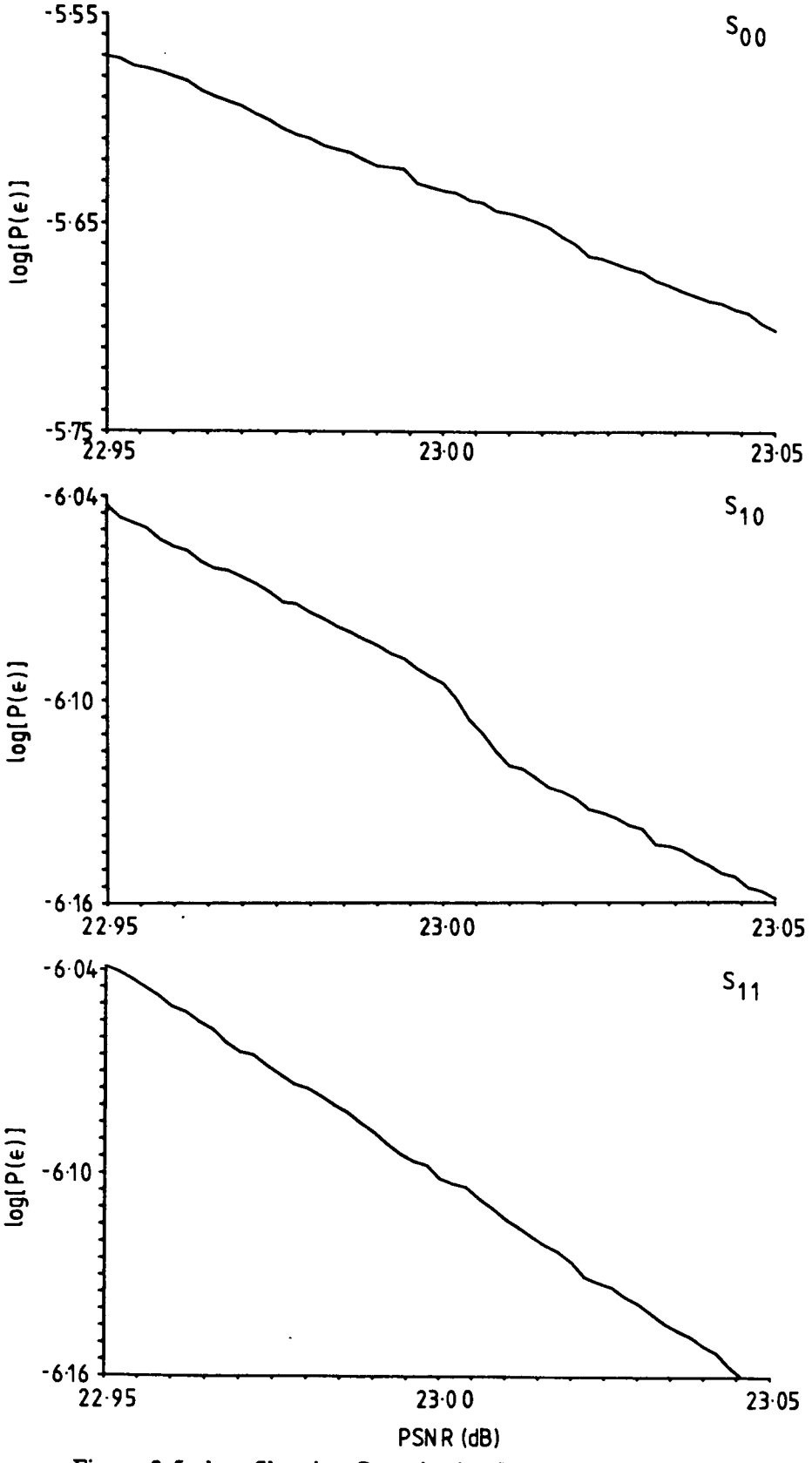


Figure 3.5a,b,c Showing Quantisation Effects for S_{00}, S_{10}, S_{11}

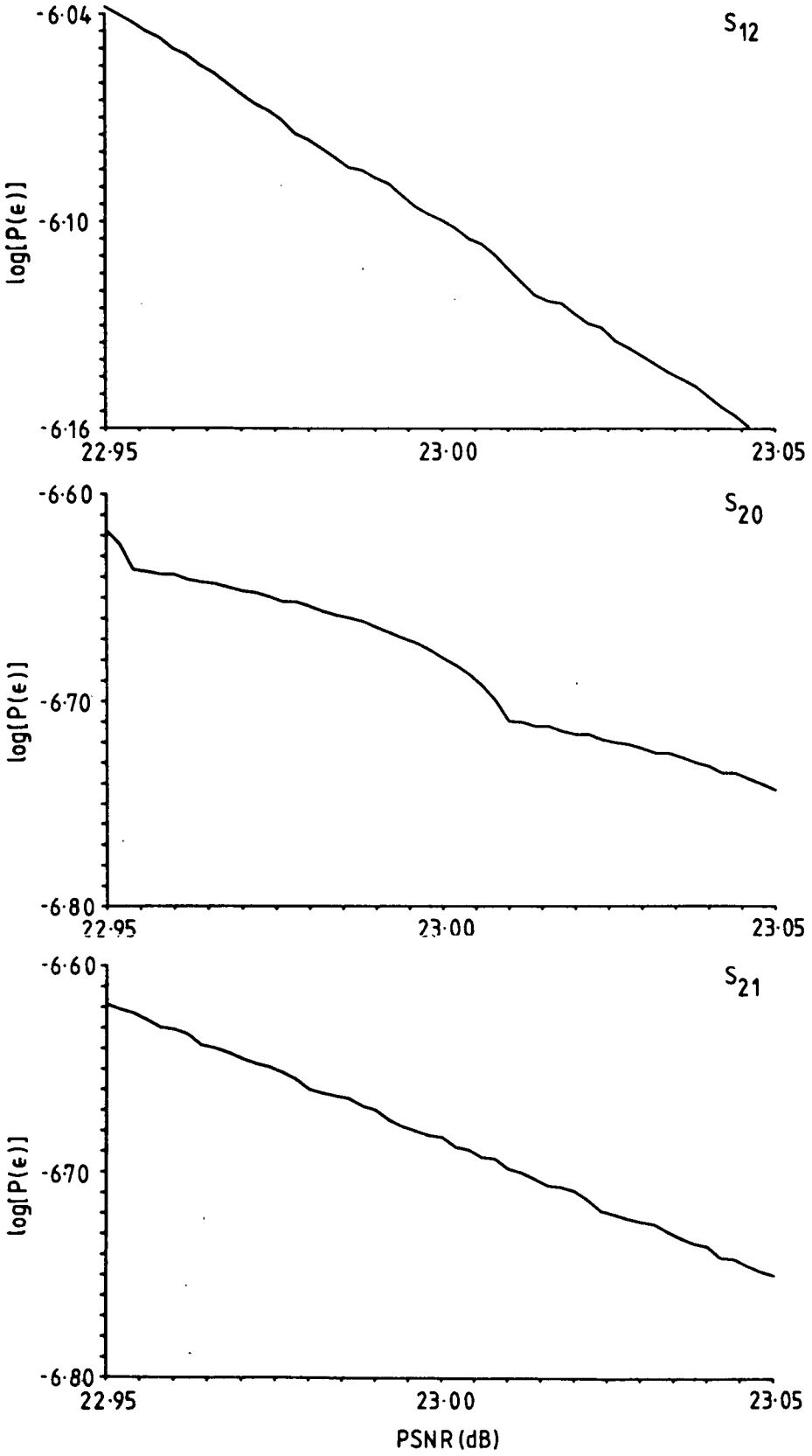


Figure 3.5d,e,f Showing Quantisation Effects for S_{12}, S_{20}, S_{21}

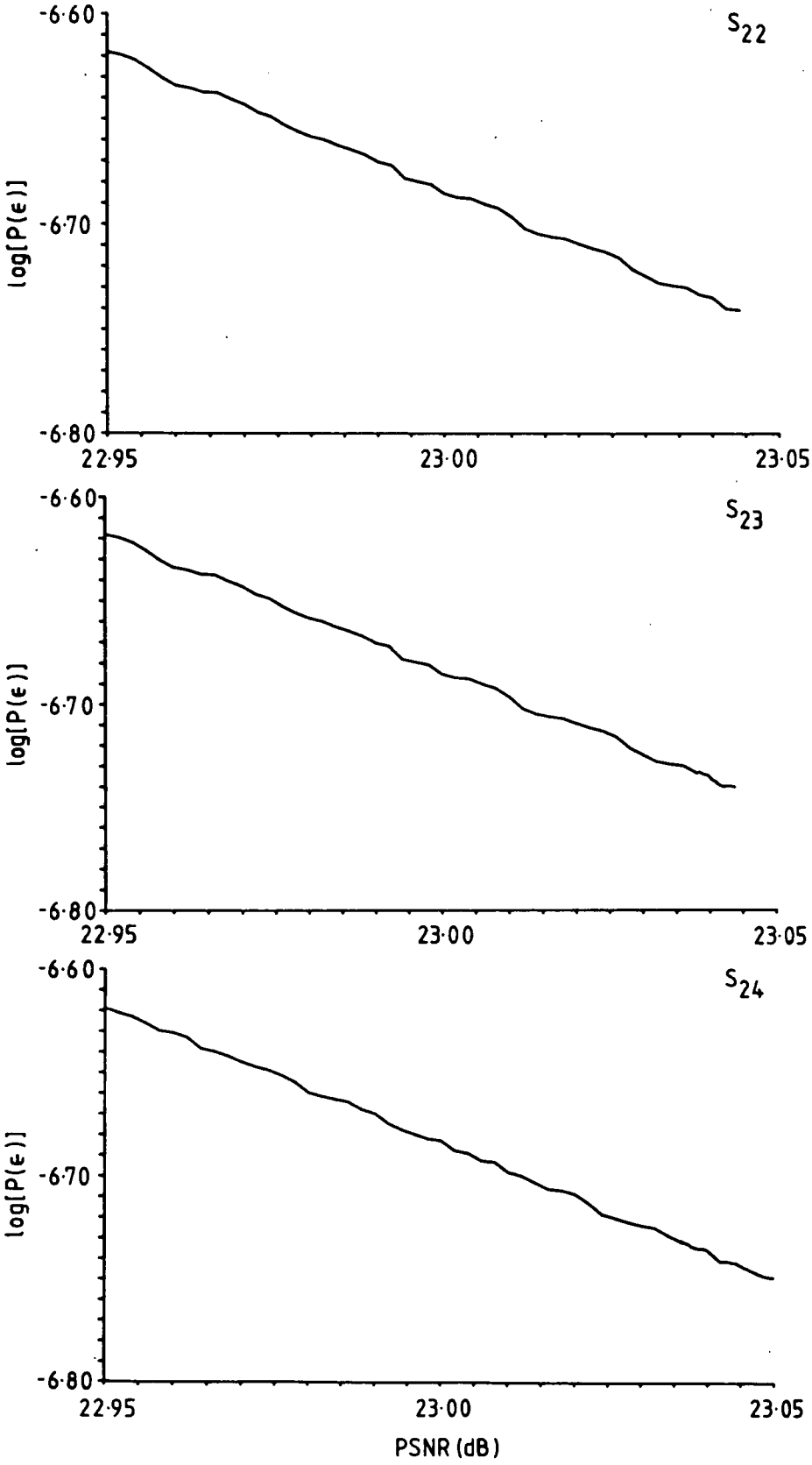


Figure 3.5g,h,i Showing Quantisation Effects for S_{22}, S_{23}, S_{24}

is circularly symmetric, 22% of the elements in the table correspond to regions of the PDF outwith the 6.44σ limit decided in §3.4.2.2. These elements could be left as they are. However, it was noted that a significant reduction in the run-time of the analysis programs could be achieved, simply by setting these extraneous elements to zero and using this information to terminate the accumulation of a row of elements on encountering one with zero value. The extra accuracy which these elements contributed was not sufficient to justify their retention.

As a check on the overall generation of the table, the sum of all the elements in the table was calculated, multiplied by four to give the volume under the whole PDF, subtracted from one and the difference compared with the expected value of 10^{-9}. In order to minimise the effects of roundoff error due to the finite precision of the values, a copy of the table was made, mapped as a one dimensional array and sorted. The sort was done in two phases: in the first pass, all zero valued elements were moved to the end of the array; in the second, the resulting reduced length array was sorted using a simple selection sort [64]. A faster running technique could have been used, but the program only needed to be run once (after checking for correctness) and the development time was negligible. The elements of the sorted reduced array were then added up, by taking the sum of the smallest pair and inserting this result back into the array, keeping it sorted.

The result of subtracting the final sum from unity was 1.09×10^{-9}, from which it may be concluded that the total absolute error in calculating the value of individual table elements was one order of magnitude less than the truncation error (being of the same sign).

3.5. Analysis of C1-5-10S

The C1-5-10S constellation shown in figure 3.1a, with its minimum distance decision boundaries, is the form of the C1-5-10 constellation studied in the literature. It consists of a unit circle (L_2) of ten equispaced symbols, a concentric circle of five symbols at a radius of 0.522 (L_1) and a zero energy symbol at the centre (L_0). It may be noted that one set of authors [54] used $L_1 = 0.5$, commenting that this was not perfect but that the difference was small; others [53] did not give precise values for the signal coordinates, while the third group [44] equated nearest neighbour distances, which

gives $L_1 = 0.526$. The value of 0.522 was arrived at by using ERPART with a varying radius to minimise $P(\epsilon_s)$, taking account of the increased significance of errors from an outer ring symbol, there being twice as many of these as inner ring symbols. The change from 0.522 to 0.500 produces an increase in $P(\epsilon_s)$ of 4% at 19dB PSNR and 53% at 23dB.

The performance of C1−5−10S with a receiver making independent decisions on the phase and magnitude of a received symbol was examined, as this was relevant to the development of the SDPLL. The optimum values for the radii of the two magnitude thresholds was determined as $T_{01} = 0.272$ (threshold between L_0 and L_1) and $T_{12} = 0.759$ (threshold between L_1 and L_2) by variation over several runs of ERPART. This was in conjunction with $L_1 = 0.522$ which is not optimum for this type of detection, but was used to make the comparison with minimum distance detection valid. The difference in performance is equivalent to 0.6dB SNR, or a factor of 4 in the $P(\epsilon)$ around 23db PSNR, as may be seen from figure 3.6.

While C1−5-10S only has symbols at 15 phases, these lie on a subset of 20 eqi-spaced radii, so that a carrier recovery circuit would be required to generate a twenty-fold phase ambiguity. Since it was felt desirable to reduce this, subsequent attention was focussed on the modified constellation C1−5−10N (figure 3.2) which has only 10 phases.

3.6. Analysis of C1−5−10N

3.6.1. Symbol Error Probability

Because of the increased proximity of symbols, brought about by the alignment of inner and outer rings, the previously determined value of L_1 is not optimum for C1−5−10N. By repeating the procedure used for the staggered constellation, the optimum value of L_1 for minimum distance detection of C1−5−10N was found to be 0.500. The minimum distance $P(\epsilon_s)$ characteristic as a function of PSNR is shown in figure 3.7a, which also includes curves for the $P(\epsilon)$ of individual (distinct) symbols. Notice that much of the degradation in performance of the non-staggered constellation can be attributed to the preference given to the five $S_{2[1,3,5,7,9]}$ symbols at the expense of the others.

The final configuration optimised was $I\Phi-M$ detection of the non-staggered constellation. The results of this are presented in full detail, as they are the most pertinent to the development of the SDPLL. The starting point was taken as

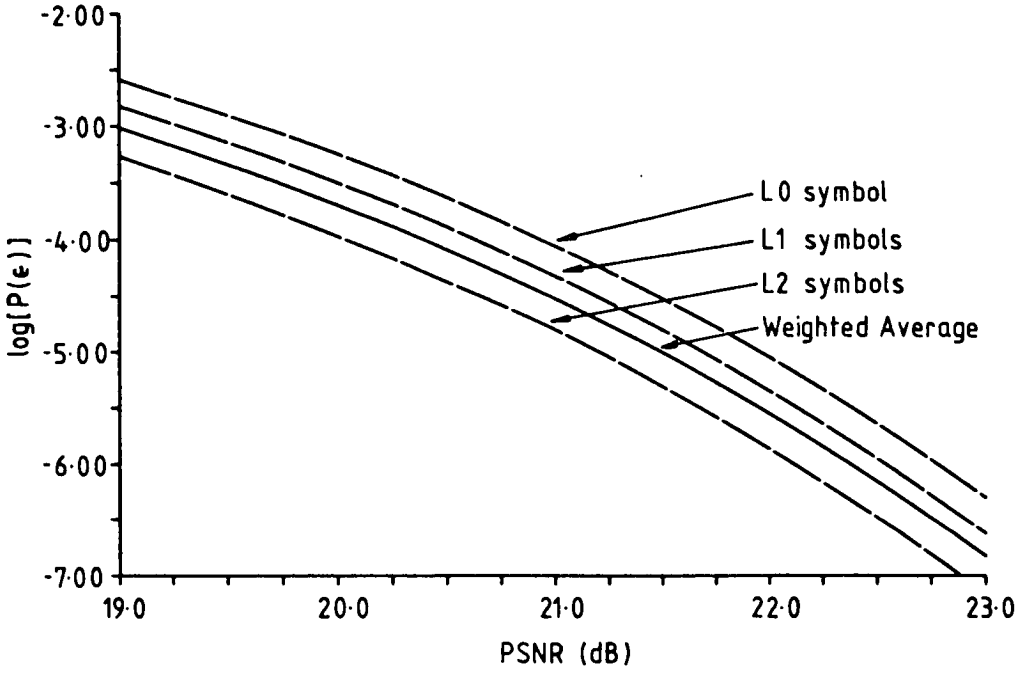


Figure 3.6a Minimum Distance $P(\epsilon_s)$ for C1−5−10S

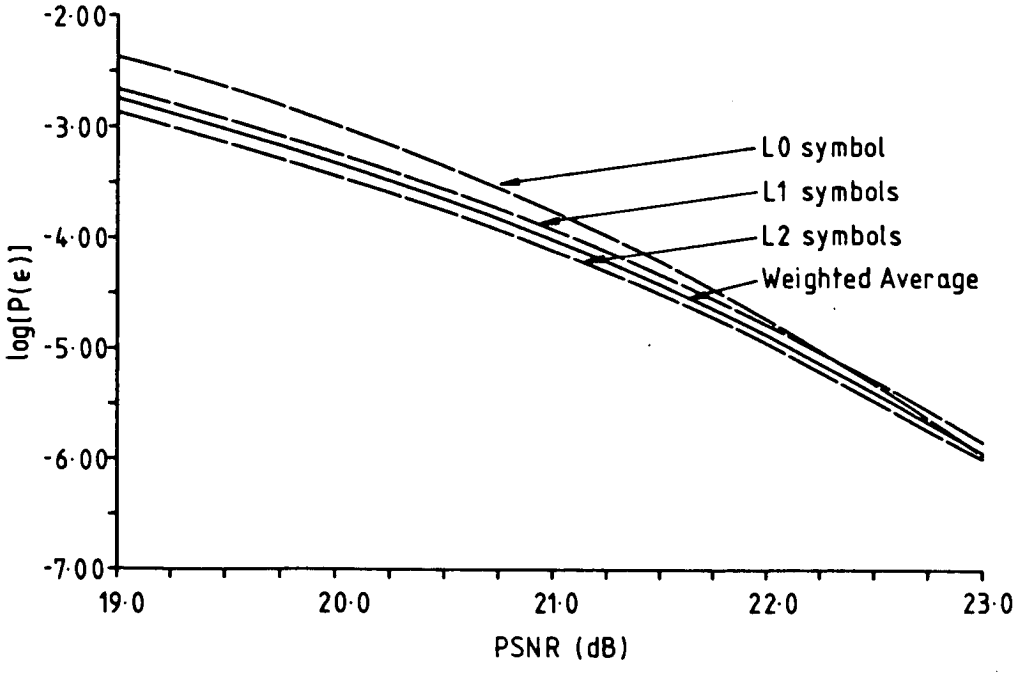


Figure 3.6b Independent Phase-Magnitude $P(\epsilon_s)$ for C1−5−10S

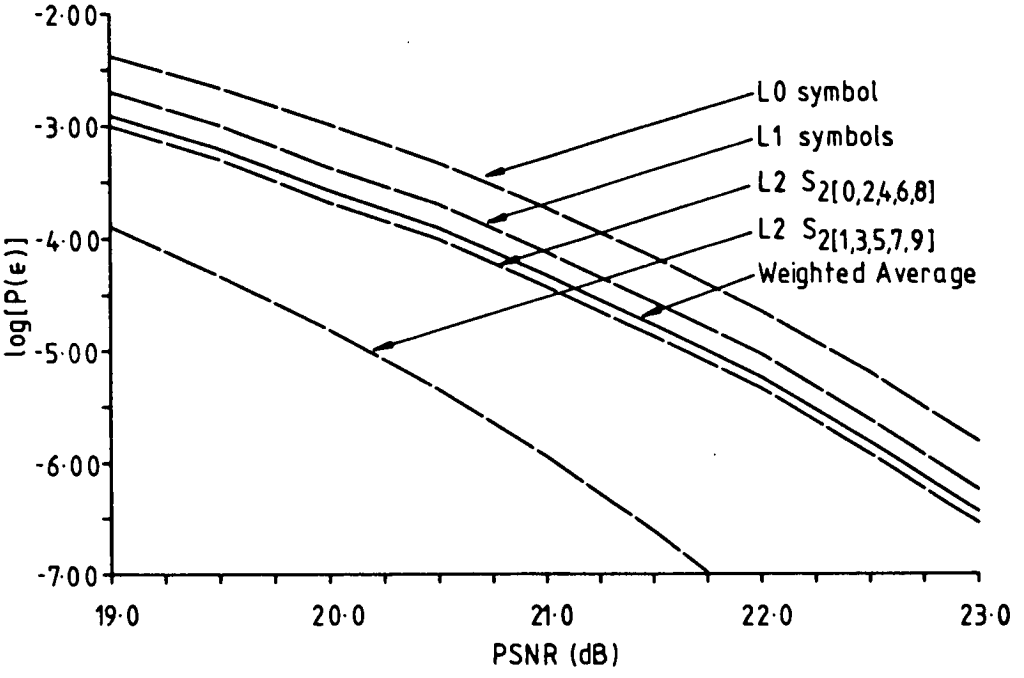


Figure 3.7a Minimum Distance $P(\epsilon_s)$ for C1−5−10N

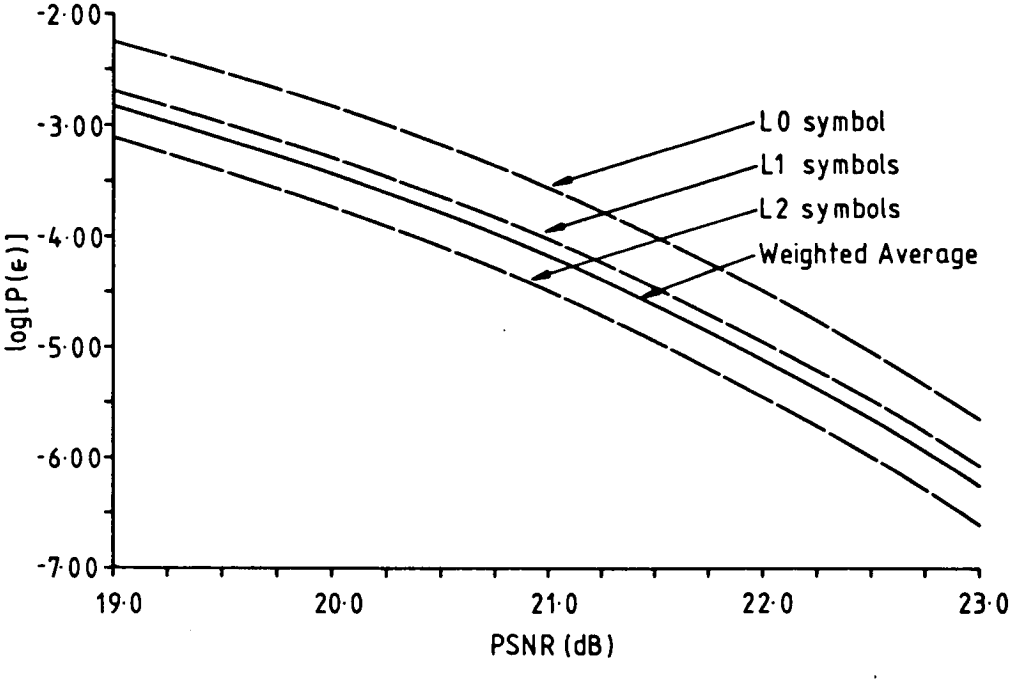


Figure 3.7b Independent Phase-Magnitude $P(\epsilon_s)$ for C1−5−10N

$L_1 = 0.500$, $T_{01} = 0.250$, $T_{12} = 0.750$, with all calculations performed for 23 dB PSNR. In the first cycle of optimisation, ERPART was run with different values for L_1 and the results interpolated with a third-order polynomial. After plotting, the value of L_1 giving minimum $P(\epsilon_s)$ was determined from the graph and used in the determination of T_{01}. A new value for this parameter was determined in the same fashion as for L_1 and used in the calculation for T_{12}. This cycle was then repeated twice; a third repeat resulting in no change in any parameter. The final curves obtained are presented in figures 3.8[a-c]. The ultimate values for the three parameters were $L_1 = 0.498$, $T_{01} = 0.255$, $T_{12} = 0.747$ ($L_0 = 0$; $L_2 = 1$). At 23 dB PSNR this produces a mean $P(\epsilon_s) = 5.21 \times 10^{-7}$ compared with 5.80×10^{-7} produced by the initial values; an 11% difference. Having determined the optimum parameters of the constellation, $P(\epsilon_s)$ curves were calculated for the individual symbols and were averaged to give the mean $P(\epsilon_s)$. These results are shown in figure 3.7b.

The curves of figure 3.8 are useful in that they indicate the sensitivity of the constellation to maladjustment of the receiving and transmitting equipment. The tolerances allowable for a factor of two increase in $P(\epsilon_s)$ are: $L_1 = 3.6\%$, $T_{01} = 6.8\%$, $T_{12} = 2.4\%$. A further result in this category is the sensitivity of the constellation to static phase error in the receiver reference. This is not the same as the susceptibility to phase noise, which includes nonlinear effects. The constellation exhibits a tolerance of $\pm 4°$ in the phase reference, producing a factor of two increase in the $P(\epsilon_s)$ at 23 dB, with little dependence on SNR. This is illustrated by figure 3.9, which shows results from ERPART where the position of the noise table origin has been skewed from the nominal symbol position by rotation about the centre of the constellation.

A $P(\epsilon_s)$ comparison of C1–5–10N using $I\Phi-M$ detection with C1–5–10S, 16QAM and V.29 using minimum distance detection, is shown in figure 3.10. The difference in the two C1–5–10 designs varies from 0.3 dB to 0.4 dB SNR over the range shown and is small considering the significant difference in the implied decoding complexity. They are 0.7 dB and 1.1 dB better than 16QAM under the peak power constraint as shown. When a mean power constraint is applied, 16QAM and C1–5–10S are essentially equivalent, while C1–5–10N is 0.7 dB worse than 16QAM. The V29 design is poor under all circumstances, although it does have implementation advantages as previously mentioned.

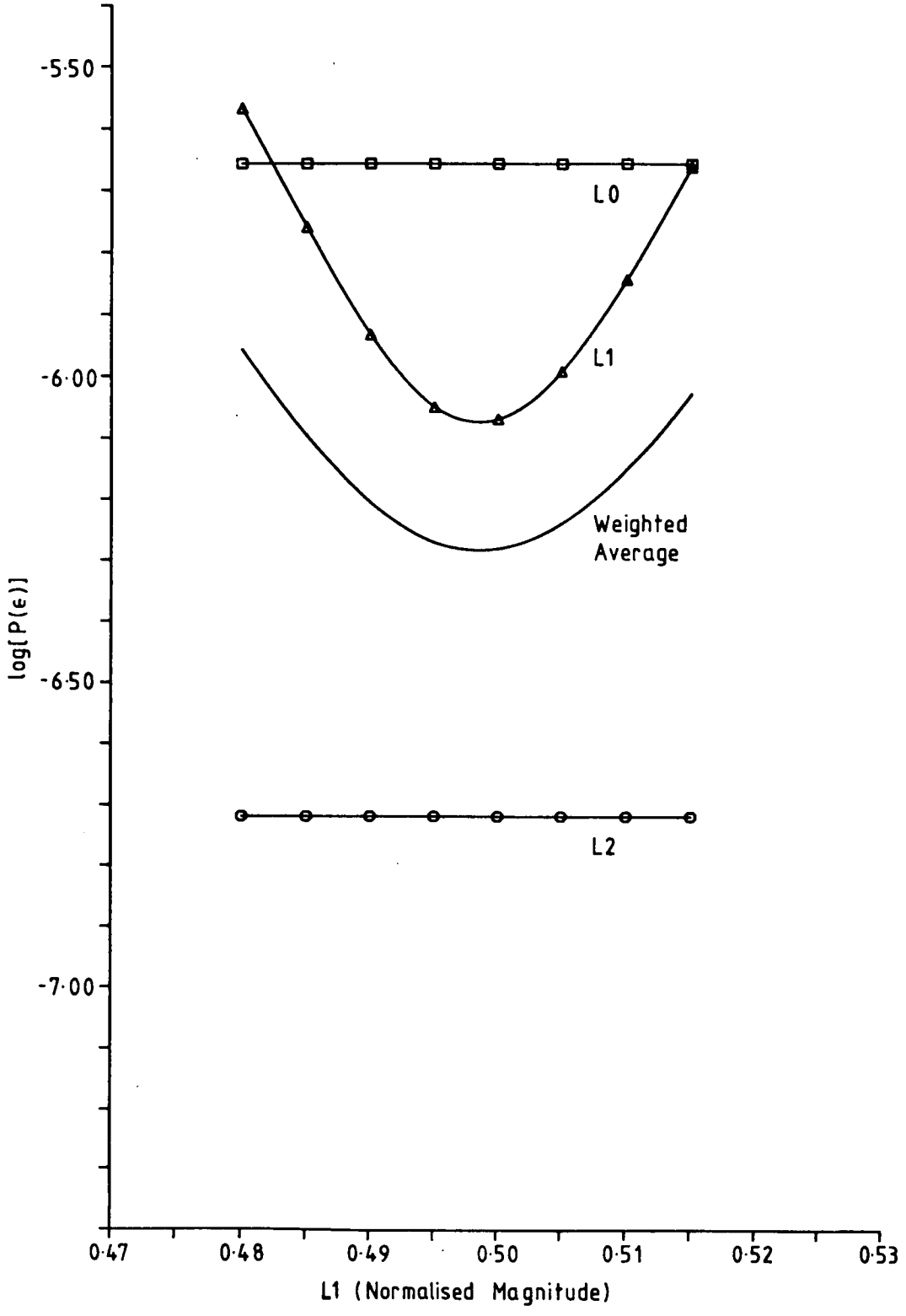


Figure 3.8a $I\Phi - M$ C1–5–10N Performance Sensitivity to L1 Variation

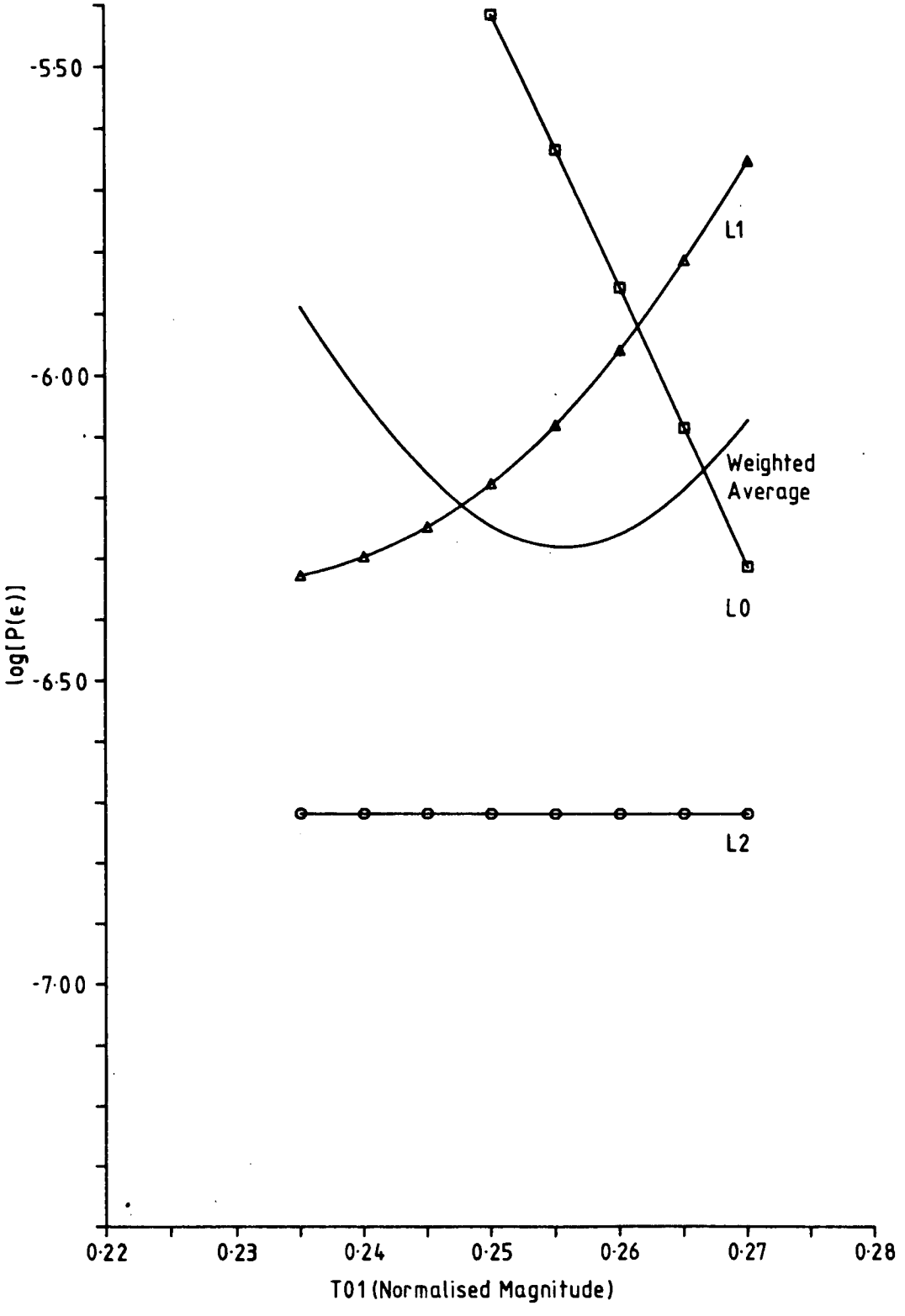


Figure 3.8b $I\Phi - M$ C1–5–10N Performance Sensitivity to T01 Variation

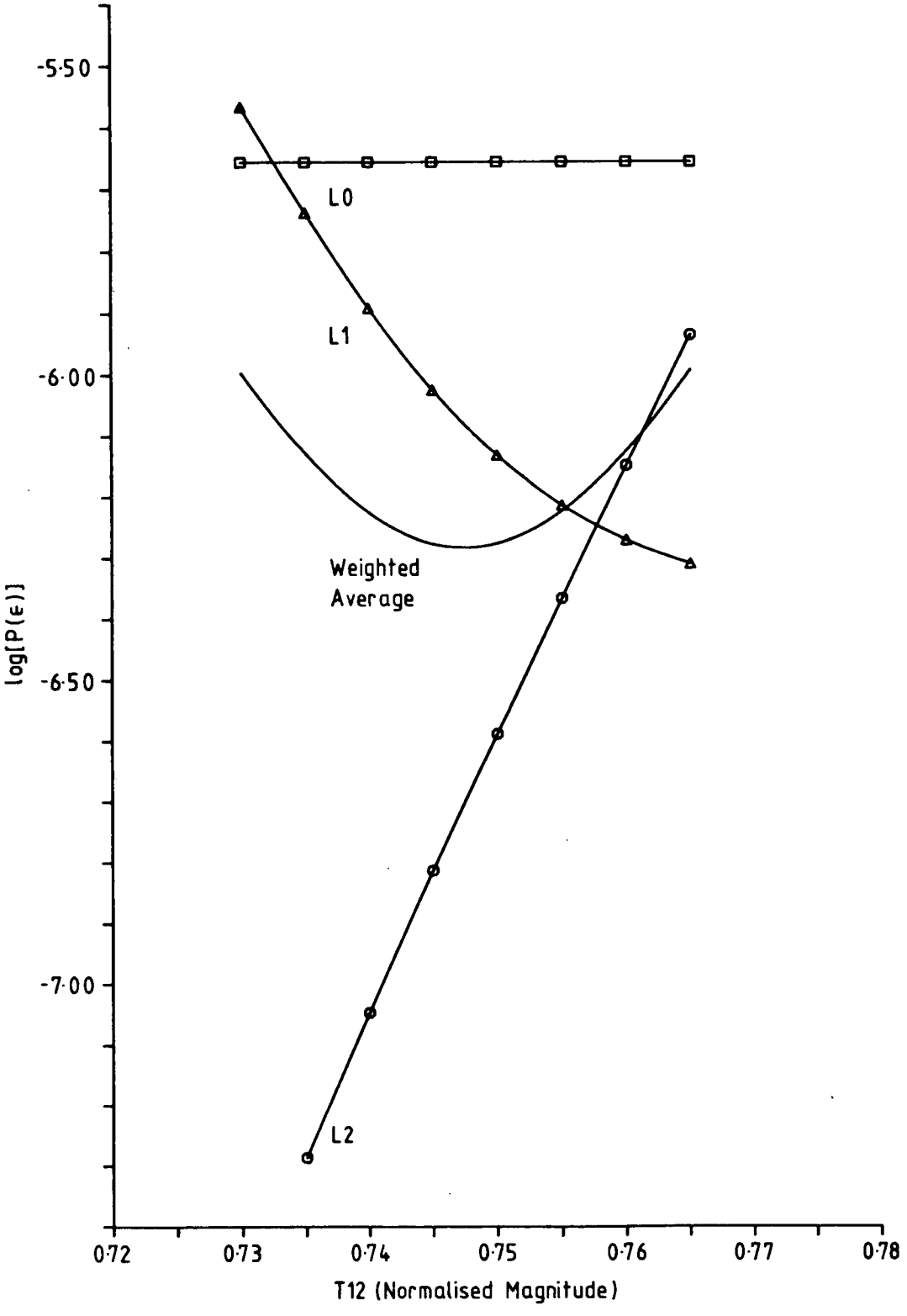


Figure 3.8c $I\Phi-M$ C1–5–10N Performance Sensitivity to T12 Variation

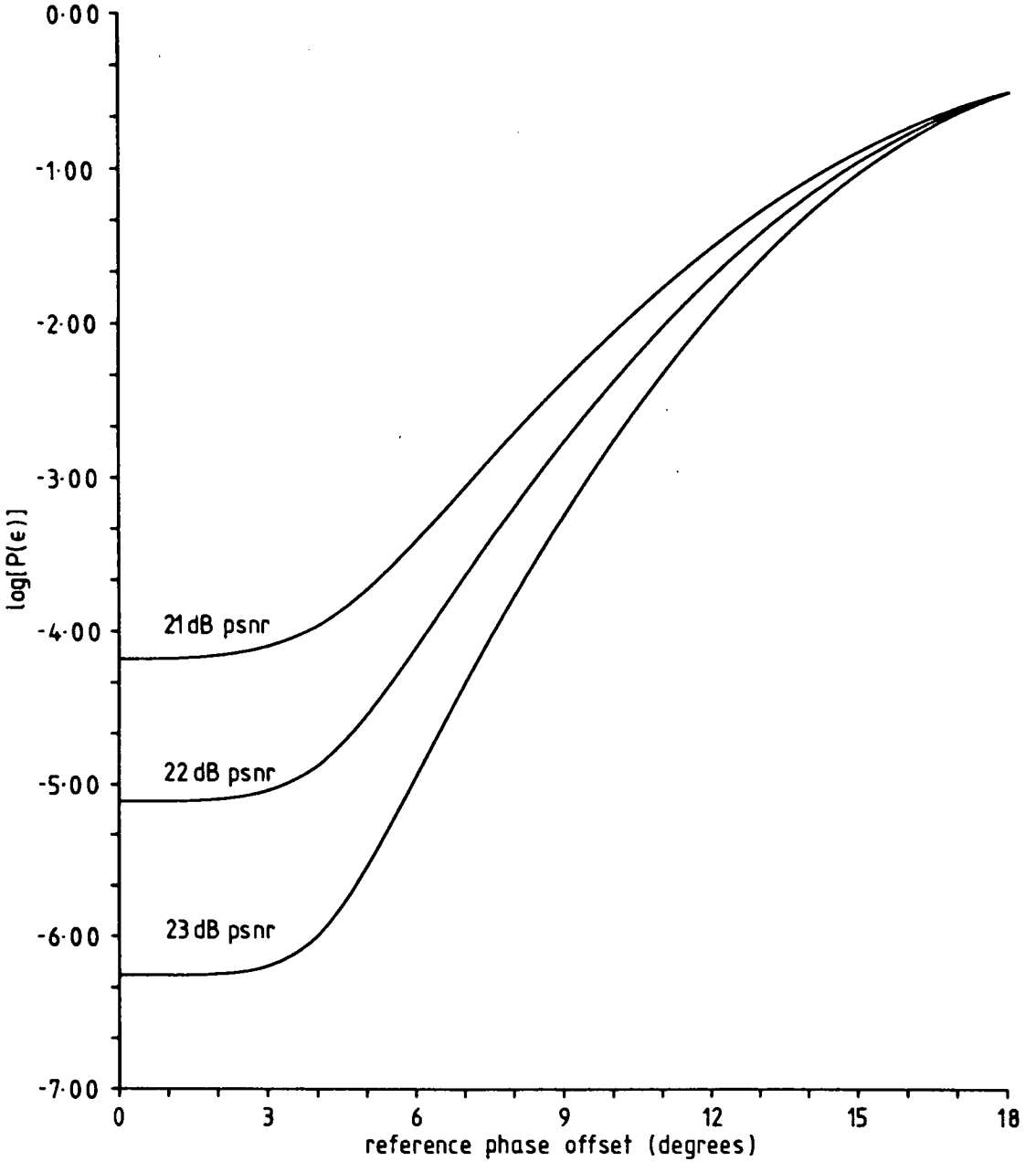


Figure 3.9 Effect of Static Phase Offset on C1−5−10N ($I\Phi - M$ boundaries)

3.6.2. Bit Error Probability

The results presented so far have all been based on symbol error probability. Since each symbol translates to a pattern of four bits, constellations with similar $P(\epsilon_s)$ may have significantly different $P(\epsilon_b)$ if they do not have equally effective grey codes. Although it is often suggested in the literature that $P(\epsilon_b)$ is the definitive measure of performance, this is not necessarily so. For example, the throughput of a data link which uses ARQ

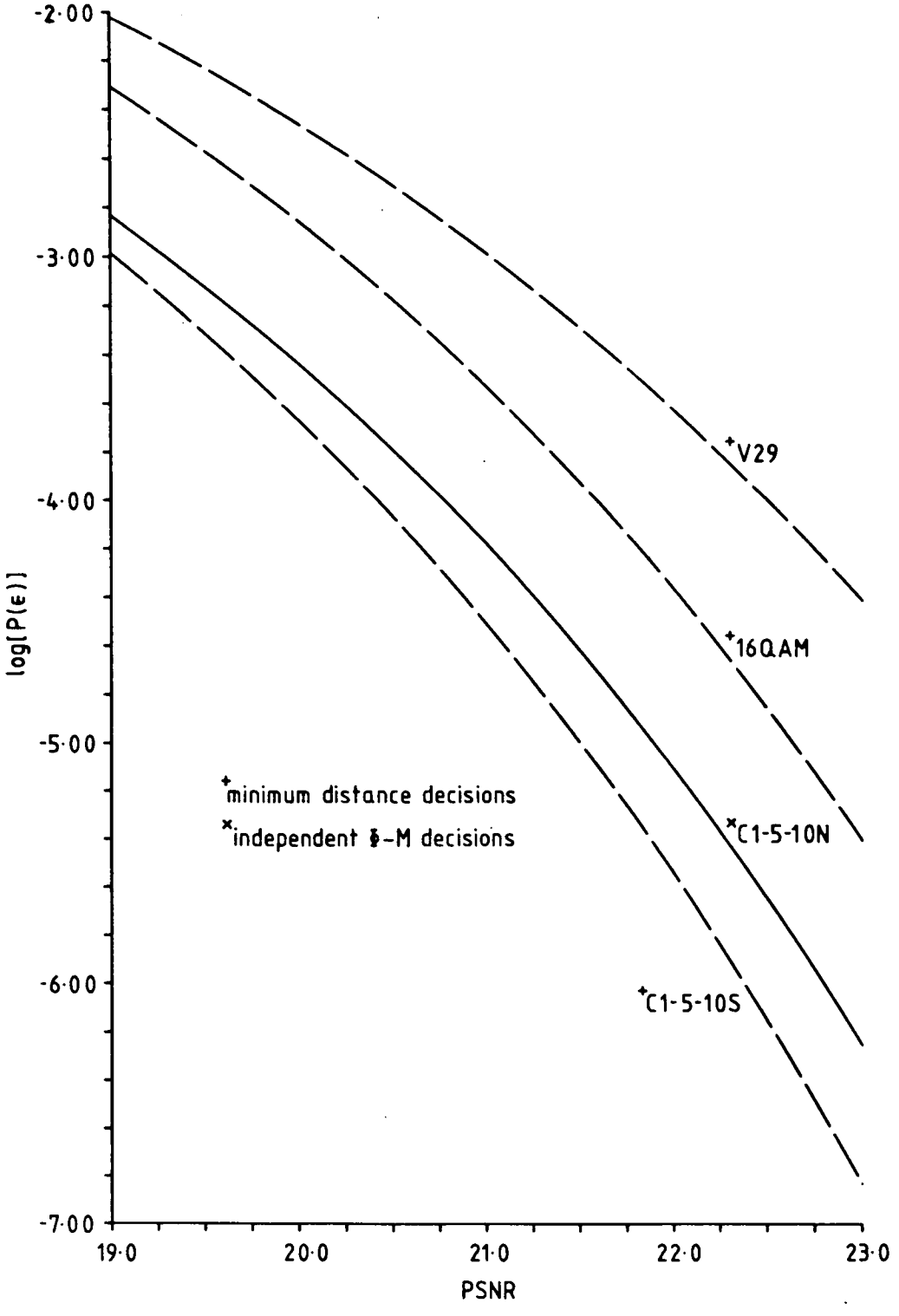


Figure 3.10 $P(\epsilon_s)$ Comparison

for error control will show no difference between a good and poor grey code, since an affected data frame will be re-transmitted whenever an error is made decoding a symbol within that frame, regardless of the number of bit errors to which this corresponds. A similar argument applies where no error recovery is used and the semantics of the transmitted data are based on groups of bits larger than are represented by a single channel symbol.

Differences in grey codes are important where FEC is employed, or where the bits of one symbol may straddle two semantically distinct words, as the power of the code limits the number of bit errors which can be corrected. It is interesting to note that ARQ is generally applicable to links with low error probability, where differences in grey codes are more pronounced; while FEC is important at higher error probability, where the difference between grey codes diminishes, as more distant-neighbour symbol errors occur.

A grey coding scheme for $I\Phi-M$ detected C1−5−10N was developed from ERPART results, starting with the observation that at high SNR most errors involve a mistaken magnitude, rather than phase angle. This observation holds for all symbols for PSNR down to 13 dB, as may be seen from the graphs of error distributions in figure 3.11. Guided by this knowledge and using a map of the sixteen binary codes, laid out as four groups of four codes, showing Hamming distance between codes and groups (figure 3.12), codes were allocated to the constellation symbols. The L_1 codes were allocated first, then the L_0 code and the L_2 codes. The technique may best be described as guided intuition. Although it is not useful in providing a rigorous model for the generation of impure grey codes for other constellations, it is preferable to a blind exhaustive search of all code permutations†. Using the data generated by ERPART, the performance of this initial code was ascertained. Then results for permutations of the five L_1 symbols were calculated (requiring 12 seconds processor time), which led to the grey code shown in figure 3.13. This has not been *proved* to be the best grey code for C1−5−10N, but is believed to be no worse than ten percent from the optimum for pure AWGN interference. Note that other types of interference, particularly reference phase offset and jitter, are likely to alter the optimum grey code from that presented here. Simon and

† A simple exhaustive search over 16 symbols would require over 66 years processor time, at 100μS per permutation.

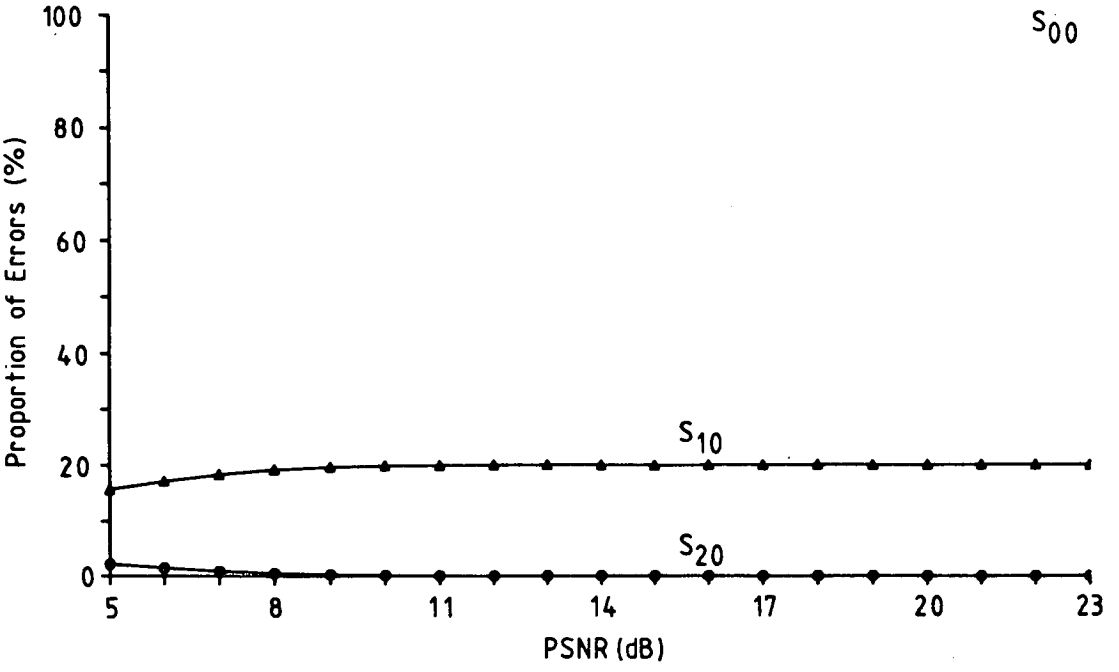


Figure 3.11a Distribution of errors from C1−5−10N S_{00}

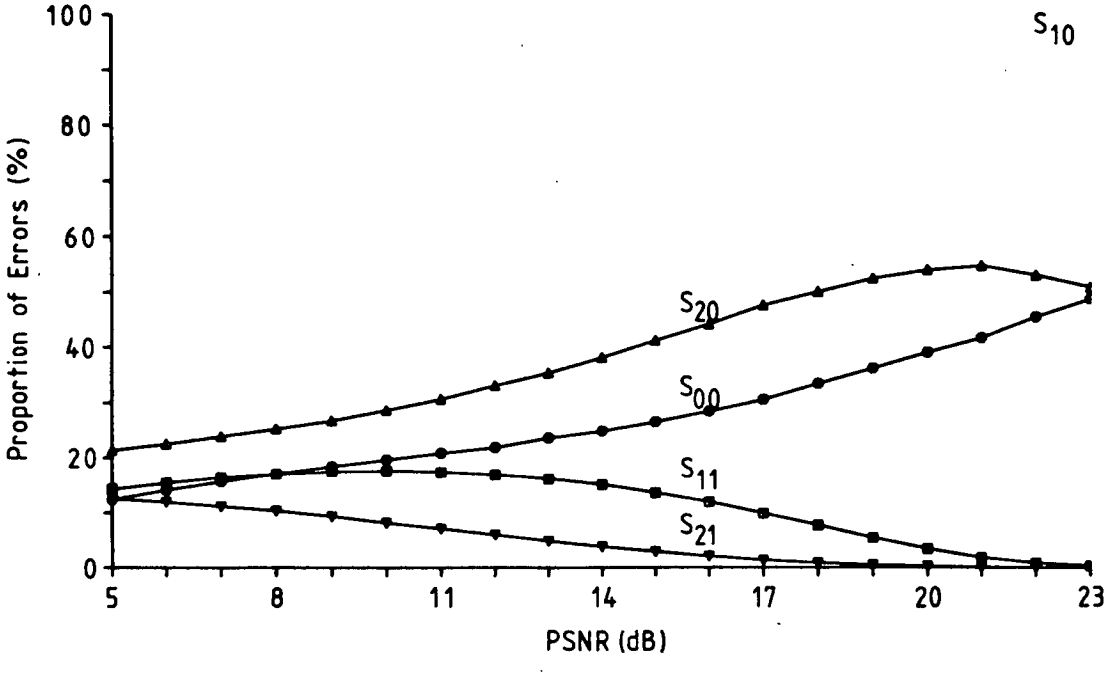


Figure 3.11b Distribution of errors from C1−5−10N S_{10}

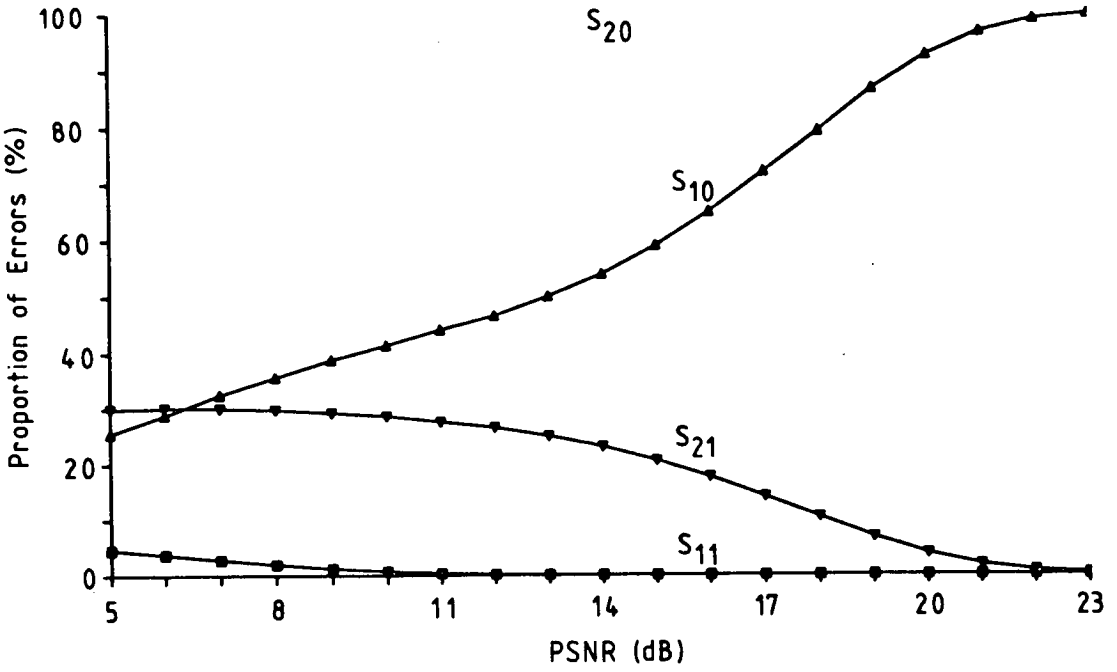


Figure 3.11c Distribution of errors from C1−5−10N S_{20}

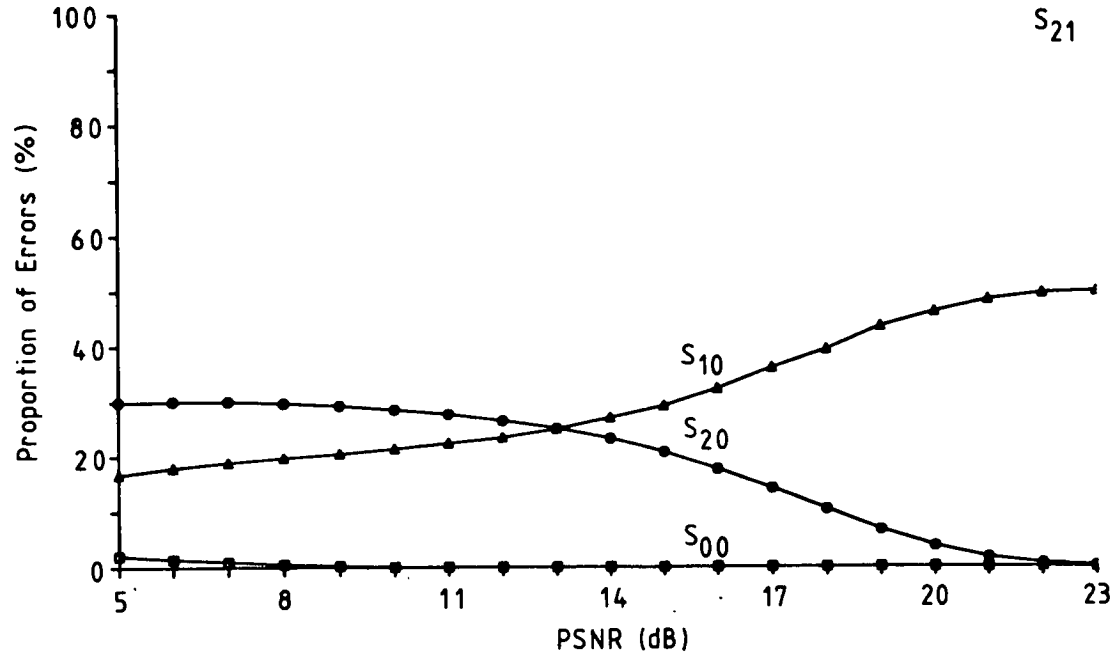


Figure 3.11d Distribution of errors from C1−5−10N S_{21}

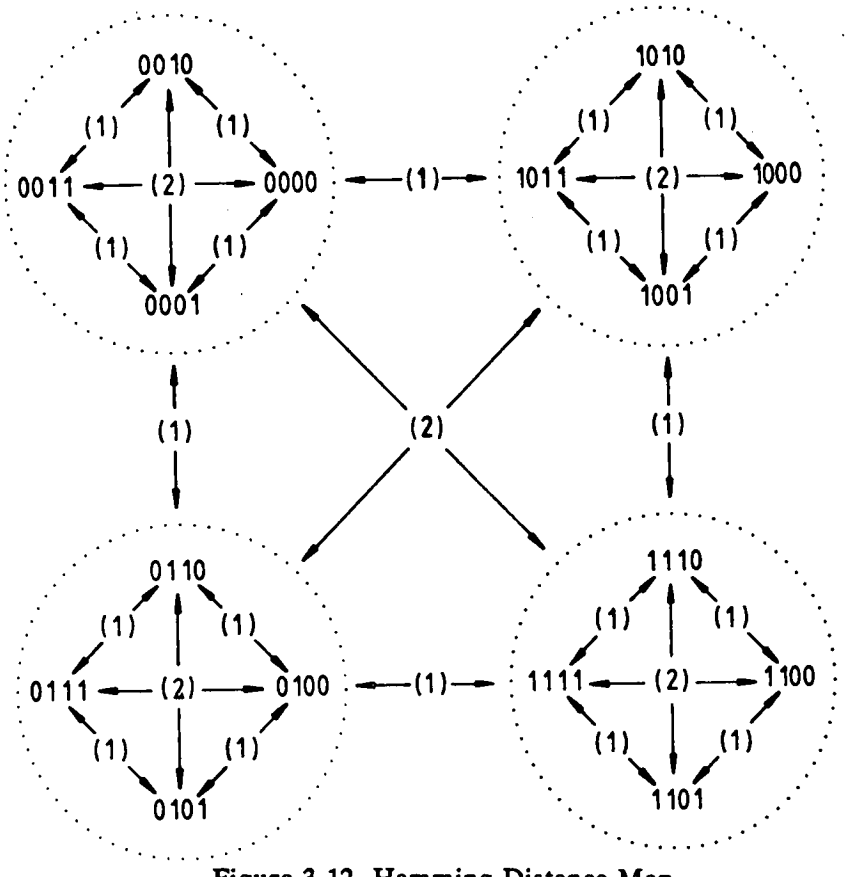


Figure 3.12 Hamming Distance Map

Smith [53] express the view that 4/3 is probably the minimum grey code penalty achievable for a C1−5−10 constellation; an alternative approach to combined differential- and grey-coding of two-dimensional APK signal sets is given by Weber [65]. The penalty for the best grey code discovered is shown as a function of PSNR in figure 3.14, along with that of the perfect grey code of 16QAM for comparison.

3.6.3. Time Domain Properties

In order to gain insight into the design requirements of the SDPLL, some of the zero-crossing properties of the C1−5−10N constellation line signal were studied. An analytic analysis of the zero-crossing properties of a gaussian random process has been made by several authors (see chapters four to six of [66]) and an analysis of a sine wave in gaussian noise is given by Rice [67]. Rather than extend this work, it was decided to make use of other work done while developing a test signal generator for the SDPLL. A digital waveform store had been constructed and programmed with a pseudo-random sequence of 8192 symbols of C1−5−10N. The sequence was derived with values from a

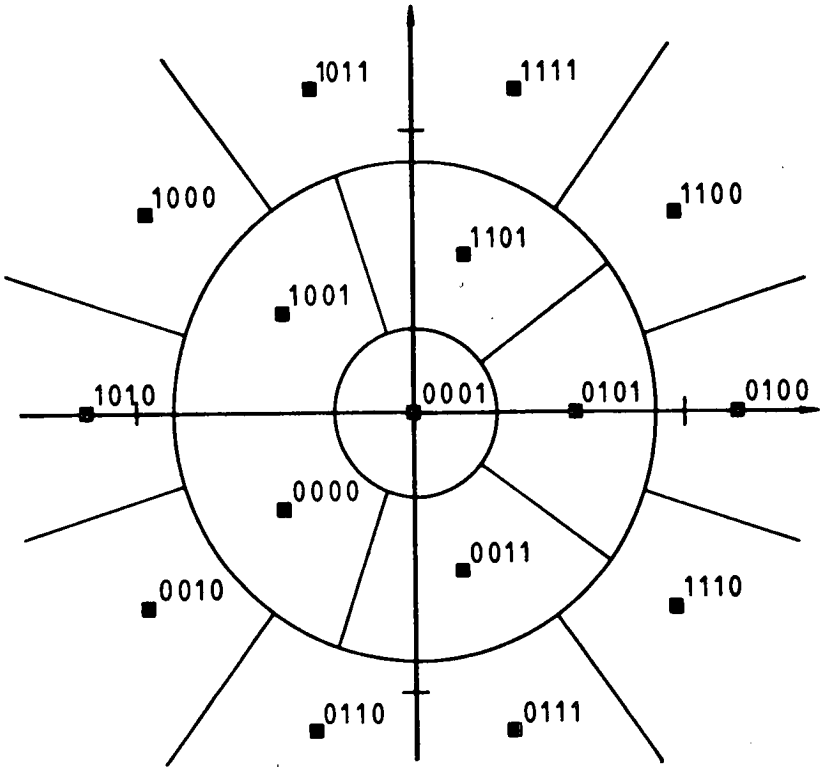


Figure 3.13 Grey Code for C1−5−10N

pseudo-random number algorithm, which was used to select, without replacement, symbols from an initial population of 512 of each symbol, thus ensuring that there was no bias toward any symbol. Adventitious correlations between symbols were tested for by using a method similar to that used by Lach [68]. The adjacent-symbol correlation diagram of the sequence is given in figure 3.15, where the area of each spot is proportional to the number of occasions on which a symbol on the x-axis is followed by the corresponding symbol on the y-axis. The matrix shows no obvious banding or structured pattern (which would indicate correlation between symbols), supporting the premise that the test pattern is representative of a general arbitrary symbol stream. By analysing this representative sequence, which has a period of 3.4 seconds at 2400 baud, three statistics were extracted and are presented as histograms in figures 3.16 to 3.18. These are the distribution of intervals between zero-crossings over an unbounded positive range (maximum observed less than 1mS), the distribution of zero-crossing instants within a repeated interval of one tenth of a carrier cycle period and the distribution of zero-crossing instants within a repeated interval of one symbol signalling period. The

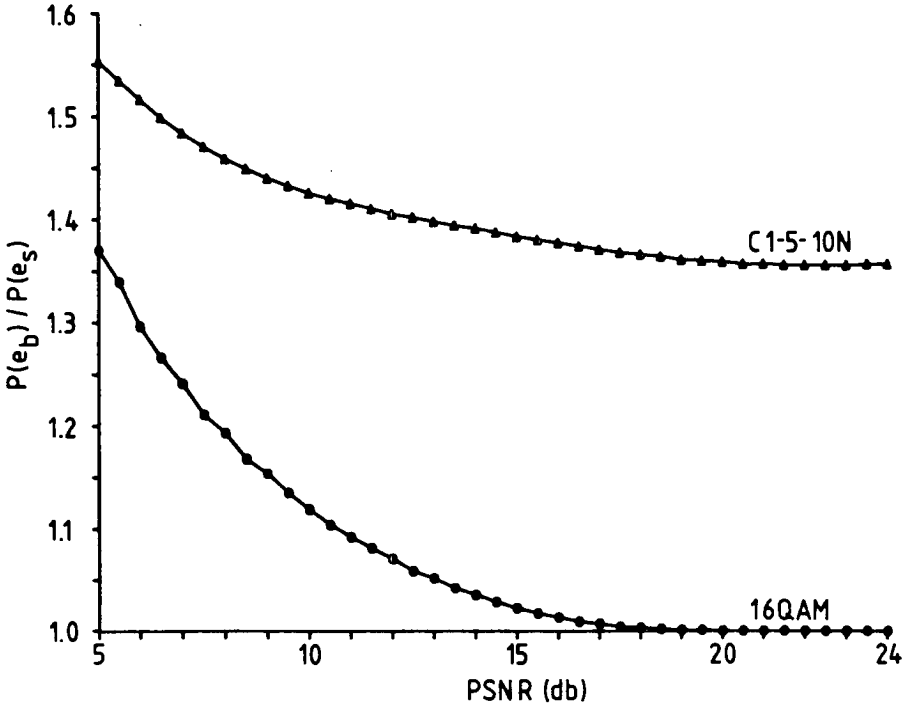


Figure 3.14 Performance of Grey Codes for C1−5−10N and 16QAM

last two show the zero-crossing distribution which would be seen by an SDPLL locked to the carrier and symbol timing waves, respectively. They are given an initial (arbitrary) phase so as to put the peak of the distribution near to the centre of the histogram. Note that this analysis does not take account of noise, but does include the effect of bandpass filtering. The filter used was a 64 point finite impulse response design (linear phase) the frequency transfer characteristic of which is shown in figure 3.19; the spectrum of the filtered sequence is shown in figure 3.20.

The histogram of figure 3.16 shows a pronounced peak with a mean value somewhat less than half of one carrier period; there is a distinct spike at exactly one half-carrier period, on the right shoulder of this main peak. Other distinct spikes are visible, with offsets from the half-carrier spike corresponding to multiples of the difference of the half-carrier and signalling period, taking into account the differences in period occasioned by the frequency of the signalling being 4/3 that of the carrier. Note that in the sequence studied, the carrier and symbol transitions are initiated simultaneously. In general, there is no reason to expect this, nor any static relationship between the carrier and symbol timing; this sequence represents a worst case from some aspects, in that it provides

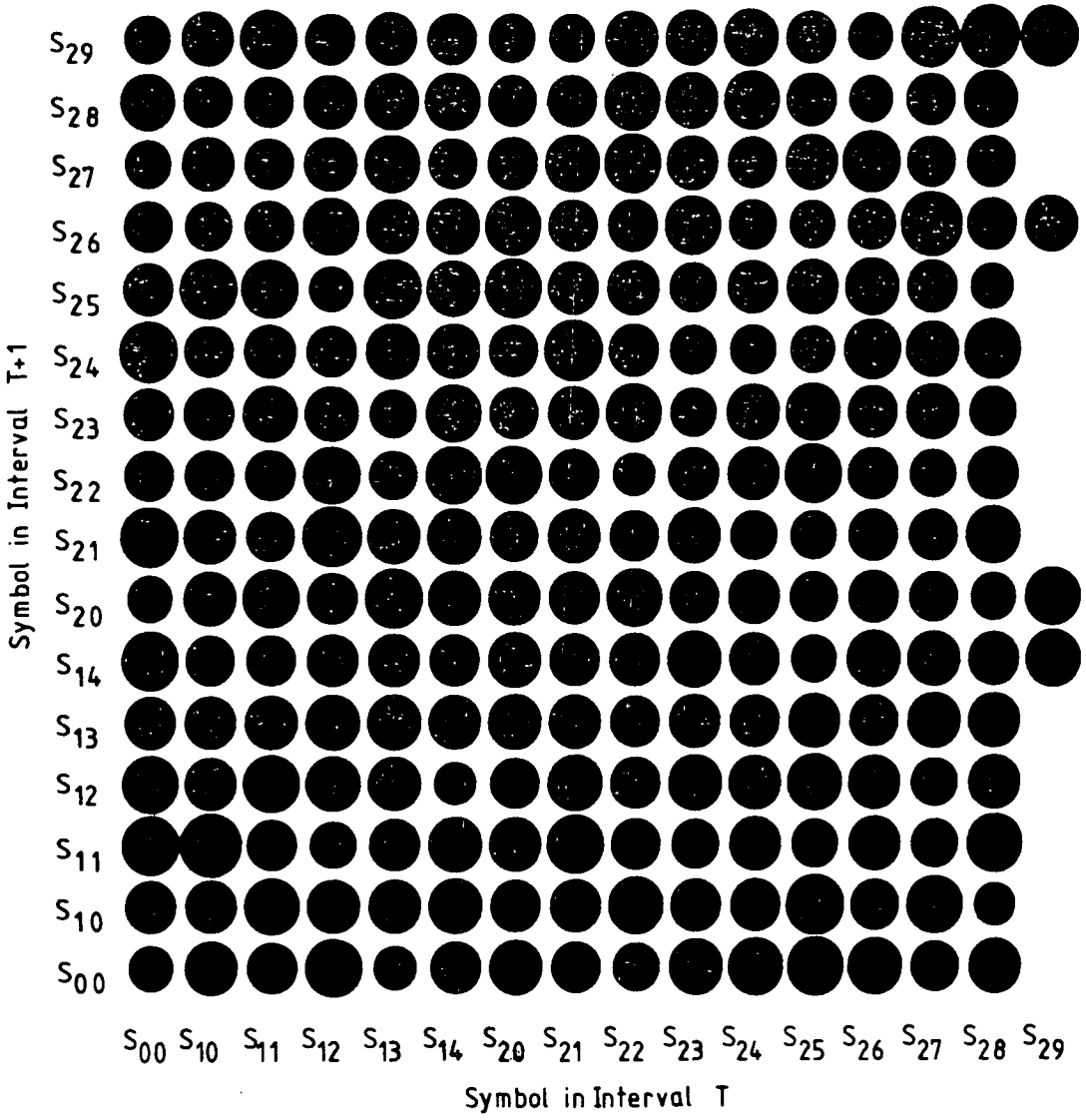


Figure 3.15 Correlation Diagram of Test Sequence

maximum opportunity for false locking of carrier and symbol timing recovery circuitry, with the associated lock stability problems.

The carrier referred zero-crossings of figure 3.17 show a well defined, although asymmetric, peak. There is a high background level due to signalling transitions and the peak is fairly broad, suggesting that carrier reconstruction from zero crossing information under these circumstances is by no means trivial. Further investigations have shown that the situation improves slightly when the carrier and signalling frequencies are not so simply related. Changing their frequency ratio from 3/4 to

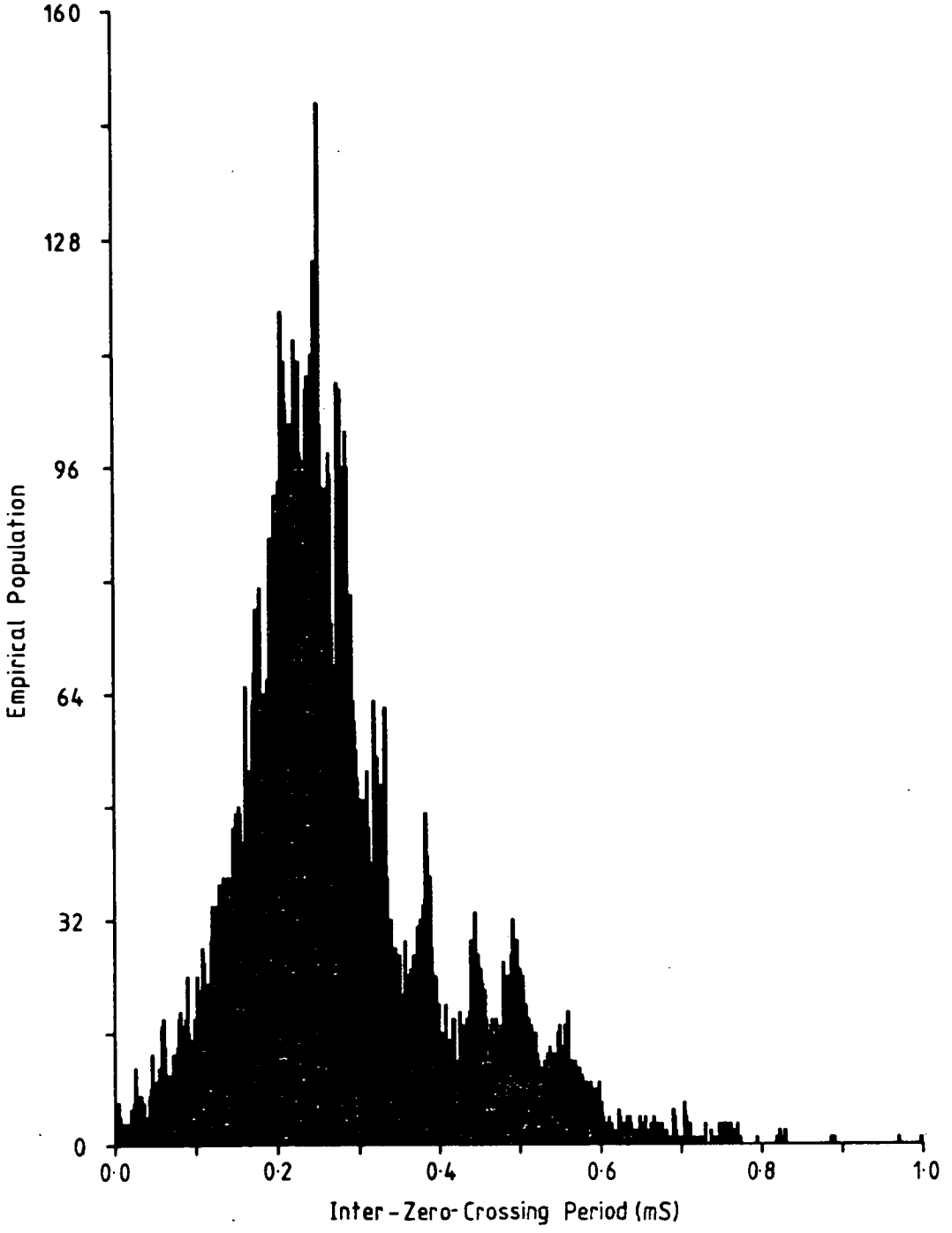


Figure 3.16 Distribution of Intervals Between Zero-Crossings

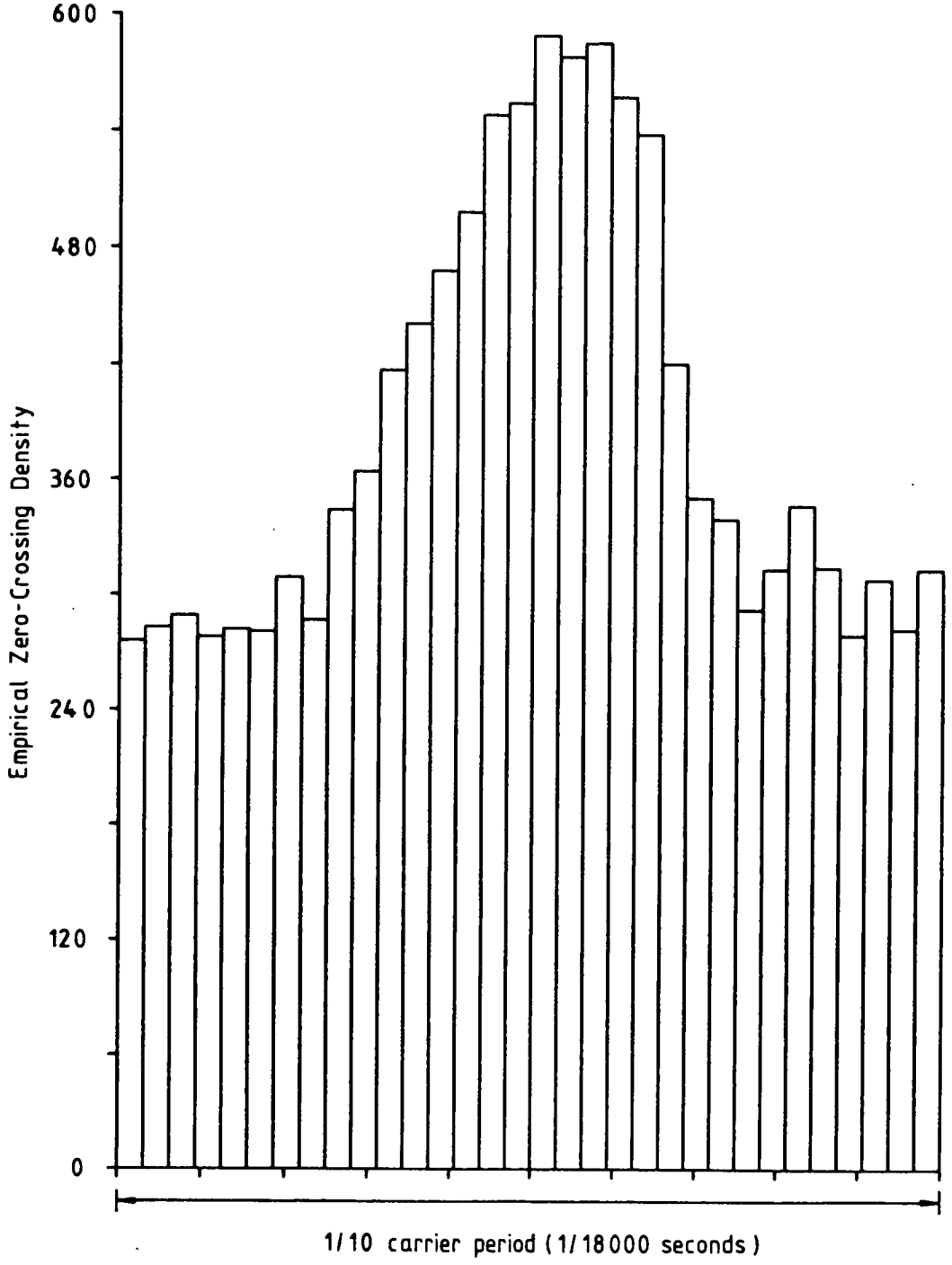


Figure 3.17 Carrier Synchronised Distribution of Zero-Crossings

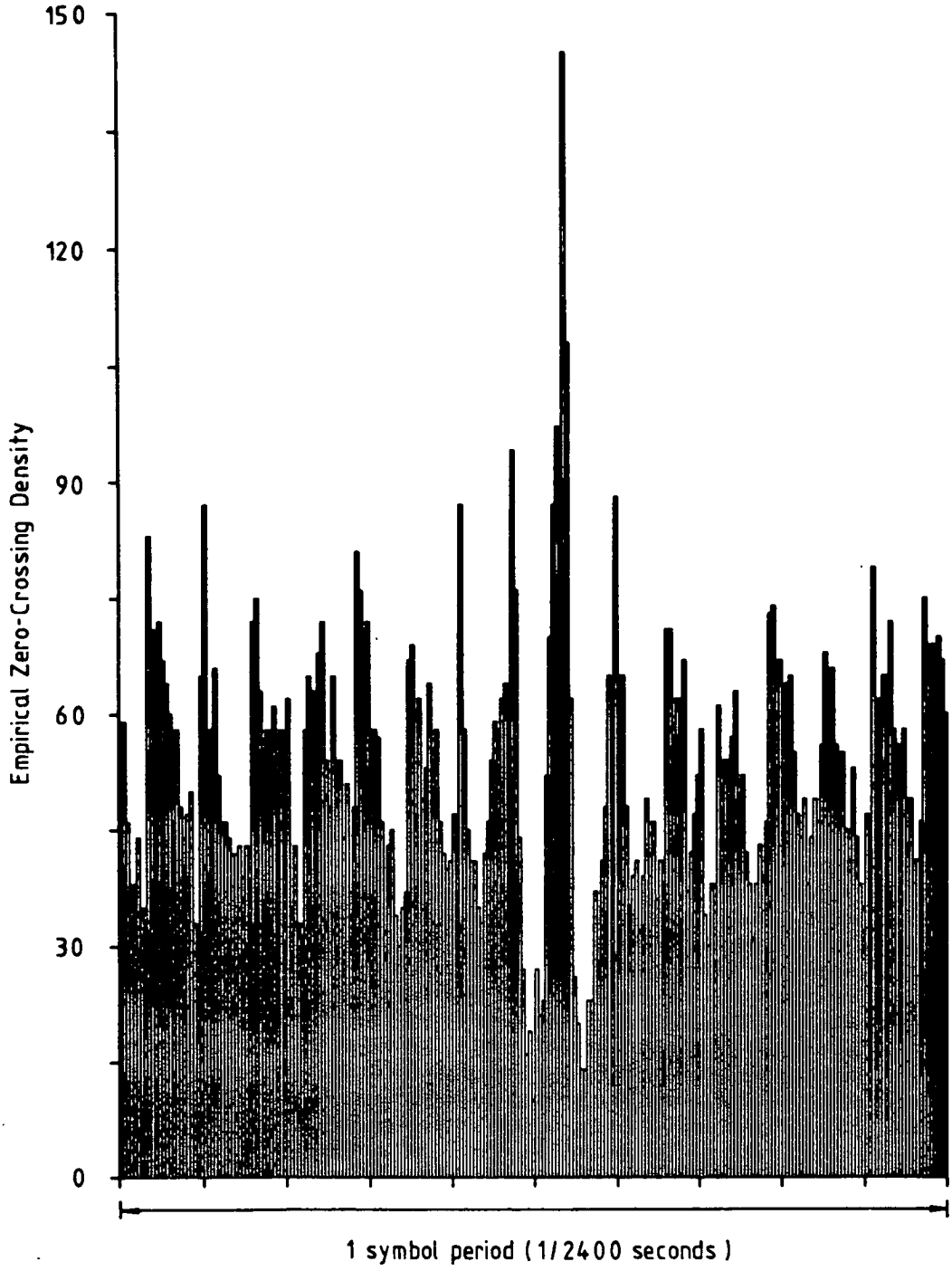


Figure 3.18 Symbol Synchronised Distribution of Zero-Crossings

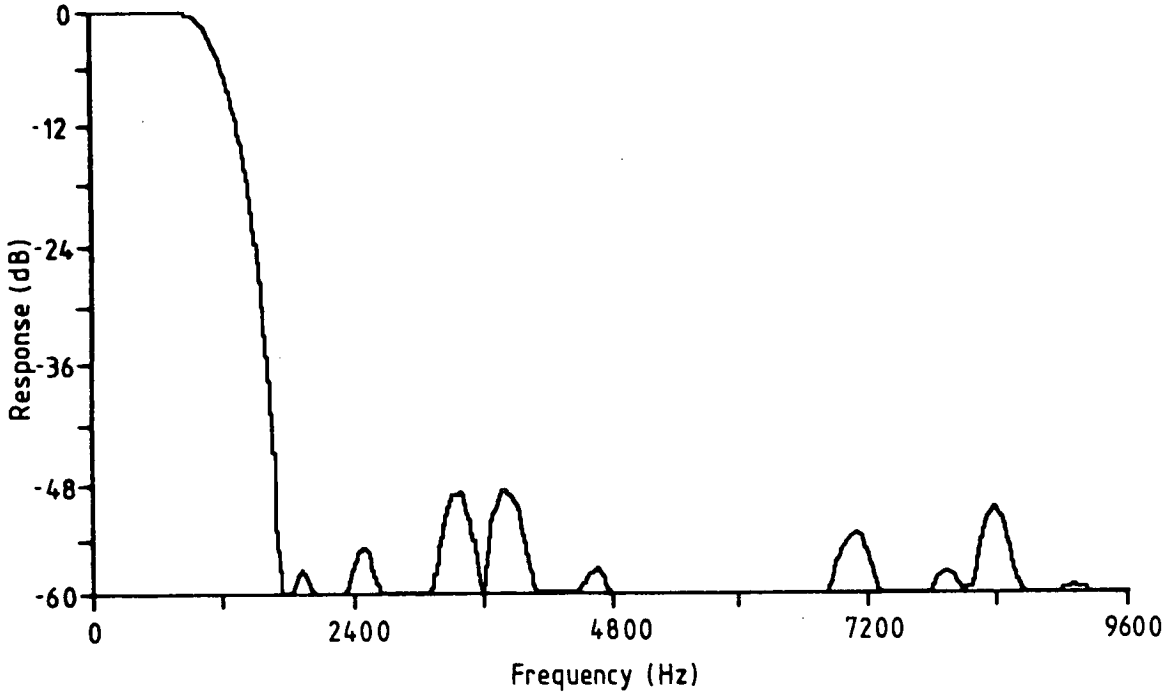


Figure 3.19a Test Signal Generator: (baseband) Filter Response

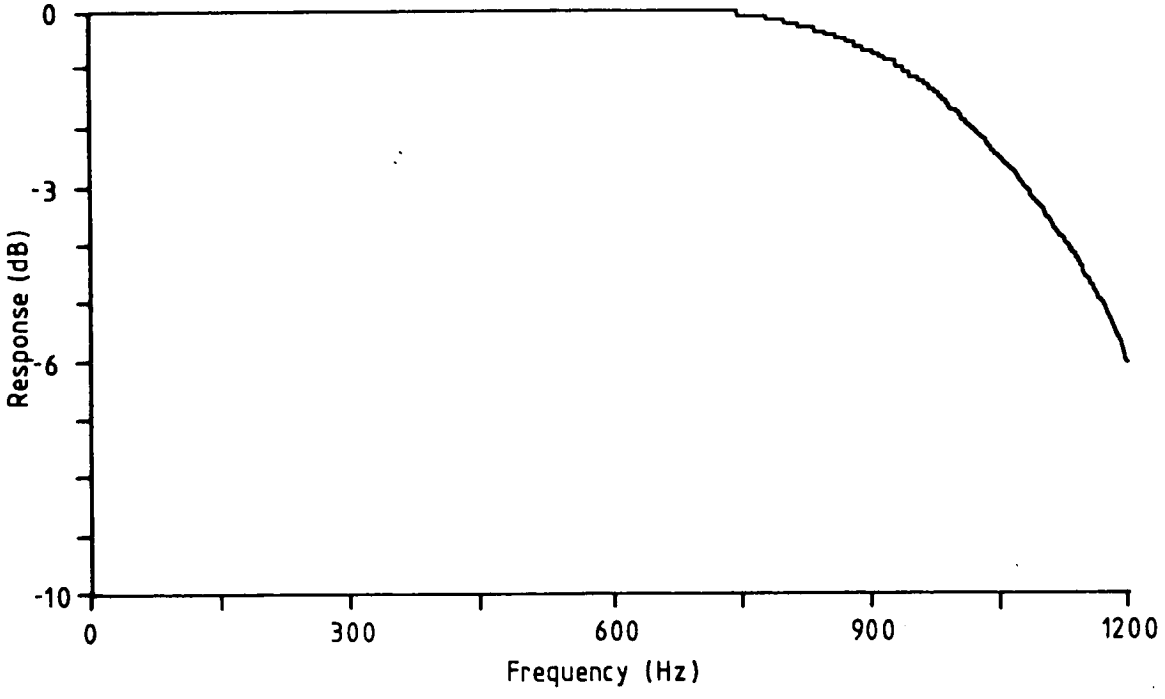


Figure 3.19b Test Signal Generator: (baseband) Filter Response

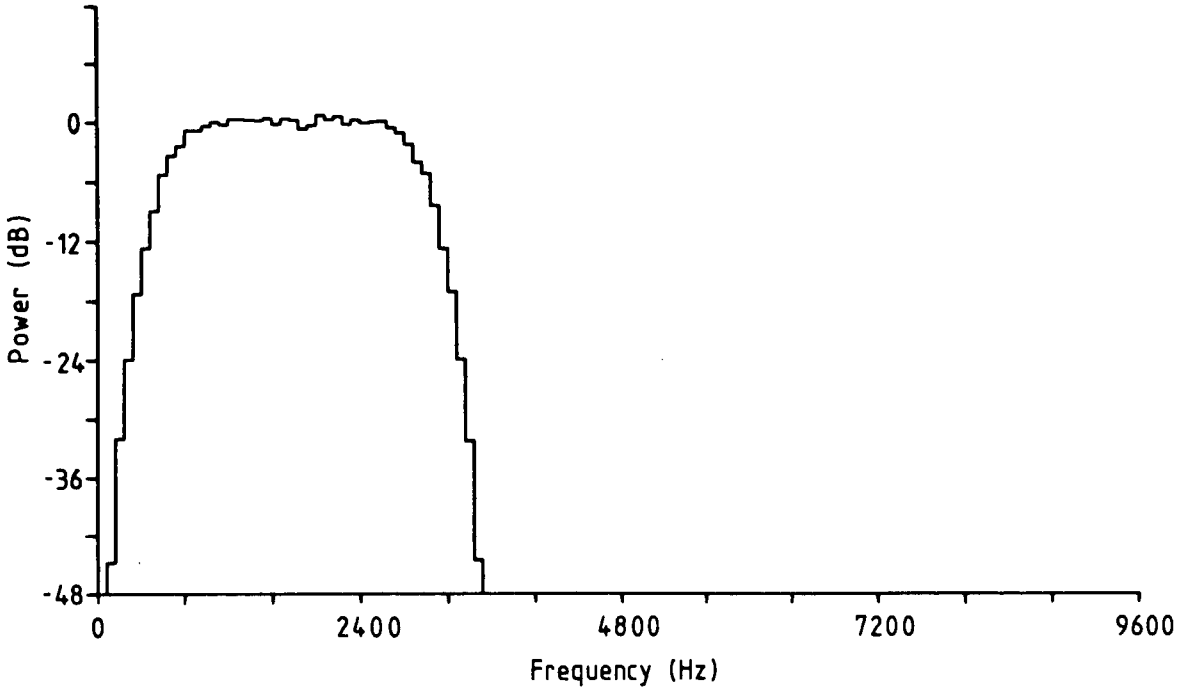


Figure 3.20 Test Signal (passband) Spectrum

6145/8192 narrows the peak by 25% (full width at half maximum) and increases its level above the background by 15%.

The symbol referred crossings in figure 3.18 show a well defined peak, with a pronounced notch in the background level, on both sides. This would indicate that direct recovery of symbol timing from the line signal is feasible, although the multiple distinct spikes in the background could produce false lock phenomena. However, the identity of the signalling period peak is considerably weakened when its ratio with the carrier period is altered, as described above.

3.7. Review of C1-5-10N

It has been shown that the C1-5-10N constellation has several desirable characteristics and fulfills the initial aims of its design. However, its practical usefulness in general modem applications is marred by certain properties which were not of immediate concern in testing the SDPLL. The major difficulty with this constellation is the presence of the zero energy symbol, which poses problems for receiver synchronisation and decoding.

3.7.1. Synchronisation Problems

Given an arbitrary data stream, there is a finite probability that a string of zero energy symbols, sufficiently long to exceed the receiver's capability to maintain synchronisation of its locally generated carrier and symbol timing references with the transmitter, could be encountered. Problems could also arise when AGC is required. In general, the only way to avoid this problem would be to add redundant data to the transmitted stream, effectively reducing the performance of the constellation. If the data stream is known to have particular inherent redundancy, it might be possible to exploit this. Specifically, many applications of high speed synchronous data links transmit data in blocks, delimited by protocol dependent control information. If the occurrence of this control information could be guaranteed to limit the length of the troublesome symbol strings, no further action would be needed. For situations where such fortuitous properties of the data could not be relied upon, one alternative would be to increase the data rate in the channel over that used end to end through the link and periodically insert synchronising symbols, which would be stripped off by the receiving modem. There is also the possibility of using C1−5−10N as a sub-constellation of a hyper-dimensional signal set incorporating redundancy coding.

3.7.2. Decoder Problems

The presence of a zero energy symbol means that differential decoding of the constellation cannot be used, unless a suitable method can be found to assign a phase value to that symbol. Thus, the receiving modem must establish an absolute reference with the transmitter, to be maintained for the duration of a session. While this is not an impractical goal in many instances, it does increase the complexity of the modem, particularly when provision must be made for re-synchronisation in mid-session (as opposed to re-starting a session).

3.7.3. Alternative Proposals

In the light of the foregoing difficulties, the C4−12N constellation appears to be a preferable alternative to development of C1−5−10N. The performance difference in AWGN between C4−12S and C1−5−10S is 0.2dB PSNR ($1.6 \times P(\epsilon_S)$) and 0.4dB MSNR ($3.0 \times P(\epsilon_S)$) [54] and there is no reason to suppose that the two non-staggered constellations would show significantly greater differences.

In the period since the start of this investigation, the increase in the amount of digital processing power which it is possible to incorporate in a modem, has greatly enlarged the range of constellations and decoding techniques which could reasonably be utilised. This diminishes the importance of specialised hardware such as the SDPLL and hence the constellation work reported here. However, current costs of such advanced modems are still high enough to make alternative solutions viable.

4. Carrier Recovery and The Signal Driven Phase-Locked Loop

In the early days of digital data transmission the selection of a modulation format was based largely on its resistance to noise and distortion, the principal use of coherent demodulation being to reduce the effect of additive noise. Phase modulation formats were few and those which did not use differential demodulation (making coherent demodulation mandatory) used large phase spacing between symbols (e.g. 4PSK), rendering them relatively immune to reference carrier jitter. Consequently, the ability to create an essentially perfect reference carrier at the receiver was a reasonable assumption. However, as the use of suppressed carrier data transmission has increased, with a trend toward dense signalling formats with closely spaced phase values, this assumption has become less realistic and is generally invalid for many of the signalling formats now gaining wide acceptance (e.g. 16QAM). This has resulted in keen interest being focussed on carrier regeneration, the problem of providing a very stable reference carrier based on the modulation sidebands, by using knowledge of the modulation characteristics.

A subject closely related to carrier regeneration is the regeneration of a clock signal to mark the boundaries of transmitted symbols. This is generally referred to as bit-timing regeneration or symbol synchronization. The theory of carrier and bit-timing regeneration, while well established for certain specific cases, is not yet well generalised. Still, there are many books and tutorial papers available which cover the subject (e.g. [36,69,70]) and which present aspects of the problem which are widely applicable to cases of practical interest.

It is important to make the distinction between carrier *regeneration* techniques and those which isolate a pilot carrier or tone(s) from the transmitted signal. The former rely on knowledge of the sideband structure for their operation and are needed for a fully suppressed carrier signal. This chapter is concerned only with the regeneration approach and presents a brief overview of techniques used in the general carrier/bit-timing

recovery problem, followed by the results of an investigation into the design and performance of one particular solution, utilising a novel design of phase-locked loop.

4.1. Noise

There are two sources which introduce jitter into the recovered carrier or clock: noise from the channel and *self-noise* produced by the synchroniser, which depends on the pattern of the modulating symbol stream. Jitter arising from the channel may be further divided into that resulting from the additive noise components and that from multiplicative sources. This latter source is often highly structured; it can arise, for example, from power supply noise in line equipment and an approach to combating this source of interference is described by Harvey [71]. One aspect of regeneration which may not be immediately obvious is that a regenerated carrier which is totally jitter free is not always desirable. If the original source of the signal carrier has appreciable jitter, it is advantageous for the regeneration system to track it. This also applies to some effects generated during the signal's passage through the channel and so this aspect is sometimes amalgamated with channel equalisation.

In many applications employing dense signalling formats, the signal to noise ratio is high and the relative bandwidth narrow, so that the systematic (pattern dependent) jitter is the dominant corrupting influence. This systematic jitter can be enhanced by the effects of channel distortions, which may unbalance the relationship of the signal sidebands. While much of this distortion can be eliminated by equalisation, this should be applied with caution as reducing intersymbol interference without careful control of the resulting pulse shape can lead to increased systematic jitter in the symbol timing extractor [72].

An important feature of self-noise which provides the key to its elimination is its phase relationship to the recovered carrier or clock. Unlike uncorrelated additive noise, the in-phase and quadrature components of self-noise have very different spectra and significance. Briefly, the quadrature noise spectrum tends to have a minimum at the carrier frequency, while the in-phase component tends to be essentially flat. This behaviour has been noted in theoretical studies of specific cases [73] and also in practical measurements covering a range of parameters [74]. Since the in-phase component does not contribute to the jitter in a perfectly aligned synchroniser [73] it is clear that any excess bandwidth in the regeneration process will have an exaggerated effect on the

overall performance, due to the slope of the quadrature component. This feature of self-noise highlights the severity of the effect of frequency offset or sideband imbalance on the line signal, which result in cross-coupling of essentially quadrature signals.

The design of the transmission format may also have a significant impact on the performance of the carrier regeneration circuitry. Rhodes [75] has shown that staggering the signalling instants in the quadrature channels of QPSK can give rise to 3dB SNR performance improvement in the synchroniser. Since the detection loss due to inaccurate carrier reconstruction may be quite appreciable, particularly where carrier frequency stability is poor, such modifications in favour of synchronization performance may often outweigh any attendant reduction in resistance to additive noise.

4.2. Carrier Regeneration Techniques

The motivation for suppressed carrier transmission was covered in chapters one and two; briefly, transmission efficiency is maximised by making a signal noise-like. The property which makes carrier regeneration possible for MPSK and other suppressed carrier signals is that they are *cyclostationary* [69]. In particular, signal components mirrored about the carrier frequency are correlated. Therefore, although it is not possible to synchronise to the noise-like line signal *per se*, synchronization to the periodic variations of one of the moments of the signal is feasible. This does not require any knowledge of the information being transmitted, but only the method by which it is impressed on the carrier.

There are three general techniques for implementing carrier regeneration from phase modulated signals. Their common principle is the multiplication of the signal with itself or some derived signal, with the aim of making several distinct input phases ambiguous at the output. In terms of polar coordinates, the effect of multiplying two signals together is to multiply their magnitudes and add their phases. By multiplying any symbol of an MPSK signal with itself the requisite number of times, a product can be created whose phase is always an integer multiple of 2π and therefore indistinguishable from the product of any of the other discrete phase values. This product may then be filtered out and used as the local reference†. The filter may either be an appropriate fixed frequency

† Since this phase reference bears an arbitrary phase relation to the original source, it must either be brought into a known relation by use of a data preamble, or differential modulation must be used.

bandpass design, or a tracking filter (phase-locked loop). Tracking filters have the advantage of operating with narrow bandwidths and at the same time, allowing a wide tolerance on the nominal centre frequency. Their disadvantage is that they are sometimes prone to locking at an incorrect frequency, or occasionally taking an excessive time to lock on to the signal. Static filters do not suffer these locking problems, but require a higher order of frequency stability of the transmitter, receiver and channel.

The three principal approaches to carrier regeneration: use of N^{th} power law nonlinear devices, the Costas Loop and its extensions and the Data Aided and Remodulation Loops, will now be surveyed.

4.2.1. N^{th}-Order Nonlinear Devices

This is the most straightforward technique, but is only applicable to signalling formats where the phase values of the symbols are regularly spaced. It is illustrated in figure 4.1.

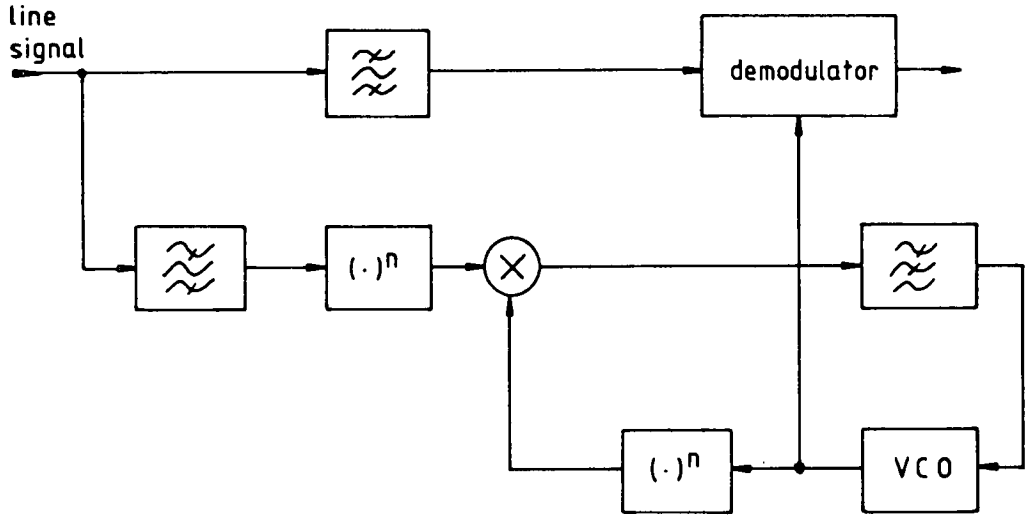


Figure 4.1 The Nonlinear Device Carrier Regenerator with Tracking Filter

The line signal is filtered to remove extraneous noise and passed through a device with a non-linear transfer characteristic. Analysed as a power series, the non-linearity is tailored to enhance the coefficient corresponding to the number of regularly spaced phase values. Thus for 2PSK, a pure square law is desirable, whereas best results for 4PSK are obtained with a transfer characteristic which follows a pure fourth power law. The output from the nonlinear device is passed through a filter to recover the component at

$N \times f_o$, where N is the number of regularly spaced phase values on which the signal is based and f_o is the nominal carrier frequency. Finally, the frequency of the extracted signal is divided by N, to produce a reference which has an arbitrary, but constant, phase relationship to the original carrier.

The main difficulties with this procedure are the production of a suitable non-linearity, particularly above fourth-order, so that sufficient power is generated at $N \times f_o$ to maintain a usable SNR, rejecting noise on the line signal without causing excessive loss of sideband power and rejecting unwanted products of the non-linearity which may lie close to the wanted carrier. The bandwidth of the line signal filter is a critical parameter, as it controls the total noise power which is included in the nonlinear processing. However, it must not be so narrow as appreciably to exclude the modulation sidebands from which the carrier is to be regenerated. The ideal line signal filter for narrowband modulation has been determined by Didday and Lindsey [76].

The chief advantage of this approach is simplicity of implementation; it is the technique almost universally used for symbol synchronization on baseband channels and is widely used for carrier recovery from 2PSK and 4PSK formats. A secondary advantage is that it is not implicitly a feedback procedure, so that problems associated with phase-locked loops (*v.i.*) are not an inseparable feature.

4.2.2. The Costas Loop

The Costas loop, illustrated in its simplest form in figure 4.2, was originally designed for the demodulation of analogue double-sideband suppressed carrier modulation [77]. Referring to figure 4.2, the local oscillator demodulates the line signal in two quadrature arms, whose outputs are then multiplied, filtered and used to control the local oscillator frequency. In the configuration shown, it will lock onto and demodulate a 2PSK signal or any antipodal MASK signal. When locked onto a signal, the output from one arm provides the demodulated data, while the output of the other is zero. Therefore, the output from the error signal multiplier is zero when the loop is locked, otherwise there is an output from both arms and a control signal of appropriate sign is generated, filtered by the loop filter and used to adjust the local oscillator to bring the loop into lock.

The principal purpose of the arm filters is to eliminate double frequency terms from the output of the line signal multipliers. In some applications the characteristics of these

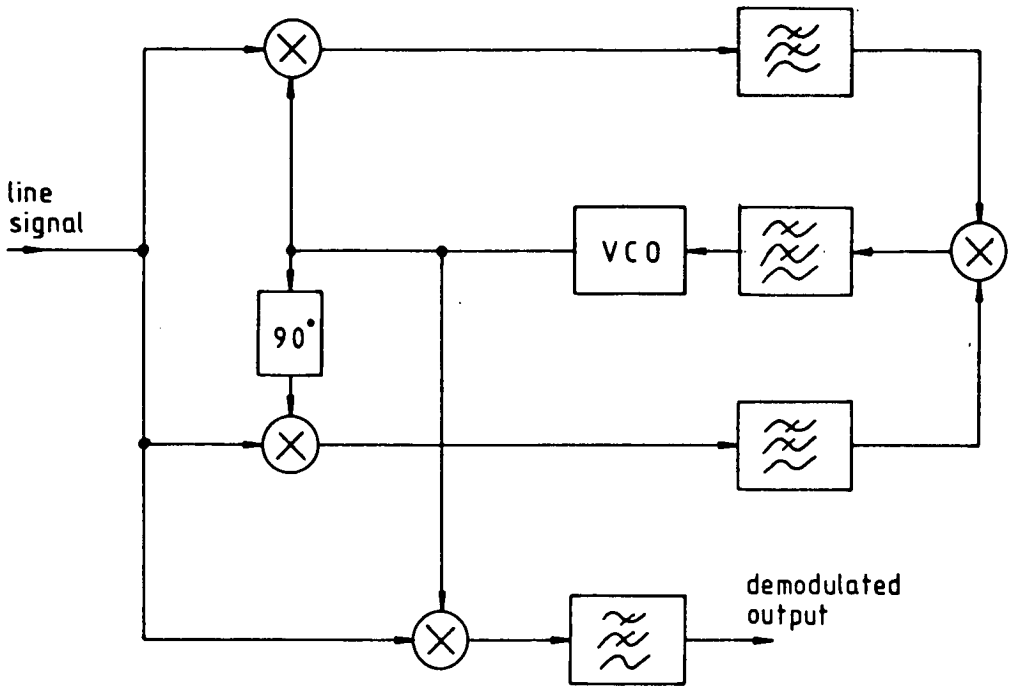


Figure 4.2 The Basic Costas Loop

filters are not suitable for ideal demodulation of the signal†, in which case a separate demodulating arm is employed. The number of arms may be increased to accommodate MPSK signals, so that generally there may be **M** arms fed with the local reference, each one phase shifted by successive integer multiples of $\pi \div \mathbf{M}$, although this number can be reduced if the symmetry of the constellation permits [78].

A comparison of the gaussian noise performance of the Costas loop and a square-law device followed by a tracking filter has been made by Didday and Lindsey [76]. They conclude that the two are equivalent, differing only in the manner by which the squaring operation is performed on the line signal. From a practical point of view, the difficulty of producing a pure square (or higher) law device tends to diminish this similarity, as power loss and the presence of additional unwanted products of the nonlinearity take effect. On the other hand, the Costas loop rapidly becomes cumbersome as the number of phases involved becomes large and has the usual problems of uncertain lock time and

† These filters are in the path of the phase control loop, so their characteristics are constrained by the overall dynamics required of the loop.

false locking associated with feedback systems. A significant advantage of the Costas loop is that it can be used with multi-level QAM signals (e.g. 16QAM) where the irregular phase spacing of the individual channel symbols defeats the nonlinearity approach. Also, technological factors related to the frequency band involved may make it attractive [78].

4.2.3. Data Aided Loops

There are two distinct approaches to data aided carrier recovery/demodulation systems. The first, illustrated in figure 4.3, is to take decisions on the output of each baseband arm of a Costas loop and use the resultant ideal signal to multiply the raw signal from the other arm. The difference of the two products is then used as input to the filter which generates the local oscillator control signal. When the $P(\epsilon)$ of the detection process is of the order of 10^{-2} or less this provides improved rejection of noise as only one input to each of the baseband multipliers carries interference.

The second approach, illustrated in figure 4.4 for a 16QAM system, also takes decisions on the received signal, but then uses the result to modulate the local reference. This remodulated signal is compared with the (suitably delayed) received line signal and the resulting error signal is filtered and used to control the frequency of the local reference. An example of this technique is given by Miyauchi et al. [79] who discuss its application to tandem modulation schemes, where a single carrier is passed through successive modulators whose clocks may or may not be isochronous and to parallel schemes, where the carriers are not required to be isochronous.

The essential difference between this approach and the data aided Costas loop is that the Costas loop derives its control signal at baseband, whereas remodulation derives it at passband. The drawback of all decision feedback systems is the inevitable delay associated with decision processing. This increases the difficulty of achieving suitable dynamic properties for the closed loop [80] and may also aggravate false lock problems.

4.3. Phase-Locked Loops

The phase-locked loop, or tracking filter, is a widely used structure in communications systems and has been analysed in great detail in its analogue, digital and hybrid forms. It is the subject of many books [81,82,83] and the review articles by Gupta [84] and Lindsey and Chie [85] provide a comprehensive survey of the field and have extensive

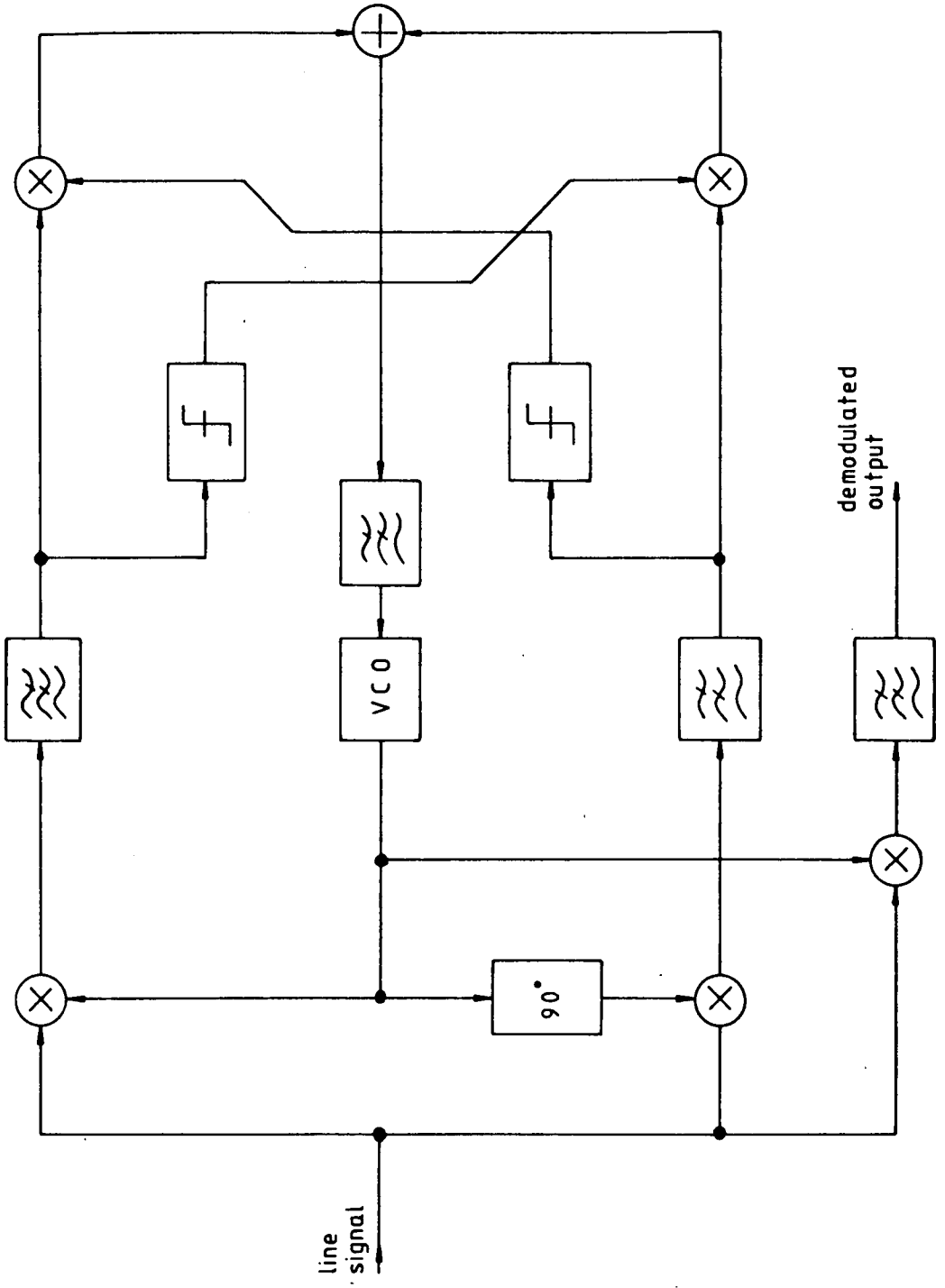


Figure 4.3 Costas Loop With Hard Limiting

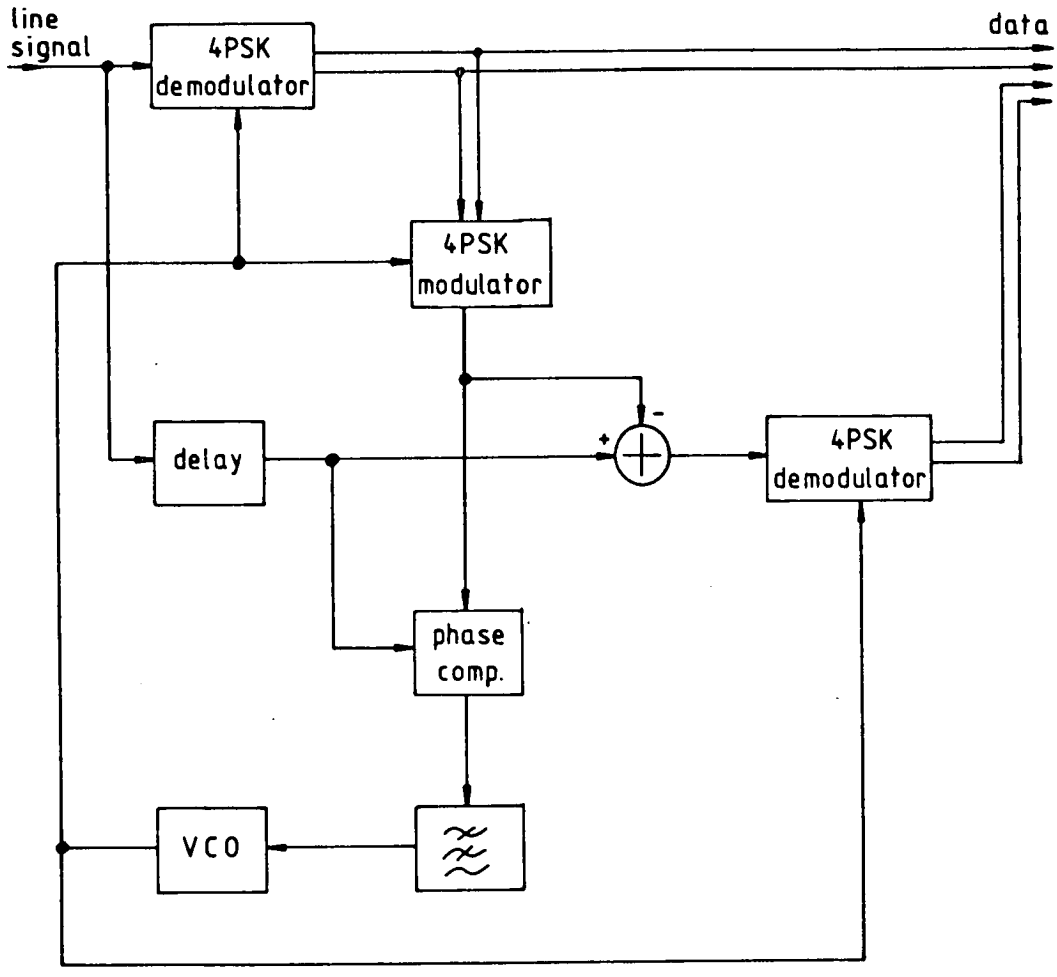


Figure 4.4 16QAM Remodulation Carrier Regeneration

bibliographies.

Figure 4.5 is the block diagram of a phase-locked loop; a feedback control system which attempts to maintain two signals in synchronization. One signal is a local reference, whose frequency can be varied by a control signal, the other is a remote input whose phase is to be tracked. These two signals are applied to a phase detector, generating an error signal which is then processed by a lowpass filter whose output provides the control signal for the local reference. This control signal acts on the local reference to keep its mean frequency equal to that of the remote input.

The exact nature of the phase comparison and hence the error signal, depends on both the phase detector itself and the two input waveforms. Analogue PLLs commonly

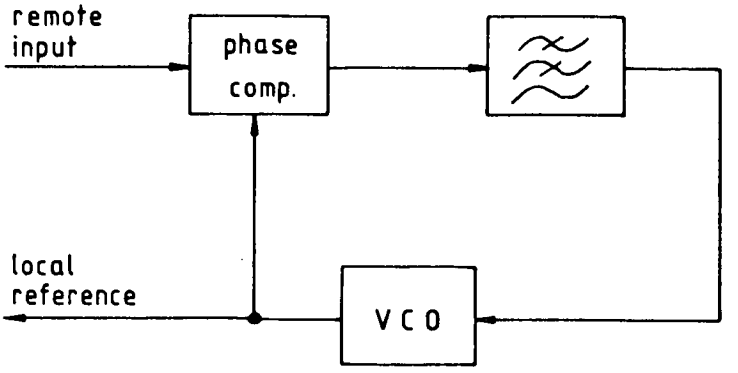


Figure 4.5 The Phase-Locked Loop

employ a linear multiplier as the phase detector, which produces a sinusoidal phase error characteristic or *S-curve* when the two input signals are sinusoidal. For small phase errors, i.e. when the loop is locked and tracking the remote signal, this may be approximated as a linear function. Phase detectors possessing other *S*-curves have been investigated, such as the tanlock detector [86] and the sawtooth detector [87]. These offer advantages such as improved lock range and acquisition time, at the expense of other properties such as threshold performance. Modification of the input waveform itself, in both shape and frequency, has been studied by Stiffler [88]. One notable conclusion of his is that performance may be improved by operating the PLL with a harmonic image of the input signal, the selected harmonic being dependent on the signal to noise ratio. The phase noise energy of the input signal is then spread over a wider frequency range, so that the variance of the output of the loop filter is correspondingly reduced. Hence, when the output from the PLL is divided down to the original frequency a reduction in phase noise is obtained, while the same loop dynamics (lock time &c.) are retained. Alternatively, output variance may be maintained permitting changes in loop behaviour. Locking to a harmonic of the input is just the situation which arises with the nonlinear carrier regeneration technique (§4.2.1).

The loop filter is used to control both the static and dynamic behaviour of the loop and usually falls into one of three classes: straight gain, a single (undamped) pole plus zero or a single (damped) pole plus zero. In addition to the contribution of the loop filter, the phase integrating action of the VCO provides an undamped pole. A PLL is described in

terms of its order and type, the order of the loop being the total number of poles in the system and the type of loop the number of undamped poles. Hence, the first class of filter results in a first-order type I loop, the other two classes give rise to a second-order type II loop and a second-order type I loop, respectively.

The order of a loop determines the extent to which it can track phase variations of the input. A first-order loop can reduce the steady state phase difference between the input and local reference signals to zero if they are at the same frequency, but will leave a phase offset if there there is a frequency difference. A second-order loop can eliminate both phase and fixed frequency differences, but will generate a constant phase offset when trying to track a frequency ramp.

Although third and higher-order loops might be expected to show further improvements in tracking ability, these expectations are seldom realised and many authors conclude that the extra complexity is not justified by the returns [89,90]. This lack of improvement is mainly the result of the increased difficulty of stabilising higher-order loops, particularly where any appreciable phase perturbation is encountered, with the result that the transient response and threshold performance are degraded. Additional poles and zeros are frequently incorporated in other feedback control systems, but usually as compensation for other, uncontrollable, elements in the system [91].

Three parameters which describe the tracking capabilities of a PLL are the seize range, pull-in range and hold-in range. The seize range defines the frequency band within which the PLL will acquire the signal without the phase error going through full scale (cycle skipping). This is narrower than the pull-in range, for which a period of cycle skipping occurs as the frequency of the local reference is brought within seize range of the remote input. In modem applications, particularly those operating in half-duplex mode, the seize range is the important parameter as the time required to lock by cycle skipping is usually too great to be useful. When locked, the PLL will track the remote signal over the hold-in range, which is wider than the seize range and of the same order as the pull-in range. The exact limits of these parameters depend on the particular PLL design.

4.3.1. Digital Phase-Locked Loops

One of the earliest published descriptions of a PLL incorporating digital elements is given by Westlake [92], who replaced the oscillator of an analogue PLL with a sample-and-hold, an analogue to digital converter and a digitally controlled oscillator. A survey of techniques used in digital implementations of the PLL has been published by Lindsey and Chie [85] and covers the varied techniques applied by different researchers in converting the analogue subsystems of the PLL into digital alternatives. The authors of [85] divide DPLLs into four groups, based on the implementation of the phase detector. They note two approaches to the realisation of the loop filter, conventional linear digital designs and statistical designs which do not have analogue counterparts, as well as noting several variations of digitally controlled oscillator. In addition to cataloguing the varied published implementations, the paper outlines the different methods used to analyse the behaviour of the DPLL. In summary, they conclude that the general design process for the DPLL follows and should be a quantised realisation of, that for the APLL. They then note that sampling rate variation can be used to alter loop bandwidth, without affecting certain dynamic properties.

The chief restriction on DPLL design which does apply to the APLL, involves the loop gain which cannot exceed certain limits without the steady state phase error becoming oscillatory. This makes the control of loop gain all the more important in a DPLL. Variation of loop gain can be brought about by some types of phase detector, whose error signal output is a function of input signal magnitude as well as phase.

The effects of quantising and sampling on jitter performance in the DPLL have been studied by Hurst and Gupta [95]. Their results indicate that, for FM demodulation, a three or four bit quantisation of phase error is sufficient resolution for most situations and that the variation of phase noise with sampling ratio has a pronounced knee, such that oversampling by more than a factor of two gives practically no further improvement.

4.3.2. False Locking in Loop Structures

Because of the symmetries possessed by most signal sets, carrier regeneration loops may lock at any one of a number of phases with respect to the original carrier. While this essential ambiguity may be resolved in a modem by the use of agreed data preamble sequences, or rendered irrelevant by differential encoding, carrier regeneration loops

also commonly exhibit lock at spurious phase values and at frequencies offset from the nominal carrier. These false lock points are not usually of concern while the loop is tracking a signal, as the phase variance of the line signal noise process is normally insufficient to cause any significant probability of the loop jumping mode. The main concern is during acquisition, when it is very probable that the loop will sweep through one of these points, some of which may have a lock potential as little as 3.5 dB down on the true locking modes [96].

In order to ensure a unique lock point, the ideal phase detector *S*-curve should go through zero in only one place within the phase interval defined by the intentional ambiguity of the overall system. Distortion of this basic characteristic may be introduced by the nature of the signal constellation via the baseband processing used to generate the loop error signal. In the case of the QASK constellations, the unequal phase separation of symbols results in an *S*-curve having multiple points of inflection, which may be analysed as the superposition of the phase shifted *S*-curves attributable to each distinct symbol successively crossing one or more decision boundaries [97]. The optimum choice of baseband processing in this case, to control the *S*-curve shape and eliminate unwanted axis-crossings, has been studied by Leclert and Vandamme [98] who arrive at an *S*-curve not having the undesirable zero points of [97].

Delay in the feedback path may also give rise to false lock points, quite apart from any stability problems it may cause. This is a good argument for keeping the filters used for error signal processing and data detection separate, when the added complexity can be tolerated. As well as the delay component introduced by the arm filters in, for example, a Costas loop, the band-limiting itself can also give rise to false lock points where the local and remote carriers are separated by integer multiples of half the signalling rate [96]. In this case, the assumption that the error signal has no steady component, except when the local and remote signals are isochronous, is invalidated by the combination of phase modulation and band-limiting and it has been shown that in such cases an increase in the suppression of the false lock points (in decibels) proportional to the factor increase in arm filter bandwidth, can be obtained [96]. For the Costas loop containing a hard limiter in one arm, Simon [99] has shown that eliminating the filter in the other arm can lead to significant reduction in the tendency to false lock, albeit with an increase in jitter on the regenerated carrier.

Digital implementations introduce a further dimension to the false lock problem, as the influence of sampling is added to the problems already mentioned. A sampled data version of the Costas loop has alias lock points at odd integer multiples of half the sampling rate, in addition to those lock points to be expected in its continuous counterpart. Furthermore, according to Simon and Woo [100], a Costas loop which employs hard limiting in the in-phase arm will tend to false lock at any rational multiple of half the sampling rate.

4.4. The Signal Driven Phase-Locked Loop

The rest of this chapter and the next describes a novel structure for a sampled data phase-locked loop called the Signal Driven Phase-Locked Loop (SDPLL) [101]. The principal feature of the SDPLL which distinguishes it from other designs, is its use of the zero amplitude transitions of the line input signal to trigger the sampling of a local oscillator, rather than the oscillator driven approaches of conventional sampled data phase-locked loops. It may be implemented in analogue, hybrid, or digital form and its intended use is in the carrier recovery and demodulation of PSK and APK signals, particularly for applications requiring low complexity. The SDPLL may also be applied to clock regeneration from NRZ binary data streams, which is the application which originally prompted the idea [102]. This chapter provides a basic system description, followed by an explanation of the SDPLL's behaviour in tracking a PSK signal and a survey of the requirements of each of the system components. Chapter five will cover more detailed design and present some results from the investigation of a working digital prototype. A brief report on a hybrid implementation is given by Bandason [103].

4.4.1. SDPLL Configuration

A block diagram of the basic SDPLL is given in figure 4.6 and shows the four main components: a line signal bandpass filter, limiter and latch; a loop filter, a numerically controlled clock and a partitioned counter. This description is appropriate for a digital implementation of the SDPLL. In an analogue implementation, a sample and hold takes the place of the digital latch and the NCC is a voltage or current controlled oscillator, generating a sawtooth (or other appropriate shape) waveform. The *S*-curve defining function of the lower half of the partitioned counter and mapping ROM is not required, being subsumed by the choice of oscillator waveform. The frequency division function of

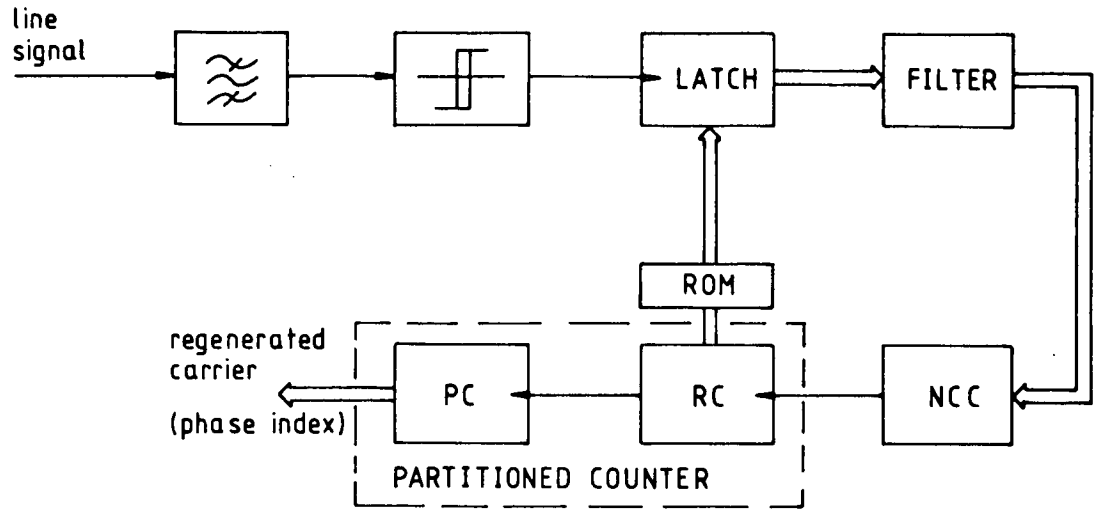


Figure 4.6 The Signal Driven Phase-Locked Loop (SDPLL)

the partitioned counter may be performed by a second, harmonically locked PLL or by digital frequency division.

The basic variation required to accommodate different orders of PSK is a change in the length of the top half of the partitioned counter, there is no need for additional signal paths as in the extended Costas loop. If the phase demodulating function is not required, the top half of the partitioned counter merely serves to divide the loop operating frequency down to the nominal line signal frequency.

4.4.2. Functional Description

The line signal enters through the bandpass filter, before being applied to the limiter, which creates the sampling clock for the loop. Each active edge of the output from the limiter causes the loop filter to acquire a new error signal sample from the low half of the partitioned counter (the ramp counter, RC) which may optionally be transferred via a mapping rom in order to modify the *S*-curve from the sawtooth default. The output from the limiter also triggers a latch on the upper half of the partitioned counter (the phase counter, PC) to generate the detected phase output. The SDPLL may operate on either or both edges of the limiter output, depending on the application. The modified error signal from the output of the loop filter is applied to the numerically controlled clock (NCC), which generates a variable frequency pulse stream to drive the partitioned counter. The overflow pulses from the PC are available as a signal at the mean carrier

frequency of the line signal input. In some implementations, e.g. where a presettable scaler or accumulator rate multiplier (ARM) is used, the NCC and RC may be combined in a single unit.

The combination of PC, RC and NCC forms a variable frequency local oscillator which, when locked, cycles at the carrier frequency of the input signal. The RC and latch combined, act as the phase detector so that the output from the latch constitutes the error signal, with the optional mapping rom being used to impose a particular *S*-curve.

The signal driven sampling arrangement, combined with phase ambiguities generated in the phase detector, enables the loop to reconstruct a carrier from an MPSK signal and to perform the subsequent demodulation.

4.4.3. Synchronization to PSK signals

Figure 4.7 illustrates the mechanism whereby the SDPLL synchronises to PSK signals and since the line signal is hard limited, this also extends to APK signals. The example is considerably simplified in order to focus attention on the basic principle and is for a signal which can adopt any one of four phases (e.g. 4PSK). The output from the combined PC and RC is depicted as a long staircase ramp. For 4PSK, the PC counts modulo-4, thus the output of the RC is depicted as a series of short ramps, four per cycle of the input signal. One possible phase of the input signal has been emphasised and the corresponding zero-crossing instants marked.

For the situation where the limiter produces pulses only at the negative going transitions of the input signal, it may be seen that the RC is sampled at every fourth cycle, in the centre of the ramp. The other three possible phases of input signal would each sample one of the other three RC cycles. Since, as far as the loop feedback signal is concerned, the PC does not exist, all RC ramps are equivalent and the resultant error signal is the same as it would be for any of the other three phases. The four-fold phase ambiguity thus created enables the SDPLL to remain locked to the signal carrier frequency, maintaining a reference phase, as the line signal switches between phases. The PC output which is not fed back around the loop, provides the (arbitrary) discrimination amongst the four phases and may be used to demodulate the line signal. Since the number of states of the PC may be any positive integer, any order of PSK may be accommodated. Interference and lack of synchronization deviate the line signal zero-crossings from the

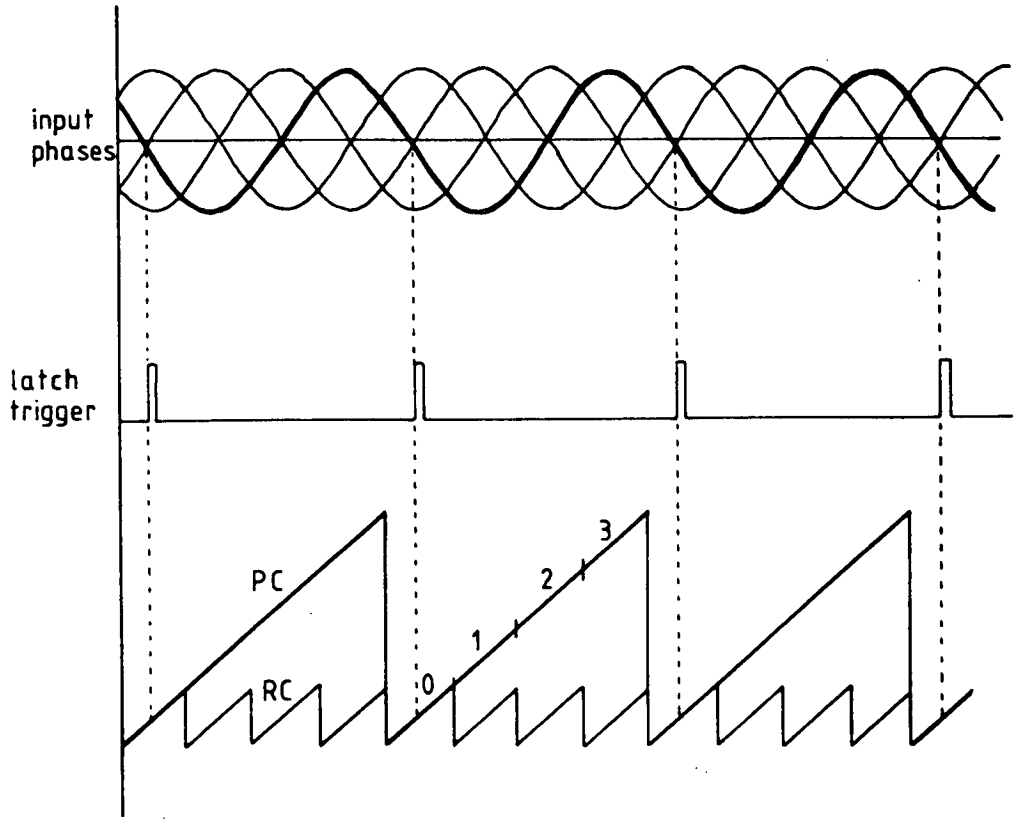


Figure 4.7 SDPLL Synchronization Mechanism

RC zero point, producing an error signal which is linearly proportional to the timing error and independent of the line signal waveform or amplitude.

Implicit in figure 4.7 is the assumption that the binary output from the RC is interpreted as a signed integer. With appropriate feedback polarity the loop will lock such that zero-crossings of the input signal occur half way through a RC cycle, where the error value is zero. Depending on the bandwidth of the incoming signal, it may be desirable to use both polarities of zero-crossing, in order to increase the sampling rate of the loop. This may be done directly when N is even, with no alteration to the RC and a trivial adjustment to the value latched from the PC to compensate for the 180° phase difference. When N is odd, alternate zero-crossings fall around the discontinuous edge of the ramp, producing very large error signals of incorrect polarity. If the period of the RC is halved to provide twice as many ramps per cycle of the input signal and the most significant bit of the error output is excluded from the feedback path, the problem may be avoided.

Alternatively, the polarity of the zero-crossing may be used alternately to invert the most significant bit of the error signal.

When the introduction of modulation is considered two factors arise which reduce the accuracy of the error signal. The transition from one phase to another will, itself, induce a spurious zero-crossing with 50% probability and the influence of these extra samples will depend on the ratio of symbol to carrier frequency. In general, symbol rate will not be correlated with the carrier frequency† and so the symbol transition samples will have their effect diminished, by being spread evenly over the carrier period. Intersymbol interference will also arise and general band-limiting distortion will disturb the location of true carrier sampling instants. Note that passband equalisation which optimises the line signal for data detection is suitable here, but is not necessarily optimum for symbol timing recovery (*v.s.* §4.1 and [72]).

4.5. Design of Loop Elements

The rest of this chapter will cover the primary requirements and options of individual elements of the SDPLL, appropriate to a purely digital implementation. Detailed design aspects of the prototype loop are discussed in the following chapter.

4.5.1. The Bandpass Filter, Limiter and Latch

The function of the bandpass filter preceding the limiter is somewhat more important than for other types of sampled data loop. In the nyquist rate PLL and zero-crossing PLL [85] a filter may be used to reduce the total noise power on the line signal in order to ease the dynamic range requirements of the analogue to digital converter. In addition, for the nyquist rate PLL, the filter is required to ensure adherence to the limits of the sampling theorem. Because the SDPLL uses a hard limiter, performance will be degraded if the total noise power on the input signal is comparable with the signal power, notwithstanding the spectral distribution of the noise. The action of a hard limiter on a sinusoid in gaussian noise is well understood [104], as is the case of two sinusoids [105]. In the low SNR limit the noise suppresses the signal by 1dB without introducing any

† With the increasing prevalence of modems which derive the symbol and carrier sources from a common clock and the introduction of digital trunk telephone circuits which do not introduce frequency shifts, this assumption requires careful consideration for many applications.

phase shift and in the high SNR limit, the signal suppresses the noise by 3dB. While there is no general result applicable to other circumstances, this result is usually taken to be reasonably representative of the general case. It should be noted that in the applications envisaged for the SDPLL the SNR will normally be high, so that some benefit might be expected from employing a limiter. However, these results only properly apply to a limiter *followed* by a bandpass filter and so they should be treated with caution in respect of the SDPLL. Probably the major advantage of hard limiting is the stabilisation of loop gain, which has already been noted to be of particular importance for digital PLLs and which is guaranteed in the SDPLL by the nature of the phase detection process.

An important function of the line signal filter in the SDPLL, which applies to a lesser extent to other types of PLL, is to prevent the loop from locking to signals at unwanted frequencies. Due to the nature of its design, the SDPLL is capable of locking to signals in a large number of quite closely spaced frequency bands and it is likely that in most situations a line signal filter will be mandatory.

The normal transfer characteristic of a limiter exhibits hysteresis, which prevents minor perturbations of a signal at the nominal threshold from causing excessive output switching. In configurations where only a single polarity of zero-crossing is being used to drive the SDPLL, the active transition threshold may be set exactly on zero and an arbitrary amount of hysteresis applied by offsetting the inactive threshold. The adjustment of the active transition is not usually critical, since its only effect is to add a small delay in the path of the remote signal, correspondingly affecting the phase between the remote and local signals when the loop is in lock. When both polarities of zero-crossing are being used, adjustment is more critical. It is important that the two thresholds are balanced about zero. Lack of balance will cause the mean period from positive to negative zero-crossings and vice versa to differ, generating a phase offset and excess jitter which will degrade performance.

The phase discriminator latch, in its simplest form, is just a latch circuit which retains the value of the RC from the instant of the most recent active transition of the limiter. It may also include circuitry to synchronise the effect of the limiter's active transitions with the clock driving the RC and also the filter, if the filter has an independent clock. In conjunction with the RC, the latch forms the phase discriminator of the SDPLL,

generating a sawtooth *S*-curve unless modified by a mapping read only memory (ROM) on the RC output. The sampling times of the latch are dictated by the input signal, which in turn drives the basic timing for the rest of the loop. For a typical line signal, the period between successive samples is extremely variable, as may be seen from the results presented in the previous chapter (figure 3.16) and in particular, there is a significant probability of very short inter-sample periods. These pose problems for the loop filter, which may not be able to accept samples as fast as they are acquired, in which case the latch may be required to provide some queue buffering or arbitration amongst closely spaced samples.

4.5.2. The Loop Filter

As has been noted, the arrival time of samples at the loop filter input is subject to extreme variability. Consequently, the filter design must cope with bursts of closely spaced samples and also deal with the generally irregular spacing of samples. These problems have different significance, depending on the class of filter employed. Three classes of filter relevant to this application are the analogue equivalent filter with variable sampling, the analogue equivalent with constant sampling and the sequential or statistical filter. The two analogue equivalent types cover those filters based on continuous linear pole/zero designs. Since only simple zero and first order designs are required, the rate at which the filter can process samples is primarily a function of the implementation technology, rather than the choice of filter design or hardware complexity. The third type of filter generates an output which does not bear a linear relationship to the input and is not readily producible in analogue form. The hardware is of comparable complexity to the other two types and sample throughput is also similar.

The variable sampling analogue equivalent filter acts on each input sample just once, generating an updated output value for the NCC and then idling until the next update of the phase error latch. Because of this strategy, the clocking rate is variable and consequently the bandwidth of the filter is hard to define. The severity of this problem is partially related to the application as it is a function of the line signal bandwidth, which itself determines the variability of the line signal zero-crossing rate and hence of the sampling. In some telephone modem applications, very wideband line signals can result in raw sample rates near the Nyquist limit for the channel (although the loop error signal remains comfortably oversampled, being of narrower bandwidth). The variability

of the sampling may be reduced by decimation, but this is only suitable for highly oversampled (narrowband line signal) systems, where it is needed least. Another drawback of decimation is the implied delay and filtering which it introduces into the feedback loop; a tutorial review of sample rate conversion treated as a filtering problem is given by Crochiere and Rabiner [106]. Non-uniform sampling is a common feature of sampled data loops, but is usually ignored when analysing their behaviour in the tracking mode [94]. This may be justified when it is considered that the clock source for most other sampled data PLLs is the local reference which, while attempting to track the zero-crossing points of the input signal, is restricted by the loop bandwidth to relatively slow, small variations. During acquisition and other transient states, the variations may become significant and rapid changes in frequency of the local reference may be observed. In this situation, prototype measurements or simulation are often used to obtain insight into the behaviour of the system. In the case of the SDPLL, the sampling may be expected to be highly variable in all circumstances, since the smoothing effect of the closed loop bandwidth does not intervene.

The constant sampling filter type utilises a continuously running sample clock, independent of the phase error latch, to determine when a new sample is to be accepted. Consequently, its parameters are well definable, but problems arise in matching the filter sample rate to the sample rate from the phase error latch. When it is possible to clock the filter at a rate much greater than the arrival of error samples, the phase error latch acts as a first order hold, interpolating samples so that the filter simulates an ideal continuous (analogue) design. The problem here is that the weight given to a sample is then a function of the period to the next sample, which is still variable. Therefore, if an unrepresentative sample were followed by a significant period when no new samples became available (which is quite likely with signals such as C1−5−10) the loop would rapidly move off frequency under its continuing influence. When the rate at which the filter can process samples approaches the mean sample rate from the phase error latch, jitter is introduced by the need to synchronise the transfer and there is the rate overload problem already mentioned.

The third type of filtering, generally possible only for digital implementations, is statistical filtering. A statistical filter is one which tracks some property of the input samples such as mean value, generating an output and possibly changing internal state

when some threshold is exceeded. The delay between samples is not significant and the timing of a new output sample is a function of the input data values. Because statistical filters have no explicit time dependence, they are not affected by irregular sampling. For the same reason, they cannot be described by conventional parameters such as bandwidth or impulse response, which places the design of PLLs using them outwith the usual techniques of control system theory. Two types of statistical filter which have been used in the PLL are the "M before N" and "random walk" designs, which appear to have similar performance [89,107,108].

4.5.3. The Numerically Controlled Clock

Methods of providing a numerically controlled variable frequency source can be grouped into those which use a combination of analogue and digital circuitry and those which use purely digital techniques. The hybrid approach has the advantage of producing a single spectral line output. A fully digital oscillator can only change frequency in fixed increments, so that in general, the instantaneous frequency must be dithered between two or more alternatives in order to generate an arbitrary mean frequency within a given range. One consequence of this dither is that the output consists of multiple spectral lines. In applications where this type of output is not acceptable the NCC may be followed by a narrowband analogue PLL, to produce a single spectral line. Since one aim of this research was to produce a design which could be fabricated as a single integrated circuit, the hybrid approach was rejected and a study made of techniques based on a fixed frequency source followed by digital processing (digital rate synthesis).

Because the basis of the NCC is a fixed frequency oscillator it is relatively easy to achieve high long term stability, while short term stability and spectral purity are a function of the operational details of the NCC. It is also possible to achieve ideal static linearity of frequency control, within the limitations of quantisation. However, in application to a feedback system such as the DPLL, it is equally important that the dynamic response of the NCC to variations in the control word is suitably reliable. Other key parameters of importance in DPLL applications are the fineness of frequency quantisation and the available frequency range, in terms of both range of control and upper limit with respect to the source clock rate. Several different methods of implementing an NCC will now be summarised, each based on one of two underlying principles: the multiplexing of several pulse streams of differing frequency, or the

deletion of selected pulses from a single stream.

4.5.3.1. Pulse Stuffing

Pulse stuffing is the process of adding or deleting single pulses from a pulse train and is essentially a phase (rather than frequency) control technique. The circuitry retains no state information, having no sequential logic except possibly that required to synchronise the control input with the pulse train. Consequently, the basic implementation is small and simple. Pulse stuffing is most often employed where there is little or no frequency offset for which to compensate and is widely used in digital baseband synchronisers. Therefore, the delete and add signals usually have a low duty cycle, so that optimum spacing for low jitter is not critical. The jitter performance cannot be defined without reference to the characteristics of the loop filter, which also controls the linearity and available range of control. The technique has proved popular, particularly for DPLLs employing statistical loop filtering [89,107,108] and is used in the 74297 TTL integrated circuit DPLL.

4.5.3.2. Binary (Counter/Scaler) Rate Multiplication

The binary rate multiplier (BRM) is a circuit which, given an input pulse train of mean rate **R**, generates an output train of mean rate **R***:

$$R* = \frac{M}{N} \times R \tag{4.1}$$

where **M** and **N** are integers and $\mathbf{M} \leq \mathbf{N}$ (usually). **N** is commonly a power of two, for convenience of implementation, but this is not mandatory for a general rate multiplier [109]. The BRM consists of a scaler combined with gating, to produce several non-overlapping pulse trains, each train having a mean rate of one half of the previous train. The value of the control word selects which trains are applied to the 'or' gate which generates the final pulse stream. The circuitry for a six stage BRM is available as a single TTL integrated circuit, the 7497, which may be cascaded.

The BRM gained attention during the development of digital computers, particularly those designed as direct digital realisations of analogue computer architectures such as the digital differential analyser [110]. A detailed study of the BRM has been made by Martin [111], who notes that the greatest drawback to the inclusion of the BRM in

frequency synthesis for signal processing applications is its highly irregular short term output rate. An ideal NCC produces output pulses which are as near evenly spaced as possible, having regard to the period of the (assumed evenly spaced) source pulses. Thus, the peak jitter on the ideal output pulse stream should be no greater than one half of the period of the source stream, but in general, it is greater for the BRM due to the BRM's principle of operation.

A further problem related to this irregularity of output is the behaviour of the BRM when its control word is varied. Over any given period when the control word to the NCC is varying, it is desirable that the total number of output pulses should be equal to the total number of source pulses multiplied by the average value of the (normalised) control word, to within ± 1 pulse. That is, the accumulated NCC output should represent an ideal time integration of the control word input. The BRM does not possess this property, partly due to the rate irregularity already mentioned and investigated by Dunworth and Roche [112]. More importantly, particular combinations of control word transition and state within the BRM may cause several pulses to be spuriously emitted or omitted. An extreme but illustrative example of this is given by Peatman [113], in which the BRM produces no output pulses for a particular coincidence of state and (non-zero) control word sequences. This problem may be avoided if the updating of the control word is synchronised with the cycle time of the BRM state, but this places the same unwelcome restrictions on the control word as are required by the variable length scaler technique (*v.i.*).

4.5.3.3. Multiple Switched Oscillators

This method is a direct implementation of the underlying principle of the NCC mentioned in the preamble to this section, but using several independent reference sources, which may be separated by arbitrarily narrow or wide (and possibly irrational) frequency differences. It has not been widely reported in the published papers on DPLLs, but has been incorporated in a design for a subcarrier tracking loop [114] and an FSK demodulator [115]. In the first of these implementations there are two oscillators, closely spaced and offset on either side of the nominal subcarrier frequency. In an attempt to minimise jitter introduced by the act of switching between the two asynchronous sources, the switching instants are constrained to occur in the periods when zero-crossings of the two sources are closely coincident. Possibly as a side-effect of this constraint, the overall

output jitter has a substantial low frequency component, which was filtered out in the cited application [114]. This filtering reduced the jitter power by approximately 9 dB. Overall jitter power may be reduced by running the basic NCC at several times the required frequency (where this is feasible) and then scaling the output, which smooths the jitter by averaging.

A limitation of the multiple switched oscillator technique is the tradeoff between frequency separation of the two sources (and hence range) and the output jitter. Jitter can be alleviated by employing several oscillators, but this entails a considerable hardware overhead. The second of the cited papers describes a multi-oscillator design and goes into the analysis and design procedure in great detail. It is an interesting approach in that it combines the filter and NCC functions of the DPLL in a single unit. The oscillators are used to clock several equal length scalers, where each scaler transfers its state to the next at a given signal, the first scaler being cleared. The number of oscillators is governed by the number of poles and zeros in the loop filter, while the frequency of the oscillators is governed by the desired output frequency range, the length of the scalers and the loop filter pole/zero locations. The main disadvantage of this technique is the general one applying to all scaler techniques, i.e. limited maximum frequency. In order to reduce quantisation effects the scalers should be long, but each bit added to a scaler halves the maximum usable frequency in a given implementation technology. There is also an inherent cause of spurious locking in the DPLL incorporating this approach, which may only be avoided by specifying a minimum length for the scalers. The overall transfer characteristic for this NCC is linear.

4.5.3.4. Variable Length Scaling

The technique of loading a scaler with an externally supplied value (rather than clearing to zero) at the end of each sequence, is simple to implement and is a standard feature of many integrated circuit scalers. Unfortunately, it has several disadvantages when used to implement the NCC of a DPLL. First, there is the inherent frequency reduction directly proportional to the required fineness of quantisation, which severely restricts the maximum achievable frequency.

Then there is the nonlinearity of the transfer function. The nonlinearity arises because it is the period (rather than frequency) of the scaler which is directly modified by the

control word. Nonlinearity is common in analogue VCOs, but is of more consequence in a DPLL application, where there is generally less tolerance to the resulting variation in loop gain than in the analogue counterpart. While it is possible to minimise the nonlinearity by limiting the range of frequency control, this further restricts the usefulness of the resulting NCC. A further potential source of trouble is that the nonlinearity is asymmetric about the nominal centre frequency and may be a source of bias in a closed loop system.

Finally, any change in the control word only takes effect at the end of each counting sequence, so that a significant and variable latency is introduced. Furthermore, the control word cannot change faster than the cycle time of the scaler without the effect of some changes being lost completely.

Although widely used in other applications, the presettable scaler has not been used commonly as a NCC in the DPLL. One example of its use may be found in a paper by Holmes and Tegnelia [116], but it is more commonly used in frequency synthesiser applications where its control input is not incorporated in the feedback loop [117].

4.5.3.5. Accumulator Rate Multiplication (ARM)

The basis of this type of NCC is described in a paper by Butaev and Romashkan [118], with the modification that the reset action which they describe is eliminated. The operating principle is unrelated to the BRM, having more in common with the presettable scaler. The ARM may be viewed as a fixed length scaler which has a variable increment. Its principal component is a register which accumulates the value of the control word at each pulse of the source clock. Each time the register overflows, which it does in a cyclic fashion such that the excess is retained, an output pulse is generated. Therefore, a control word of zero value results in a zero frequency output, while a full scale control word produces an output at the source frequency; with intermediate frequencies linearly proportional to the control word. Unlike the presettable scaler, the maximum rate of change of the control word is limited by the source clock frequency and unlike the BRM, the output is maximally smooth and is an ideal integral representation of the control word. The fineness of control of the ARM may be increased by simply lengthening the accumulator, without directly affecting the maximum usable frequency. If the accumulation is effected with a straight parallel adder, there is an indirect reduction in

maximum usable frequency due to the increased carry propagation time over the longer wordlength. The latency of such a full parallel implementation is less than one period of the source clock. Where it is acceptable to increase this latency, the maximum operating frequency may be raised by partitioning the addition and pipelining the sub-sections to form a diagonally timed accumulator [119].

4.5.3.6. Virtual Oscillator

The increasing availability of programmable digital signal processors makes the implementation of a NCC exemplified by Garodnick et al. [90] likely to gain widespread acceptance. Rather than implementing a real NCC in hardware, they use an algorithm to determine the state of a software 'oscillator' based on all preceding control inputs. This takes advantage of the even sampling property of their system; an output value is only required at the sampling instants and takes the value +1 or -1. The carrier phase tracking performance of a complete 'software receiver' is described in a paper by Cahn and Leimer [120].

5. An SDPLL Prototype

This chapter reports some of the work done while developing and testing a digital hardware prototype of the SDPLL. The prototype is designed to regenerate the carrier from a ten phase signal and is intended to form part of a C1−5−10N modem.

Since a detailed description of the prototype hardware would be lengthy and largely irrelevant to the general properties of the SDPLL, only salient features of the development are described here. There follows a presentation of test results using an unmodulated sinewave in gaussian noise, with an analysis and discussion of the observed behaviour.

5.1. Design of the Prototype

One of the motives behind the development of the SDPLL is the potential which it offers for providing a low-complexity solution to the implementation of communication links utilising high density, polyphase transmission formats. The principal goal is a hardware solution destined for an application specific integrated circuit, as opposed to an algorithm to be applied to a programmable digital signal processor. The primary constraint on the hardware is that it be simple, with the emphasis of the resulting modem's performance being as much on bandwidth conservation as noise immunity.

While the complexity envisaged for the ultimate implementation of each element of the SDPLL is limited, the prototype itself was designed to incorporate a reasonably large degree of flexibility. For example, the prototype loop filter element was constructed around an eight bit microprocessor. Initially, the filter was built from registers, adders &c., but the problems encountered in reworking it to incorporate even minor changes, led to the adoption of the microprocessor solution for the prototype. It is envisaged that a final design would utilise custom circuitry.

From a somewhat different perspective, inclusion of the ARM as a possible final implementation of the NCC was sanctioned by its design flexibility. Although the

prototype design is relatively large physically, this is due to the limited functionality of the small scale TTL circuits used and does not accurately reflect the design complexity, nor size when totally integrated. The high degree of regularity in the ARM structure makes it a good candidate for parallel or serialised designs, permitting wide-ranging compromise between size, clock speed and latency. The prototype could have been made smaller by using a serial architecture, but the parallel form was chosen in a deliberate attempt to minimise latency and hence reduce the number of uncontrollable factors in the SDPLL.

To summarise the development philosophy, the prototype is constructed to be a test-bed rather than a working model. Individual elements are designed to allow independent control or elimination of particular influences, but the merit of a given design is influenced strongly by the potential for minimising its complexity.

5.1.1. The Limiter and Latch

The line filter on the prototype consists of cascaded second-order highpass and lowpass butterworth sections, this particular design being chosen for its ease of adjustment and acceptable phase linearity. The measured 3 dB bandwidth is 2.76 kHz centred on 1.8 kHz, with a noise bandwidth of 2.39 kHz $\pm$ 150 Hz; plots of the frequency and phase response of the filter are shown in figure 5.1. The output of this filter feeds an LM311 comparator, which has hysteresis variable up to 10% of the input signal level (135mV rms fixed).

When both polarities of zero-crossing are being utilised it may be anticipated that a balanced application of hysteresis to the trigger point of the limiter is desirable, in order to eliminate the jitter which would arise from asymmetrically located positive-and negative-going thresholds†. An effect noticed during the adjustment of the prototype deserves special mention, as it is quite widely applicable. It was seen that the output phase noise did not vary symmetrically about the (single) minimum as the hysteresis balance was adjusted and that the minimum itself did not correspond to equally spaced

† In the case of a phase quantiser having a mid-tread characteristic [121] a small imbalance may prove beneficial; the resulting dither superimposed on the pre-quantised error enabling a reduction in dead-zone and quantisation noise, after suitable filtering. The prototype detector has a mid-riser characteristic, which excludes dead-zone problems.

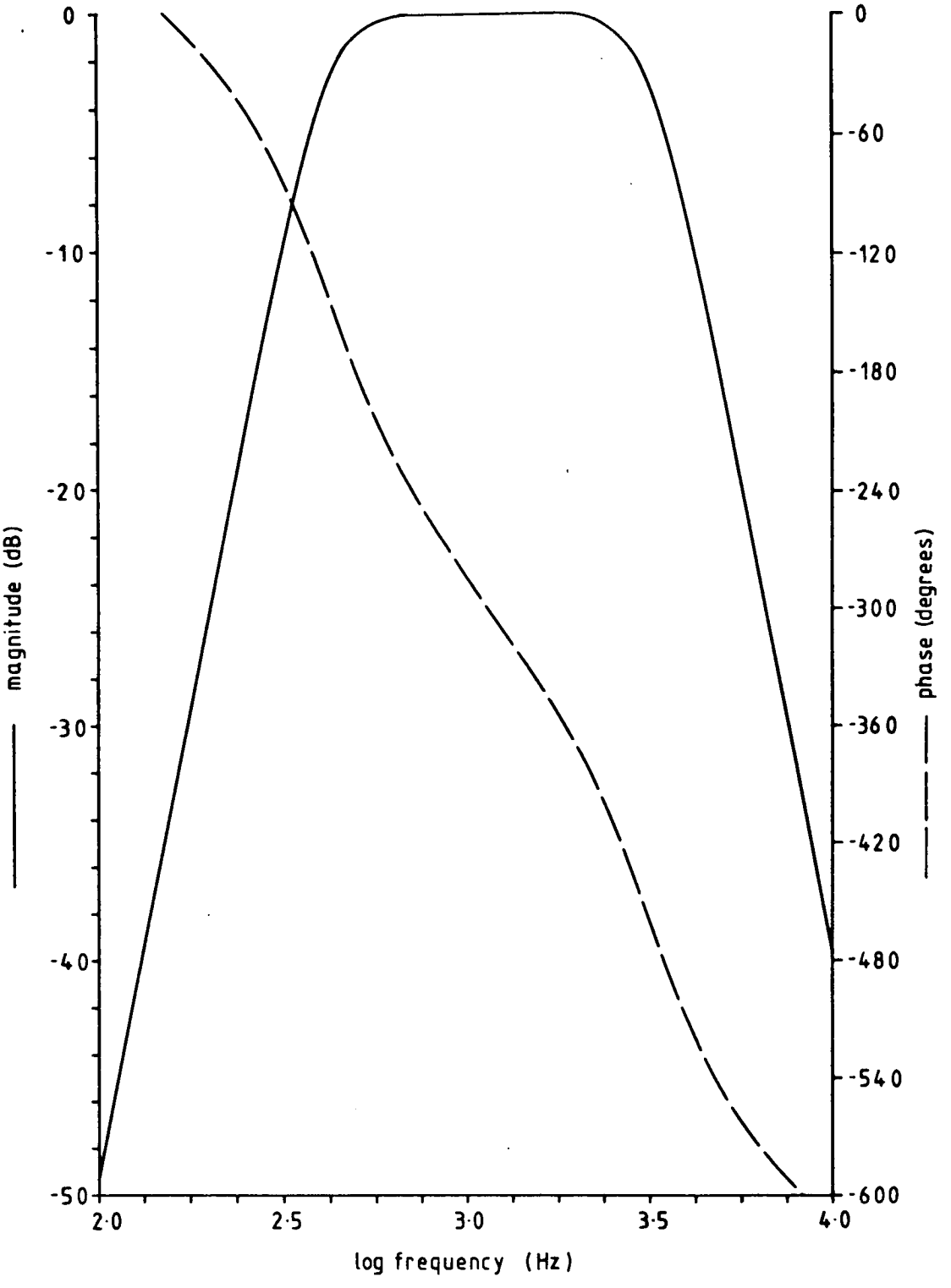


Figure 5.1 Line filter frequency and phase response

positive and negative thresholds. Moreover, the position of the minimum varied with CNR. The clue to this behaviour came when, in the course of investigating the design of the limiter, an extra (inverting) stage of buffering was included before it. This had the effect of inverting the polarity of the bias, despite the signal being purely a.c. coupled. The cause was traced to the noise source, which was based on a maximal length digital pseudo-random binary sequence (PRBS) generator. The binary sequences from such sources when *appropriately* filtered, can produce an output having a flat spectrum over a certain range and an amplitude PDF which is gaussian out to several standard deviations. This signal is quite widely accepted as being a white gaussian noise signal. Such a description is misleading and is widely disseminated without caveat, e.g. [122]. In fact, if the ratio of the "noise" bandwidth to the clock rate of the PRBS generator is less than 20 [123], the amplitude PDF becomes skewed; a fact which has been re-discovered several times [123,124,125,126]. Such skewed noise sources exhibit large spikes preferentially of one polarity and when added to a sinusoid, produce a waveform having a mean period from a positive-going to a negative-going transition which differs from the mean period from a negative-going to a positive-going transition. Thus, for such a waveform the output from a balanced limiter has an uneven mark to space ratio (on average). This gives rise to noise on the phase error signal of the SDPLL, which can be reduced by unbalancing the positive and negative-going thresholds to restore an even (average) mark to space ratio on the limiter output. Note that this problem only arises in the SDPLL when both polarities of zero-crossing are being utilised, but may also affect the behaviour of other systems which attempt to track the mean frequency of a signal whose amplitude PDF is not symmetrical.

An edge to pulse conversion circuit, activated by both edges from the limiter, feeds 150nS wide pulses to logic which synchronises the active edge of each with the clock driving the RC, in order to ensure reliable transfer of the RC state to the sampling latch. In general, the retiming of a totally asynchronous pulse train is complicated by the inability to guarantee setup and hold times at at least one point in the circuit. This can lead to unpredictable behaviour, such as the occasional loss or duplication of a pulse. However, the probability of such an occurrence is low and in the case of the SDPLL, negligible in comparison to the effects of noise on the line signal. Indeed, given an RC with sufficiently low bit skew, or a grey coded RC, the retiming circuit could be

eliminated.

An option which is available on the prototype, but which has not yet been extensively studied, is the application of a blanking signal to eliminate pulses occurring in certain time windows. Specifically, zero-crossings due to phase switching at a symbol boundary may be eliminated once the timing of symbols has been established.

5.1.2. The Loop Filter

The prototype implementation of the loop filter is based on a general purpose, eight bit microprocessor, for reasons already explained. Once again, input and output signal retiming is necessary, as the filter clock is independent of that driving the other subsystems. The use of a common clock for all subsystems was investigated, but was not possible due to the timing specifications of the microprocessor. Signals are exchanged between the filter and the sample latch and NCC through standard parallel interfaces on the microprocessor bus. In addition to the data paths, the microprocessor bus interfaces provide handshaking signals which are used to implement a lock, preventing the contents of the sample latch from being changed until the filter has accepted the current sample. At the same time as the sample latch is updated, an interrupt is sent to the microprocessor to initiate a filter cycle.

A side effect of this read/write interlock is a blind spot in the filter, which causes potential updates to the sample latch to be lost if they occur during the time taken for the filter to service the interrupt caused by the preceding sample. Following a period of quiescence in the filter, the blind spot lasts for approximately 4μS, i.e. 7% of the period of the RC. If the filter is active when an update occurs, the blind spot can last for up to 80μS, i.e 1.4 RC periods. This may give rise to effects associated with the preferential selection of early samples from a closely spaced burst caused by noise, signal distortion or symbol transitions, a problem which is discussed further in §5.4.2.

The filter is programmed for a single pole/zero pair (lead-lag) design†, consisting of a minimally delayed attenuating path in parallel with a second, independently attenuated path incorporating a unit delay with accumulation. In order to stay within the limitations imposed by the ultimate hardware implementation, the values of the attenuation

† N.B. This description is used advisedly; the SDPLL is not an evenly sampled system

coefficients are restricted to powers of two and are effected by shifting operations. This is a not uncommon restriction [90,120] and only has a slight effect on the operability of the loop, due to the generally non-critical nature of phase-locked loop optima [81]. It brings the advantage that the microprocessor is not required to perform time consuming division operations and so lessens the delay in the feedback path. The direct channel parameter is referred to as α and is the number of right shifts performed, i.e. $\alpha = 2$ corresponds to the direct channel passing 0.25 of its input. The integrating channel parameter is referred to as β and is interpreted correspondingly. Arithmetic within the filter is performed to sixteen bit fixed point precision, eight bits being extensions below the binary point. Intermediate values of the integral channel are stored to sixteen bit precision, with results rounded towards zero or hard limited as necessary. The input quantisation is seven bits for the results reported here. This very fine quantisation is a by-product of the RC length required to allow a variable length scaler implementation of the NCC. Although well in excess of that which would be required with an ARM NCC, it serves to eliminate quantisation effects from this initial study. The output precision has similar dependencies and is eight bits, with the LSB at the same significance as that of the input LSB.

5.1.3. The Numerically Controlled Clock

Two options were investigated for this element: the presettable scaler and the accumulator rate multiplier, although the results presented in this chapter relate only to the prototype using an ARM.

5.1.3.1. The Presettable Scaler Implementation

In the initial design of the prototype the presettable scaler was felt to be the preferred method of frequency control, as it can readily be combined with the ramp counter function, minimising circuitry. However, apart from the problems raised in §4.5.3, straightforward truncation of the RC count sequence has the side effect of shifting the output value corresponding to the centre of the ramp (i.e. zero phase error). To compensate for this, the prototype incorporates circuitry to subtract one half of the preload value from the output of the RC, effectively truncating the RC at both ends of its range. This detracts considerably from the original simplicity of the approach, nullifying much of its advantage over the ARM technique. More detailed investigation of different

counter structures may provide a better solution, or alternatively, the phase offset produced may prove to be acceptable in some applications. The prototype used an eight bit counter clocked at 4.5MHz, to give a frequency resolution at 1800Hz of 7.2Hz per LSB for small variations.

5.1.3.2. The Accumulator Rate Multiplier Implementation

The ARM NCC is constructed in full parallel form, using twelve bit registers and a parallel adder (ripple carry) in LSTTL logic, which permits a source clock frequency of 10MHz (10.5MHz maximum). This allows a total post-scaling factor of 2^{-9}, which is divided between a seven bit RC and a two bit RC prescaler. The resulting frequency resolution at 1800Hz is 0.48Hz per LSB, linear over the full range. Reducing the RC length by only one bit, to six bits and hence the source clock frequency to 5MHz, would allow an increase in the accumulator length up to 32 bits, with a corresponding frequency resolution of 0.00000046Hz at 1800Hz. While it is doubtful that such a resolution would be useful, given the limitations of purely digital frequency synthesis outlined in the previous chapter, the example does illustrate how easily the resolution of the ARM technique can exceed that of the presettable scaler. Changing the ARM architecture from a parallel to a pipelined serial one and still using discrete LSTTL packages, the maximum source frequency would be limited to 40MHz, largely by the setup and hold requirements of the register latches. Total integration in a suitable silicon-based technology could raise this to 100MHz, permitting a ten phase SDPLL utilising a four bit RC to operate at carrier frequencies in excess of 500kHz.

5.2. Testing

Perhaps the most significant result of this investigation is the demonstration that a fully digital version of the SDPLL is feasible and that it will lock onto a C1−5−10N signal, particularly since there was considerable uncertainty at the start of this project as to whether the SDPLL would lock-in at all. The results presented in this chapter relate only to a test signal consisting of an unmodulated carrier in gaussian noise, although a modulator for the C1−5−10N constellation, based on the Texas Instruments TMS32010 digital signal processor chip, was developed during a period of work at the premises of Racal Research Ltd. However, time did not permit the collection of a set of controlled results using this source. A digital signal store was also constructed, capable of storing a

3.4 second waveform segment, sampled at 19600 samples per second. Software was written to generate and analyse patterns for this generator and some of this work has already been reported in the chapters on signal design, but once again, time did not permit the collection of modulated signal test results using this facility.

The objective of the carrier in noise tests was to establish the stability of the SDPLL and to provide data for comparison with a mathematical model. The operating state of the loop was set up for the tests at 18 dB CNR by adjusting the two loop filter parameters, α and β, to give a critically damped response and a lock-in time of about 50mS. Phase error samples of the recovered carrier were taken, both when the loop was in lock (steady state) and when reacting to a phase step (transient state), with data being recorded at two values of CNR, 18 dB and 11 dB†. The CNR was changed by keeping the carrier level fixed and varying the noise level, emulating a system with perfect AGC. This was done in order to keep the relationship between the carrier and the limiter constant. The 18 dB CNR was chosen as being at the lower limit of normal operation of a C1–5–10N modem and provides a baseline of performance when testing with a modulated carrier. Input jitter at 18 dB CNR is of suitable magnitude to provide a small–signal fingerprint of the SDPLL. In contrast, 11 dB CNR was chosen to mimic the jitter resulting from waveform distortion and intersymbol interference and as a general comparison with the 18 dB results. The problem of effects produced by phase changes during the transition between symbols (data derived noise) is not directly addressed here, but awaits tests using the signal source now available.

The carrier used for the tests was generated from the digital signal store driven by a crystal oscillator, bringing two benefits: stability and a straightforward means of acquiring phase data to compare with the recovered carrier from the SDPLL. The stability aspect is important, as it proved impossible to find an 1800 Hz general purpose oscillator with sufficiently low phase noise that it did not mask the effects of the SDPLL. The phase data was obtained by arranging for a scaler, driven from the same clock source as (and synchronised to) the signal store, to be sampled by a latch triggered from the carrier recovered by the SDPLL. This is similar to a reversed arrangement of the SDPLL

† Steady state measurements were made at 11 dB, but it proved difficult to obtain transient response data below 12 dB, as the noise induced cycle skipping when the loop was closed close to the $\pm 18°$ phase boundary.

phase detector, except that the variation in sampling period is very much less as it is the narrow bandwidth recovered carrier which determines the sampling instants, as opposed to the wideband line signal. The alternative approach of using a second latch to sample the RC value, sampling being determined by the jitter free reference, was not adopted as it was not readily adaptable to the presettable scaler NCC implementation.

5.2.1. Data Error Estimates

Approximate limits on the error originating from the data acquisition technique may be estimated for long and short term perturbations, by an inductive approach which assumes that the information derived from such data are valid.

The long term error over a given block is dominated by the probability of cycle skip within the block. Assuming that this probability is sufficiently small that the loop may be considered still to be tracking the input, it may be calculated as one half of the phase error PDF at the transition boundary (±18° relative to 1800Hz) [89]. From the phase error distribution presented in figure 5.4 (11dB CNR worst case), the probability of any given (independent) sample showing the maximum phase deviation (±18°) is 0.003. The mean block length used to determine the spectrum is acquired over 0.91 seconds (16K samples at 18000 samples per second). By inductive argument, assuming that the spectrum of figure 5.10 is valid, the bandwidth of the recovered carrier is approximately 16Hz, permitting

$$16 \times 2 \times 0.91 = 29$$

independent phase samples per block. Therefore, the probability of a single cycle skip in any block is

$$1 - (1 - 0.003 + 2)^{29} = 0.043$$

One skip in a block produces a percent timebase error of

$$\frac{100}{18000 \times 0.91} = 0.006\%$$

in that block. The probability of multiple cycle skips within a block (having an aggregate effect greater than one skip) is not significant. With the final spectrum averaged over 160 blocks, it is apparent that on the readings presented, this source of error is negligible.

The estimated maximum effect of short term perturbation is calculated by replacing the phase noise signal with a triangular waveform of equal variance, at a frequency equal to the upper band edge of the noise spectrum. Again, information from figures 5.4 and 5.10 is used, giving a 16Hz waveform with a peak phase deviation of 6.3°. The piecewise-linear phase slew of this waveform is 400° per second, suggesting an error from this source in the region of 1Hz, although as with the previous case it is to be expected that this effect will be diminished by the averaging over many blocks of data. The value of 1Hz corresponds with brief monitoring of the NCC control word, which showed a maximum variation confined to the two least significant bits and represents less than 10% of any of the critical frequencies in the results.

5.2.2. Input Phase Noise Characteristics

After passing through the line filter, the noise added to the test carrier has a gaussian PDF and a flat spectrum for 1kHz either side of the nominal carrier frequency. Because the test signal is a pure carrier, the noise properties of the output produced by the phase error latch of the SDPLL can be determined, when in lock, by mapping the line signal amplitude deviations to phase deviations via the sine function. This is shown in figure 5.2, which illustrates the situation of a segment of sinusoid passing through zero amplitude at an instant corresponding to the central (zero error) state of the RC in a synchronised SDPLL. Localised amplitude perturbations of the sinusoid, caused by additive noise, result in a shift of the zero-crossing instant, producing phase noise as seen by the SDPLL which determines phase on the basis of the zero amplitude location. The probability of any given phase disturbance less than 90° in magnitude depends on the probability of the noise magnitude equaling the sinusoid magnitude at that phase. For small phase deviations, the sinusoid may be approximated by a linear function, the error from this approximation being +1.7% at ±18°. Also, provided that the point at which the input noise PDF maps across the ±18° phase boundary is sufficiently far out from the centre, the fold-over effect on the transformed noise PDF will be negligible. Hence, it is reasonable to model the input phase noise during these tests† as a linear mapping of the line signal amplitude noise, i.e. gaussian and white up to approximately 1kHz, providing that the input noise amplitude is small with respect to the carrier amplitude.

† Note that this argument is invalid when the carrier is modulated and subject to distortion.

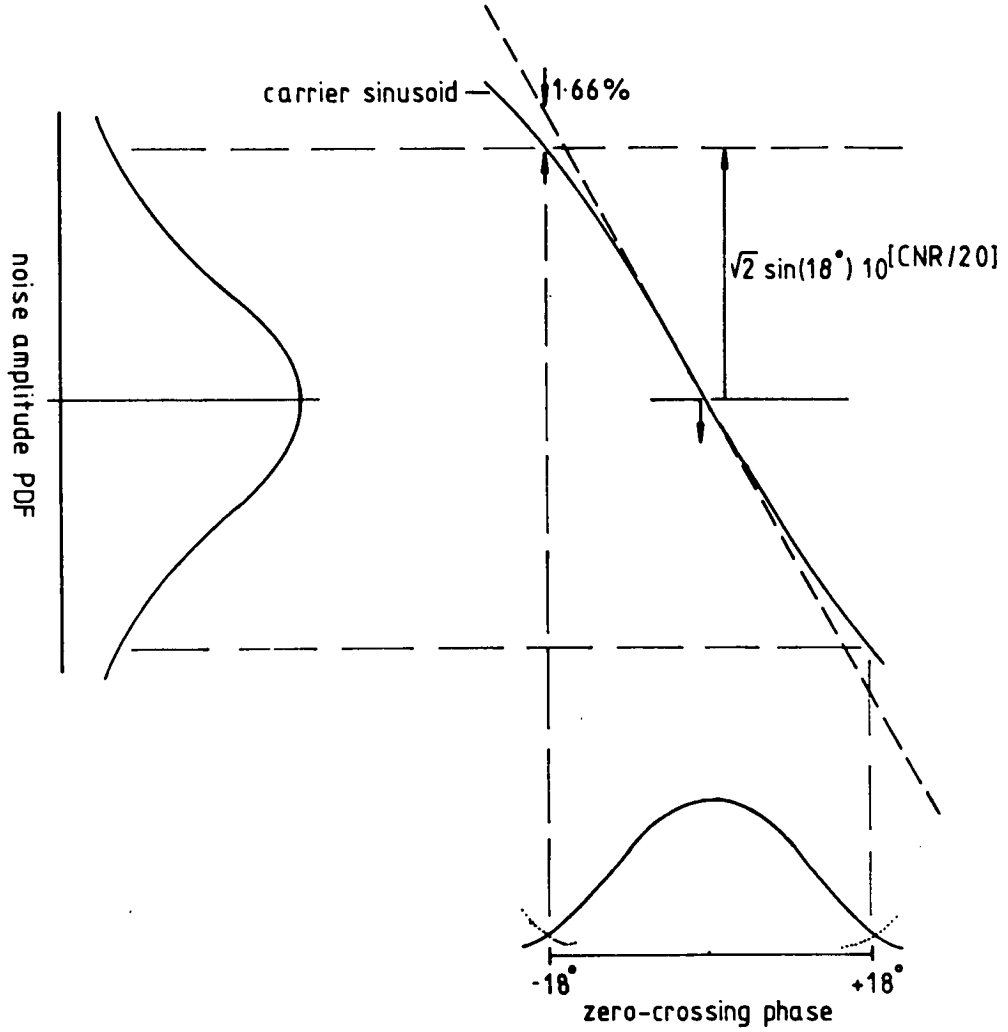


Figure 5.2 Mapping of Line Noise Amplitude PDF to Zero-Crossing PDF

The point at which the input noise distribution is wrapped round on itself as a result of the ±18° boundary of the RC, can be calculated from the expression:

$$BDRY_{\sigma} = \sqrt{2}\ sin(18°)\ 10^{[CNR/20]} \qquad (5.1)$$

At a CNR of 18dB, the gaussian input distribution maps across the phase detector boundary at 3.47 standard deviations, dropping to 1.55 standard deviations at 11dB. The 18dB case is clearly acceptable (the probability of a boundary transition is extremely low), but the 11dB case must be treated with caution.

5.2.3. Output Noise Statistics

As a test of the assumptions made in the preceding section, samples of the loop output noise were taken and analysed. The PDF histograms of these samples are shown in figures 5.3 and 5.4 (normalised to unit area) and represent the accumulation of samples over forty 3.64 second intervals at 18000 samples per second.

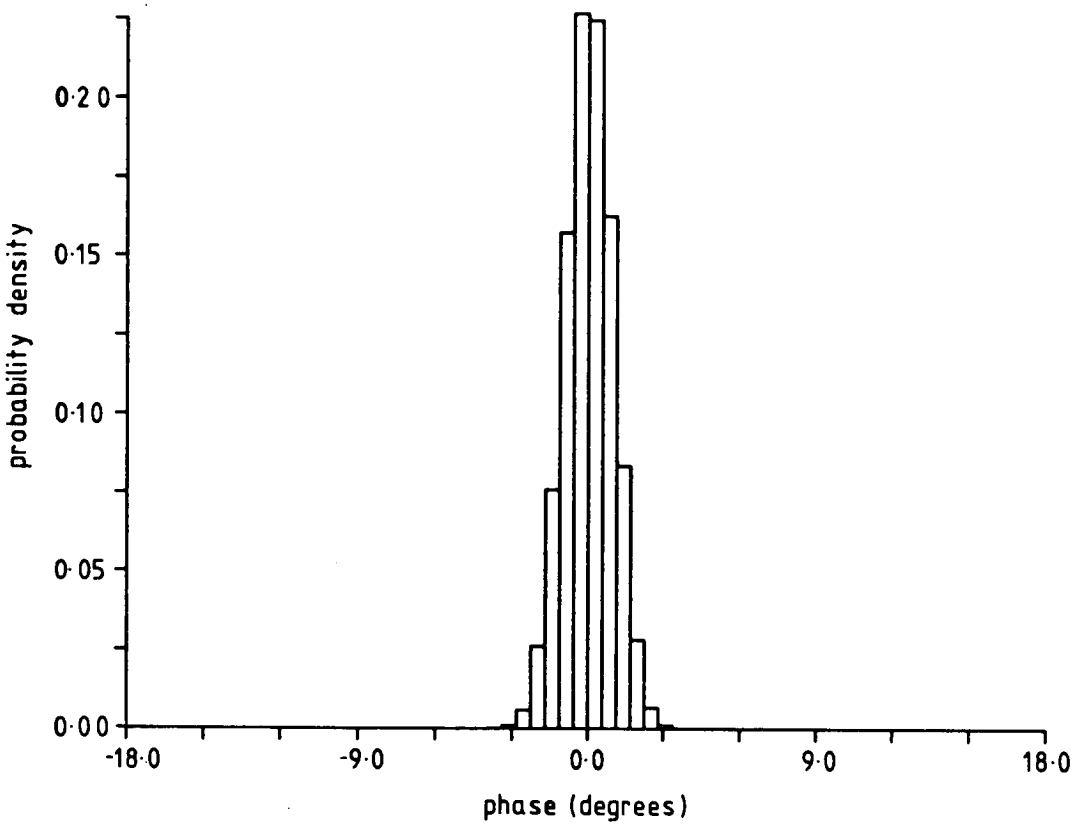


Figure 5.3 Output Phase Noise PDF (18 dB)

In order to compare these PDFs with the normal distribution, the following procedure was adopted. For each empirical distribution, the cumulative frequency distribution (CDF) was computed and compared with the CDF of the normal distribution. This established a correspondence between values on the two abscissae (phase error and standard deviation, respectively) having the same ordinate (cumulative probability). When these corresponding pairs of values, *location on the normal distribution* and *phase error*, are plotted one against the other, the graphs of figures 5.5 and 5.6 result. A (truncated) gaussian phase error distribution plotted by this method generates a straight

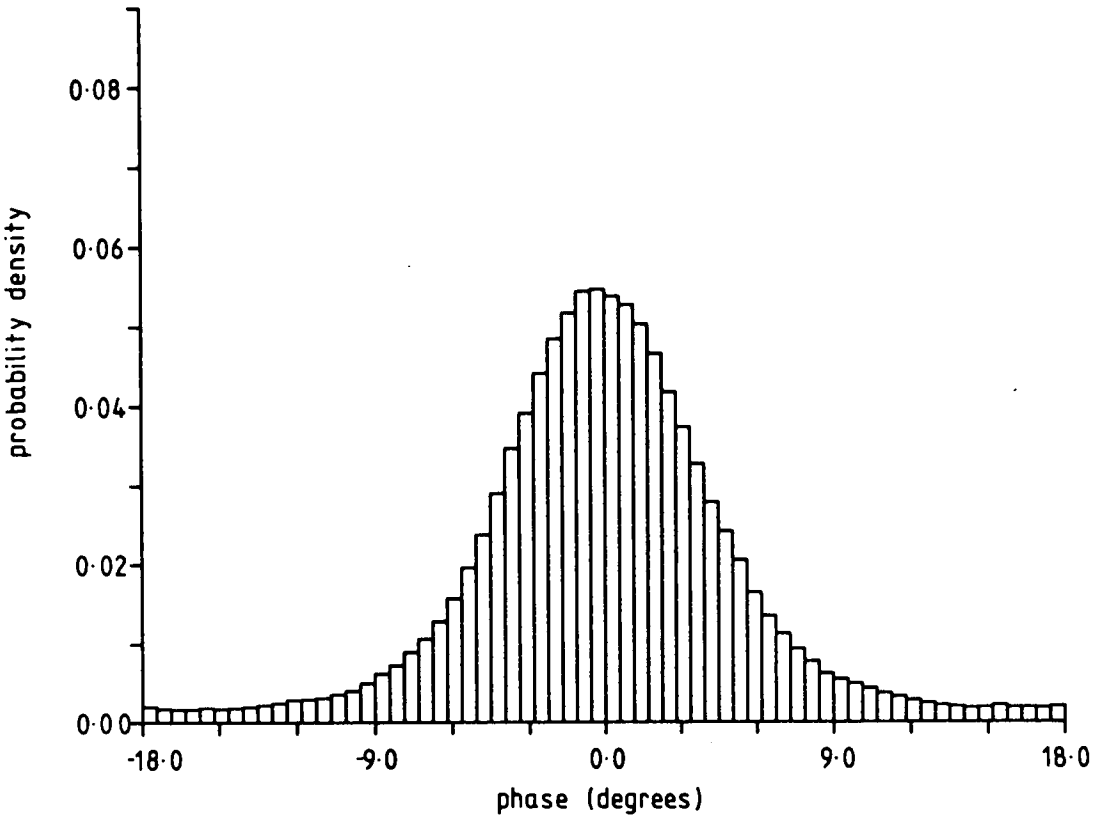


Figure 5.4 Output Phase Noise PDF (11 dB)

line segment, whose slope is the reciprocal of the number of degrees in a standard deviation of the noise distribution. From figure 5.5 it can be seen that at 18 dB CNR a gaussian PDF with a spread of 0.93° per standard deviation is a very good fit to the phase noise on the recovered carrier. The noticeable deviation of the two extreme points may be ascribed to the small number of samples available at these phase values.

Figure 5.6 is more complicated. Between $\pm 6.8°$ ($\pm 1.55\sigma$) it shows a slope of 0.23, corresponding to a gaussian distribution with a spread of 4.38° per standard deviation. The $\pm 6.8°$ breakpoints correspond with the locations on the line noise PDF which cross the $\pm 18°$ RC phase boundary, as determined in the preceding section. Beyond these breakpoints, the slope changes to 0.10; the PDF apparently still gaussian, but having a broader spread of 9.82° per standard deviation. The eight most extreme points show yet another change, but there are insufficient data to estimate reliably details of the change, except to note that the slope increases.

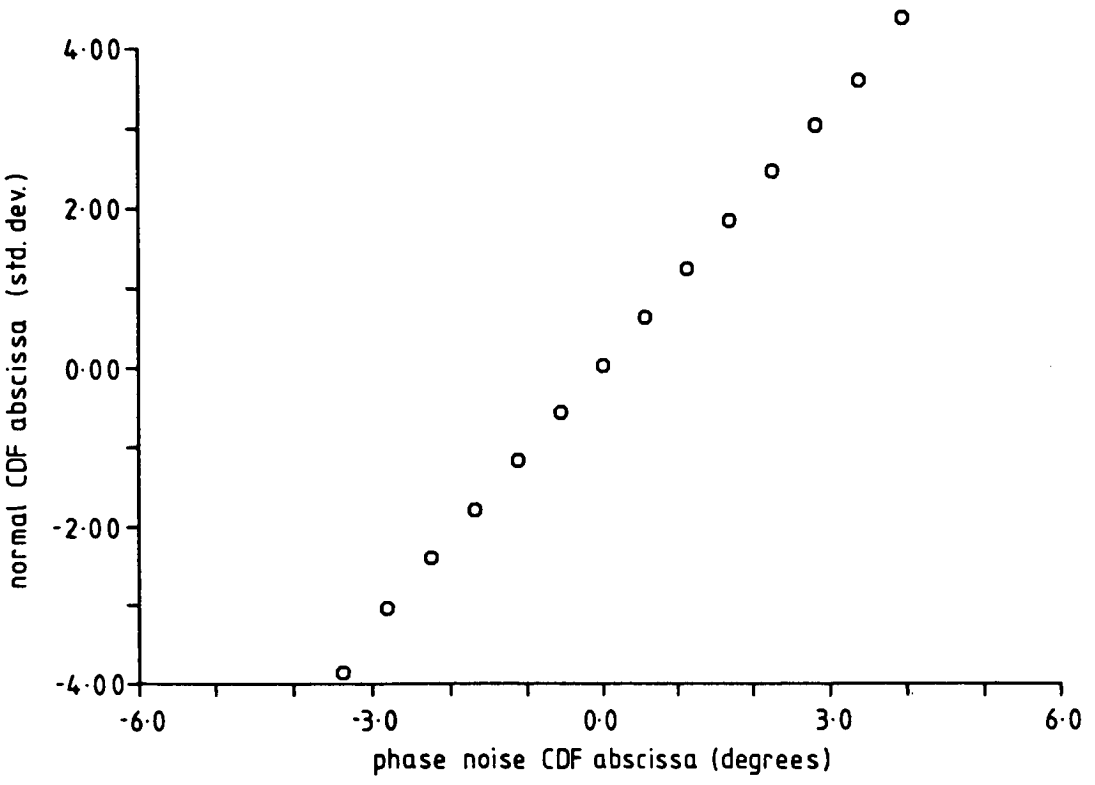


Figure 5.5 Output Noise Cumulative Frequency Ratio (18 dB)

Careful inspection of the second pair of breakpoints shows them to occur close to ±16°. At 11 dB CNR, the line signal noise PDF intersects the ±18° phase detector boundary at 1.55σ, giving it a spread of 11.6° per standard deviation. When rescaled to 9.82° per standard deviation, the image of the boundary intersection lies at ±15.5°, which is close to 16° within the uncertainty of interpretation of figure 5.6. The shape of the curve about the first pair of breakpoints suggests that the measured PDF is the sum of two separate gaussian PDFs; a narrow distribution with a spread of somewhat less than 4.38° per standard deviation and a wider one of 9.82° per standard deviation. One hypothesis which accounts for this combination is that the narrow distribution represents the residual line noise lying within the bandwidth of, and hence tracked by, the SDPLL, whereas the wide distribution represents leakage of almost untracked input noise, which occurs when the loop temporarily loses lock and skips a cycle. The loop does not pass such noise entirely unaffected, its distribution is narrowed slightly by the efforts of the loop to regain lock.

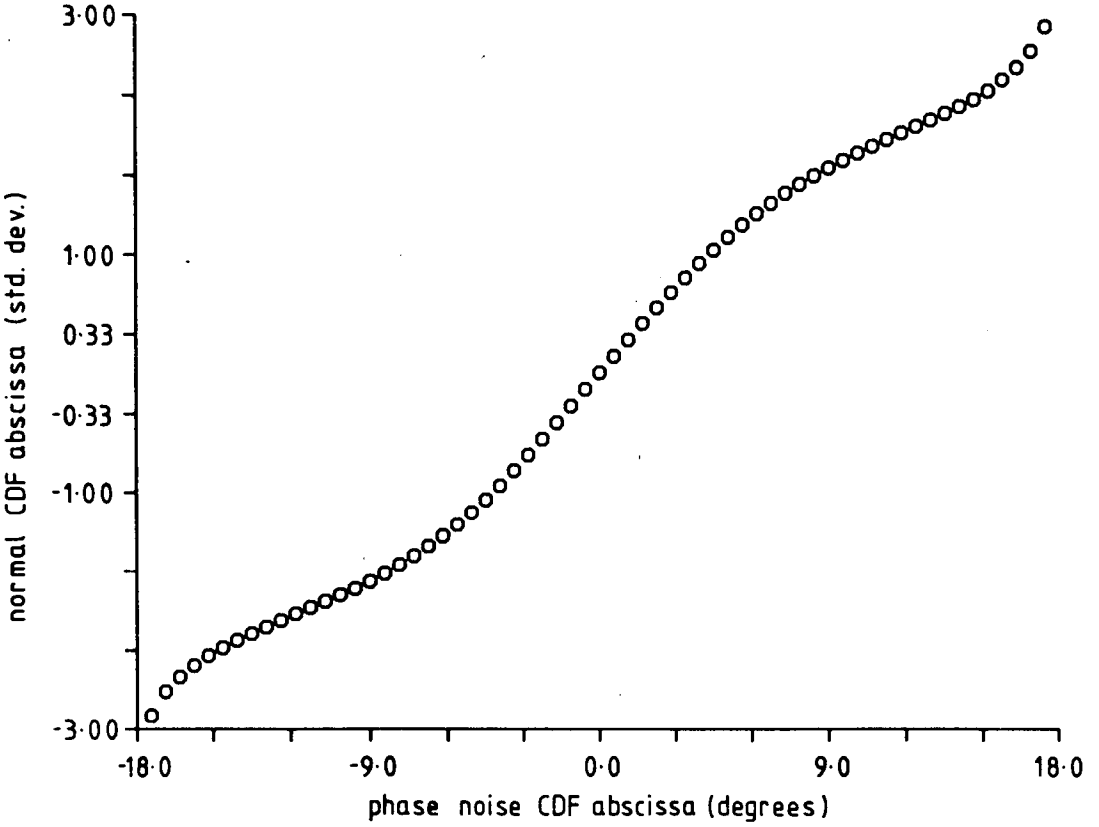


Figure 5.6 Output Noise Cumulative Frequency Ratio (11 dB)

5.2.4. Transient Response

The response of the prototype SDPLL to a phase step in gaussian noise was measured at 12 dB and 18 dB CNR. The value of 12 dB was chosen as at 11 dB CNR it was difficult to obtain a trajectory starting near to 18° which did not include a cycle skip induced by noise. In order to obtain the step response, the loop filter was reset with its output held frozen at zero, while a lissajous figure derived from the RC and the clock driving the signal store was observed. Both signals contributing to the lissajous figure were noise free and oscillating at a nominal frequency of 18 kHz. By adjusting the clock source to the NCC, the loop frequency was offset by 1 part in 10^5 to produce a figure which rolled once in about 10 seconds. When the lissajous figure indicated that the nominal zero-crossing instant of the line signal had just coincided with the $\pm 18°$ phase detector discontinuity, the filter was released, allowing it to begin processing samples and adjusting the NCC frequency to synchronise the loop. Just before taking the first

sample, the filter triggered the data logger to begin recording a block of 65536 samples (at 18000 samples per second) of the phase relationship between the line and recovered carriers. When a block was completed, the recorded transient was displayed on an oscilloscope to confirm that the initial phase error was correct and that the loop had not skipped a cycle. After which, it was accumulated with other recordings to average out the effect of noise. In this way, forty blocks were accumulated at 18 dB and ninety at 12 dB, the averaged transient responses being shown in figures 5.7 and 5.8.

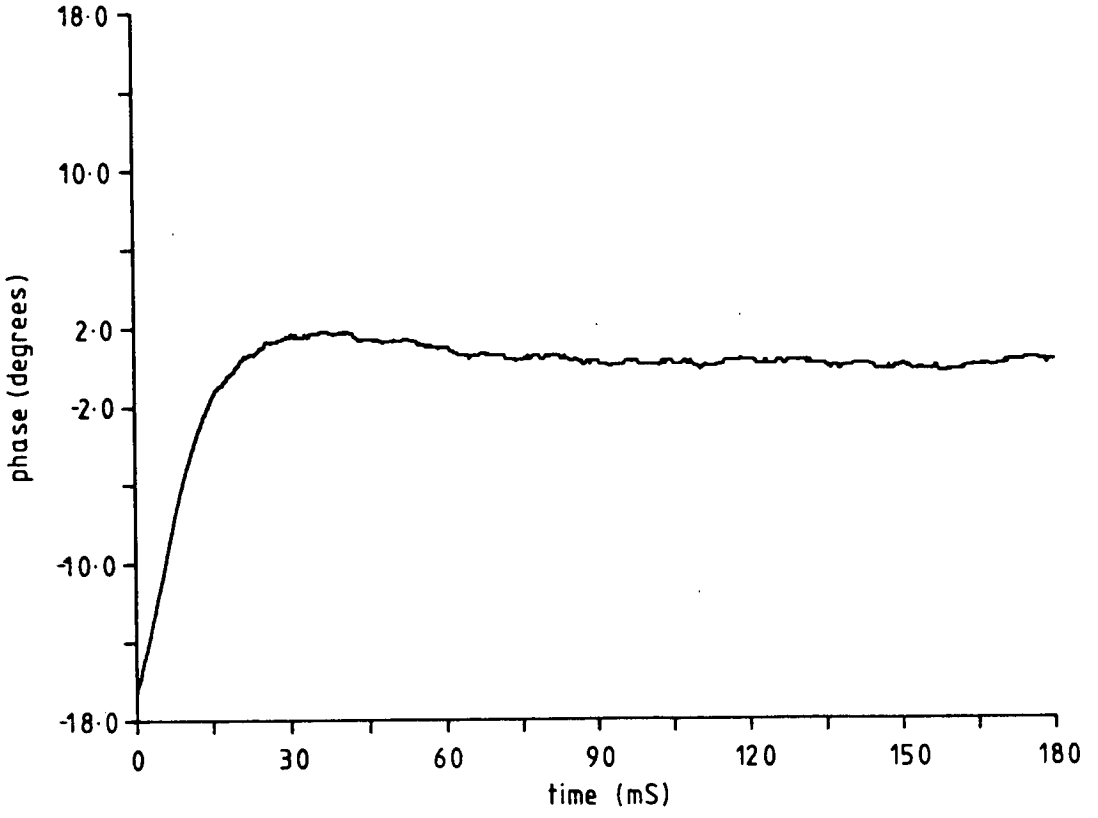


Figure 5.7 Step Response (18 dB)

From the figures, it is apparent that the loop dynamic behaviour is a function of CNR. As the CNR is reduced the loop lock time increases, as does the overshoot and ringing. Measuring the peak overshoot in figures 5.7 and 5.8 gives 9% at 18 dB CNR and 13% at 12 db CNR, with peak times of 36mS and 73mS, respectively. If it is assumed that the prototype is behaving as a second-order system, these values correspond to a damping ratio and natural frequency of 0.62 and 17.7 Hz at 18 dB and 0.55 and 8.1 Hz at 12 dB.

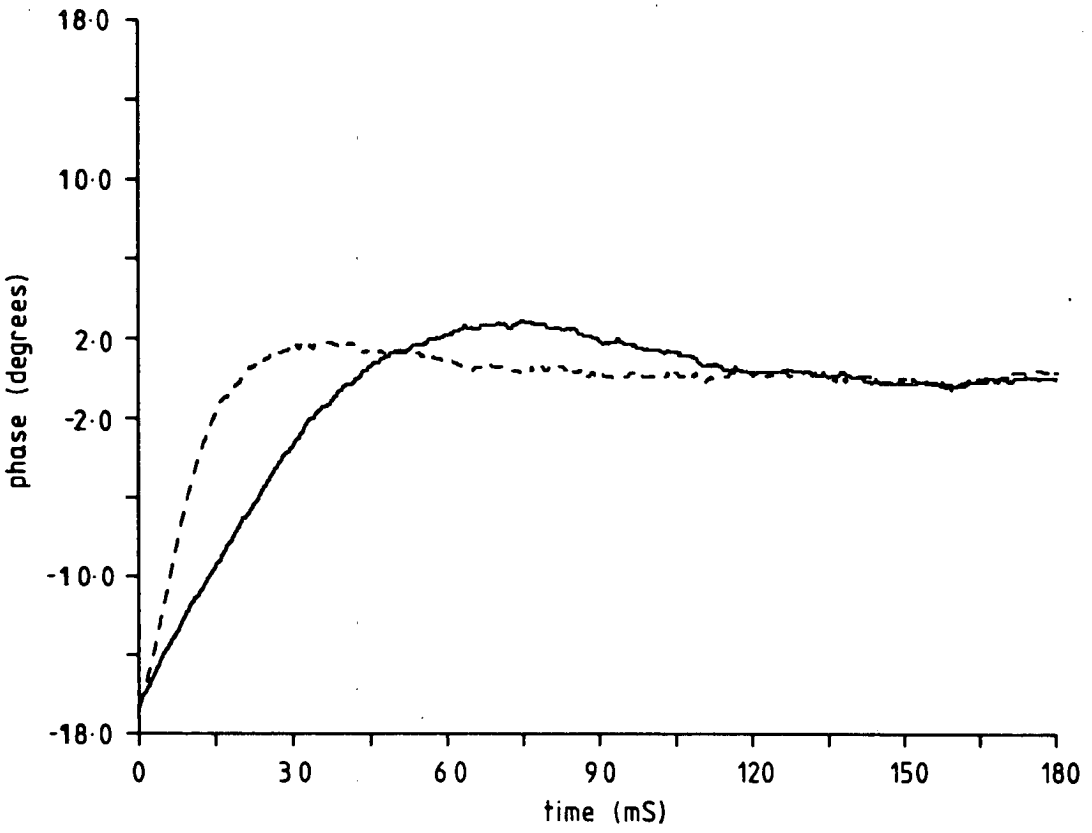


Figure 5.8 Step Response (12 dB)
(*18 dB CNR measurement shown dotted, for comparison*)

5.2.5. Frequency Response

In order to characterise the small signal (steady-state) behaviour of the loop, the spectrum of the phase noise on the recovered carrier was measured. It has already been shown that the input phase noise can be represented by a linear mapping of the input amplitude noise and these spectral measurements are based on an extension of this correspondence to the frequency domain, with a shift appropriate to the carrier frequency. Therefore, the spectrum of the output phase noise is expected to reflect the transfer function of the SDPLL acting on the input noise, whose spectrum is flat out to 1 kHz. Note that, since the input signal is an unmodulated carrier, there is no self-noise component involved in these measurements.

To generate a spectrum, the same time series sample blocks as were used to determine the noise PDF are sub-divided and mapped into the frequency domain using a fast fourier

transform. These raw spectra, 160 at each CNR, are then averaged to produce figures 5.9 and 5.10.

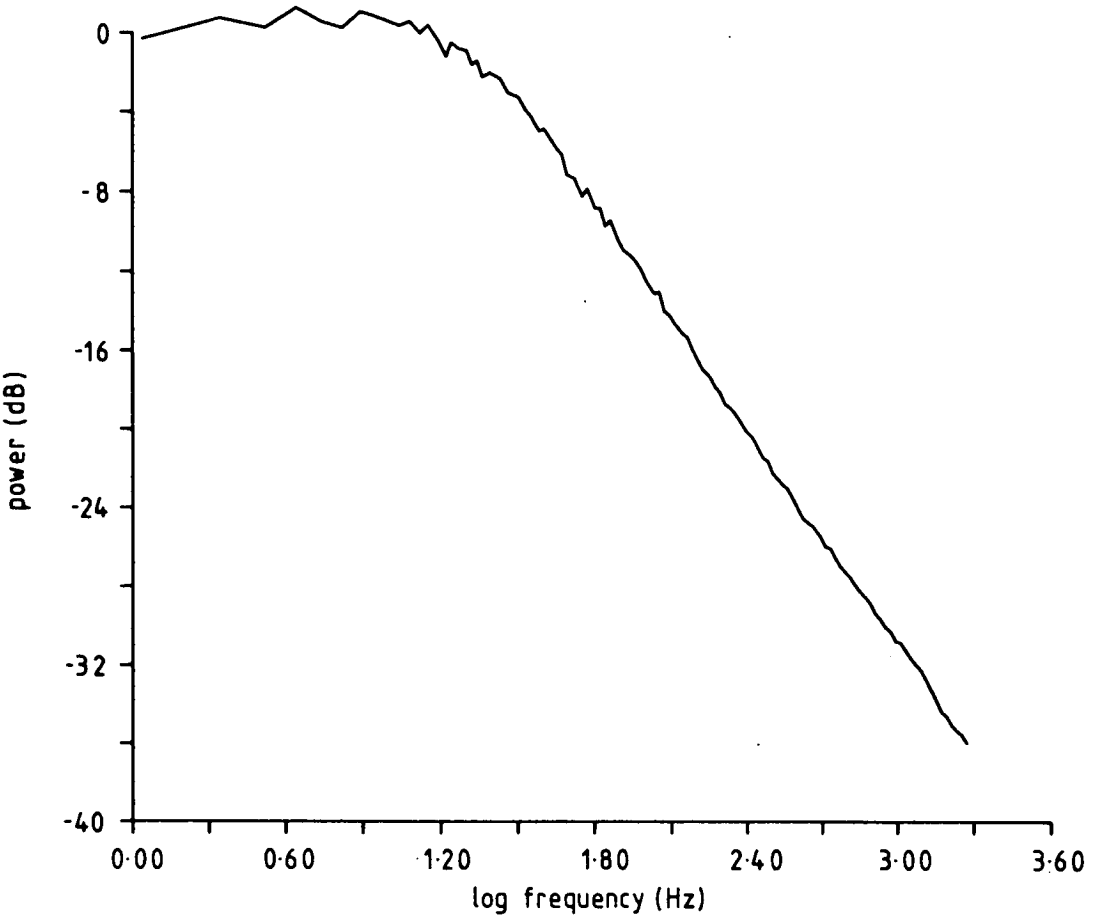


Figure 5.9 Output Noise Spectrum (18 dB CNR)

In contrast to the preliminary interpretation of the transient response, both spectra superficially suggest a first-order loop; they exhibit a single break frequency and asymptotic slopes beyond cutoff of -18.5 dB per decade at 18 dB CNR and -17.5 dB per decade at 11 dB CNR. However, they do confirm the unexpected behaviour at low CNR, where the bandwidth of the SDPLL appears to narrow considerably. The reduction in bandwidth, from 48 Hz at 18 dB CNR to 16 Hz at 11 dB CNR, is sufficiently large to reject the suggestion that it could be explained by inaccuracy in the measuring technique, as estimated in §5.2.1.

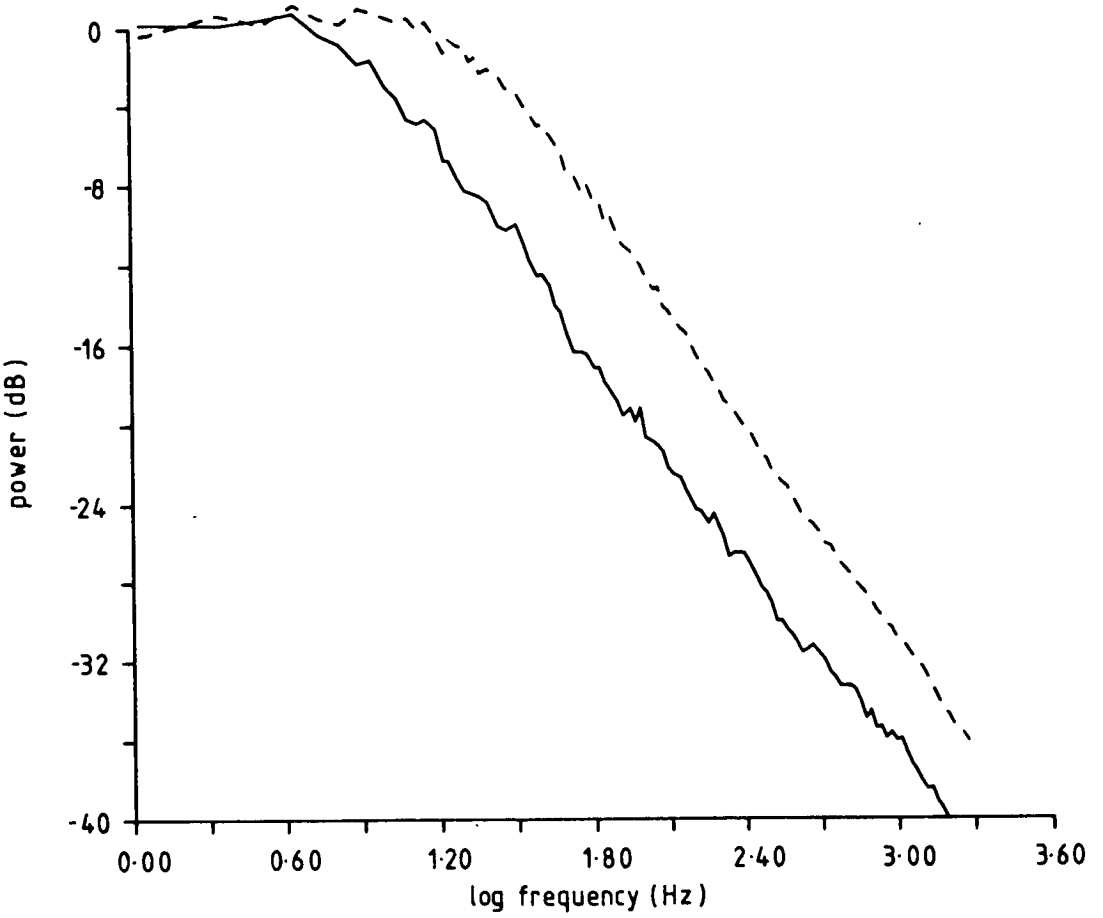


Figure 5.10 Output Noise Spectrum (11dB CNR)
(*18dB CNR measurement shown dotted, for comparison*)

5.3. The Quasi-Even Sampling Model

Although the SDPLL is not an evenly sampled system, it was felt that a Z-plane model of the prototype might provide some insight into its behaviour and explain the apparently first-order noise spectrum. A diagram of the prototype loop Z-transform model is shown in figure 5.11. The model assumes a constant sampling period equal to the mean sampling period and includes a delay element representing the latency of the loop filter. The factor of 10 on the input path and divisor of 10 on the output correspond to the ten-fold phase ambiguity which raises the frequency within the feedback loop.

The other elements are common to most second-order loop models; a factor $\mathbf{K_p}$ representing the scaling of angle to raw error value by the phase detector, a filter represented by a direct attenuating channel in parallel with an attenuated perfect

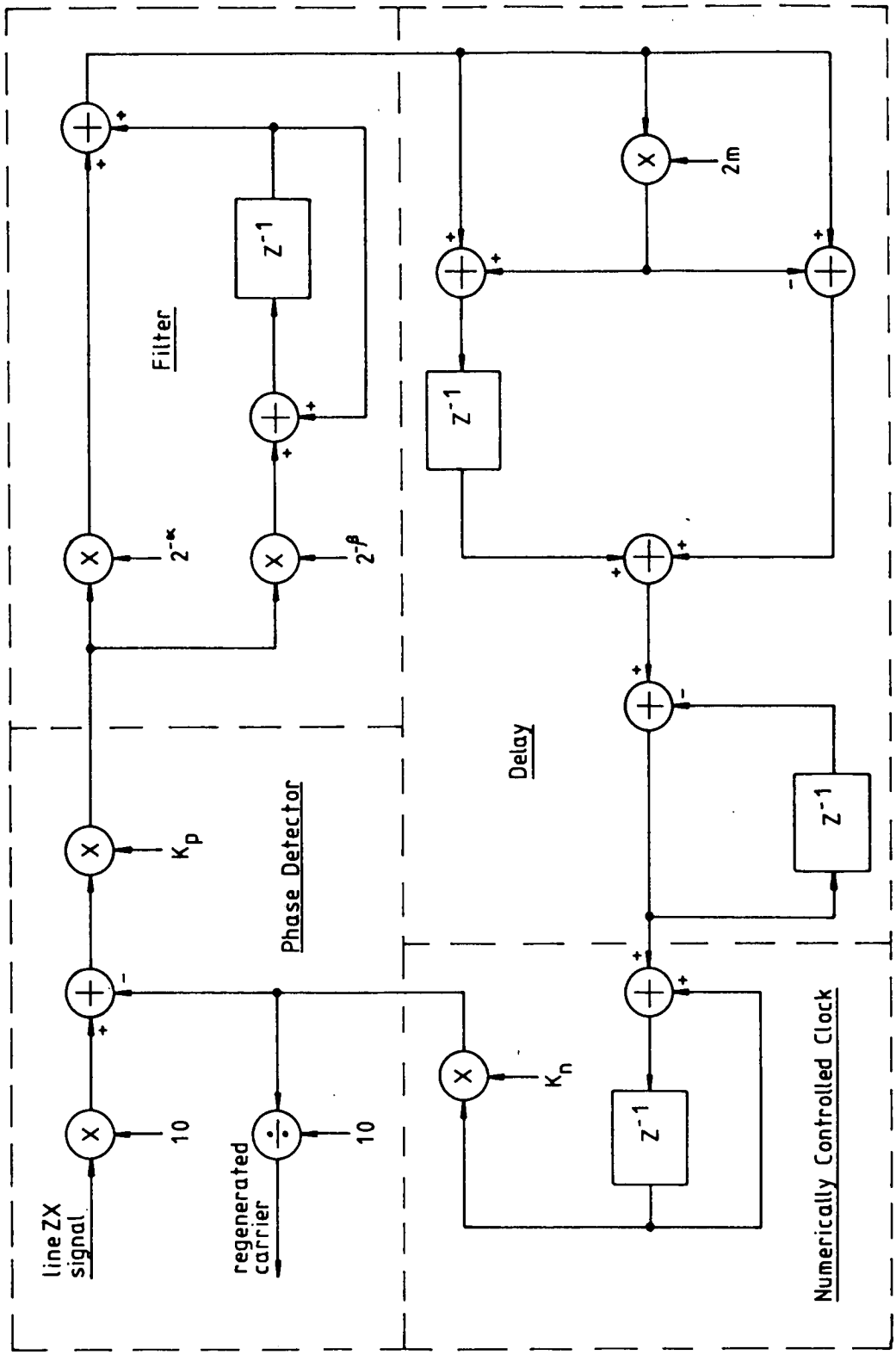


Figure 5.11 (ARM) Prototype SDPLL, Z-transform Model

integrator and an oscillator represented by a perfect integrator in tandem with an error value to angle scale factor $\mathbf{K_n}$.

5.3.1. Loop (filter) Delay

The delay is modelled as a first-order approximation, representing a *fixed* fraction of the *mean* sample period. It does not take account of the extra delay incurred at the filter by secondary samples within a sample burst, which may be significant when the input is modulated or for the 11dB CNR tests. Neither does it allow for the possibly different delay behaviour of other elements in the system, apart from the loop filter, assuming them all to be negligible. In particular, this model would not be appropriate for tests using the presettable scaler NCC implementation.

Let the delay relative to the inter-sample period be:

$$m = \frac{\tau}{T} \ll 1 \tag{5.2}$$

where τ is the absolute delay and T is the mean inter-sample period. Then, in terms of the Laplace transform variable, the delay may be represented as:

$$e^{-mTs} = \sum_{n=0}^{\infty} \frac{(-1)^n (mTs)^n}{n!} \tag{5.3}$$

Truncating after first-order, we have:

$$e^{-mTs} \simeq 1 - mTs \tag{5.4}$$

Transforming the S-plane to the Z-plane via the bilinear mapping:

$$s = \frac{2(z-1)}{T(z+1)} \tag{5.5}$$

produces the approximation:

$$z^{-m} \simeq 1 - mT\frac{2(z-1)}{T(z+1)} \tag{5.6}$$

which may be re-arranged to give:

$$\frac{[1-2m]\left(z + \left[\frac{1+2m}{1-2m}\right]\right)}{(z+1)} \tag{5.7}$$

Thus, the effect of delay in the loop is modelled as a single pole-zero pair and a modulation of the loop gain. For values of relative loop delay **m** less than 10% of the sampling period **T**, the magnitude of the error incurred as a result of this approximation is less than 0.5%. In the prototype the latency of the loop filter, from receipt of a new input sample to emission of a new output sample, is 28μS, which gives a value of **m** = 0.1008 of the mean inter-sample period.

5.3.2. Phase Detector

The phase detector characteristic is represented by the function:

$$P(\phi) = K_p(10\phi_i - 10\phi_r) \tag{5.8}$$

where the values of the input and output phase variables ϕ_i and ϕ_r, are associated with the corresponding signals at the carrier frequency (1800Hz). In the prototype investigated, the optional mapping ROM (figure 4.6) was not included, leaving the phase detector characteristic as a sawtooth function with no nonlinearity to be included or approximated. The constant $\mathbf{K_p}$ scales the phase difference, referred to the *loop* operating frequency (18kHz), into an error value based on the number of states in the RC. For the prototype loop it has the value:

$$K_p = \frac{L_r}{2\pi} = 20.37 \tag{5.9}$$

where $\mathbf{L_r}$ = 128 is the number of states of the ramp counter.

5.3.3. Loop Filter

Leaving aside the uneven sampling, the loop filter is of orthodox design, contributing a pole-zero pair and variable attenuation to the loop characteristic:

$$F(z) = \frac{2^{-\alpha}\left(z + [2^{\alpha-\beta} - 1]\right)}{(z - 1)} \tag{5.10}$$

5.3.4. Numerically Controlled Clock

The numerically controlled clock (ARM implementation) provides a single pure integrator and a scaling factor:

$$\phi_r = \frac{K_n}{(z - 1)} P(\phi)\, F(z) \tag{5.11}$$

The factor K_n is calculated from:

$$\frac{f_x}{f_s}.\frac{1}{N_a}.\frac{1}{S}.\frac{2\pi}{L_r} = 0.0083224 \qquad (5.12)$$

where $f_x = 10^7$ is the frequency of the (fixed) reference source, $f_s = 3600$ (high CNR) is the mean sampling rate of the loop, $N_a = 4096$ is the number of states in the accumulator, $S = 4$ is the post-ARM division and $L_r = 128$ is the number of states of the ramp counter.

5.3.5. The Characteristic Function

Combining the above elements, we obtain the open loop characteristic function:

$$G(z) = \frac{K_p K_n 2^{-\alpha}[1 - 2m]\left[z + (2^{\alpha-\beta} - 1)\right]\left[z + \dfrac{1+2m}{1-2m}\right]}{(z - 1)(z + 1)(z - 1)} \qquad (5.13)$$

This describes a third-order type III system. The poles of the closed loop transfer function are found as the roots of the cubic equation:

$$0 = (z-1)^2(z+1) + K_n K_p 2^{-\alpha}[1-2m]\left[z + [2^{\alpha-\beta} - 1]\right]\left[z + \frac{1+2m}{1-2m}\right] \qquad (5.14)$$

The root loci for this model are shown in figure 5.12 for values of open loop gain in the range zero to one, with the locations of the poles at a gain of

$$K = K_p K_n 2^{-\alpha}[1 - 2m] = 0.0338 \qquad (5.15)$$

indicated. Note that in order to render all its features visible, this diagram is not drawn to scale. From figure 5.12 it may be seen that there is a strong pole-zero cancellation effect, explaining the measured frequency transfer characteristic of the loop. An equivalent model filter for the closed loop response, derived from equation (5.14), along with its frequency and step response is given in figures 5.13 to 5.15.

5.4. Behaviour of the Model

5.4.1. Comparison with 18 dB CNR Measurements

The match between the theoretical (figure 5.14) and 18 dB CNR empirical frequency response (figure 5.9) is very good. The most noticeable discrepancy occurs close to the

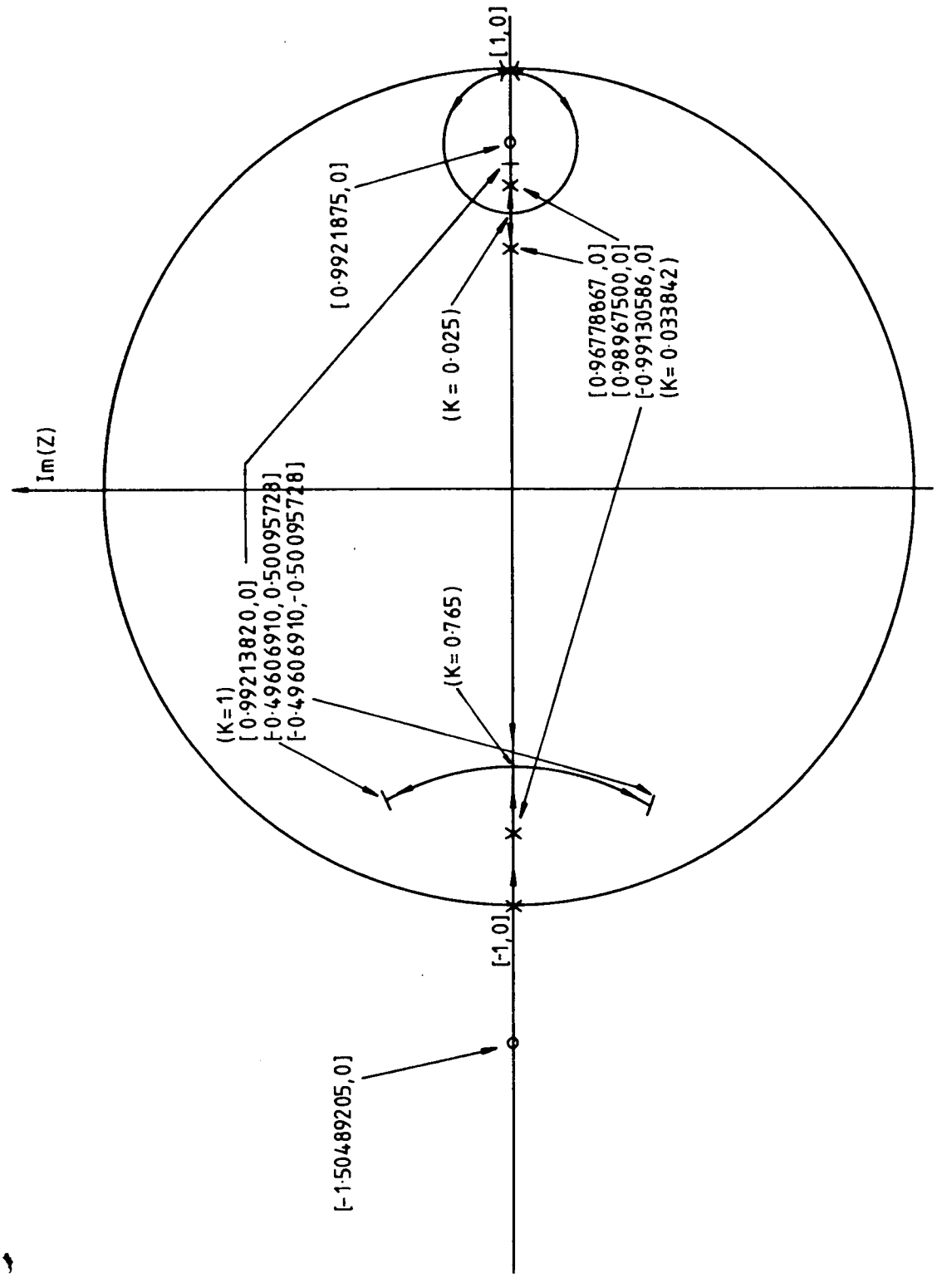


Figure 5.12 SDPLL Model, Z-plane Root Loci (*not to scale*)

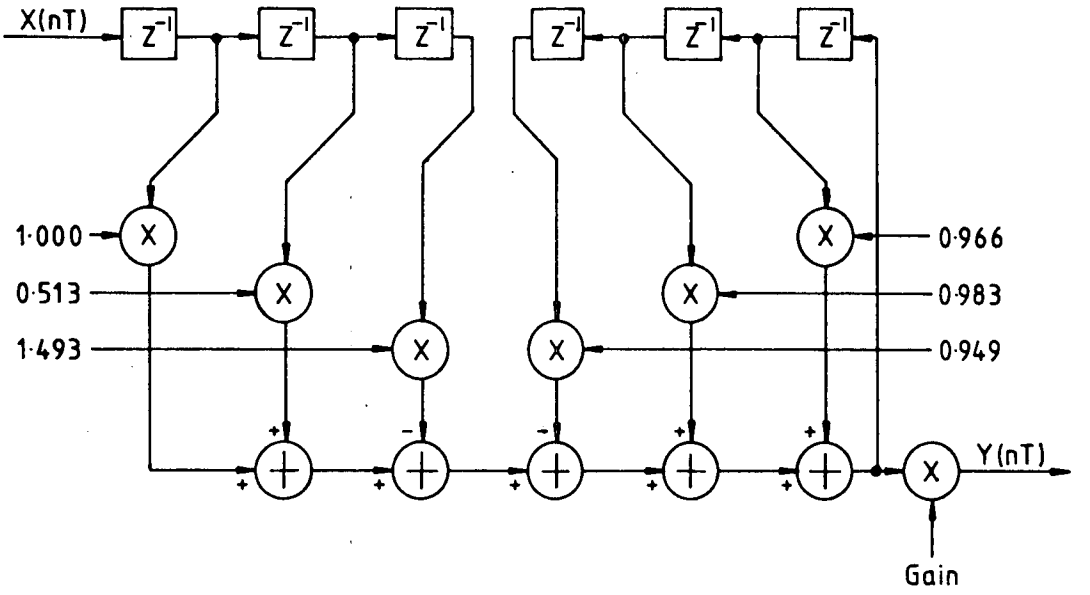


Figure 5.13 SDPLL Model, Equivalent Filter

nyquist frequency, where the shortcomings of the first-order approximation to the delay function result in the spurious prediction of a frequency response peak. It should be remembered that the equivalence between the frequency response of the loop and the output phase noise spectrum is only expected to hold up to 1kHz, beyond which the input noise spectrum cannot be considered to be flat. The predicted step response (figure 5.15) is also in fairly close agreement with the measured behaviour at 18dB CNR (figure 5.7), but exhibits a slightly faster rise time and larger overshoot, the two trajectories conjoining after 40mS. Overall, the model seems to be a good reflection of the prototype at 18dB CNR, although tests with other values of α and β are needed before this conclusion is confirmed.

5.4.2. Comparison with 11dB and 12dB CNR Measurements

The drastic change in the observed behaviour of the prototype at the low test CNRs is not easily explained. Inaccuracy in the data cannot be ruled out purely on the basis of the analysis presented in §5.2. However, observations of the loop error signal, made with a digital to analogue converter and analogue oscilloscope, agree with the output phase step response of figure 5.8. Assuming, therefore, that the data are good, what other explanations are there for the observed results?

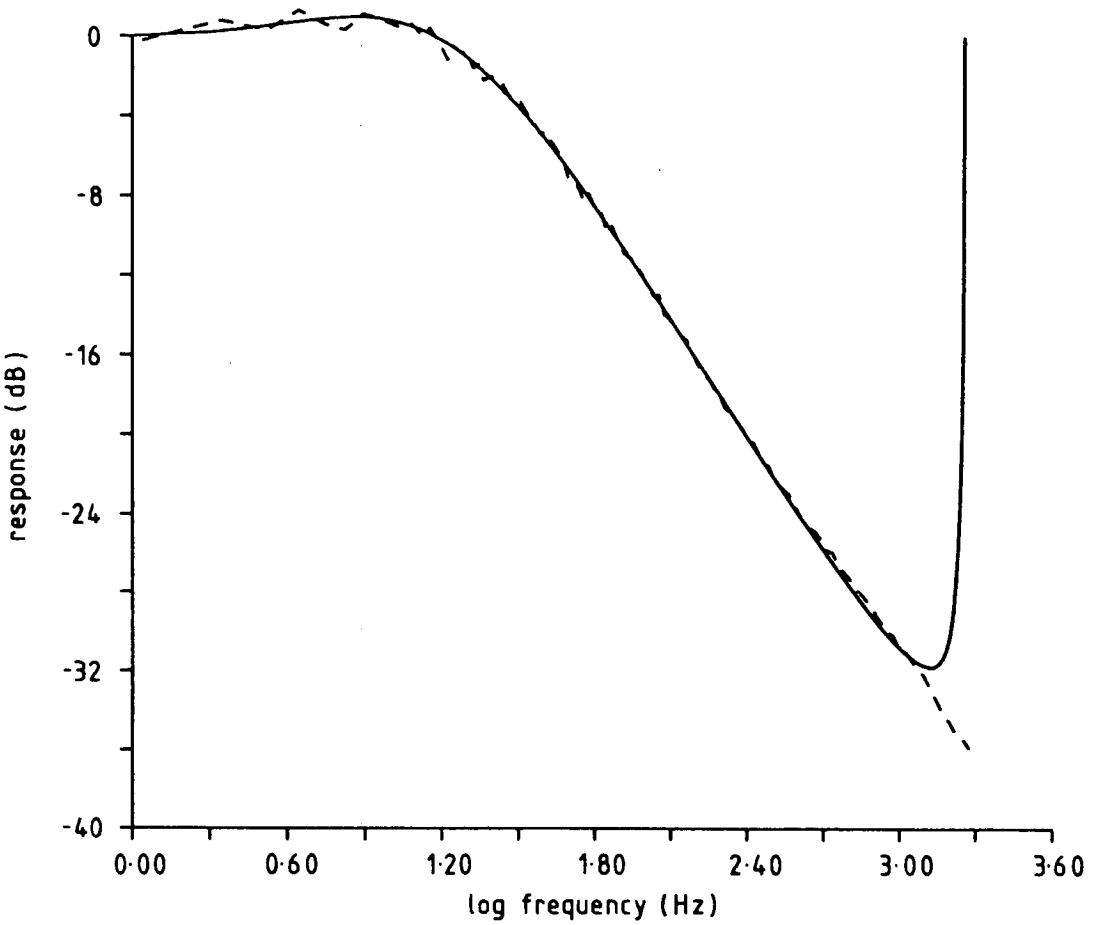


Figure 5.14 SDPLL Model, Equivalent Filter Frequency Response
(*18dB CNR measurement shown dotted, for comparison*)

An immediately obvious parameter to examine is the change in mean sampling frequency. This increases as the CNR is lowered, the raised level of line noise causing occasional triple, quintuple &c. bursts of threshold-crossing on the limiter input. Two factors combine to make it unlikely that this is the direct cause of the observed changes. The magnitude of the change, of the order of 10%, is not sufficiently large and the direction of change is wrong, an increase in sampling frequency would be expected to widen the bandwidth of the loop filter and hence the whole loop, whereas a decrease is observed.

There is an indirect effect of the increase in sampling frequency on the open loop gain, through its incorporation in the NCC scaling factor (equation 5.12) and indirectly through the filter delay parameter **m** (equation 5.7), both of these variations resulting in

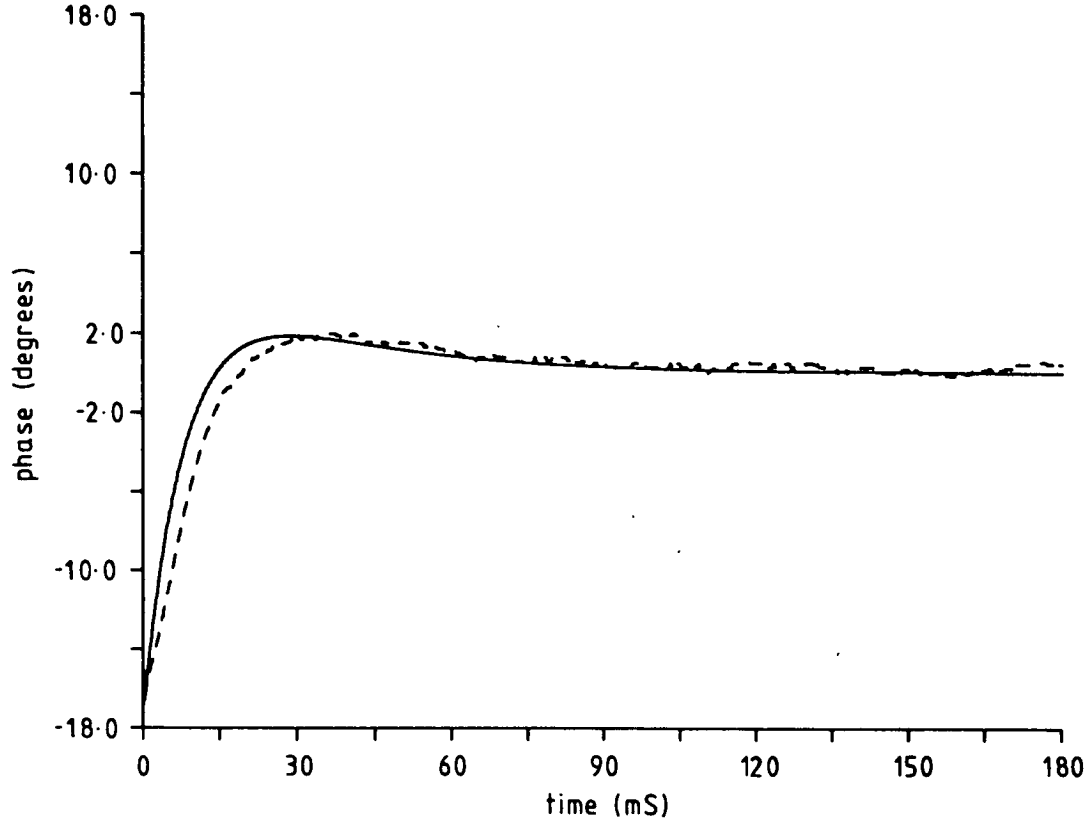


Figure 5.15 SDPLL Model, Equivalent Filter Step Response
(*18dB CNR measurement shown dotted, for comparison*)

a reduced gain. Straightforward manipulation of the open loop gain of the model made it clear that this could only be a partial explanation. A glance at the root loci (figure 5.12) shows that as the gain is reduced from its value calculated for high CNR, the dominant pole does indeed move right, towards the breakaway point, producing a small reduction in bandwidth, but at the same time the partially cancelled pole moves left. The result is a frequency response which narrows slowly, but rapidly becomes more second order in appearance, with steeper roll-off and increased peaking near the break frequency.

This prompted the hypothesis that any variation in open loop gain would also have to be accompanied by movement of the location of the dominant zero, assuming that the general form of the model remains valid. This corresponds to a differential variation in gain between the direct and integral channels of the loop filter (equation 5.9), one plausible explanation for which is based on a feature of the loop filter program.

The delays caused by the loop filter input interlock have been explained in §5.1.2. Of particular importance is the first (4μS) blind spot which covers 2.6° of phase at 1800Hz, corresponding to 0.22σ on the raw phase PDF derived from the line signal. This is sufficiently short to give a high probability that in the event of a sampling triplet being produced by the limiter, the second sample would not occur until after the 4μS blind spot had passed and so would not be lost. The third sample, however, would be locked out unless it occurred more than 30μS (19°) after the second and so the probability of ignoring such a sample is relatively high. Thus, for a triplet of samples, which for this argument are assumed to be evenly balanced about zero phase, only the first two will be seen, the second value having a much smaller magnitude than the first, on average. The loop filter algorithm reads the first sample for use in the direct channel calculation, releasing the interlock after the 4μS interrupt latency. However, the value obtained is not stored for use in the integral channel calculation. The integral channel value is obtained by a second reading of the sample latch, which occurs 30μS after the direct channel reading and is therefore quite likely to pick up the value of the second sample of a triplet. After a further processing delay of approximately 44μS, the filter cycle is completed and the interrupt raised by the second sample is acknowledged. Assuming that the third sample was locked out (and hence lost) the filter will go through a second cycle, but this time both channels will process the same (second) sample value.

Overall then, the direct channel has processed one large and one small sample, while the integral channel has processed two small samples. Hence the apparent gain of the integral channel, relative to the direct channel, has been reduced. The effect of this on the Z-plane diagram is to move the dominant zero to the right, closer to the unit circle, resulting in a general compression of the root loci in that region towards the point {1,0}, with an associated reduction in bandwidth. Heuristic application of this to the model, in conjunction with a variation of open loop gain, produced the frequency response shown in figure 5.16, which corresponds with the 11dB CNR observations (figure 5.10). The reduction in open loop gain was by a factor of 0.42, with the difference in gain between the direct and integral channels increased by a factor of 4. Having regard to the number of minor variations possible in the foregoing hypothesis and to the combinations possible for the distribution of the two variations postulated, it seems inadvisable to continue speculation. Furthermore, although a match to the frequency response has been found,

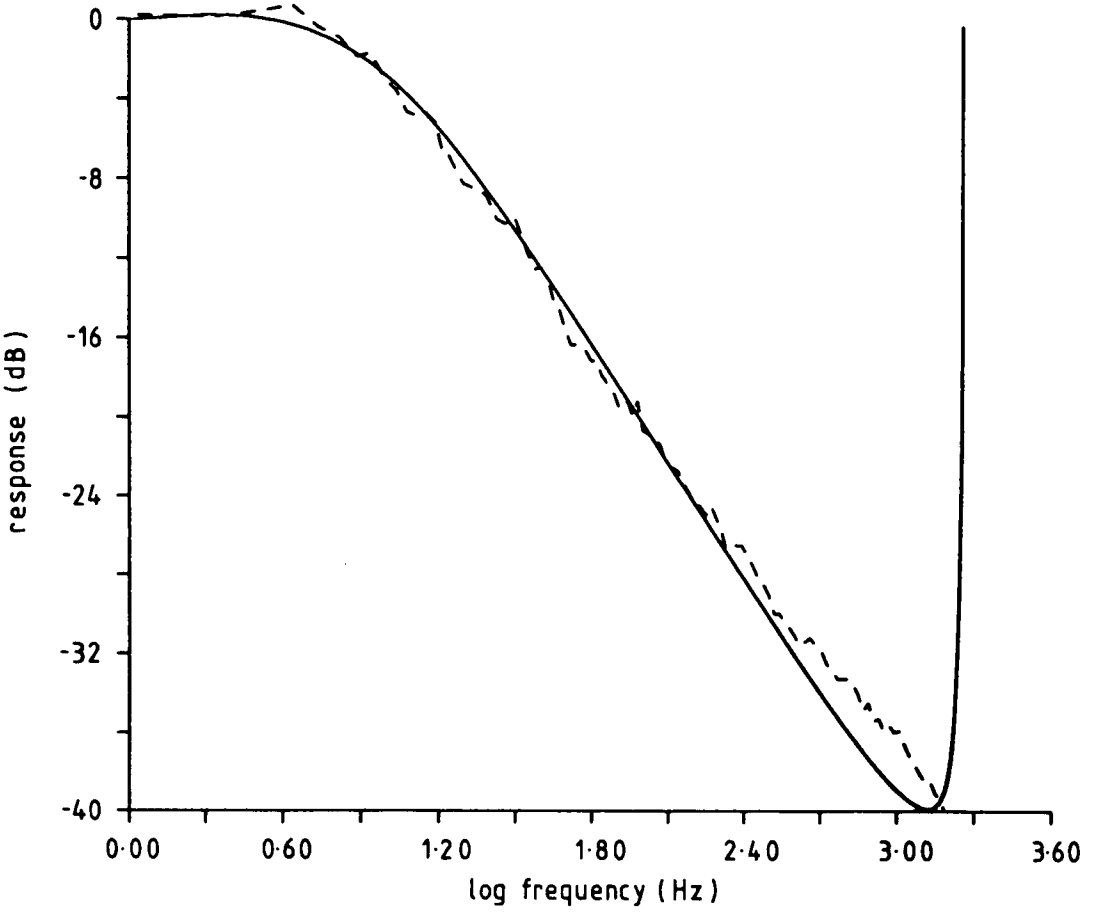


Figure 5.16 Modified Model Frequency Response
(*11dB CNR measurements shown dotted, for comparison*)

the same does not apply to the step response, shown in figure 5.17. This compares poorly in overall form with figure 5.8 (12dB CNR) which is likely to be very similar to the behaviour at 11dB. It may be that a continued search would discover a different combination of open loop gain and filter channel differential gain, which satisfies both frequency- and time-domain responses. However, considering the irregularity of sampling at this CNR, which makes a linear transformation between time- and frequency-domain responses unlikely, a better course of action, rather than further manipulation of the model, would be to modify the filter algorithm and observe any change in the loop behaviour.

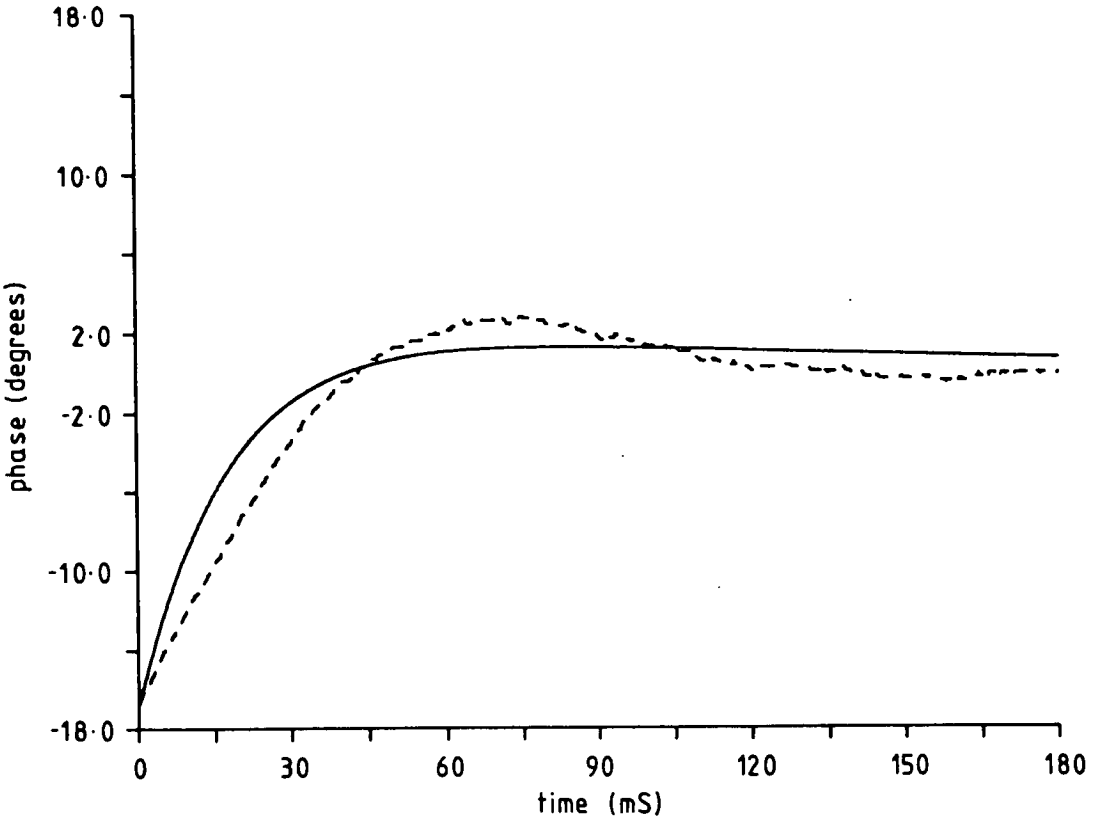


Figure 5.17 Modified Model Step Response
(*12dB CNR measurement shown dotted, for comparison*)

6. Conclusion

This closing chapter will attempt to summarise the work presented so far, offering suggestions for future directions of development. At this point it may be noted that in the course of this research many ideas have been tried in a superficial manner, in order to gauge their potential. They have not been explicitly mentioned in this thesis for want of data gathered under controlled conditions. However, they have provided a useful source of context within which to pursue those aspects which have been described.

6.1. Summary

This thesis has presented research into two elements of modem design. After a general introduction to the transmission of data and the description of channels and modems in chapter one, chapters two and three have described the principles of signal set design and the development of a signal set suitable for use over voiceband telephone channels. After a necessarily brief survey of the extensive fields of carrier recovery and phase-lock loop techniques in chapter four, chapter five has presented a hardware approach to the carrier recovery and phase demodulation of the signal set described in chapter three.

The hardware design is based on a feedback-loop structure, the Signal Driven Phase-Lock Loop, believed to be unreported previously and the subject of a British patent application†. It is suitable for the regeneration of carrier and symbol-timing from the line signal of a PSK or APK transmission, particularly where low complexity is an important criterion. Although the SDPLL structure may be implemented in analogue, hybrid or fully digital forms, only the latter has been investigated; the intention is to analyse the effects of non-uniform sampling inherent in the SDPLL and to gather information on suitable methods for implementing elements of the SDPLL in a custom (digital) integrated circuit. The construction of a test-bed prototype has proved the basic

† Number 8228009.

viability of a fully digital SDPLL. Preliminary measurements (§5.2) have revealed that moderate sampling irregularity, of a degree which may be expected to arise from channel noise and bandlimiting effects, does not substantially affect the behaviour of the SDPLL when compared to a mathematical model (§5.3) using average values for sample rate and loop delay. However, unexpected behaviour has been observed when the sampling irregularity is of a degree which would be caused by signalling. Hypotheses to account for this behaviour have been put forward (§5.4), but these have not yet been tested. Time-domain filtering (windowing) has been proposed as a method of countering the effects of signalling, with information on the C1−5−10N constellation being presented to permit the efficacy of such a procedure to be estimated.

Development of the SDPLL provided the impetus for the development of a signal set, suitable for use with and illustrating the advantages of the SDPLL. This led to the investigation and adoption of the C1−5−10N signal design (§3.6), which is similar to the C1−5−10S designs published elsewhere, but has implementation advantages over that design. A detailed description of the C1−5−10N constellation has been presented, defining the location of symbols and decision boundaries in signal space, showing its symbol error behaviour in AWGN and the effects of maladjustment of phase and magnitude thresholds when demodulation is performed by taking independent decisions on those parameters. An efficient grey code has been formulated for use with $I\Phi-M$ demodulation of C1−5−10N. Certain potential failings of the constellation have been noted (§3.7) and their relevance to real applications discussed.

6.2. Comparison of the SDPLL with Other Designs

In order to ascertain the originality of the SDPLL concept, an initial literature search was made using the Lockheed "Dialog" information retrieval service. This provided a great many references to publications likely to be relevant to the SDPLL. Many of these references could be rejected as irrelevant, simply on the basis of their titles; of those remaining which were reasonably accessible, none turned out to be descriptions of loops similar to the SDPLL.

In addition to the machine-aided approach, a more conventional search of Electronics Abstracts was continued in parallel with development of the SDPLL prototype. This eventually† brought to light a paper by Hentinen et al. [127], that describes a phase

† This reference was found after all the SDPLL development to which it bore

detection method bearing many similarities to that used in the SDPLL. This description is the closest so far found to the SDPLL, but only describes a phase detector and does not incorporate it in a PLL or any other feedback structure. Furthermore, it is more restricted than the SDPLL in its ability to handle closely spaced zero-crossings of the line signal and has more in common with the charge-pump type of phase detectors (*v.i.*) seen in many other DPLLs. The paper does contain measurements of some quantisation parameters relevant to the SDPLL, particularly the time-domain windowing of samples to exclude those most likely to be perturbed by noise (both additive noise and self-noise due to signalling), a function essential to the phase detector described and beneficial to the operation of the SDPLL.

The key feature of the SDPLL, which distinguishes it from other PLL designs, is the use of zero-crossings of the line signal waveform to provide the sampling clock for the loop; samples of the local reference oscillator being used to generate the loop error signal. Other PLLs may be broadly classified as zero-crossing-tracking types e.g. [128], where the local reference of the loop causes samples of the line signal to be taken when it is anticipated that a zero-crossing of the line signal will occur; or charge-pump types e.g. [129], where the interval between consecutive zero-crossings of the line signal and local reference is used to generate the error signal by switching the output of a current source (analogue) or pulse generator (digital). In the first type, the output of the phase comparator is a function of the line signal wave-shape, as well as its relative phase. Many fully digital PLL designs are of the latter type [107,130,131], for which the wave-shape is of less significance, particularly when the loop is out of lock.

The phase detection process of the SDPLL can be viewed as an inverted zero-crossing-tracking type, in which the line signal causes samples to be taken of the local oscillator waveform. This brings the same degree of control over the relationship between raw error signal and line signal wave-shape as is possessed by the charge-pump type of phase comparator, at the cost of greater irregularity of sampling‡. The advantage over the

---

similarities had been completed. It is not referenced in either of the two review papers [84,85].

‡ The sampling irregularity of the charge-pump type of phase detector is influenced by details of design, but it may be arranged that the nominal sample instants coincide with the zero-crossing instants of the (slowly varying) local reference, rather than the line signal.

charge-pump phase comparator is that an error sample is provided at every line signal zero-crossing, without the need for a trigger from the local reference to complete the measurement. This allows steps to be taken to avoid the inherent bias of using only the information of one sample, when several precede a trigger from the local reference.

The analogy to the inverted zero-crossing-tracking PLL also provides insight into the sub-harmonic locking properties of the SDPLL. Just as the more conventional types of PLL may lock to signals bearing a harmonic relationship to the local reference, by virtue of those signals' possession of zero-crossings coincident with the zero-crossings of the fundamental, so the SDPLL locks to sub-harmonics, its local reference being an under-sampled harmonic of the line signal. When running the SDPLL at the N^{th} harmonic frequency of the line signal, the local reference is sampled once only in every N cycles. Since it matters not which cycle of the local reference is sampled, there are N possible equally spaced phases of line signal which result in equivalent loop error signals.

6.3. Future Directions

Despite the reservations which have already been expressed regarding the status of specialised hardware design in the present environment of programmable digital signal processing, continuation of the work presented here may still be justified, so long as there still exists the possibility of a large scale application where a simple low-cost bandwidth-efficient modem is required. One such area is the design of long-line drivers for high speed data transmission over twisted copper pairs having lengths of more than a few hundred yards. Baseband pulse transmission over such a medium at rates of more than a few hundred baud is generally impractical, due largely to pulse distortion. While equalisation is one possible solution to this problem, replacement of current line-driver technology (based on manchester coding) with a dense-constellation modem based on the SDPLL, may offer increased throughput at equivalent cost.

6.3.1. Theoretical Modelling

There are two areas of modelling which are of immediate importance to future work based on that presented here. They are the description of signals as characterised by zero-crossing location and the description of irregularly sampled systems

The theoretical description of signal characterisation by zero-crossing location has been mentioned only superficially in this thesis. Its importance to work on the SDPLL derives

from the use of the line signal as the fundamental clock of the SDPLL. As such, a general description is necessary both for the influence it has on loop parameterisation, i.e. the influence of irregular sampling as might be applied to any sampled-data system and also as a means for deriving strategies for optimum utilisation of the carrier and symbol timing information inherent in the zero-crossing data. Time-domain analysis is not commonly applied to communications signals, with the exception of the equalisation problem and the prediction of signal fading. One recent paper which does take this approach is by Piwnicki [132], in which he develops a description of a perfectly bandlimited signal in terms of its intersections with a reference sinusoid. Although this paper was published too late to influence the work described here and its expressed area of application is the *modulation* process, it may provide useful material for further work related to the SDPLL.

The loop model presented in chapter five is clearly inadequate for all but the most trivial applications. Although some of the inadequacy demonstrated in the high level jitter (11/12 dB CNR) tests may be due to imperfect understanding of the behaviour of the loop filter, it seems likely that substantial work will need to be done to create a model which describes both the time- and frequency-domain behaviour of the SDPLL. Bearing in mind the stochastic nature of sampling in the SDPLL and also the problems of trying to model the behaviour of conventional PLLs in transient states, it may be that separate models for the two domains, sharing several common parameters, may be a suitable approach to the problem.

Computational modelling based on direct imaging of the physical mechanics of the loop may always be employed as a design tool, but the lack of generality which such models tend to suffer from limits their utility, particularly for development of a general understanding.

6.3.2. Constellation Design

As has been suggested in §3.7, the development of C4−12N as an alternative to C1−5−10N would bring some benefits, e.g. guaranteeing the operation of a modem under all circumstances. Alternatively, another avenue worth following would be the insertion of a pilot carrier at −13 dB into the C1−5−10N line signal. This would provide SNR performance equivalent to C4−12N, but with the additional advantage of

an absolute phase and magnitude reference. This poses the question of whether carrier regeneration *per se* would then be required; this is a subject which will be left awaiting detailed investigation, along with the changes required to demodulate such a signal.

The use of C1−5−10N as a chip-set for a hyperdimensional constellation has also been mentioned. The potential simplicity and small size of an $I\Phi-M$ implementation of C1−5−10N, exploiting the SDPLL in a highly parallel hardware architecture, could provide a suitable rival to equivalent QAM based constellations using In-phase/Quadrature demodulation. At present, it seems unlikely that more exotic constellations, such as those based on crystallographic-derived or non-regular structures [49,52,53,54] will provide any real alternative to either QAM or regular-concentric designs.

Even the justification of the pursuit of an alternative to the QAM designs for the type of channel considered in this work is questionable, given the current rapidly advancing state of programmable digital signal processors, the fairly straightforward nature of In-phase/Quadrature demodulation and the generally marginal $P(\epsilon)$ improvements offered by other designs. It must, therefore, be restated that the development of C1−5−10N was subservient to the primary object of this research, i.e. development of the SDPLL. Where other classes of channel are considered, the effectiveness of different constellations may be much altered. For example, the effectiveness of pure PSK at very low SNR has already been mentioned.

6.3.3. Development of the Prototype

Although the prototype described is of a wholly digital implementation, it should be emphasised that the SDPLL may also be realised as an analogue sampled data PLL, thus relieving some of the frequency limitations inherent in the design presented here. Hence, a prototype of an analogue SDPLL operating at high signalling rates (e.g. 50Mbaud) is a potential development area for the investigation of problems and techniques not covered by the current prototype. Leaving that option aside, there are several things to be done with the current prototype which are of immediate concern. These are, confirmation of the apparent loop behaviour under conditions of large jitter amplitude, identification of this behaviour as either a general loop property or the result of the particular loop filter implementation and investigation of the presettable scaler as an implementation of the

NCC.

A method of confirming the validity of the data collected so far has already been described in §5.2, namely by collecting data based on a clock derived from the digital signal store, rather than one based on the output of the SDPLL reference oscillator. Simultaneous monitoring of the value of the control word sent to the NCC could also be used to confirm the expected accuracy of the data already gathered.

An alternative, somewhat more deterministic method of collecting information would be to inject a line signal carrier having low-level swept frequency phase jitter, while monitoring the effect on the NCC control word in order to probe the frequency and phase response of the loop.

The loop filter program should be modified to ensure that the same sample value is always used for both the direct and integral channels on each cycle. The initial reason for making the filter behave in its present fashion was an attempt to minimise latency as far as possible. It is apparent now that this does not pose as serious a problem as was first envisaged. There are several options for extending the present filter design. These include an investigation of methods for arbitrating amongst several closely spaced zero-crossings (i.e. occurring within one filter cycle period) and optimising the filter design for minimal hardware implementation. There is also the class of statistical filters to investigate, e.g. the 'random walk' and 'M before N' types described in other DPLL designs [89,107,130].

When the doubts concerning the current filter implementation have been satisfactorily dealt with, progress may continue towards full testing with a modulated line signal and analysis of the optimum technique for reducing self noise. This will involve the extraction of symbol-timing information from the line signal, a problem for which the SDPLL is also expected to be a suitable solution.

An estimate of the maximum nominal carrier frequency which could be handled by the SDPLL using current technology has been given in §5.1.3.2 as 500 kHz. The assumptions used in deriving this estimate, based mainly on extending the application of the prototype as described herein, may not be applicable in many cases. In particular, for application to narrowband systems, the line signal may be heterodyned to a lower frequency provided that the absolute bandwidth is not so great as to prevent the nominal carrier frequency from being brought within range of the SDPLL. It may be possible to avoid

the additional mixers and filters implied by this frequency translation by making use of the work of Piwnicki [132], which is essentially a time-domain description of the heterodyne process, to incorporate frequency translation into the main zero-crossing detector of the SDPLL. Division of the frequency tracking function between the SDPLL local reference and the heterodyne oscillator is a further possibility which may be helpful in extending the tracking range of the loop.

Frequency translation of high frequency narrowband signals is also possible by the conventional approach of under-sampling, as commonly used in the back end of radios which incorporate passband digital signal processing. In the case of the SDPLL, this simply requires the zero-crossing detector to discard a certain fraction of the line signal zero-crossings, rather than using them to trigger the SDPLL sample latch. Using this approach, direct tracking and demodulation of e.g. a 100 kHz bandwidth PSK signal at a nominal carrier frequency of 100 MHz should be possible, with the carrier frequency limit being set principally by the speed of the limiter.

Conversely, superheterodyning a voice-band signal may be a means of providing additional zero-crossings per symbol for the application under consideration in this work. Although this thesis has presented no data regarding problems of nominal sample rate in the SDPLL, for application to a voice-band modem the SDPLL is operating very close to the Nyquist rate if zero-crossings are to be used for phase demodulation. Consequently, an increased sample rate may prove useful in the course of further development.

References

1. R.V.L. Hartley, "Transmission of Information," *BSTJ* **VII** pp. 535-563 (Jul.1928).

2. G. Ungerboeck, "Channel Coding with Multilevel/Phase Signals," *Trans. IEEE Inform. Theory* **IT-28**(1) pp. 55-67 (Jan.1982).

3. A. Digeon, "On Improving Bit Error Probability of QPSK and 4-Level Amplitude Modulation Systems by Convolutional Coding," *Trans. IEEE Commun.* **COM-25**(10) pp. 1238-1239 (Oct.1977).

4. J.D. Brownlie and E.L. Cusack, "Duplex Transmission at 4800 and 9600 bit/s on the General Switched Telephone Network and the Use of Channel Coding with a Partitioned Signal Constellation," *Proc. Int. Zurich Seminar on Digital Comms.*, IEEE, (Mar.1984). Paper G1

5. S.G. Wilson, H.A. Sleeper, and N.K. Srinath, "Four-Dimensional Modulation and Coding: An Alternate to Frequency-Reuse," *Proc. IEEE Int. Conf. on Comms.*, pp. 919-923 IEEE/Elsevier Science Publishers, (May.1984).

6. A. Gersho and V.B. Lawrence, "Multidimensional Signal Design for Digital Transmission Over Bandlimited Channels," *Proc. IEEE Int. Conf. on Comms.*, pp. 377-380 IEEE/Elsevier Science Publishers, (May.1984).

7. H.K. Thapar and L.F. Wei, "On the Real-Time Performance of Trellis Coding for High-Speed Voiceband Modems," *Proc. IEEE Int. Conf. on Comms.*, pp. 381-385 IEEE/Elsevier Science Publishers, (May.1984).

8. CCITT, "Recommendation V.32," in *CCITT Book of the VIIIth. Plenary Assembly*, International Telecommunication Union, Geneva (In preparation).

9. H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *Trans. AIEE* **47**(2) pp. 617-644 (Apr.1928).

10. D. Slepian, "On Bandwidth," *Proc. IEEE* **64**(3) pp. 292-300 (Mar.1976).

11. C.E. Shannon, "A Mathematical Theory of Communication," *BSTJ* **XXVII**(3) pp. 379-423 (Jul.1948). Parts I and II

12. C.E. Shannon, "A Mathematical Theory of Communication," *BSTJ* **XXVII**(4) pp. 623-656 (Oct.1948). Part III

13. H.J. Landau and H.O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty-III: The Dimension of the Space of Essentially Time- and Band-Limited Signals," *BSTJ* 41(4) pp. 1295-1336 (Jul.1962).

14. F.Westall, *private communication*, B.T. Research Laboratories, Martlesham, U.K., (1983).

15. P. Monsen, "Fading Channel Communications," *IEEE Commun. Soc. Mag.* 18(1) pp. 27-36 (Jan.1980).

16. D.H. Morais, A. Sewerinson, and K. Feher, "The Effects of the Amplitude and Delay Slope Components of Frequency Selective Fading on QPSK, Offset QPSK and 8PSK Systems," *Trans. IEEE Commun.* COM-27(12) pp. 1849-1853 (Dec.1979).

17. K.D.D., "Are Group Delay Time and/or Phase Delay Time Useful Parameters for Defining Low Distortion Transmission?," *Trans. IEEE Commun.* COM-21(12) pp. 1446-1448 (Dec.1973).

18. H. Mar, "Comments on 'Are Group Delay Time and/or Phase Delay Time Useful Parameters for Defining Low Distortion Transmission?'," *Trans. IEEE Commun.* COM-22(8) p. 1148 (Aug.1974).

19. V.C. Chohan, "On the Usefulness of Group Delay and/or Phase Delay Parameters for Low Distortion Transmission," *Trans. IEEE Commun.* COM-22(8) pp. 1147-1148 (Aug.1974).

20. P.H. Wittke, S.R. Penstone, R.J. Keightley, and J. Yau, "Measurements on Telephone Circuits of Parameters Pertinent to High-Speed Full-Duplex Voiceband Data Transmission," *Proc. IEEE Int. Conf. on Comms.*, pp. 365-368 IEEE/Elsevier Science Publishers, (May.1984).

21. T. Maseng, "On the Characterization of a Bandpass Nonlinearity by Two-Tone Measurements," *Trans. IEEE Commun.* COM-26(6) pp. 746-754 (Jun.1978).

22. W.R. Vincent, "Examples of Signals and Noise in the Radio-Frequency Spectrum," *Trans. IEEE Electromagn. Compat.* EMC-19(3) pp. 241-253 (Aug.1977).

23. J.H. Fennick, "Amplitude Distributions of Telephone Channel Noise and a Model for Impulse Noise," *BSTJ* 48(10) pp. 3243-3263 (Dec.1969).

24. G.F. Gott, S. Dutta, and P. Doany, "Analysis of HF interference with application to digital communications," *Proc. IEE Pt.F* **130**(5) pp. 452-458 (Aug.1983).

25. D. Middleton, "Statistical-Physical Models of Electromagnetic Interference," *Trans. IEEE Electromagn. Compat.* **EMC-19**(3) pp. 106-127 (Aug.1977).

26. B. Dunbridge, "Asymmetric Signal Design for the Coherent Gaussian Channel," *Trans. IEEE Inform. Theory* **IT-13**(3) pp. 422-431 (Jul.1967).

27. M.P. Shinde and S.N. Gupta, "Signal Detection in the presence of Atmospheric Noise in Tropics," *Trans. IEEE Commun.* **COM-22**(8) pp. 1055-1063 (Aug.1974).

28. D. Middleton, "Procedures for Determining the Parameters of the First-Order Canonical Models of Class A and Class B Electromagnetic Interference," *Trans. IEEE Electromagn. Compat.* **EMC-21**(3) pp. 190-208 (Aug.1979).

29. A.D. Spaulding and D. Middleton, "Optimum Reception in an Impulsive Interference Environment-Part I: Coherent Detection," *Trans. IEEE Commun.* **COM-25**(9) pp. 910-923 (Sep.1977).

30. J.D. Oetting, "A Comparison of Modulation Techniques for Digital Radio," *Trans. IEEE Commun.* **COM-27**(12) pp. 1752-1762 (Dec.1979).

31. E. Olcayto and T. Lesz, "Class of linear cyclic block codes for burst errors occurring in one-, two- and three-dimensional channels," *Proc. IEE Pt.F* **130**(5) pp. 468-475 (Aug.1983).

32. H. Yamamoto, "Advanced 16-QAM Techniques for Digital Microwave Radio," *IEEE Commun. Soc. Mag.* **19**(5) pp. 36-45 (May.1981).

33. P.F. Adams, "Adaptive Filters in Telecommunications," pp. 216-255 in *Adaptive Filters*, ed. C.F.N. Cowan and P.M. Grant, Prentice-Hall (1985).

34. CCITT, "Recommendation V.29," pp. 143-153 in *CCITT Book of the VIth. Plenary Assembly*, International Telecommunication Union, Geneva (1976). (Orange Book)

35. R.W. Bigg, "A New Modem for the Datel 2412 Service: Datel Modem No. 12B," *POEEJ* **70**(3) pp. 185-193 (Oct.1977).

36. J.G. Proakis, "Modulation and Demodulation for the Additive Gaussian Noise Channel," pp. 193-203 in *Digital Communications*, McGraw-Hill, Tokyo (1983).

ISBN 0-07-066490-0

37. C.P. Tou and D.A. Roy, "On Efficient Spectrum Utilisation from the Standpoint of Communication Theory," *Proc. IEEE* 68(12) pp. 1460-1465 (Dec.1980).

38. E. Bedrosian, "Spectrum Conservation by Efficient Channel Utilisation," *IEEE Commun. Soc. Mag.* 15(2) pp. 20-27 (Mar.1977).

39. I.S. Reed and R.A. Scholtz, "N-Orthogonal Phase-Modulated Codes," *Trans. IEEE Inform. Theory* IT-12(3) pp. 388-395 (Jul.1966).

40. J.D. Ralphs, "The application of m.f.s.k. techniques to h.f. telegraphy," *The Radio and Electronic Engineer* 47(10) pp. 435-444 (Oct.1977).

41. J.D. Ralphs, "An improved 'Piccolo' m.f.s.k. modem for h.f. telegraphy," *The Radio and Electronic Engineer* 52(7) pp. 321-330 (Jul.1982).

42. J.M. Wozencraft and I.M. Jacobs, *Principles of Communication Engineering*, John Wiley & Sons, New York (1965).

43. N.J.A. Sloane, "Tables of Sphere Packings and Spherical Codes," *Trans. IEEE Inform. Theory* IT-27(3) pp. 327-338 (May.1981).

44. G.J. Foschini, R.D. Gitlin, and S.B. Weinstein, "On the Selection of a Two-Dimensional Signal Constellation in the Presence of Phase Jitter and Gaussian Noise," *BSTJ* 52(6) pp. 927-965 (Jul-Aug.1973).

45. C.L. Weber, "New Solutions to the Signal Design Problem for Coherent Channels," *Trans. IEEE Inform. Theory* IT-12(2) pp. 161-167 (Apr.1966).

46. M.P. Ristenbatt, "Alternatives in Digital Communications," *Proc. IEEE* 61(6) pp. 703-721 (Jun.1973).

47. J. Salz, J.R.Sheehan, and D.J. Paris, "Data Transmission by Combined AM and PM," *BSTJ* 50(7) pp. 2399-2419 (Sep.1971).

48. C.R. Cahn, "Combined Digital Phase and Amplitude Modulation Communication Systems," *Trans. IRE Comm. Systems* CS-8(3) pp. 150-155 (Sep.1960).

49. L.H. Zetterberg and H. Brändström, "Codes for Combined Phase and Amplitude Modulated Signals in a Four-Dimensional Space," *Trans. IEEE Commun.* COM-25(9) pp. 943-950 (Sep.1977).

50. R.A. Scholtz and C.L. Weber, "Signal Design for Phase-Incoherent Communications," *Trans. IEEE Inform. Theory* IT-12(4) pp. 456-463 (Oct.1966).

51. A.J. Viterbi, "On Coded Phase-Coherent Communications," *IRE Trans. on Space Electronics and Telemetry* SET-7 pp. 3-14 (Mar.1961).

52. G.J. Foschini, R.D. Gitlin, and S.B. Weinstein, "Optimisation of Two-Dimensional Signal Constellations in the Presence of Gaussian Noise," *Trans. IEEE Commun.* COM-22(1) pp. 28-37 (Jan.1974).

53. M.K. Simon and J.G. Smith, "Hexagonal Multiple Phase-and-Amplitude-Shift-Keyed Signal Sets," *Trans. IEEE Commun.* COM-21(10) pp. 1108-1115 (Oct.1973).

54. C.M. Thomas, M.Y. Weidner, and S.H. Durrani, "Digital Amplitude-Phase Keying with M-ary Alphabets," *Trans. IEEE Commun.* COM-22(2) pp. 168-179 (Feb.1974).

55. H. Yanagidaira, "The 12,000 - 12,800 bit/s Data Transmission MODEM for Use on Voice-Band Circuit by Modified 12Phase-2Level APSK System," *Trans. IECE Japan* E 63(4) pp. 292-293 (Apr.1980). (abstract)

56. R.W. Lucky and J.C. Hancock, "On the Optimum Performance of N-ary Systems Having Two Degrees of Freedom," *Trans. IRE Comm. Systems* CS-10(2) pp. 185-192 (Jun.1962).

57. A.J. Viterbi and J.J. Stiffler, "Performance of N-Orthogonal Codes," *Trans. IEEE Inform. Theory* IT-13(3) pp. 521-522 (Jul.1967).

58. CCITT, "Recommendation V.29bis," in *CCITT Book of the VIth. Plenary Assembly*, International Telecommunication Union, Geneva (1976). (Orange Book)

59. D. Soufflet and M. Joindot, "Étude du comportement d'une modulation numérique à 16 états d'amplitude et de phase (MAQ16) dans un canal de transmission hertzienne," *Ann. Télécommunic.* 34(7-8) pp. 401-417 (1979). (in French)

60. D. Soufflet, I. Horikawa, and Y. Saito, "16QAM Carrier Recovery with Selective Gated Phase Locked Loop," *Trans. IECE Japan* E 63(7) pp. 548-549 (Jul.1980). (abstract)

61. P.D. Stephens, *EMAS 2900: Concepts*, Edinburgh Regional Computing Centre, Edinburgh (1980).

62. M. Schwartz, "Limitations Due to Noise," pp. 377-383 in *Information Transmission, Modulation, and Noise*, ed. S.W. Director, McGraw-Hill (1980). ISBN 0-07-066547-8

63. International Computers Limited, *Mathematical Procedures Reference Manual 2900 Series*, International Computers Limited, London (Sep.1983).

64. R. Sedgewick, "Elementary Sorting Methods," pp. 94-95 in *Algorithms*, ed. M.A. Harrison, Addison-Wesley (1983). ISBN 0-201-06672-6

65. W.J. Weber III, "Differential Encoding for Multiple Amplitude and Phase Shift Keying Systems," *Trans. IEEE Commun.* **COM-26** pp. 385-391 (Mar.1978).

66. M. Rosenblatt, *Time Series Analysis*, John Wiley & Sons Inc., New York (1963).

67. S.O. Rice, "Statistical Properties of a Sine Wave Plus Random Noise," *BSTJ* **XXVII** pp. 109-157 (Jan.1948).

68. F. James, "Monte Carlo Theory and Practice," DD/80/6, Data Handling Division, CERN, Geneva (Feb.1980).

69. L.E. Franks, "Carrier and Bit Synchronization in Data Communication — A Tutorial Review," *Trans. IEEE Commun.* **COM-28**(8) pp. 1107-1120 (Aug.1980).

70. R.W. Chang and R. Srinivasagopalan, "Carrier Recovery for Data Communication Systems with Adaptive Equalisation," *Trans. IEEE Commun.* **COM-28**(8) pp. 1142-1153 (Aug.1980).

71. J.D. Harvey, "Carrier Phase Estimation and Phase Jitter Suppression for Data Modems," *IEE Colloquium on Phase Lock Techniques*, (E10) pp. 4/1-4/7 (26th Mar.1980).

72. Y. Takasaki, "Optimising Pulse Shaping for Baseband Digital Transmission with Self-Bit Synchronization," *Trans. IEEE Commun.* **COM-28**(8) pp. 1164-1172 (Aug.1980).

73. F.M. Gardner, "Self-Noise in Synchronisers," *Trans. IEEE Commun.* **COM-28**(8) pp. 1159-1163 (Aug.1980).

74. A.M. Abdulsatar and G. Maral, "Self-Noise Spectral Measurements in Clock Synchronisers," *Trans. IEEE Commun.* **COM-31**(10) pp. 1204-1207 (Oct.1983).

75. S.A. Rhodes, "Effect of Noisy Phase Reference on Coherent Detection of Offset-QPSK Signals," *Trans. IEEE Commun.* **COM-22**(8) pp. 1046-1055 (Aug.1974).

76. R.L. Didday and W.C. Lindsey, "Subcarrier Tracking Methods and Communication System Design," *Trans. IEEE Commun.* **COM-16**(4) pp. 541-550 (Aug.1968).

77. J.P. Costas, "Synchronous Communications," *Proc. IRE* **44**(12) pp. 1713-1718 (Dec.1956).

78. C.R. Hogge, "Carrier and Clock Recovery for 8PSK Synchronous Demodulation," *Trans. IEEE Commun.* **COM-26**(5) pp. 528-533 (May.1978).

79. K. Miyauchi, S. Seki, and H. Ishio, "New Technique for Generating and Detecting Multilevel Signal Formats," *Trans. IEEE Commun.* **COM-24**(2) pp. 263-267 (Feb.1976).

80. S.F. Wetenkamp and K.J. Wong, "Transportation Lag in Phase-Locked Loops," *Watkins-Johnson Company Tech-notes* 5(3) pp. 2-14 (May/Jun.1978).

81. F.M. Gardner, *Phase Lock Techniques*, Wiley, New York (1979). 2nd ed.

82. W.C. Lindsey, *Synchronization Systems in Communication and Control*, Prentice-Hall, Englewood-Cliffs, NJ USA (1972).

83. J. Klapper and J.T. Frankle, *Phase-Locked and Frequency Feedback Systems*, Academic Press, New-York (1972).

84. S.C. Gupta, "Phase-Locked Loops," *Proc. IEEE* **63**(2) pp. 291-306 (Feb.1975).

85. W.C. Lindsey and C.M. Chie, "A Survey of Digital Phase-Locked Loops," *Proc. IEEE* **69**(4) pp. 410-431 (Apr.1981).

86. J.J. Uhran, Jr. and J.C. Lindenlaub, "Experimental Results for Phase-Lock Loop Systems Having a Modified nth-Order Tanlock Phase Detector," *Trans. IEEE Commun.* **COM-16**(6) pp. 787-795 (Dec.1968).

87. C.J. Byrne, "Properties and Design of the Phase-Controlled Oscillator with a Sawtooth Comparator," *BSTJ* **41** pp. 559-602 (Mar.1962).

88. J.J. Stiffler, "On the Selection of Signals for Phase-Locked Loops," *Trans. IEEE Commun.* **COM-16**(2) pp. 239-244 (Apr.1968).

89. J.R. Cessna and D.M. Levy, "Phase Noise and Transient Times for a Binary Quantised Digital Phase-Locked Loop in White Gaussian Noise," *Trans. IEEE Commun.* COM-20(2) pp. 94-104 (Apr.1972).

90. J. Garodnick, J. Greco, and D.L. Schilling, "Response of an All Digital Phase-Locked Loop," *Trans. IEEE Commun.* COM-22(6) pp. 751-764 (Jun.1974).

91. R.C. Dorf, *Modern Control Systems*, Addison-Wesley, Reading, Massachusetts USA (1976).

92. P.R. Westlake, "Digital Phase Control Techniques," *IRE Trans. Comm. Sys.* CS-8 pp. 237-246 (Dec.1960).

93. G.S. Gill and S.C. Gupta, "First-Order Discrete Phase-Locked Loop with Applications to Demodulation of Angle Modulated Carrier," *Trans. IEEE Commun.* COM-20 pp. 454-462 (Jun.1972).

94. A. Weinberg and B. Liu, "Discrete Time Analysis of Nonuniform Sampling First- and Second-Order Digital Phase Locked Loops," *Trans. IEEE Commun.* COM-22(2) pp. 123-137 (Feb.1974).

95. G.T. Hurst and S.C. Gupta, "Quantizing and Sampling Considerations in Digital Phase-Locked Loops," *Trans. IEEE Commun.* COM-22(1) pp. 68-72 (Jan.1974).

96. G.L. Hedin, J.K. Holmes, W.C. Lindsey, and K.T. Woo, "Theory of False Lock in Costas Loops," *Trans. IEEE Commun.* COM-26(1) pp. 1-11 (Jan.1978).

97. M.K. Simon and J.G. Smith, "Carrier Synchronization and Detection of QASK Signal Sets," *Trans. IEEE Commun.* COM-22(2) pp. 98-106 (Feb.1974).

98. A. Leclert and P. Vandamme, "Universal Carrier Recovery Loop for QASK and PSK Signal Sets," *Trans. IEEE Commun.* COM-31(1) pp. 130-136 (Jan.1983).

99. M.K. Simon, "The False Lock Performance of Costas Loops with Hard-Limited In-Phase Channel," *Trans. IEEE Commun.* COM-26(1) pp. 23-34 (Jan.1978).

100. M.K. Simon and K.T. Woo, "Alias Lock Behaviour of Sampled-Data Costas Loops," *Trans. IEEE Commun.* COM-28(8) pp. 1315-1325 (Aug.1980).

101. I.A.B. Lindsay and J.H. Dripps, "A Novel Phase-Locked Loop for PSK Carrier Recovery," *Digital Processing of Signals in Communications*, pp. 173-177 IERE, (Apr.1985). IERE publication number 62

102. J.H. Dripps, *An Optimum Modulation Method for Digital Data Transmission on H.F. Ionospheric Channels*, University of Strathclyde, Glasgow (1977). PhD Thesis

103. G.T. Bandason, "A Sampling Phase Locked Loop QPSK Demodulator," *Department of Electrical Engineering BSc Honours Project Report* HSP 297The University of Edinburgh, (May.1982).

104. W.B. Davenport Jr., "Signal-to-Noise Ratios in Band-Pass Limiters," *J. Appl. Phys.* **24**(6) pp. 720-727 (Jun.1953).

105. C.R. Cahn, "A Note on Signal-to-Noise Ratio in Band-pass Limiters," *Trans. IRE Inform. Theory* IT-7(1) pp. 39-43 (Jan.1961).

106. R.E. Crochiere and L.R. Rabiner, "Interpolation and Decimation of Digital Signals — A Tutorial Review," *Proc. IEEE* **69**(3) pp. 300-331 (Mar.1981).

107. R.M. Horton, "The Development of a Digital Phase-Locked Loop Integrated Circuit," *Proc. IEEE Nat. Aero. Electronic Conf. (NAECON)*, pp. 938-947 (1976). Ohio USA

108. Y. Ogawa, M. Sengoku, and T. Matsumoto, "Stuffing Jitter Suppression Using a Digital Phase-Locked Loop," *Trans. IECE Japan* E **60**(7) p. 364 (Jul.1977). (abstract)

109. A.D. Cox, "Digital Lead Compensator Using a B.C.D. Rate Multiplier," *Electron. Letters* **6**(23) pp. 757-759 (Nov.1970).

110. T.R.H. Sizer, *The Digital Differential Analyser*, Chapman and Hall, London (1968).

111. J.D. Martin, "Signal Processing and Computation using Pulse-rate Techniques," *The Radio and Electron. Engineer* **38**(6) pp. 329-344 (Dec.1969).

112. A. Dunworth and J.I. Roche, "The Error Characteristics of the Binary Rate Multiplier," *Trans. IEEE Computers* C-18(8) pp. 741-745 (Aug.1969).

113. J.B. Peatman, "Pulse Rate Algorithms," pp. 350-362 in *The Design of Digital Systems*, McGraw-Hill Kogakusha Ltd., Tokyo (1972).

114. K. Ohtake, M. Taka, and N. Kuroyanagi, "Jitter Characteristics of Digital Phase Locked Loop Using Quantised Clock Rates," *Trans. IECE Japan* E **60**(7) p. 354 (Jul.1977). (abstract)

115. G. Pasternack and R.L. Whalin, "Analysis and Synthesis of a Digital Phase-Locked Loop for FM Demodulation," *BSTJ* **47**(10) pp. 2207-2237 (Dec.1968).

116. J.K. Holmes and C.R. Tegnelia, "A Second-Order All-Digital Phase-Locked Loop," *Trans. IEEE Commun.* **COM-22**(1) pp. 62-68 (Jan.1974).

117. P. Atkinson and A.J. Allen, "Design of type 2 digital phase-locked loops," *The Radio and Electron. Engineer* **45**(11) pp. 657-666 (Nov.1975).

118. G.M. Butaev and V.S. Romashkan, "Linear Code to Pulse-Frequency Converter," *Measurement Techniques*, pp. 194-196 (Feb.1968).

119. P.M. Thompson and A. Bélanger, "Digital arithmetic units for a high data rate," *The Radio and Electron. Engineer* **45**(3) pp. 116-120 (Mar.1975).

120. C.R. Cahn and D.K. Leimer, "Digital Phase Sampling for Microcomputer Implementation of Carrier Acquisition and Coherent Tracking," *Trans. IEEE Commun.* **COM-28**(8) pp. 1190-1196 (Aug.1980).

121. A. Gersho, "Quantisation," *IEEE Commun. Soc. Mag.* 15(5) pp. 20-29 (Sep.1977).

122. P. Horowitz and W. Hill, "Digital Meets Analog," pp. 438 in *The Art of Electronics*, Cambridge University Press (1980). ISBN 0-521-29837-7

123. R.P. Gilson, "Some Results of Amplitude Distribution Experiments on Shift Register Generated Pseudo-Random Noise," *Trans. IEEE Electronic Computers* **EC-15**(6) pp. 926-927 (Dec.1966).

124. J.H. Lindholm, "An Analysis of the Pseudo-Randomness Properties of Subsequences of Long m-Sequences," *Trans. IEEE Inform. Theory* **IT-14**(4) pp. 569-576 (Jul.1968).

125. G.D. Weathers, E.R. Graf, and G.R. Wallace, "The Subsequence Weight of Summed Maximum Length Sequences," *Trans. IEEE Commun.* **COM-22**(8) pp. 997-1004 (Aug.1974).

126. G.H. Tomlinson and P. Galvin, "Analysis of skewing in amplitude distributions of filtered *m* sequences," *Proc. IEE* **121**(12) pp. 1475-1479 (Dec.1974).

127. V.O. Hentinen, P.P. Laiho, and R.M. Särkilahti, "A Digital Demodulator for PSK Signals," *Trans. IEEE Commun.* **COM-21**(12) pp. 1352-1360 (Dec.1973).

128. C.P. Reddy and S.C. Gupta, "A Class of All Digital Phase Locked Loops: Modelling and Analysis," *Trans. IEEE Indust. Electron. and Control Inst.* **IECI-20**(4) pp. 239-251 (Nov.1973).

129. F.M. Gardner, "Charge-Pump Phase-Lock Loops," *Trans. IEEE Commun.* **COM-28**(11) pp. 1849-1858 (Nov.1980).

130. H. Yamamoto and S. Mori, "Performance of a Binary Quantized All Digital Phase-Locked Loop with a New Class of Sequential Filter," *Trans. IEEE Commun.* **COM-26**(1) pp. 35-45 (Jan.1978).

131. A.J. Allen and P. Atkinson, "Settling Time of Type-2 Digital Phase-Locked Loops," *IEE Colloquium on Phaselock Techniques*, (publication E10) pp. 9/1-9/10 (Mar.1980).

132. K. Piwnicki, "Modulation Methods Related to Sine-Wave Crossings," *Trans. IEEE Commun.* **COM-31**(4) pp. 503-508 (Apr.1983).

Appendix A

Publication:

I.A.B. Lindsay and J.H. Dripps, "A Novel Phase-Locked Loop for PSK Carrier Recovery," *Digital Processing of Signals in Communications*, pp. 173-177 IERE, (Apr.1985) IERE publication number 62

A NOVEL PHASE-LOCKED LOOP FOR PSK CARRIER RECOVERY

Summary

A novel sampled data phase-locked loop, designed for carrier recovery applications, is presented. The loop derives information from signal zero crossings; when implemented in digital form it does not require an ADC. The characteristics of the loop create a simple method for the demodulation of signals with many discrete phase states. An example is given, illustrating the demodulation of a four phase PSK signal. Characteristics of key elements of the loop are discussed. Notes on implementation are included, contrasting analogue and digital versions of the loop. Results from a digital prototype have been used to identify critical features of performance and implementation.

1 Introduction

This paper reports a novel structure for a sampled data phase-locked loop. The main feature of the loop is that it uses the input signal to trigger the sampling of a local oscillator, rather than vice-versa. For this reason, it will be referred to as a Signal Driven Sampling Phase-Locked Loop (SDSPLL). Conventional digital sampling phase-locked loops are generally of two kinds [1,2]. In one (fig.1), the local oscillator is used to trigger an ADC to take a sample of the incoming signal. These samples are filtered and used to control the local oscillator, making it track the nominal zero crossing point of the input waveform. Alternatively, the incoming signal is hard limited and then combined with a square wave from the local oscillator, e.g. by using an exclusive-or gate or flip-flop, to generate the error signal.

In the SDSPLL (fig.2), the input signal is hard limited, then either one or both edges of the resulting pulses are used to latch a sample from the phase error detector (v.i.), whose output is a digital word. This sampling arrangement, combined with phase ambiguities generated in the phase detector, enable the loop to reconstruct a carrier from a multi-level phase shift keyed (PSK) signal, and to perform the subsequent demodulation. Any level of PSK may be demodulated, there is no restriction to powers of two. In principle, it is as easy to demodulate a signal with ten phase states as one with two. The SDSPLL is also well suited to clock regeneration from non-return-to-zero (NRZ) binary data streams, this is the problem for which it was originally designed [3].

2 SDSPLL Configuration

Figure 2 shows the principal elements of the loop, when implemented digitally. The analogue input signal enters via a limiter and associated circuitry, which generates pulses at the zero crossing instants. These pulses are used to trigger

a latch which samples the output of a counter, the ramp counter (RC). The RC is clocked by a numerically controlled oscillator (NCO), and is itself used to clock a second modulo-N counter, the phase counter (PC). The RC and PC may be viewed as one counter from which only the least significant bits are taken on to the latch. The output of the RC may optionally be transferred to the latch via a mapping rom. The combination of PC, RC and NCO forms a variable frequency local oscillator which cycles at the carrier frequency of the input signal. The RC and latch combined, act as the phase detector. The outputs from the latch constitute the error signal; the optional mapping rom being used to impose a particular phase detector characteristic. The error signal is fed to the loop filter, and thence to the NCO, closing the feedback loop.

The key features of the loop are: the absence of an ADC; the sampling of the RC (local oscillator) by the zero-transitions of the input signal; and the partitioning of the local oscillator/phase detector output so that only the least significant bits are sent round the loop as the error signal.

3 Synchronisation to PSK Signals

Figure three illustrates the manner in which the SDSPLL synchronises to PSK signals. The example is for a signal which can adopt any one of four phases. The output from the PC and RC, considered as a whole, is depicted as a continuous long ramp. The PC counts modulo-4, thus the output of the RC is depicted as a series of short ramps, four per cycle of the input signal. One possible phase of the input signal has been emphasised, and the corresponding zero crossing instants marked.

Considering the situation where the limiter produces pulses only at the negative going transitions of the input signal, it may be seen that the RC is sampled only at every fourth cycle. The other three possible phases of input signal would each sample one of the other three RC cycles. Since as far as the loop feedback signal is concerned, all RC ramps are equivalent, the resultant error signal is the same for all four phases. The four-fold phase ambiguity which has been created enables the SDSPLL to remain locked to the signal carrier frequency, maintaining a reference phase, as the signal switches between phases. The PC output, which is not fed back round the loop, provides the arbitrary phase reference which may then be used to demodulate the incoming signal. Since the number of states of the PC may be any positive integer, any order of PSK may be accommodated.

If the output from the RC is interpreted as a signed integer, the loop will lock so that zero crossings of the input signal occur half way through a RC cycle, where the error value is zero. Noise and lack of synchronisation deviate the signal zero crossings from this point, producing an error signal which is linearly proportional to the timing error, and independent of the i/p signal waveform or amplitude. When the input signal is wideband, it may be desirable to use both positive

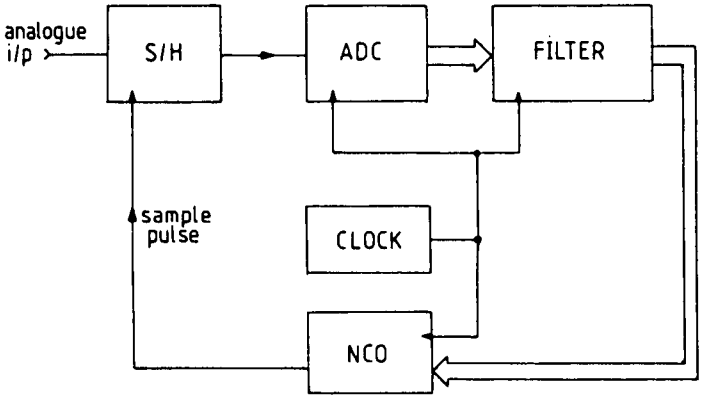


Figure 1 Conventional Sampling DPLL

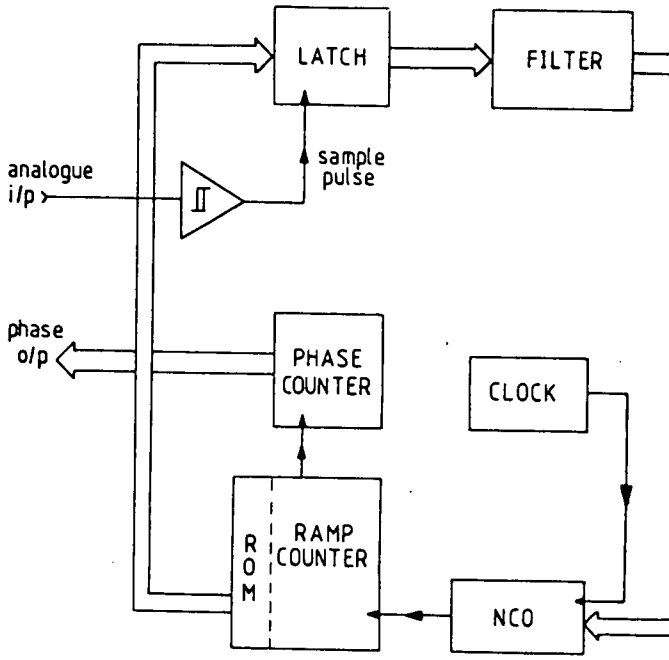


Figure 2 Signal Driven Sampling DPLL

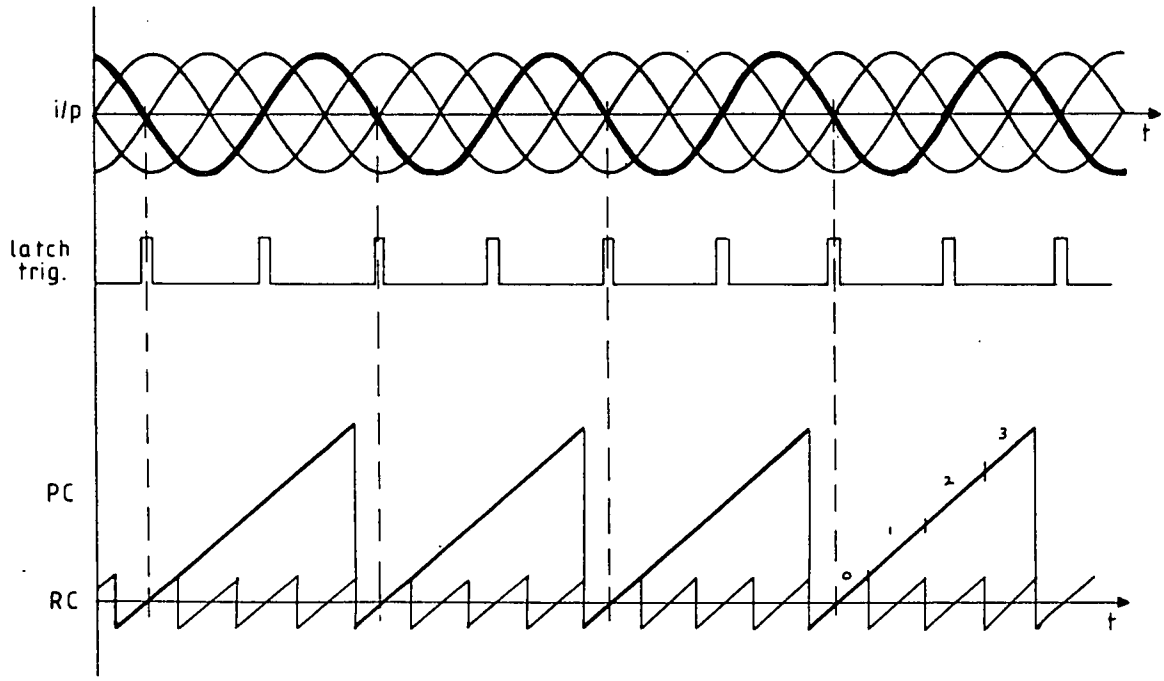


Figure 3 Illustrating the Error Signal Partitioning for 4 Phase PSK

and negative going zero crossings, to increase the sampling rate of the loop. When N is even, this may be done directly (fig.3). When N is odd, some phases would attempt to synchronise to the discontinuous edge of the ramp. This may be avoided in one of two ways. The RC may be run at 2N ramps per cycle of the input signal, with the most significant bit being excluded from the feedback signal. Alternatively, the polarity of the zero crossing may be used to alternately invert the most significant bit of the RC to produce a phase shift of half a ramp.

4 Loop Elements and Implementation

4.1 The Limiter and Latch

The limiter provides the analogue-digital interface for the loop, it may be thought of as a one bit ADC. The use of a limiter implies a performance degradation approaching 2dB at low signal to noise ratios (SNR), which would normally preclude its use in low SNR coded data applications. However, for a large class of PSK systems, where multi-level keying is used to overcome bandwidth restrictions, the SNR is good and there is a 3 dB. advantage to be gained by using a limiter. As would normally be the case, the input signal should have been passed through a filter matched to the PSK symbol rate, to remove out of band noise, before being applied to the limiter. The transfer characteristic of a limiter normally exhibits a small amount of hysteresis. In applications where only a single polarity of zero crossing is being used to drive the SDSPLL, a small amount of hysteresis will move the phase reference very slightly from the zero transition of the signal. The switching thresholds may be set to avoid this, but the resultant static phase offset is almost always insignificant. When both polarities of zero crossing are being used, the static phase offset is again negligible, but in this case the timing offset resulting from the hysteresis should be symmetrical about zero. If it is not, the period from a positive zero crossing to a negative zero crossing, and vice-versa, will differ. The resulting jitter may degrade performance if it is comparable to a clock period of the RC. Because the sampling times of the latch are dictated by the input signal, successive loop sampling periods are not uniform. Non uniform sampling is a common feature of sampled data loops when in transient states. In the case of the SDSPLL, the sampling may be expected to be non uniform in all circumstances; except for the special case of an input signal with uniformly spaced zero crossings. This makes the analytical prediction of loop performance difficult.

4.2 Loop Filter

The filter controls the transient behaviour of the loop, and the phase noise on the recovered carrier when locked. Loop filters rarely exceed first order, resulting in a second order closed loop response when the integrating action of the local oscillator is included. The result of going to a third order loop does not often justify the complication, and in the case of some sampled data filters, has been shown to produce inferior performance

[4]. Consequently, only the first and second order SDSPLL are being investigated at present.

For a first order loop, the filter is required to produce simple attenuation. If the attenuation coefficient is restricted to a power of two, and the multiplication is effected by binary shifts, the hardware required may be kept simple. The performance loss produced by this restriction in the value of the coefficient is negligible. The generally non-critical nature of phase-locked loop optima [5] allows this useful approximation. For a second order loop, a filter providing a pole and a zero allows independent control of bandwidth and damping. Once again, binary shift approximations to optimum multiplier coefficients are used. Although the terms 'pole' and 'zero' have been used here, the usual techniques of S and Z plane analysis are not directly applicable, due to the non uniform sampling. The effects of this irregular sampling on loop dynamics and stability are still being assessed. Some techniques aimed at stabilising the loop filter characteristics have been proposed:

(i) Decimation of the incoming samples, in order to smooth out the incoming sample rate.
(ii) Clocking the loop filter at a fixed rate, independent of the incoming sample rate.
(iii) Statistical filtering.

In its simplest form, decimation involves accumulating the values of a fixed number of samples and then using the result, renormalised if required, as an input sample. As the number of samples accumulated is increased, so the jitter on the timing of the decimated samples is reduced. For a review of decimation techniques see [6]. The drawback of this technique is the additional delay which it introduces. As with any feedback system, long delays within the loop have a detrimental effect on stability. Having the filter driven by an independent clock is a natural choice when RC truncation is being used as the frequency control technique. Since the RC only accepts a filter output sample once per cycle, frequency control samples will be missed if they occur more frequently. There is still a variable delay factor, due to timing jitter on the input samples, but the effect of this is less serious. If filter clock rate is significantly greater than the mean input sample arrival rate, this becomes an interpolation technique [6]. Statistical filters are not characterised by measures such as bandwidth. The timing of output samples is not explicitly dependent on input sample rate, but depends on a certain threshold being exceeded by some statistical measure of the input samples. An example is the 'random walk' filter, which generates an output sample when the mean of the input samples exceeds a set bound. This type of filter has already found application in phase locked loops [7]. In all three cases, if RC truncation is being used as the means of frequency control, it is important that the filter does not generate new outputs at a rate greater than the cycle period of the RC. If two outputs are generated within the period

of one RC cycle, only the later one will be loaded for the following RC cycle.

4.3 NCO, RC and PC

The RC and PC perform the quantisation in the loop and, in conjunction with the latch, provide the phase discrimination function. The quantisation precision is limited by the length of the RC, which in turn is limited by the rate at which the NCO can be clocked. This depends on the technology used and the manner in which the NCO is implemented. Current work is being focussed on purely digital NCO techniques, where the clock origin is a crystal reference. One technique being investigated for this application will be discussed: truncation of the RC cycle period to implement the NCO as part of the RC.

4.3.1 RC Truncation Frequency Control

The RC is clocked directly by the crystal source, and has a parallel preload port. The RC output is made to agree with the preload input at the clock pulse which would cause the counter to recycle. An advantage of this method of frequency control is that it is simple to implement, and requires little additional hardware. A static phase error is introduced by truncating the ramp at one end. For a first order loop, this error often may not be significant. However, in a second order loop it must be compensated for, otherwise the integrating action of the filter, when fed with the static offset, will cause the loop to latch up. Therefore a variable offset of one half of the previous preload should be subtracted from the error sample before it is fed to the filter, to remove the static phase error which would otherwise result from truncating the ramp at one end. For first order loops with narrow bandwidths, and therefore low gain, this offset usually will not be significant. It may well be exceed by the normal static phase offset incurred due to frequency offset on the incoming signal. Correcting the offset increases the complexity and produces a small shift in the positions of the open loop poles, but does not change the order of the loop.

Because the frequency of the local oscillator is being controlled by varying its period, the transfer characteristic is non-linear. This non-linearity, in conjunction with the varying number of quantisation levels in a ramp, causes the loop gain to vary with input signal frequency, although the variation is small for shifts of less than 20% from the nominal centre frequency. A disadvantage of this approach to frequency control is the length of RC needed to give the required precision in adjustment. This is the dominant restriction on the maximum operating frequency of the loop, the centre frequency becoming lower as the number of phase states of the input signal (length of the PC) becomes larger. For a centre frequency of 1800 Hz, 10 phase states and a clock source of 20MHz, the nominal RC length is approximately ten bits. This allows a frequency resolution of 0.1% for small variations.

Tests using a carrier in noise have indicated that for a ten phase loop, the

variable delay and potential loss of some update samples encountered when using ramp truncation as the method of frequency control, produce no significant difference in performance when compared with a frequency control technique which does not have these properties. This result may be directly extended to narrowband PSK signals. It has yet to be determined if the result holds for wideband signals.

5 Analogue/Digital Realisation

This description has concentrated on aspects of a digital implementation of the SDSPLL. A purely digital version has the advantage of being better suited to silicon microfabrication than an analogue version. The analogue loop has the advantage of a higher operational frequency. The frequency limitation has not been a problem in the current study, which is based around a voice band modem. Where higher operating frequencies are required, the analogue SDSPLL uses a sawtooth VCO in place of the RC etc. and the latch is replaced by a sample and hold. The upper frequency limit will probably then be determined by the limiter or the sample and hold.

To prove the concept, an analogue version of the loop was built. This locked to, and demodulated, a four phase PSK signal. The loop used positive zero crossings only, and operated satisfactorily at symbol rates up to one half of the carrier frequency. Having established the viability of the initial idea, work is now proceeding on a digital version. A voice band modem is being constructed where two versions of the loop are being used. One provides synchronisation to, and demodulation of, a ten phase amplitude and phase shift keyed signal, while the second is used to derive symbol timing from the same signal. Because of the wideband nature of the signal, zero transitions caused by the phase modulation form a significant fraction of the total number. Likewise, carrier transitions interfere with the symbol timing estimation. This problem may be counteracted by using the output from each loop to cancel a proportion of the spurious pulses fed to the other.

This phase-locked loop structure is the subject of a British patent application, number 8228009.

References

1. Gupta, S.R.,"Phase-Locked Loops", Proc IEEE, vol.63 No.2, Feb 1975, pp 291-306.

2. Lindsey, W.C., and Chie, C.M., "A Survey of Digital Phase-Locked Loops", Proc. IEEE, vol 69 No.4, Apr 1981, pp 410-431.

3. Dripps, J.H., "An Optimum Modulation Method for Digital Data Transmission on H.F. Ionospheric Channels". Ph.D. thesis, University of Strathclyde, January 1977.

4. Garodnick, J., Greco, J., and Schilling, D.L., "Response of an All Digital Phase-Locked Loop". IEEE Trans. Commun., vol COM-22 No.6, Jun 1974, pp 751-64.

5. Gardener, F.M., Phaselock Techniques. New York: Wiley, 1966.

6. Crochiere, R.E., and Rabiner, L.R., "Interpolation and Decimation of Digital Signals - A Tutorial Review", Proc. IEEE, vol 69 No.3, Mar 1981, pp 300-331.

7. Cessna, J.R., and Levy, D.M., "Phase Noise and Transient Times For a Binary Quantised Digital Phase-Locked Loop in White Gaussian Noise", IEEE Trans. Commun., vol COM-20 No.2, Apr 1972, pp 94-104.