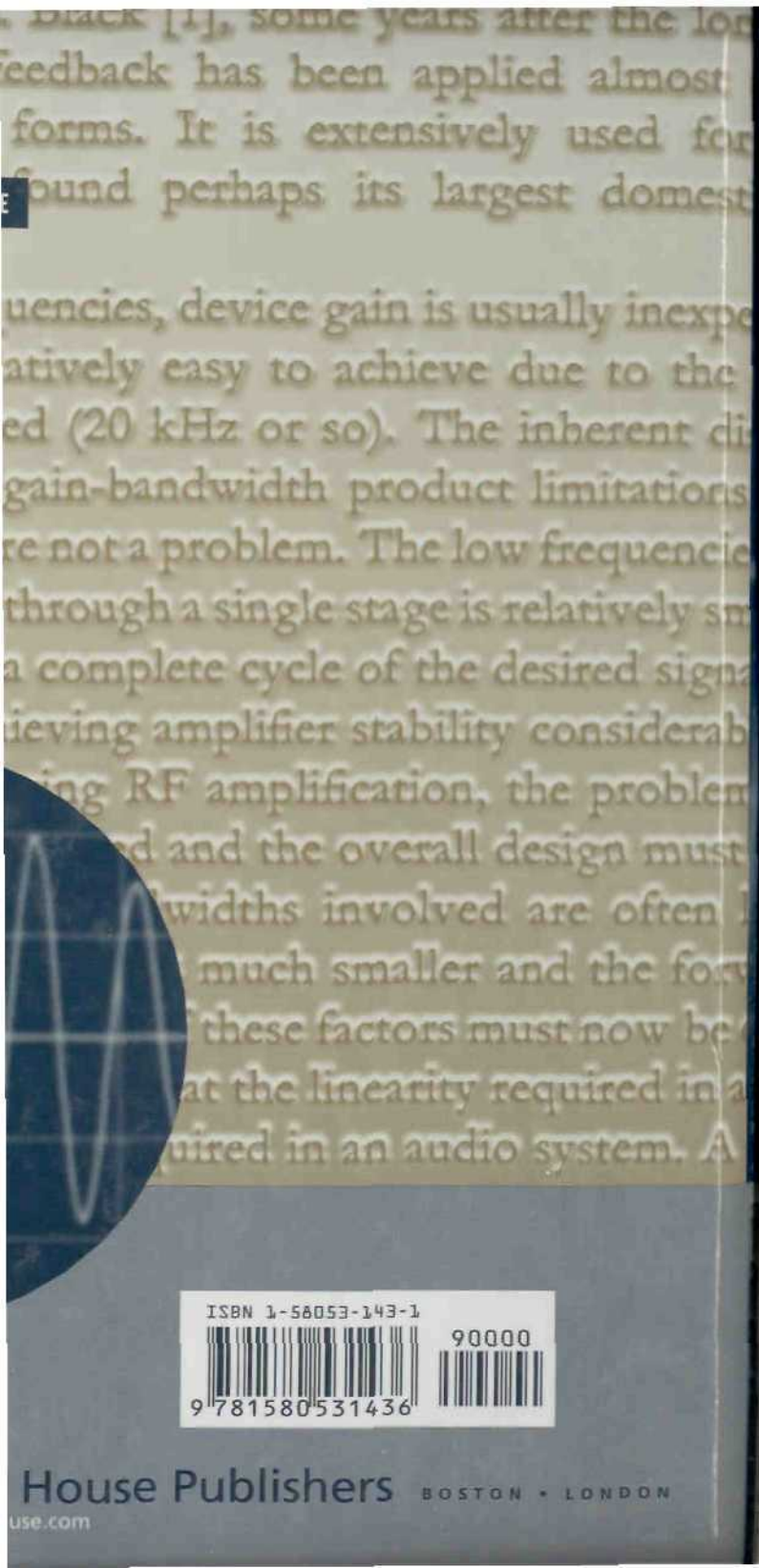
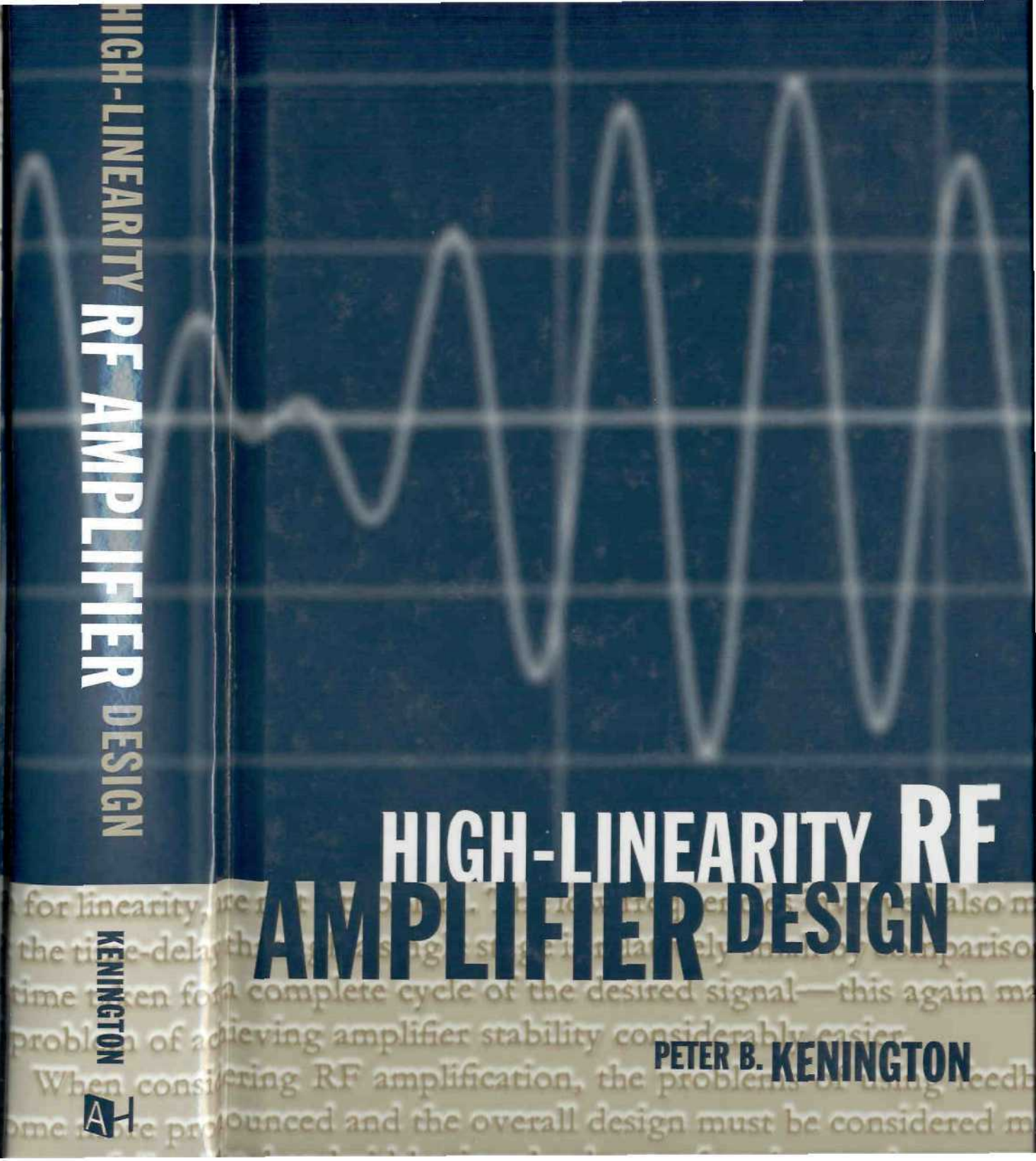


ISBN 1-58053-143-1
9 781580 531436 90000

House Publishers BOSTON • LONDON
use.com

HIGH-LINEARITY RF AMPLIFIER DESIGN

KENINGTON



HIGH-LINEARITY RF AMPLIFIER DESIGN

PETER B. KENINGTON

High-Linearity RF Amplifier Design

For a listing of recent titles in the *Artech House Microwave Library* turn to the back of this book.

High-Linearity RF Amplifier Design

Peter B. Kenington

Artech House
Boston • London
www.artechhouse.com

Library of Congress Cataloging-in-Publication Data
Kenington, Peter B.
High-linearity RF amplifier design / Peter B. Kenington.
p. cm. — (Artech House microwave library)
Includes bibliographical references and index.
ISBN 1-58053-143-1 (alk. paper)
1. Amplifiers, Radio frequency. 2. Phase distortion (Electronics)—Prevention. 3. Radio—Transmitters and transmission. 4. Mobile communication systems. I. Title. II. Series.

TK6565.A55 K45 2000 00-059401
621.3815'35—dc21 CIP

British Library Cataloguing in Publication Data
Kenington, Peter B.
High-linearity RF amplifier design. — (Artech House microwave library)
1. Amplifiers, Radio frequency 2. Phase distortion (Electronics)—Prevention 3. Radio—Transmitters and transmission
I. Title
621.3'81535

ISBN 1-58053-143-1

Cover design by Gary Ragaglia

685 Canton Street
Norwood, MA 02062

International Standard Book Number: 1-58053-143-1
Library of Congress Catalog Card Number: 00-059401

10 9 8 7 6 5 4 3 2

Table of Contents

Acknowledgements

I would like to gratefully acknowledge the help and support received from friends and colleagues, both past and present, without whom this book would have been a much poorer offering. In particular, I would like to thank Ross Wilkinson, Kieran Parsons, David Bennett, Adrian Mansell, Majid Boloorian, Simon Hetzel, Wyc Slingsby, and Andrew Bateman for their helpful comments and discussions over the years. In addition, I would like to thank Simon Jones for proofreading the manuscript and for his general encouragement in this endeavour, and also to thank my colleagues at Wireless Systems International Ltd. for help in producing many of the practical results presented here.

Finally, I would like to thank my wife Gay for her understanding and devotion throughout the many evenings and weekends taken to prepare this book, and my parents for their encouragement and support in my early electronic endeavours.

1

Introduction

1.1 Distortion

The Oxford English Dictionary defines distortion as:

Distortion *n*: Change in form of signal during transmission etc. usually with impairment of quality.

All amplifiers possess this property of distorting the signals they are required to amplify. The existence of distortion and hence nonlinearity in audio amplifiers is very displeasing to the ear and high fidelity amplifiers have been designed and refined over the years to reduce it to levels considered to be inaudible by the human ear. The advent of feedback correction by H. S. Black has enabled this to be achieved with relative ease.

When considering radio frequency amplifiers, the resolved audio fidelity of the transmitted signal is still of importance, but is no longer the only consideration. Spectral efficiency, interference and the need to be considerate to other users of the spectrum all become important, along with signal vector error considerations for the signal itself.

1.2 The Requirement for Linearity

All radio systems are required to cause the minimum possible interference to other radio users; they must therefore maintain their transmissions within the bandwidth allocated to them and not radiate significant energy outside of it. Nonlinearities within the system components of the radio equipment

cause distortion of the transmitted signal and result in the generation of signals outside of the intended frequency channel or band. These unwanted distortion products are potential interfering sources to other radio users and must be reduced to a level where both systems can operate satisfactorily.

In the case of a high power broadcast transmitter this requirement becomes acute, as the distortion products, although many times smaller than the main output signal, may still be quite large in absolute terms and hence cause interference.

Fortunately, radio frequency systems can employ filtering successfully to reduce harmonic distortion components to an acceptable level (unlike their audio-frequency counterparts). Thus transmitters radiating a single carrier signal with constant envelope modulation (such as an FM radio station) can be restricted to radiate only within their allocated bandwidth, by means of filtering alone. Transmissions incorporating envelope variations (including AM, SSB, and many filtered digital schemes) can still use filtering to eliminate harmonics, but will also produce *intermodulation distortion* (IMD) products close to the wanted channel and these cannot usually be eliminated by filtering.

It is possible, therefore, to use highly nonlinear amplifiers to transmit constant envelope modulation schemes and to rely on lowpass or bandpass filtering to remove the unwanted harmonic distortion products (see Figures 1.1 and 1.2). The fidelity of the recovered audio signal will be unaffected by any amplitude nonlinearity within the radio frequency (RF) parts of either the

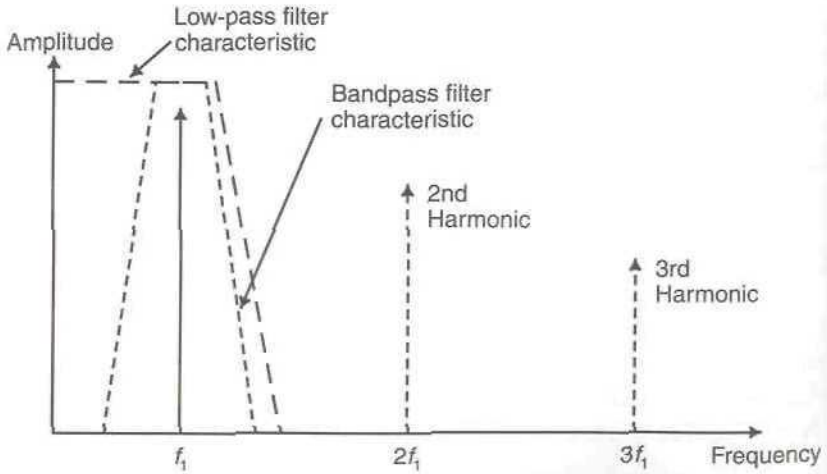


Figure 1.1 Use of filtering to eliminate harmonic distortion components.

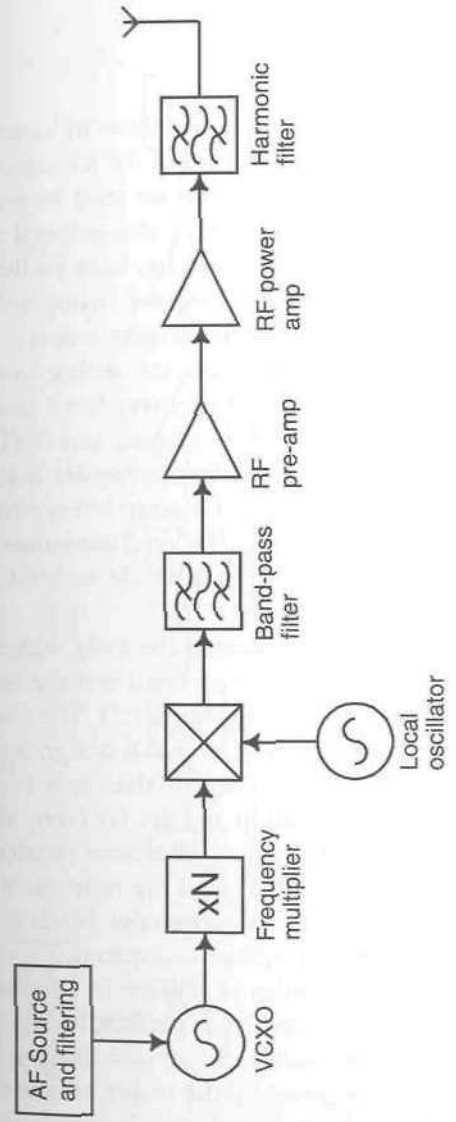


Figure 1.2 Typical FM transmitter architecture.

transmitter or the receiver. The greater efficiency[1] and simplicity of highly nonlinear amplifiers therefore renders them ideal for constant envelope transmitter applications.

1.2.1 Linear Modulation Schemes

Linear modulation schemes can be defined as those in which information is transmitted in both the amplitude and phase of the RF signal. The envelope of the RF signal thus varies with time and hence must be preserved in order to preserve the full information content of the original message signal. Recent interest in linear modulation schemes has been fuelled by the need to use the radio spectrum more efficiently, both for analog voice transmission (e.g., telephone conversations) and for data transmission.

Single-sideband is one popular choice for analog voice transmission and a number of technically advanced derivatives have been used, such as: ACSSB—amplitude companded single-sideband, and TTIB—transparent tone-in-band. Many data transmission schemes require both the amplitude and phase of the RF signal to be preserved. Recent interest has been centered around *quadrature amplitude modulation* (*QAM*) and *quadrature phase shift keying* (*QPSK*) schemes with specific interest in 16-QAM and $\pi/4$-shift QPSK for mobile radio systems.

The major drawback which prevented the early widespread adoption of many linear modulation techniques (e.g., SSB) was the requirement for a linear power amplifier in the transmitter (Figure 1.3). The traditional form of RF linear amplifier has been a class-A or class-AB design and these are both inefficient and not particularly linear. They are thus only suitable for certain point-to-point forms of communication and are far from ideal for use in a mobile radio environment. Battery life considerations restrict the use of such amplifiers from an efficiency standpoint, and the near–far effects of mobile propagation (Figure 1.4) can result in unacceptable levels of interference to other users, due to their inherent levels of distortion.

The use of a true linear modulation scheme in a mobile environment therefore requires a highly linear amplifier and hence, ideally, the use of some form of linearisation technique, such as those described in this book. The fact that spectrum efficiency is generally the major reason for employing a linear scheme means that the channel bandwidth is often narrow. The linear amplifier in the transmitter may therefore only be required to operate over a small instantaneous bandwidth; however, that bandwidth could appear at

[1] Efficiency may be defined in various ways and the assumption made in this statement is that of DC supply to RF power conversion efficiency (see Section 1.3 for further details).

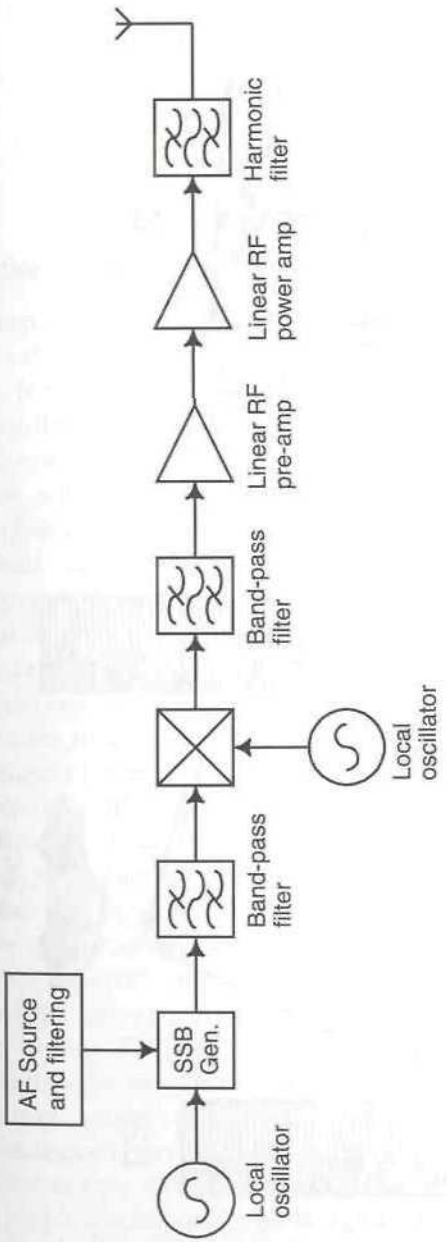


Figure 1.3 Typical (traditional) linear transmitter architecture.

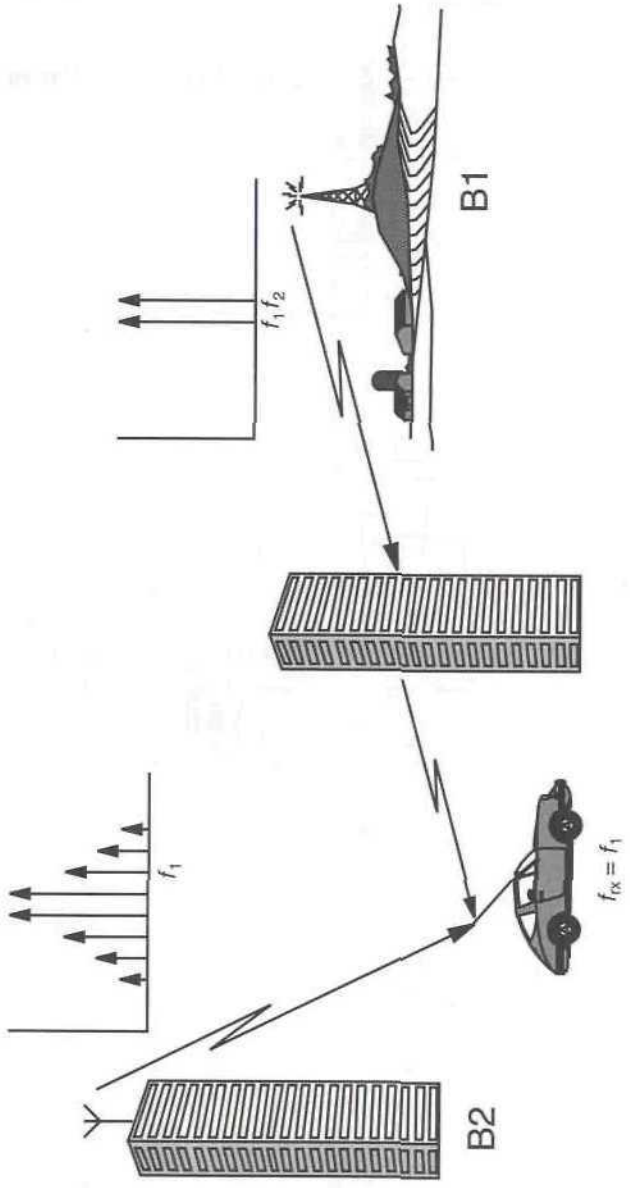


Figure 1.4 Near–far effect encountered in a mobile environment.

any point in the frequency allocation of the service. A broadband linearisation technique, such as feedforward or RF predistortion, is therefore not required (although either could be used, if desired). Cartesian loop and adaptive predistortion are currently popular for this application, although LINC and EE&R may begin to challenge this view in the near future.

Wider-band, quasi-linear schemes are also popular, such as direct-sequence CDMA, and in such cases true wideband linearisation must be employed.

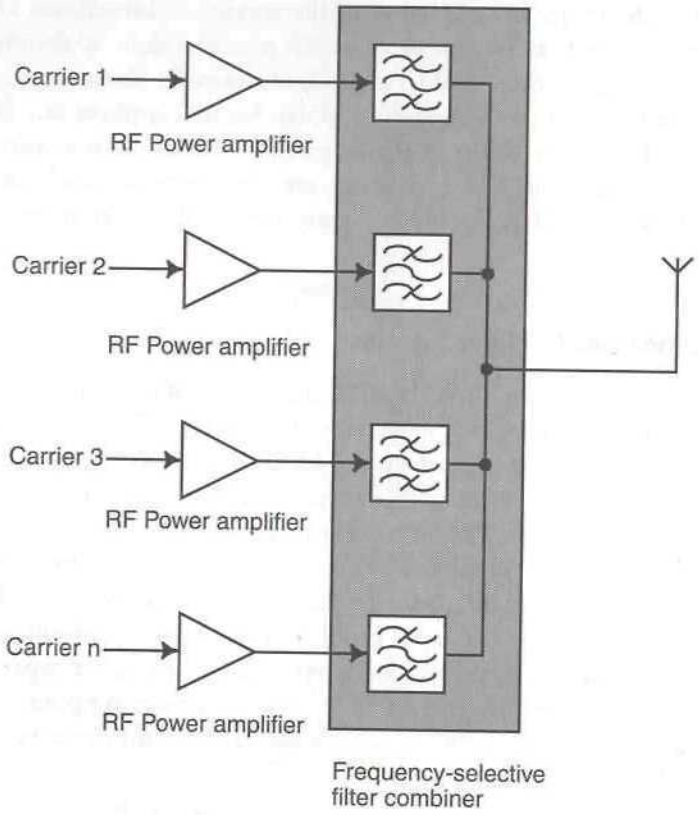
1.2.2 Multicarrier Amplifier Systems

The ability to amplify multiple channels simultaneously is a desirable, if not essential characteristic of many systems. For example, successive satellite programs have, for many years, relied on travelling wave tube amplifiers (TWTAs) as broadband multicarrier transmitters for the many traffic signals involved in such systems. The linearity constraints in this case are such that a class-A amplifier with suitable output back-off is quite adequate; intermodulation products in the order of 25 dB down on the wanted signals are often sufficient. The principal problem with output back-off is that of inefficiency: a back-off level of 10 dB means that an amplifier capable of 10 times the desired output power must be used. The power supplies, heatsinks and output devices must all be considerably larger than the actual output power level might suggest.

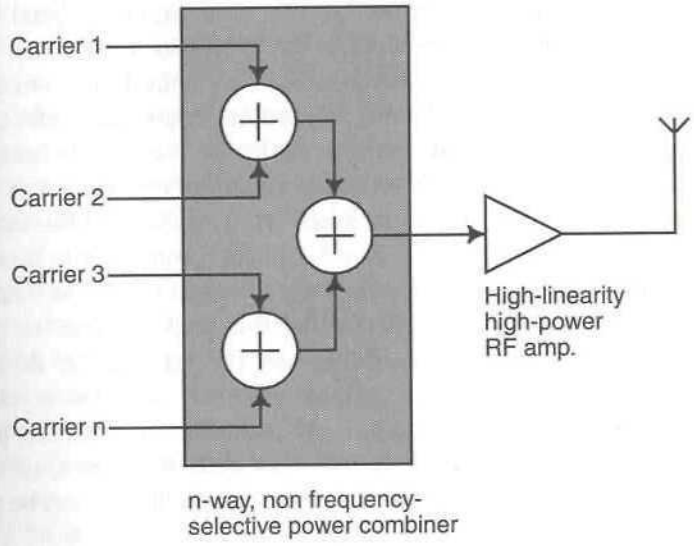
Current trends in satellite systems are, however, in the direction of increasingly stringent linearity specifications and these may soon be similar to those of terrestrial systems. This will place increasing importance on the linearity and efficiency of this type of system.

Since the degree of linearity necessary in the satellite field is generally not large (at present), a variety of simpler linearisation techniques have been applied in order to reduce the level of output back-off required. Simple intermediate frequency (IF) predistorters [1] have been used and adaptive bias techniques [2] employed, resulting in an improved efficiency for the overall amplifier system. The goal in each case is principally that of efficiency improvement (and hence weight saving) and not linearity improvement as such. The saving of weight on the satellite in turn yields huge savings in launch costs, and hence is very attractive to the satellite community.

The situation is very different for a mobile radio system, as efficiency (in the base station) is less important (although far from unimportant as will be demonstrated below), and absolute linearity is usually of much greater importance. This difference arises because of the near–far effect encountered in mobile propagation, and this is illustrated in Figure 1.4. The signal from



(a)



(b)

Figure 1.5 Low (b) and high (a) level combining for a mobile radio base-station.

the nearer base station (B1) is being blocked by the large building and hence is greatly attenuated before reception by the mobile. A spurious signal (intermodulation product, for example) from an adjacent-cell base station (B2) could therefore be of an equivalent or greater strength as it may not suffer the same attenuation. The intermodulation distortion specification of a cellular base station is therefore very tight, with the unwanted products needing to be at least 60 dB lower (or more in many systems) than the main signal output. This is true even if multi-path propagation is not present, although multi-path will generally increase the number of locations at which problems may be encountered. Thus, if a linear amplifier is to be used in a cellular or PCN base station, the design of the linearisation system must be considered very carefully. Currently the most popular choice for this broadband case is feedforward, with low-level broadband power combination employed at the input.

The configuration of a low-level combination system is shown in Figure 1.5, where it is compared with the traditional approach of using a tuned power combiner. The use of tuned power combination has many disadvantages, beside the inherent power loss, which occurs at the point where it can least be afforded, that is, at the output of the base station [3]. The use of dynamic channel allocation is described below and such techniques cannot easily be applied with a tuned combiner (often called a *multicoupler*). An alternative form of high-power combining, using high-power hybrid combiners, overcomes the frequency inflexibility disadvantages of cavity combining, but at the expense of significant combiner loss (3 dB per pair of channels). The current interest in broadband linear amplification is thus very great.

1.2.3 Multicarrier Modulation Formats

The use of multiple narrow-band carriers is emerging as a method for transmitting very high bit-rate data without the need for an equaliser. The

primary requirement for this technology, at present, is in digital broadcast systems, such as digital audio broadcast (DAB) [4] and digital video broadcast (DVB) [5]. These schemes utilise orthogonal frequency division multiplexing (OFDM) as their modulation format and involve many hundreds (or in some cases thousands) of individual carriers, such that each can be treated as an individual signal within the coherence bandwidth of the channel. It is this use of sufficiently narrow-band carriers which removes the need for an equaliser.

The carriers are, ideally, arranged in such a manner as to minimise the required peak-to-mean ratio of the overall channel to be transmitted. Without this encoding, theoretically the full peak-to-mean ratio of this number of carriers would be presented to the transmitter. A very high power, linear amplifier would then be required and this would almost certainly be impossible to realise for any sensible level of mean power (in a broadcast context). Fortunately, the statistical likelihood of this peak power being reached is small and hence a degree of clipping can be tolerated. This fact is relied upon in many OFDM systems.

Multicarrier modulation has also been proposed for the same purpose in narrower band systems, such as SMR (specialised mobile radio) in the United States. One example uses six narrowband, linearly-modulated carriers, employing 16-QAM on each, within a narrowband channel. This is again designed to eliminate the requirement for an equaliser, but does, of course, still require a linear transmitter—both for the multicarrier format of the modulation as well as for the amplitude and phase modulations inherent in the use of 16-QAM on each carrier.

1.2.4 Dynamic Channel Allocation

The use of frequency synthesisers in the base station transceiver units of a mobile radio system should allow an individual transceiver to operate on any channel within the allocated system bandwidth. This should therefore permit the distribution of channels within the entire network to be reallocated to meet demand at various points during the day; this is known as *dynamic channel allocation* [6,7]. At present, however, the use of tuned power combiners in many systems (to reduce loss) places an additional restriction on the frequency agility of the system. A given transceiver may only be used within the set (narrow) bandwidth of the cavity resonator to which it is connected. Alteration of the operational frequency of the transceiver therefore not only requires reprogramming of the synthesiser but also necessitates retuning of the cavity combiner. It is this particular operation which often requires direct human intervention and hence greatly increases the time taken and therefore the expense of the operation.

The combiner manufacturers are attempting to address this problem by developing remotely-tuneable combiners utilising either mechanical servo-motor control systems or varactor diodes. The former technique is obviously slow and has the potential for unreliability, particularly given the infrequent operation predicted for the system. It will also increase the already considerable size of the combiner still further.

The varactor-based systems [8,9] overcome the speed and reliability problems of the mechanical system, but decrease the intermodulation performance of the combination process and restrict the power handling. They are thus removing arguably the two greatest strengths of the cavity combination system, and no commercial high-power systems are currently known to be in widespread use.

Both of the above variable tuning systems are limited in their bandwidth of operation and could not hope to provide total flexibility. At best they may provide a 'stop-gap' solution in some systems but should ultimately be replaced by broadband amplification techniques.

The application of a broadband amplifier to a base station scenario allows total flexibility. Not only can any transceiver be used on any channel within the system bandwidth, but additional channels may simply be added, without significant modification to the system (assuming that the power amplifier rating is large enough).

The system is, ideally, independent of the modulation scheme employed and should even permit a number of different schemes to be utilised simultaneously at the same base station. This could be of considerable benefit during the transition from an old system to a newer one, such as in the transition from the current GSM system to, for example, GSM EDGE. The financial savings which may thus be made have prompted considerable investment in research into broadband amplification techniques.

1.3 Power Efficiency

1.3.1 Single-Carrier Applications

The issue of power efficiency of amplifiers and modulation schemes has been the subject of considerable debate, in particular between the advocates of constant envelope and linear modulation schemes. The use of class-C amplification has, for many years, been justified on the grounds of power efficiency and has of itself been used to justify the continued service of constant envelope modulation schemes. For a constant envelope signal,

there is no doubt that the efficiency of a class-C stage is very high compared to the more traditional forms of linear amplification (class-A and class-AB); however, what is perhaps more important is the issue of battery drain.

The increasing use of hand-portable equipment has focused attention on methods of decreasing the size and weight of the handheld unit and consequently on reducing the size and weight consumed by the battery. A class-C stage in a unit utilising a constant envelope modulation scheme will consume its full current from the battery no matter what level or form the input signal takes. Thus, for a unit designed to transmit analog voice the battery drain will be identical on the voice peaks as it is in the gaps between words. By comparison, a unit designed around a linearised class-C module (using a Cartesian loop, for example) and a linear modulation scheme, will only draw a battery current corresponding to the level of the modulation [10]. Thus, very little current will be drawn during the interword gaps and the overall current drawn will be comparatively low, due to the low duty cycle of normal speech.

The above is perhaps a rather simplistic argument, particularly as it takes no account of companding on the voice signal, but serves to illustrate a point: when considering the efficiency of an amplifier, its raw CW (carrier wave) efficiency is only part of the problem; the intended application must also be considered. It is no longer the case that a linear amplifier is, almost by definition, an inefficient amplifier, as many of the configurations described in this book prove.

1.3.2 Multicarrier Applications

The use of linearised power amplification in a multicarrier application, such as a cellular base station, can also lead to significant power savings, despite the more modest efficiencies available from the high-linearity systems required in this case.

Power efficiency is of prime importance in a base station PA, since both capital and running costs are influenced by it. The requirement for air conditioning (or otherwise) and its maintenance and running costs are an additional drain on resources, over and above that required for the equipment itself. With the increase in market penetration of mobile networks, it is predicted that 8 or 16 carrier base stations will be common and even dominant in the future. The power consumption of a base station amplifier resulting from using the techniques detailed in this book could be roughly one-third (or less) of that of a conventional hybrid combiner amplifier solution (based on the ETSI GSM product developed by Wireless Systems International Ltd. [11]). Across a complete network of around

10,000 base sites (based on a major EU country and a DCS1800 system), with an average supply requirement per site of 2.5 kW (dominated by the RF section and based on 8 carriers with conventional combining), this leads to an *annual power saving* of 164 million kilowatt hours. This would clearly have a significant impact on network running costs (approximately £10M/annum).

The impact of this technology on future generations of communications networks (e.g., 3G and GSM EDGE) is predicted to be even more dramatic, since these will be using modulation formats with inherent envelope variations. The conventional approach of combining multiple single carriers will therefore become significantly less efficient than was assumed above. Coupled with this, the number of base sites is predicted to increase in order to support a larger number of users and significantly greater bandwidth demands (such as rapid data transfer or internet browsing). Some estimates predict that the required number will be roughly five times that of a conventional DCS1800 network (i.e. 50,000 for an EU country). Based on these figures, the power consumption benefit across a network can now lead to a saving of up to £150M/annum. This is clearly a substantial sum.

1.3.2.1 Base Station Example—GSM EDGE

GSM EDGE is emerging as a '2.5G' standard providing a relatively straightforward upgrade path from current GSM networks. Many of the system parameters are the same or similar to those of GSM (unlike those of the wideband CDMA standard adopted for 3G) with the principal difference being the adoption of a filtered 8-PSK modulation format. The filtering applied to the modulation results in an envelope variation for the modulated carrier, creating a 3.2 dB peak-to-average ratio (PAR). The consequence of this for the amplifier is in the imposition of a linearity requirement. This linearity requirement is difficult to meet (particularly in a production system, over a wide temperature range) by a conventional unlinearised (but still linear) single-carrier power amplifier (SCPA).

Thus, *most* EDGE implementations will need a linearised power amplifier, in order to achieve a reasonably sensible power efficiency, of between 15% and 25% (for the amplifier alone). This is rather less than is achieved in some comparable GSM installations and hence the overall efficiency of an EDGE base station is likely to be less than that of a conventional GSM site.

Signal vector error (SVE) is also an important parameter for GSM EDGE and is affected by the linearity of the power amplifier (in addition to local oscillator phase noise, and modulator accuracy). SVE requirements can, in some cases, place a more stringent requirement on amplifier linearity than does adjacent channel performance.

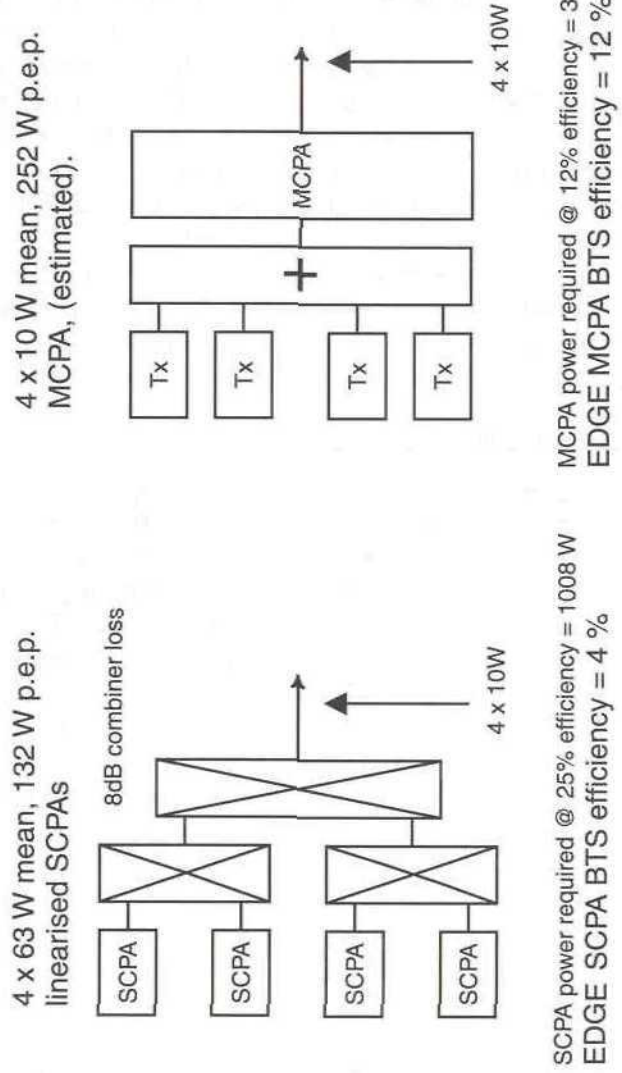


Figure 1.6 *Comparison of SCPA and MCPA approaches to an EDGE base station architecture.*

Figure 1.6 clearly shows the significant energy saving resulting from the adoption of an MCPA-based approach to an EDGE base station architecture, when compared to an SCPA-based approach employing passive power combining. The figures shown can be considered as 'worst-case', since the SCPA efficiency figure used may be optimistic and the MCPA efficiency can probably be improved upon. In addition, this energy saving does not take into account the power consumption of the additional cooling required in the SCPA based approach (for example, air conditioning).

Whilst the above argument assumes a passive combining network for the SCPA-based approach, even the use of spatial combining or frequency-selective combining (that is, lower loss combining) results in the MCPA approach being at least as efficient. Thus the other benefits of, for example, very low EVM degradation ($< 0.1\%$), more efficient space and cooling utilisation within the base site, and improved ACP around the carriers (that is, each carrier very easily meeting its individual mask) are all achieved for 'free'. The latter benefit greatly aids efficient spectral utilisation in congested networks and hence can be particularly valuable in, for example, 900 MHz systems. In addition, the use of frequency-selective combining places severe restrictions on the use and spacing of carrier frequencies within a base station and introduces additional mechanical or electro-mechanical tuning issues.

The use of linearisation in the form of multicarrier power amplification in base station applications is therefore very desirable and is set to increase dramatically in the future.

1.4 Effect of Nonlinearity on a W-CDMA System

The new third-generation wideband CDMA (W-CDMA) standard places some interesting constraints on the linearity of both the handset and base station transmitters [12]. The envelope variations inherent in the modulation format lead to an adjacent channel issue, in much the same way as for any narrowband system—however, in the case of W-CDMA, this is generally an issue for another network operator and not one of 'self-interference' as in most other mobile radio systems. Thus, to take the scenario to its extreme, a given operator *could specify low-linearity equipment, with poor adjacent* channel performance, and could still operate his network with little or no loss of capacity. A neighboring network operator (in frequency terms) would, however, suffer a severe loss of capacity, particularly close to his rival's site. An adjacent channel power (ACP) standard is therefore required for both the handset and the base station and this must be based on an overall loss of capacity for all of the networks to be deployed and not just for a single

Table 1.1
Adjacent channel power requirements for a 3G handset [13]

Frequency offset from carrier Δf	Minimum requirement	Measurement bandwidth
2.5–3.5 MHz	$-35-15*(\Delta f-2.5)$ dBc	30 kHz (Note 1)
3.5–7.5 MHz	$-35-1*(\Delta f-3.5)$ dBc	1 MHz (Notes 2 and 3)
7.5–8.5 MHz	$-39-10*(\Delta f-7.5)$ dBc	1 MHz (Notes 2 and 3)
8.5–12.5 MHz	−49 dBc	1 MHz (Notes 2 and 3)

Notes:

1. The first and last measurement position with a 30 kHz filter is 2.515 MHz and 3.485 MHz.
2. The first and last measurement position with a 1 MHz filter is 4 MHz and 12 MHz.
3. The lower limit shall be −50 dBm/3.84 MHz or which ever is higher.

operator. At the present time, this has been set to the values outlined in Table 1.1 (for handsets) [13]. A similar table exists for base station equipment [14].

As the number of users on the 3G networks increase, it is possible that improved linearity specifications will be required in order to realise the necessary capacity. This will require cooperation between different network operators and also the upgrading of equipment to a higher linearity standard.

1.5 Requirement for Linearity in Adaptive Antenna Systems

An adaptive antenna system may be deployed in a cellular network in order to provide capacity enhancement without the need for further cell sites. The available frequencies may be reused in a number of beams and these beams steered to follow the required traffic pattern, so long as they are not required to overlap, as serious co-channel interference would then result. As a consequence of the operation of an adaptive antenna system, each of the antenna elements must be capable of radiating all of the available channels and the most obvious (and generally lowest cost) solution is therefore a multicarrier power amplifier for each array element.

In this configuration, there are two effects of nonlinearity within the amplifier on system performance:

1. Intermodulation distortion causes adjacent channel interference to users in another cell, or even users served by another 'beam' in the same cell.

Table 1.2
Effect of intermodulation distortion on adaptive antenna performance [15]

Intermodulation level (dBc)	Beamwidth (degrees)	Sidelobe level (dB)	Null depth (dB)	Change in null direction (degrees)
−10	11.7	−8.7	−35.7	9.9
−17	13.3	−12.6	−29.0	7.4
−20	14.1	−15.0	−26.2	6.2
−27	15.6	−21.4	−24.6	2.6
−30	15.9	−23.1	−25.8	1.7
−75	16.4	−29.6	−42.5	0

2. Intermodulation products cause a degradation in the antenna amplitude and phase weightings and hence a degradation in the antenna beam pattern and null depth.

This latter effect has been studied and the results presented in [15]. The effect of different levels of intermodulation distortion on beamwidth, sidelobe level, null depth and the change in null direction were quantified by means of a computer simulation and the results are summarised in Table 1.2. These results are based on an 8-element linear array, with Dolph–Chebyshev weighting for a sidelobe level of 30 dB below the main lobe.

The results clearly show that the sidelobe level of the array is degraded by high levels of intermodulation distortion and that the change in null direction is also marked. This has implications for the array's ability to reject interferers by steering nulls to reduce their effect.

An IMD level of −75 dBc is also shown in Table 1.2 for reference purposes. This shows that although most of the parameters approach acceptable levels at −30 dBc IMD, null depth is still significantly affected and this could have serious implications for system performance.

A curious effect is also evident from Table 1.2, in that the null depth appears to improve as the IMD level degrades. This is reported to be due to the 30 dB Dolph–Chebyshev weighting becoming distorted. All of the other parameters degrade, hence rendering this 'benefit' useless.

1.6 Organisation of the Text

The text is divided into eight chapters covering all aspects of linear radio frequency amplifier design. The main sections may be summed up as follows:

1. Introduction.
2. Distortion in Amplifiers. The concepts of noise and distortion in both small and large signal amplifiers are covered including a discussion of the various distortion measurement techniques and performance parameters. Amplifier modelling techniques are also discussed.
3. RF Amplifier Design. All of the standard power and small-signal RF amplifier classes are discussed, including various forms of switching amplifier. They are compared in terms of linearity, efficiency and operational characteristics. Biasing techniques and affects are also covered in this chapter.
4. Feedback Linearisation Techniques. The use of feedback in linearising various forms of radio frequency amplifier is discussed. The problems of stability in RF feedback systems are highlighted and a method of overcoming some of these problems is described in the form of modulation feedback systems. The design of both polar and Cartesian loop systems is described and practical results and problems are discussed.
5. Feedforward Linearisation. The recent resurgence of interest in feedforward techniques for RF amplifier linearisation is due to their inherent broadband properties and stability. The theory and practice of feedforward systems is discussed for both RF and microwave applications and the problems of temperature stability and component ageing are highlighted.
6. Predistortion Linearisation. The use of RF, IF and baseband predistortion is covered, including both adaptive and nonadaptive schemes. The relative merits and applications of the various techniques covered are also discussed.
7. Linear Transmitters Employing Signal Processing. This category of linear power amplifier or transmitter technique requires the use of baseband signal processing as a fundamental part of its operation. The resulting linear amplifiers and transmitters are generally narrow-band and highly efficient (at least in theory). Envelope elimination and restoration, LINC (LInear amplification using Nonlinear Components) and LIST (LInear amplification using Sampling Techniques) techniques are described.
8. Efficiency Boosting Systems. A number of 'linearisation' techniques exist which are primarily intended to improve the efficiency of an already linear amplifier. The theory and practical limitations of these techniques are examined and their success is gauged in terms of both efficiency and linearity.

References

1. Schwarz Jr., W. J., R. P. Slade, and J. J. Kenny, "Radio repeater design for 16 QAM," in *Proc. of the IEEE International Conference on Communications*, Vol. 1, June 1981, pp. 13.5.1–13.5.7.
2. Saleh, A. A. M., and D. C. Cox, "Improving the power-added efficiency of FET amplifiers operating with varying envelope signals," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 31, No. 1, January 1983, pp. 51–56.
3. Johnson, A. K., and R. Myer, "Linear amplifier combiner," *37th IEEE Vehicular Technology Conference*, Tampa, Florida, USA, 1–3 June 1987, pp. 421–423.
4. ETS 300 401: Radio Broadcasting Systems; Digital Audio Broadcasting (DAB) to Mobile, Portable and Fixed Receivers, Edition 1–1995–02, European Telecommunications Standards Institute (ETSI).
5. ETS 300 744: Digital Broadcasting Systems for Television, Sound and Data Services; Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television, Edition 1–1996–05, European Telecommunications Standards Institute (ETSI).
6. Beck, R., and H. Panzer, "Strategies for handover and dynamic channel allocation in micro-cellular mobile radio systems," *IEEE 39th Vehicular Technology Conference*, San Francisco, California, USA, May 1989, pp. 668–672.
7. Nettleton, R. W., and R. Schloemer, "A high capacity assignment method for cellular mobile telephone systems," *IEEE 39th Vehicular Technology Conference*, San Francisco, California, USA, May 1989, pp. 359–367.
8. Kazeminejad, S., A. Mahdi, T. Zagar, A. Khanifar, and D. P. Howson, "Electronically reconfigurable cellular radio multicouplers," *IEE 5th International Conference on Cellular Radio and Personal Communications*, University of Warwick, UK, 11–14 December 1989, pp. 103.
9. B. Singh, A. Mahdi, and D. P. Howson "Transient responses of electronically tuned cavity resonator," *IEE Electronics Division Colloquium on RF Combining*, IEE, London, UK, 12 April 1990, pp. 6/1–6/7.
10. Bateman, A., R. J. Wilkinson, and J. D. Marvill, 'The application of digital signal processing to transmitter linearisation," *IEEE 8th European Conference on Electrotechnics, Stockholm, Sweden, 13–17 June, 1988*, pp. 64–67.
11. http://www.wsil.com
12. Hamalainen, S., H. Lilja, and A. Hamalainen, "WCDMA adjacent channel interference requirements," in *Proc. of the 50th IEEE Vehicular Technology Conference, Fall 99*, Vol. 5, Amsterdam, The Netherlands, 19–20 September 1999, pp. 2591–2595.
13. ETSI TS 125 101 v3.2.2 (2000–04): UMTS; UE Radio transmission and reception (FDD).
14. ETSI TS 125 104 v3.2.0 (2000–03): UMTS; UTRA (BS) FDD; Radio transmission and reception.
15. Xue, H., M. Beach, and J. McGeehan, "Non-linearity effects on adaptive antennas," *IEE 9th International Conference on Antennas and Propagation*, Vol. 1, Eindhoven, The Netherlands, April 1995, pp. 352–355.

2

Distortion in Amplifiers

2.1 Introduction

In any discussion of amplifier linearity and of methods of ensuring that linearity is maintained to a high degree, it is necessary to examine the nature of amplifier distortion in all of its various forms and to establish techniques for determining its level accurately and simply.

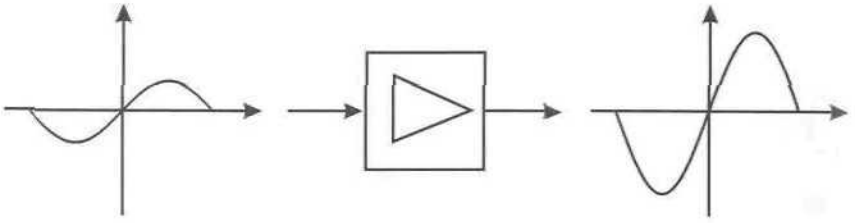
Audio amplifier distortion has been of concern for very many years and considerable design effort has resulted in its virtual elimination from modern high-fidelity amplifiers. The feedback techniques conventionally used at audio frequencies are, however, not generally applicable to many radio-frequency amplifiers due to problems of stability at high bandwidths and of cost for high gains in RF stages. As a result many RF amplifier designs need to address the compromise of linearity vs. efficiency. This chapter examines the various forms of RF amplifier distortion, together with some of the standard methods of measurement and characterisation.

2.2 Amplitude Distortion

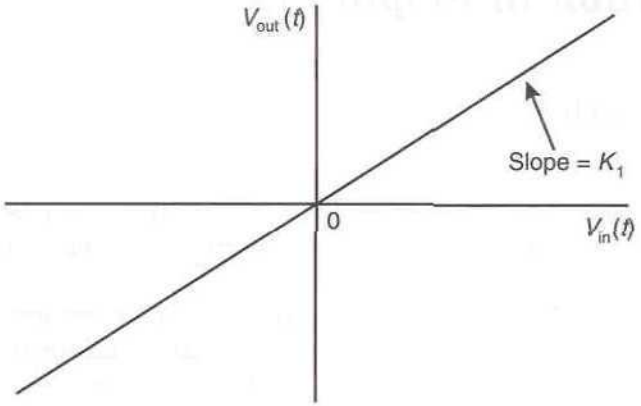
A perfect amplifier would have a linear transfer characteristic, where the output voltage would be a scalar multiple of the input voltage, that is,

$$V_{out}(t) = K_1 V_{in}(t) \tag{2.1}$$

where K_1 is the *voltage gain* of the amplifier. This situation is illustrated in Figure 2.1.



(a)



(b)

Figure 2.1 Ideal amplifier (a) with a linear transfer characteristic (b).

The output waveshape from such an amplifier will be identical to that of the input and no new (additional) frequency components will be introduced either within or outside of the amplifier bandwidth.

2.2.1 Square-Law Characteristic

The simplest form of amplitude nonlinearity may be illustrated by the addition of a second term to the transfer characteristic (2.1): a term proportional to the square of the input voltage.

$$V_{out}(t) = K_1 V_{in}(t) + K_2 V_{in}^2(t) \tag{2.2}$$

This form of transfer characteristic is referred to as *second-order* due to the power of two which has now been introduced. Figure 2.2 illustrates an

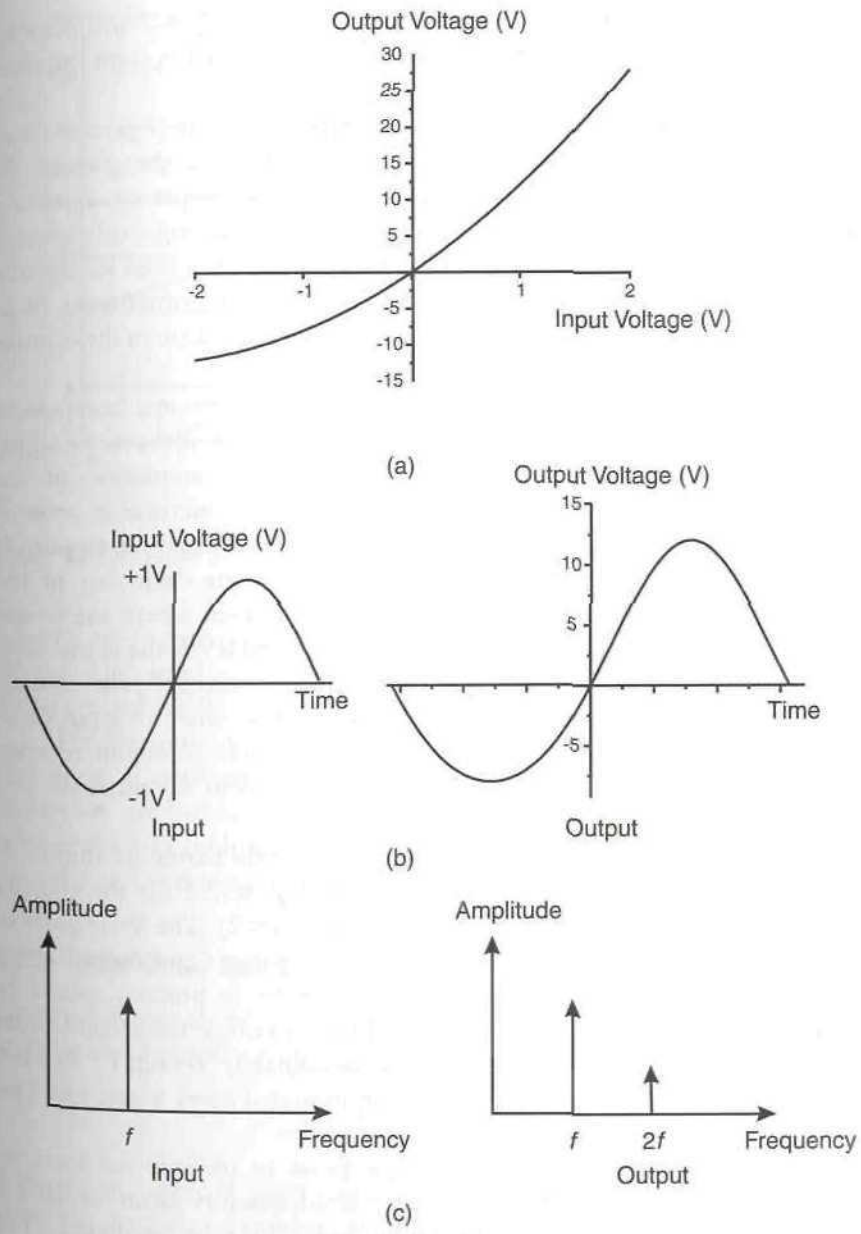


Figure 2.2 Transfer characteristic (a) and effect on a sinusoid in the time domain (b) and frequency domain (c) of an amplifier with transfer characteristic: $V_{out}(t) = 10V_{in}(t) + 2V_{in}^2(t)$.

example characteristic for the case where $K_1 = 10$ and $K_2 = 2$ and demonstrates the effect of such a characteristic on a pure sinusoid in both the time and frequency domains.

The larger the coefficient of the second-order term (K_2), the more curved the transfer characteristic will appear and hence the greater the distortion of the input waveshape. Note that in the frequency domain a second signal component has now appeared at twice the original frequency ($2f_1$) and this gives rise to the term *second harmonic distortion* used to describe the form of nonlinear distortion introduced by the second-order term. Note further that a DC term also results from the second-order term in the transfer characteristic.

Examination of the amplitude of the second harmonic component indicates that it will increase in proportion to the square of the input signal (and also in proportion to the constant, K_2). The amplitude of the fundamental frequency component, however, will only increase in proportion to the fundamental gain, K_1. As a result, it is evident that the amplitude of the second harmonic will increase at a greater rate than that of the fundamental component. A point can thus be envisaged where the fundamental and second harmonic components are of equal level; the signal level at which this would occur is termed the *second-order intercept point*, usually expressed as a power in dBm. This may be quoted as either an input or an output intercept point; the former is most commonly found in receiver front-end specifications and the latter is the usual form for medium- and high-power amplifiers.

The characteristics of the fundamental and second harmonic amplitude levels, with varying input level, are shown in Figure 2.3 for the transfer characteristic illustrated previously ($K_1 = 10$ and $K_2 = 2$). The latter parts of the two characteristics are shown dotted since the input and output levels required to obtain these parts of the characteristics in practice would be impossible without destroying the device. In this example the second-order (output) intercept point may be quoted as approximately 50 volts (+47 dBm for a 50Ω load), corresponding to the output signal level where the two characteristics cross.

The advantage of using an intercept point to indicate the linearity performance of an amplifier is that it is a fixed quantity from which the distortion level at a particular operating point may be predicted. The percentage of harmonic distortion which is generally specified in audio amplifiers must be referenced to a particular output power level (usually the maximum for which the amplifier is rated) and gives no indication of the amplifier's performance below that level. A compromise often adopted in RF amplifiers is to de-rate them from their maximum power level, in order to

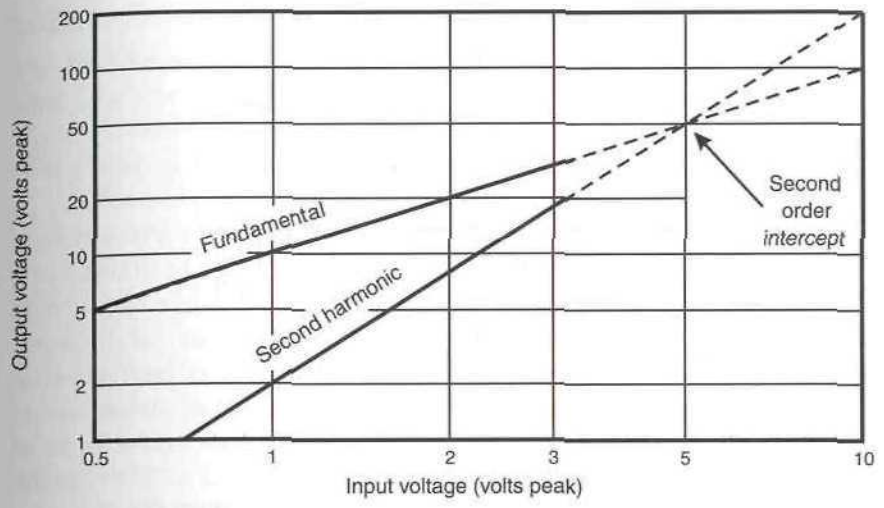


Figure 2.3 Illustration of the second-order intercept point of a nonlinear amplifier.

achieve an improved distortion performance. It would be impossible to predict the level of de-rating required from a percentage distortion measurement unless it was either tabulated or presented graphically.

Finally, note that a second-order characteristic produces harmonic distortion, as outlined above, but does not produce in-band intermodulation distortion (see below). This is an important distinction, in general, between even-order and odd-order nonlinearities: even-order nonlinearities do not generate in-band intermodulation distortion.

2.2.2 Third-Order Characteristic

A very different set of problems occur if an amplifier has a third-order term in its transfer characteristic.

$$V_{out}(t) = K_1 V_{in}(t) + K_3 V_{in}^3(t) \tag{2.3}$$

This characteristic is shown for $K_1 = 10$ and $K_3 = -3$ in Figure 2.4. Note that the output waveshape is now symmetrical above and below the horizontal axis and that a term at three times the original input signal frequency has appeared in the spectrum. This signal is the third harmonic and gives rise to its description as *third harmonic distortion*. Note further that no DC component exists for third-order distortion unlike that present with second-order distortion.

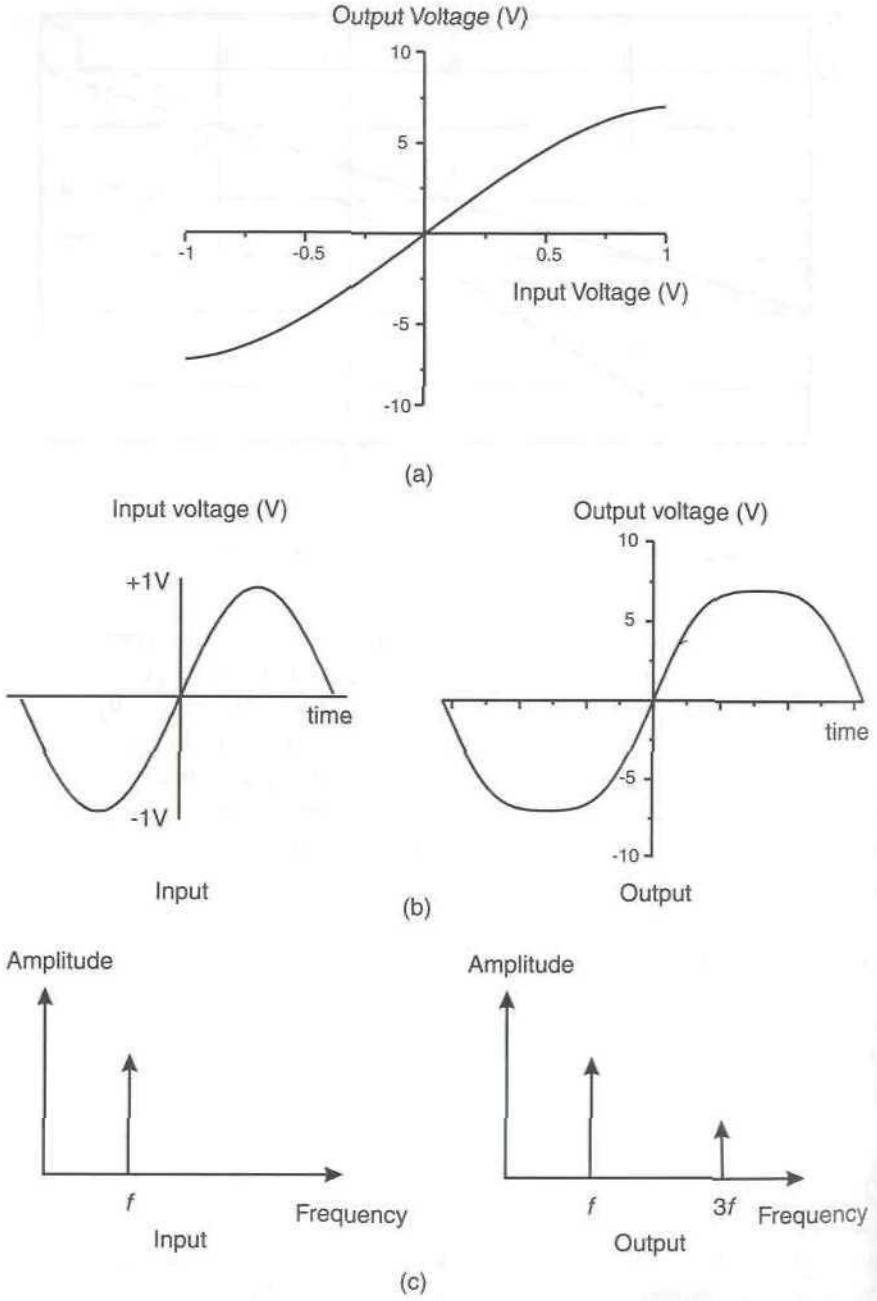


Figure 2.4 Transfer characteristic (a) and effect on a sinusoid in the time domain (b) and frequency domain (c) of an amplifier with transfer characteristic: $V_{out}(t) = 10V_{in}(t) - 3V_{in}^3(t)$.

2.2.2.1 1dB Compression Point

The *1 dB compression point* of an amplifier refers to the output power level at which the amplifier's transfer characteristic deviates from that of an ideal, linear, characteristic by 1 dB. The 1 dB compression point of the amplifier with a third-order nonlinearity is illustrated in Figure 2.5.

A further problem arises when considering two amplitude modulated carriers as the input signals, instead of the simple case of two unmodulated tones examined previously. The amount of compression experienced by a given signal will depend on the instantaneous level of the other signal being amplified. It is thus possible for the amplitude modulation appearing on one carrier to transfer to the other carrier and vice-versa. This problem is known as *cross-modulation* and can be a major problem in AM receivers when they are faced with very strong signals, as well as in transmitters operating close to saturation. This is described in more detail in Section 2.12.

Although gain compression is obviously a problem for amplifiers with a third-order nonlinearity, of greater concern are the intermodulation products appearing at $2f_2 - f_1$ and $2f_1 - f_2$. These distortion products appear 'in-band' and hence will distort the desired waveshape of the original input signal. Furthermore, since these products appear within the band of interest, it is usually impossible to filter them out, unlike the harmonic

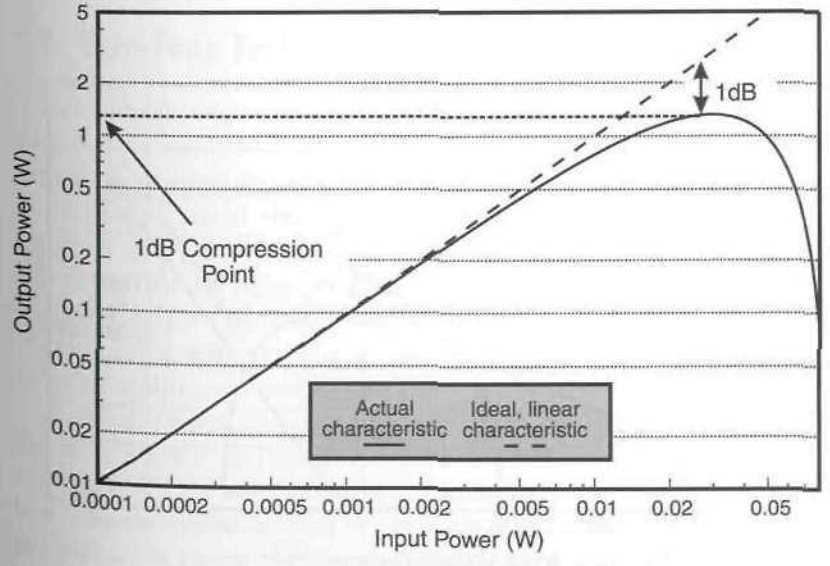


Figure 2.5 1 dB compression point of an amplifier with a characteristic given by: $V_{out}(t) = 10V_{in}(t) - 3V_{in}^3(t)$.

products at $3f_1$ and $3f_2$. For this reason advanced amplifier linearisation techniques, such as those discussed in this book, are required in order to secure their elimination.

2.2.2.2 Third-Order Intercept Point

A *third-order intercept point* may be defined in a similar manner to that of the second-order intercept point examined above; however, the form of the fundamental characteristic requires further explanation. The fundamental and third-order characteristics of an amplifier with a transfer function of the form shown in Figure 2.4 are shown in Figure 2.6. At low signal levels the magnitude of the fundamental component increases almost linearly with input signal level, however, it then begins to deviate from a linear characteristic and eventually decreases again.

This result (Figure 2.6) may be explained as follows. Suppose the input sinusoid is of the form:

$$V_{in}(t) = V_p \sin(\omega t) \tag{2.4}$$

This signal forms the input of an amplifier with a transfer characteristic:

$$V_{out}(t) = 10V_{in}(t) - 3V_{in}^3(t) \tag{2.5}$$

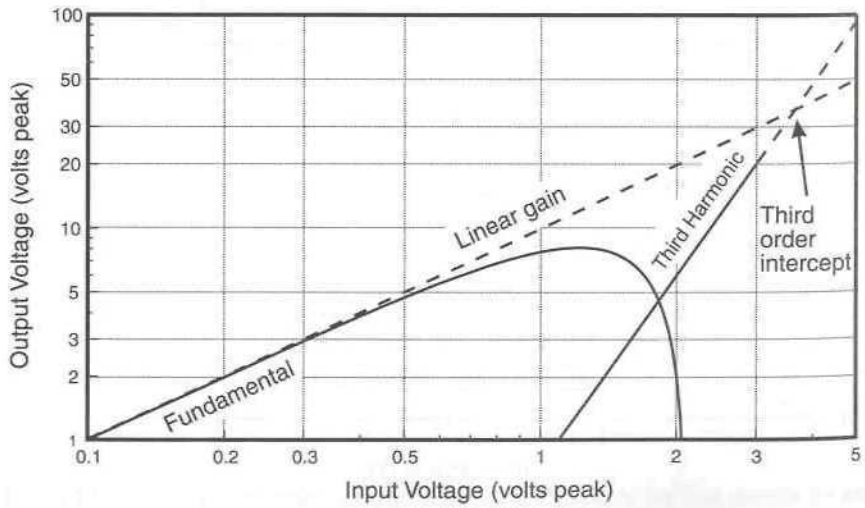


Figure 2.6 Illustration of the third-order intercept point of a nonlinear amplifier.

The resulting output signal is thus:

$$V_{out}(t) = 10V_p \sin(\omega t) - 3\{V_p \sin(\omega t)\}^3 \tag{2.6}$$

which reduces to:

$$V_{out}(t) = 10V_p \sin(\omega t) - \frac{9V_P^3}{4}\sin(\omega t) + \frac{3V_P^3}{4}\sin(3\omega t) \tag{2.7}$$

The first term represents the linear amplification of the fundamental (input signal) and the final term represents the third harmonic distortion. The middle term gives rise to the unusual shape of the fundamental characteristic in Figure 2.6 as it causes partial cancellation of the fundamental due to it appearing at the same frequency. The level of this cancelling signal is proportional to the cube of the input signal amplitude and hence it can quickly have a significant effect on the level of the fundamental in the output signal.

Thus, in an amplifier where third-order distortion predominates, the linear gain characteristic of the fundamental must be extrapolated in order to obtain the third-order intercept point. This is indicated by the dashed line in Figure 2.6.

2.3 Two-Tone Test

The two-tone test is an almost universally accepted method of assessing amplifier linearity and can illustrate both amplitude and phase distortions present in an amplifier. The effect of the two-tone test is to vary the envelope of the input signal throughout its complete range in order to test the amplifier over its whole transfer characteristic. It is thus arguably the most severe test of an amplifier's linearity performance, although it is being challenged as a 'standard' test by alternative techniques more suited to characterising nonlinearities in digital modulation transmitters (e.g., white noise or multicarrier test signals).

When viewed in the frequency domain, the spectrum of a two-tone test signal is shown in Figure 2.7. The individual signals are unmodulated carriers and if the test is carefully constructed, no other products should be in evidence within the band of interest (there will inevitably be some signals present at the generators' harmonics, but these should be small).

If the two-tone signal is viewed in the time domain, the envelope variation can clearly be seen (Figure 2.8). The signal levels should be

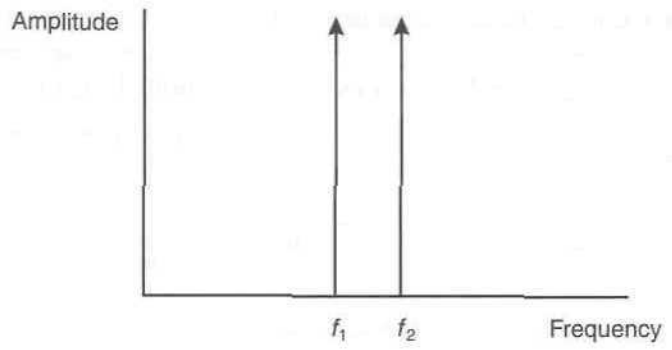


Figure 2.7 Two-tone test in the frequency domain.

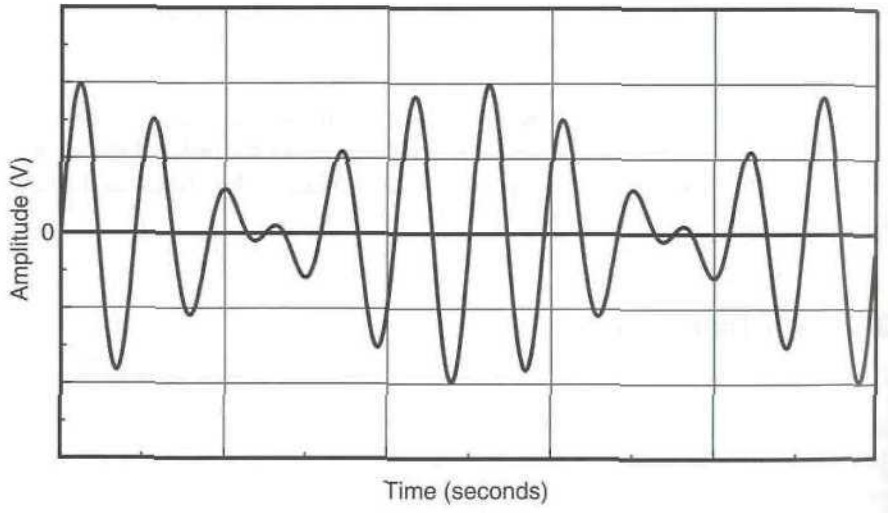


Figure 2.8 Time domain representation of a two-tone test.

arranged such that the peak envelope power (PEP) from the two-tone signal is equal to that of the full power rating at which the amplifier will be used (which may or may not be the same as its maximum CW power rating). For two unmodulated sinusoidal tones of equal level, the peak envelope power of the resulting two-tone signal is 6 dB greater than the CW power in either of the tones (the mean power being 3 dB higher than the CW power in either tone).

2.3.1 Two-Tone Test Applied to an Amplifier With a Second-Order (Square-Law) Nonlinearity

The two-tone test can be applied to an amplifier of the form discussed in Section 2.2.1. Each of the two tones will have a second harmonic and additional sum and difference frequencies will appear. These additional tones will occur at frequencies of: $f_2 - f_1$ and $f_2 + f_1$ and are known as *second-order intermodulation products* since they are created by the $V_{in}^2(t)$ term in the transfer characteristic.

In many RF applications, these distortion products are not significant as they will occur out of the bandwidth of interest, many RF applications requiring bandwidths of less than one octave. Thus the harmonics and intermodulation products falling outside of this bandwidth can be filtered out, the only penalty in this process being the insertion loss of the filter.

A square-law characteristic is often usefully applied in frequency mixers where the two 'tones' are formed by the local oscillator and incoming RF signal. A tuned circuit or monolithic filter is then used to select the required upconverted or downconverted signal.

2.3.2 Two-Tone Test Applied to an Amplifier With a Third-Order Nonlinearity

In the general case of distortion created by any order of nonlinearity when supplied with a two-tone input signal, new frequencies will be generated, as in the case of the second-order nonlinearity discussed above, and these will be of the form:

$$f_{im} = mf_1 \pm nf_2 \tag{2.8}$$

where m and n are positive integers (including zero) and $m+n$ is equal to the order of the distortion.

Thus for a third-order nonlinearity, the additional output frequencies will be:

$$\begin{aligned} f_{im1} &= 3f_1 \\ f_{im2} &= 3f_2 \\ f_{im3} &= 2f_1 + f_2 \\ f_{im4} &= f_1 + 2f_2 \\ f_{im5} &= 2f_1 - f_2 \\ f_{im6} &= f_1 - 2f_2 \end{aligned} \tag{2.9}$$

The original tones will, of course, also appear amplified at the output.

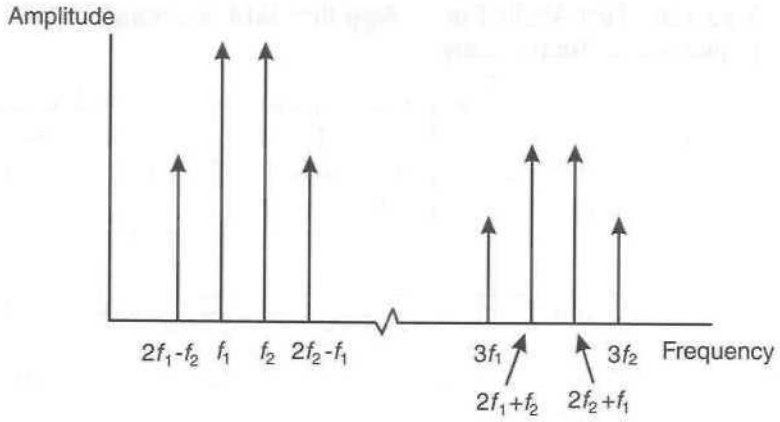


Figure 2.9 Spectrum of an amplifier with linear gain and a third-order nonlinearity.

Figure 2.9 shows these intermodulation products in the frequency domain, along with the amplified tones. The amplified tones will again not appear as large as would be expected from the linear gain portion of the amplifier characteristic (that is, the $10V_{in}(t)$ term in the example given in Figure 2.4) due to the partial cancellation of each tone as described in Section 2.2.2. This effect is termed *compression* and leads to a commonly used specification of maximum output level for an amplifier, that of the 1 dB compression point.

2.3.3 Higher-Order Nonlinearities

When applying a two-tone test to an amplifier exhibiting a number of orders of nonlinearity, a large number of harmonics and intermodulation products are generated and these are all given by the general expression of (2.18). Thus, if the amplifier contains nonlinearities up to and including a seventh-order then m and n will run between 0 and 7 in that equation (provided, of course, that $m+n \leq 7$). If the band of interest is filtered such that out-of-band products are removed, then the remaining intermodulation products will show the classic shape shown in Figure 2.10, for a well-behaved nonlinearity.

The most commonly used measure of intermodulation distortion (IMD) is the ratio of the largest intermodulation product to the amplitude of one of the two tones (assuming that they are equal in level). For a class-A amplifier, this will generally be in the range −30 dB to −35 dB. For most semiconductor amplifiers, the largest IMD product (in-band) will be the third-order product, although this is not always the case.

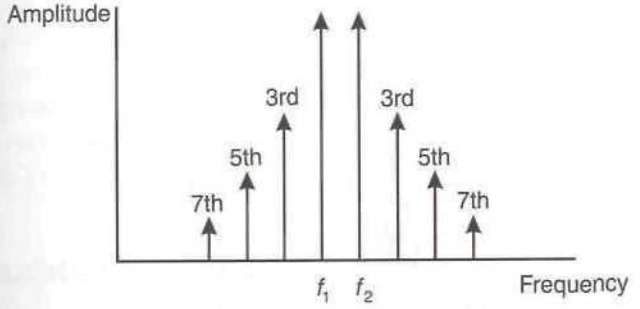


Figure 2.10 In-band intermodulation spectrum of an amplifier with up to a seventh-order nonlinearity.

2.3.4 Combined Effect of Harmonic and Intermodulation Distortion

If an amplifier with all orders of distortion up to a seventh-order is supplied with a two-tone test and the out-of-band products are not assumed to be removed by filtering, the complete spectrum will be that shown in Figure 2.11. The harmonic zones refer to the lowest order of nonlinearity which will produce a signal in that zone. Thus, for example, the second harmonic zone will contain products generated by a second-order nonlinearity (and above).

2.3.5 Effect of IMD on Carrier-to-Noise Ratio

The effect of IMD on a system also corrupted by noise will be to degrade the carrier-to-noise ratio (CNR) from that arising due to noise alone. The IMD products may therefore be considered to add to the received noise level, resulting in a new CNR level.

If the IMD products can be considered noise-like in their properties (this is not true of CW carriers, but is true of most digital modulation formats and of multicarrier signals consisting of a large number of modulated carriers), then it is possible to provide a simple mechanism for analysing the degradation in CNR resulting from the addition of IMD.

The resultant CNR ratio (in dB) for a system corrupted by both IMD and noise is given by:

$$CNR_{N\&IMD} = CNR_I - 10\log\left(1 + 10^{C_{ID}/10}\right) \qquad (2.10)$$

where CNR_I is the intrinsic carrier-to-noise ratio (in dB), C_{ID} is the difference between the carrier-to-interference ratio and the intrinsic CNR

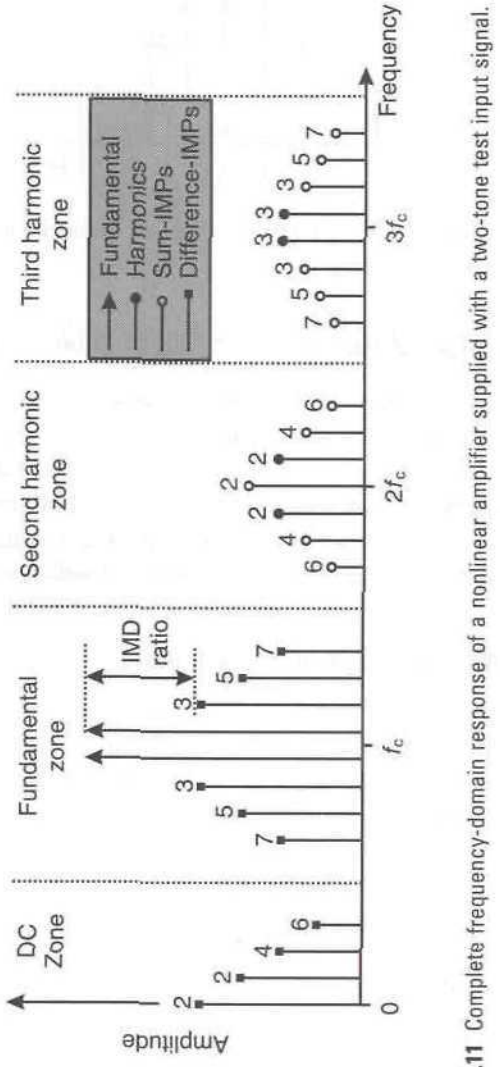


Figure 2.11 Complete frequency-domain response of a nonlinear amplifier supplied with a two-tone test input signal.

(i.e., without any IMD); it is usually negative, implying that the system noise level is higher than the IMD 'noise' level. Thus, for example, if a multicarrier FM signal requires a CNR of 18 dB per carrier (10 dB for the FM capture-effect to operate satisfactorily and 8 dB to allow for a satellite channel fading margin), then a carrier-to-interference ratio of $\geq$28 dB is required in order for the IMD to have a negligible effect (approximately 0.4 dB degradation).

2.4 Calculation of Intermodulation Distortion Ratio

2.4.1 Two-Tone Intermodulation

For an amplifier with a transfer characteristic given by:

$$V_{out}(t) = K_1 v_{in}(t) + K_2[v_{in}(t)]^2 + K_3[v_{in}(t)]^3 \tag{2.11}$$

given an input signal of the form:

$$v_{in}(t) = A_1 \cos(\omega_1 t) + A_2 \cos(\omega_2 t) \tag{2.12}$$

will yield an output signal containing both harmonic and intermodulation distortion (IMD) terms, as outlined in Section 2.3.5.

The intermodulation distortion ratio is defined as the ratio of the amplitude of the highest intermodulation product to the amplitude of one of the tones in the two-tone test. In the ideal case assumed above, with equal tone levels for the two-tone test (i.e., $A_1 = A_2$) and a simple polynomial model for the amplifier characteristic, the highest intermodulation products will generally be the third-order products and both will be equal in level.

In this case, the intermodulation distortion power will vary as the cube of the input signal power, giving:

$$P_{IMD} = (K_{IMD} P_1)^3 \tag{2.13}$$

where P_{IMD} is the power in a single third-order IMD product, K_{IMD} is a constant and $P_1 = A_1^2/2$ is the power in one of the input signal components. Thus for each change in input power of 1 dB, the IMD output power will change by 3 dB.

The intermodulation distortion ratio at the output of the amplifier is given by:

$$P_{IMR} = \frac{P_{IMD}}{P_{O,A1}} \tag{2.14}$$

where $P_{O,A1}$ is the output power in one of the wanted two output tones and the two input tone amplitudes are the same.

Combining the above equations, and recalling that the power level of a wanted tone at the output of the amplifier is linearly proportional to its input power, gives:

$$P_{IMR} = K_C^2 P_1^2 \tag{2.15}$$

where K_C is a constant.

The input third-order intercept point, P_{3rd}; illustrated in Figure 2.6, is the input power at which the IMD power (in the output spectrum) is equal to the output power contributed by the linear term $(K_1 A_1)^2/2$. It is therefore possible to relate this quantity to the IMD ratio and the input power, by realising that at the intercept point, the IMD ratio is unity (by definition), and hence:

$$1 = K_C^2 P_1^2 \tag{2.16}$$

At this point, the input power from a single tone is then the input third-order intercept point, P_{3rd}, that is, $P_1 = P_{3rd}$, and (2.16) then becomes:

$$K_C = \frac{1}{P_{3rd}} \tag{2.17}$$

Therefore (2.15) becomes:

$$P_{IMD} = \left(\frac{P_1}{P_{3rd}}\right)^2 \tag{2.18}$$

If the various values are expressed in dB, this expression may be simplified further to become:

$$P_{IMD,dB} = 2\left(P_{1,dBm} - P_{3rd,dBm}\right) \tag{2.19}$$

Thus, for example, an amplifier with an input intercept point of +30 dBm and receiving a two-tone input level of 0 dBm per tone (6 dBm PEP), would produce third-order intermodulation products:

$$P_{IMD,dB} = 2(0-30) = -60 \text{ dBc}$$

that is, −60 dB with respect to the level of each tone. The same would be true of an amplifier with an output intercept point of +30 dBm, when generating an output power level of 0 dBm per tone.

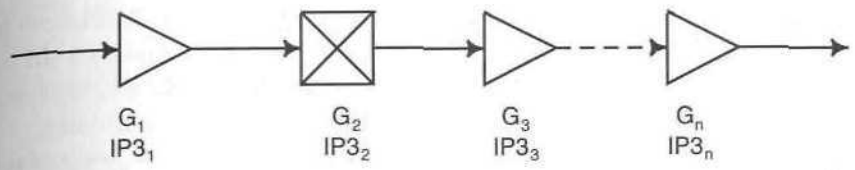


Figure 2.12 Cascaded intercept point calculation.

2.4.2 Cascaded Third-Order Intercept Point

It is possible to calculate the effective third-order intercept point of a cascade of elements (such as amplifiers, mixers) from the intercept points of the individual elements. This is useful in determining an approximate value for the intercept point of a complete receiver front-end or of a cascade of amplifier stages forming a power amplifier. The general situation is illustrated in Figure 2.12.

The third-order intercept point of the complete amplifier chain is then given by:

$$\text{IP3}_{tot} = \frac{1}{\dfrac{1}{\text{IP3}_1} + \dfrac{G_1}{\text{IP3}_2} + \dfrac{G_1 G_2}{\text{IP3}_3} + \cdots + \dfrac{G_1 G_2 \cdots G_{n-1}}{\text{IP3}_n}} \tag{2.20}$$

where each of the IP3 and gain terms is expressed in linear units (not dB), that is,

$$\begin{aligned} G_n &= 10^{G_{n,dB}/10} \\ \text{IP3}_n &= 10^{IP3_{n,dB}/10} \end{aligned} \tag{2.21}$$

It is important not to forget about the intercept point units used initially, as these will usually be specified in dBm, hence producing an intercept point in milliwatts in linear units. Note also that the above intercept points are at the input to each stage and that the overall result is therefore an input intercept point.

2.4.3 Effect of a Driver Stage on Overall Amplifier IMD

It is tempting to assume that the output IMD of an amplifier consisting of a cascade of two or more stages is approximately equal to the IMD from the

final stage, assuming that each of the previous stages has some headroom in driving the following stage. Unfortunately, the IMD produced by a driver stage can have a significant effect upon the overall IMD of the amplifier and hence the amount of 'headroom' allowed must be carefully considered.

Examining (2.20), it is clear that when the output intercept point of the driver stage is equal to the input intercept of the final stage (that is, if the stages were individually tested with low-distortion input signals, each would produce the same relative level of IMD), then the overall output IMD level will degrade by 3 dB. For example, if each of the stages individually (one driver and one output stage) was running at an IMD level of −30 dBc, then the cascaded amplifier would produce an output IMD level of −27 dBc. This can be expressed mathematically as:

$$\text{Degradation in IM3 (dB)} = 10\log\left[1 + 10^{\frac{\text{IM3(Driver)}-\text{IM3(Final)}}{20}}\right] \qquad (2.22)$$

where 'IM3(Driver)' and 'IM3(Final)' refer to the relative third-order intermodulation levels of the driver and final stages in the (two-stage) amplifier and both are expressed in −dBc. This equation is illustrated in Figure 2.13 for a representative range of relative IM3 levels.

As an example, consider a driver amplifier with an output intercept point which is 6 dB greater than the input intercept point of the final stage. It will therefore produce third-order IMD, when tested independently, at

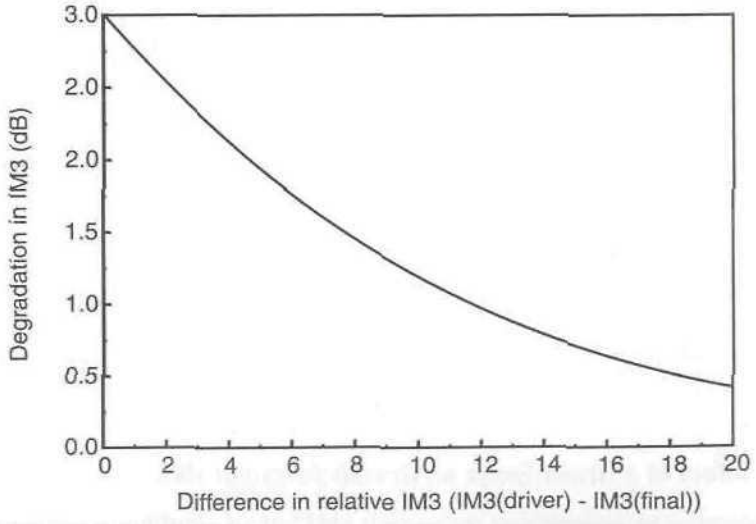


Figure 2.13 Degradation in IMD for a two-stage cascaded amplifier.

12 dBc lower than that of the final stage (i.e., IM3(Driver) −IM3(Final) = −12 in (2.22)). This results in an IM3 degradation of 1 dB for the output IMD of the final stage, when the two stages are cascaded.

2.5 Nonlinearity Measures for Multitone and Modulated Signals

There are a number of standard measurements for determining the degree of unwanted signal energy added by a nonlinear device. The two-tone test has been covered extensively above and provides a good indication of the degree of nonlinearity present over the whole of the amplifier characteristic, under simple signal conditions. Most modern systems are not, however, simple in this respect and hence a number of other measures are required; these will be outlined below.

2.5.1 Adjacent Channel Power Ratio

Adjacent channel power ratio (ACPR) is a measure of the degree of signal spreading into adjacent channels, caused by nonlinearities in the power amplifier. It is defined as the power contained in a defined bandwidth (B_1) at a defined offset (f_o) from the channel center frequency (f_c), divided by the power in a defined bandwidth (B_2) placed around the channel center frequency. The two bandwidths B_1 and B_2 need not be the same (and indeed are not for many current standards). The concept is illustrated in Figure 2.14.

2.5.2 Noise Power Ratio

Noise power ratio (NPR) is a measure of the unwanted in-channel distortion power caused by the nonlinearity of the power amplifier. This can be

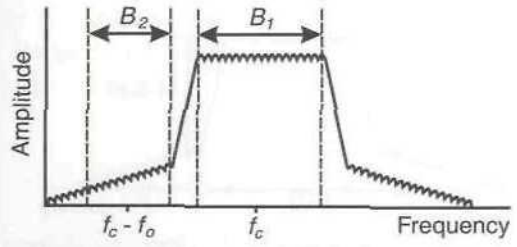


Figure 2.14 Adjacent channel power ratio.

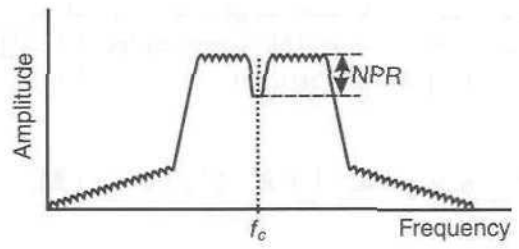


Figure 2.15 Noise power ratio.

measured by extracting a portion of the input signal, using a notch filter, and examining the level of distortion 'filling in' the space within the resulting gap. Further details of this test are provided in Figure 2.15.

NPR is defined as the ratio between the noise power spectral density of a white noise signal passing through the amplifier, measured at the center of the notch, to the noise power spectral density without the notch filter, where the amplifier is driven at the same power level in each case. The concept is illustrated in Figure 2.15.

2.5.3 Multitone Intermodulation Ratio

Multitone intermodulation ratio (M-IMR) is a measure of the effect of nonlinearity on a multicarrier signal. This could be a multicarrier modulation format (e.g., OFDM) or a multicarrier signal from, for example, a base-station transmitter. It is defined as the ratio between the wanted tone power (of one of the multiple tones) and the highest intermodulation tone power just outside of the wanted band. The concept is illustrated in Figure 2.16.

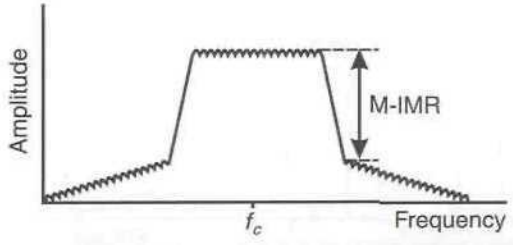


Figure 2.16 Multitone intermodulation ratio.

2.5.4 Relationship Between Two-Tone IMD and Complex Signal IMD

Based on simple third-order modelling of an amplifier characteristic, it is possible to derive empirical relationships between the above complex-signal IMD measures and their two-tone counterparts [1,2]. It is therefore possible to approximate the multitone or complex signal behavior of an amplifier based on only a simple two-tone measurement. This process has many potential areas of inaccuracy, but it does serve to provide an approximate idea of the desired response prior to more detailed measurements.

The process is based upon Volterra–Weiner theories [3,4], which state that any third-order system may be completely characterised by a three-tone test. Increasing the number of tones, therefore, will not yield any further information about the system and any additional effects seen in practice, when performing multitone testing, are due to the presence of higher-order products.

By examination of the statistics of a multitone signal with uncorrelated phases and greater than about 10 carriers, the central limit theorem indicates that the resulting waveform tends toward a narrow-band noise excitation and the response of the system therefore approximates a noise power ratio test. Examination of a multicarrier signal utilising an analytical approach can therefore be used to derive approximate measures of the three complex-signal IMD measures discussed above (ACPR, NPR, and M-IMR).

Based on the definitions below, the following equations may be derived:

$\text{IMR}_{2\text{-tone}}$	Two-tone intermodulation ratio (dBc)
IP_3	Device third-order intercept point (dBm)
P_{ave}	Mean output power from the device (dBm)
n	Number of tones
r	Adjacent channel product number ($r = 1$ for the first—closest—IMD product)
b	Number of tones on one side of the 'gap', where one or more tones is removed for NPR testing. For removal of a single tone at the center of the band, $b = f[(n-1)/2]$, where $f(\bullet)$ extracts the integer part of the expression in brackets (by rounding down).

2.5.4.1 ACPR

$$\text{ACPR}_{\text{dBc}} = \text{IMR}_{2\text{-tone}} - 6 + 10\log\left(\frac{n^3}{4A+B}\right) \tag{2.23}$$

where

$$A = \frac{2n^3 - 3n^2 - 2n}{24} + \frac{\text{mod}(n/2)}{8} \tag{2.24}$$

and

$$B = \frac{n^2 - \text{mod}(n/2)}{4} \tag{2.25}$$

Note that mod(x/y) is defined as the remainder when x is divided by y.

2.5.4.2 NPR

$$\text{NPR}_{\text{dBc}} = \text{IMR}_{2\text{-tone}} - 6 + 10\log\left(\frac{n^2}{4C + D}\right) \tag{2.26}$$

where

$$C = \left(\frac{n-b-2}{2}\right)^2 - \frac{\text{mod}[(n+b)/2]}{4} + \left(\frac{b-1}{2}\right)^2 - \frac{\text{mod}[(b+1)/2]}{4} + b(n-b-2) \tag{2.27}$$

and

$$D = \left(\frac{n-b-2}{2}\right) - \frac{\text{mod}[(n+b)/2]}{2} + \left(\frac{b-1}{2}\right) + \frac{\text{mod}[(b+1)/2]}{2} \tag{2.28}$$

2.5.4.3 M-IMR

$$\text{MIMR}_{dBc} = \text{IMR}_{2\text{-}tone} - 6 + 10\log\left(\frac{n^2}{4E + F}\right) \tag{2.29}$$

where

$$E = \left(\frac{n-r}{2}\right)^2 - \frac{\text{mod}[(n+r)/2]}{4} \tag{2.30}$$

and

$$F = \left(\frac{n-r}{2}\right) + \frac{\text{mod}[(n+r)/2]}{2} \tag{2.31}$$

For all of equations (2.23), (2.26), and (2.29), the relationship between the two-tone intermodulation ratio, $\text{IMR}_{2\text{-tone}}$, and the total *mean* output power, P_{ave}, is given by:

$$P_{ave} = \frac{-\text{IMR}_{2\text{-tone}}}{2} + \text{IP}_3 + 3\text{dB} \tag{2.32}$$

The additional factor of 3 dB (when compared with (2.19)) arises from the fact that (2.32) uses the average power of the output signal and not the power per tone considered previously.

Figure 2.17 plots these characteristics, taking the same equivalent two-tone IMD level in each case (0 dB). This allows the relative level of each measure to be judged and indicates that ACPR is the closest to a two-tone equivalent, with NPR being the furthest from the two-tone case.

2.5.5 Examples

Consider an unlinearised multicarrier power amplifier with a third-order intercept point of + 60 dBm, which is required to amplify 16 carriers, each of

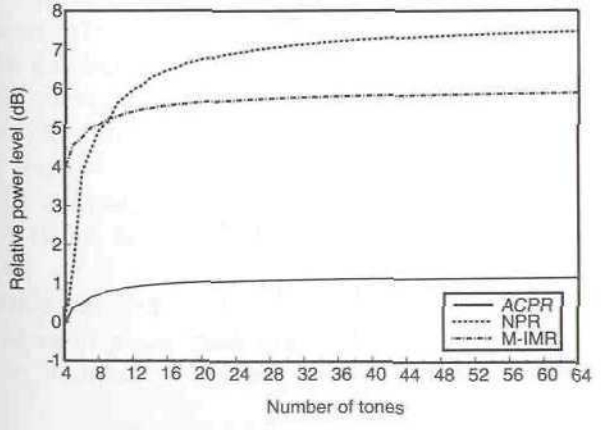


Figure 2.17 Normalised relationship between the various nonlinearity measures for multitone signals.

1W (+30 dBm). Using this as an example, it is possible to predict (and compare) the three multicarrier IMD specifications outlined above (ACPR, NPR, and M-IMR).

2.5.5.1 ACPR

Calculating the two intermediate values gives:

$$\begin{aligned} A_{16} &= 308 \\ B_{16} &= 64 \end{aligned} \tag{2.33}$$

and hence:

$$\mathrm{ACPR_{dBc}} = \mathrm{IMR_{2\text{-}tone}} - 1.023 \tag{2.34}$$

substituting for $\mathrm{IMR_{2\text{-}tone}}$ from (2.32):

$$\begin{aligned} \mathrm{IMR_{2\text{-}tone}} &= 2(IP_3 - P_{ave}) + 6 \\ &= 2[IP_3 - (10\log(n) + P_{tone})] + 6 \\ &= 2[60 - (10\log(16) + 30)] + 6 \\ &= 41.9 \text{ dBc} \end{aligned} \tag{2.35}$$

where n is the number of carriers (16 in this example) and P_{tone} is the per-carrier power of the signal (+30 dBm in this example).

Therefore:

$$\begin{aligned} ACPR &= 41.9 - 1.023 \\ &= 40.9 \text{ dBc} \end{aligned} \tag{2.36}$$

2.5.5.2 NPR

Calculating the two intermediate values gives:

$$\begin{aligned} C_{16} &= 69 \\ D_{16} &= 7 \end{aligned} \tag{2.37}$$

and hence:

$$\mathrm{NPR_{dBc}} = \mathrm{IMR_{2\text{-}tone}} - 6.456 \tag{2.38}$$

substituting for $\mathrm{IMR_{2\text{-}tone}}$ from (2.35) above gives:

$$\begin{aligned} \mathrm{NPR_{dBc}} &= 41.9 - 6.456 \\ &= 35.5 \text{ dBc} \end{aligned} \tag{2.39}$$

2.5.5.3 M-IMR

Calculating the two intermediate values gives:

$$\begin{aligned} E_{16} &= 56 \\ F_{16} &= 8 \end{aligned} \tag{2.40}$$

and hence:

$$\mathrm{MIMR_{dBc}} = \mathrm{IMR_{2\text{-}tone}} - 5.593 \tag{2.41}$$

substituting for $\mathrm{IMR_{2\text{-}tone}}$ from (2.35) above gives:

$$\begin{aligned} \mathrm{MIMR_{dBc}} &= 41.9 - 5.593 \\ &= 36.3 \text{ dBc} \end{aligned} \tag{2.42}$$

This analysis in no way represents a 'worst-case' for n carriers being amplified by a single amplifier—the phases have been assumed to be random and uniformly distributed and the nonlinearity has been assumed to be purely third-order. It can therefore be said to represent an 'average' level of intermodulation if a device were to be tested with a number of carriers a large number of times. It is therefore not a conservative estimate of IMD level (or power rating), although it does serve to illustrate, in some measure, the effect of multiple tones on a nonlinear amplifier. A more conservative 'rule-of-thumb' often used in the absence of more rigorous information or analysis, is to rate a multicarrier amplifier at a mean power level 9 dB or 10 dB below its peak power rating, for a large number of carriers (>16).

Modulation applied to the carriers, specifically that involving a significant degree of phase modulation, may serve to create the 'random' phase behavior, assumed in this analysis, as a continuous feature.

2.6 Crest Factor and Crest Factor Reduction Techniques

2.6.1 Crest Factor Definition

There are a range of terms and often contradictory definitions for '*crest factor*', '*peak factor*', '*peak to average ratio*' and '*peak to mean ratio*'. Essentially they are all methods of defining the statistics of a modulated signal in a manner which an amplifier designer can understand and interpret. The two definitions which have become standard are:

Crest factor: the ratio of the peak to r.m.s. *amplitude* of a signal [5]

$$\mathrm{CF} = \frac{|\hat{S}(t)|}{\sqrt{|S(t)|^2}} \tag{2.43}$$

Peak-to-mean ratio: the ratio of the peak *power* to r.m.s. *power* of a signal (for example, 6). Some typical values for common modulation formats are provided in Table 2.1.

$$\mathrm{PMR} = \frac{\text{Peak Power}}{\text{Average Power}} \tag{2.44}$$

Note that:

$$\mathrm{PMR} = \mathrm{CF}^2 \tag{2.45}$$

2.6.2 Crest Factor Reduction for Multicarrier Signals

There are a range of techniques for reducing the crest factor (and hence by implication the peak-to-average ratio) of a multicarrier signal and this can have a significant impact upon the power rating required of a linear or linearised power amplifier. This is a substantial (and highly mathematical) subject and a detailed treatment is beyond the scope of this book—however, a brief summary of the primary techniques is appropriate. These techniques involve either careful phasing of RF carriers to minimise the overall peak-to-mean ratio, coding of the modulation information in order to minimise signal transitions which increase the peak-to-mean ratio (e.g., those traversing the origin in the complex plane) or appropriate distribution (mapping) of the required data across the carriers (e.g., in an OFDM system). Clearly not all techniques are appropriate to all systems and much of the work in this area has been directed toward OFDM systems.

The primary methods include *Shapiro–Rudin sequences* [7,8], *Golay codes* [9,10], *maximal-length sequences* [11], *Barker codes* [12,13], *Newman phases* [14,5], *Schroeder phases* [15], *block coding* [16], *selected mapping* [6] and *partial transmit sequences* [17]. More recently, attention has turned to multicarrier CDMA and the reduction of peak-to-mean power in this type of system [18].

Table 2.1
Peak to mean ratios for some common modulation formats

Modulation scheme	Parameter	Crest factor (dB)
Wideband CDMA (see Note 1)		
	16 occupied channels	10.5
	32 occupied channels	11.1
	64 occupied channels	12.2
	128 occupied channels	13.6
π/4-DQPSK (see Note 2)		
	$\alpha = 0.20$	4.86
	$\alpha = 0.25$	4.55
	$\alpha = 0.30$	4.23
	$\alpha = 0.35$	3.87
	$\alpha = 0.40$	3.38
	$\alpha = 0.50$	3.21
16-QAM (see Note 3)		
	$\alpha = 0.20$	6.03
	$\alpha = 0.25$	5.92
	$\alpha = 0.30$	5.66
	$\alpha = 0.35$	5.40
	$\alpha = 0.40$	5.18
	$\alpha = 0.50$	4.94
GSM EDGE (8-PSK) (see Note 4)		3.21

Notes:

1. Wideband CDMA parameters: chip-rate: 4.096Mchips/s, filter type: root-raised cosine, $\alpha = 0.22$. All channels occupied with statistically-independent pseudo-random data.
2. π/4-DQPSK parameters: symbol rate: 24.3ksymbols/s, filter type: root-raised cosine, DAMPS coding applied.
3. 16-QAM parameters: symbol rate: 25ksymbols/s, filter type: root-raised cosine, no coding applied.
4. GSM-EDGE: as per ETSI SMG2 technical document WPB 386/98.

2.7 Phase Distortion

Any deviation from perfect linearity of the amplitude transfer characteristic of an amplifier will lead to distortion of the output waveshape. The same is also true of the phase characteristic for that amplifier, although the reason why is perhaps less obvious. A nonlinear phase characteristic will result if an amplifier does not delay all frequency components within the input signal by the same amount when they reach the output. The effect of such a characteristic on a 1 kHz 'squarewave', comprised of only the first three frequency components of its Fourier analysis, is shown in Figure 2.18.

The fundamental (1 kHz) has been delayed by 100 μs, the third harmonic (3 kHz) by 200 μs and the fifth harmonic (5 kHz) by 300 μs; the resultant waveshape bears little relation to that of the original squarewave. Similar distortion is present in long cable runs, such as those of the public switched telephone network (PSTN), and hence the use of direct baseband data modulation on such systems is inadvisable (FSK or M-ary QAM modems are more usual).

The relationship between time delay, τ and phase shift, ϕ is:

$$\tau = \frac{\phi}{2\pi f} \tag{2.46}$$

where f is the fundamental frequency of the waveform and ϕ is in radians. The time delay will be identical for all frequency components making up a complex waveform, if the phase-shift increases in proportion to the frequency. Thus it is the time delay imposed by the amplifier, and not its phase-shift, which must remain constant over the bandwidth of interest, to eliminate distortion of the waveshape.

If these ideas are extended to an amplifier operating at radio frequency, then it is the modulating signal's waveshape which must be preserved. As an example, consider the three-frequency 'squarewave' used above. If it is used to amplitude modulate a high frequency carrier, then the spectrum shown in Figure 2.19 will result. Since the modulation will be recovered from the carrier further on in the communication system it must be preserved in the transmitter to a high degree of fidelity. Thus the delay through the RF amplifiers for each of the frequencies including the carrier must be constant. The actual magnitude of the delay is, in general, not important.

With modern network analysis techniques it is a relatively simple matter to obtain the delay characteristics of an amplifier over the frequency range of interest. The deviation from a flat characteristic indicates the degree of phase distortion which may be expected for the amplifier.

Alternatively, the delay may be measured using a high-frequency oscilloscope and by passing a sinusoidally modulated carrier through the device under test (DUT). The envelope delay may be measured as illustrated in Figure 2.20 and this should be constant as the modulating frequency is varied. The absolute value of envelope delay is given by:

$$T_e = \frac{\Delta\phi}{2\pi\Delta f} \quad \text{seconds} \tag{2.47}$$

where $\Delta\phi$ is the phase difference between the carrier and the modulation sideband (in radians) and Δf is the frequency difference between carrier and sideband.

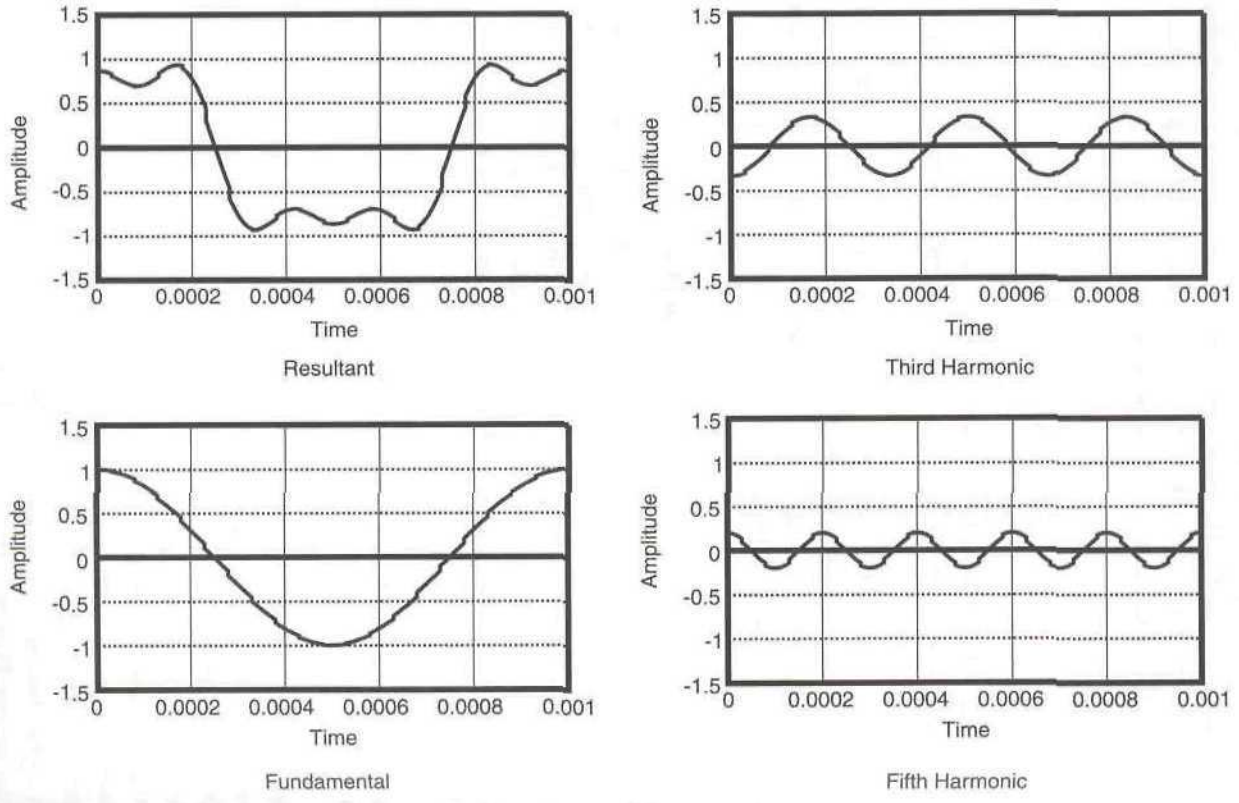
Note that envelope delay and group delay are *not* the same and are measured in a different manner. The terms are frequently used interchangeably and the error resulting from this is small in a narrow-band system, however care should be exercised to ensure that the correct term is used [19].

2.8 Practical Creation of a Multitone Test Signal

The creation of a multitone test signal suitable to fully characterise the in-band distortion characteristics of a highly linear RF amplifier is a non-trivial problem. The distortion present on the test signal must be appreciably below the anticipated distortion level from the amplifier and this may prove difficult if the amplifier is expected to perform to an IMD specification of 50 dBc or better.

The problem may be broken down into that of testing at RF and that of testing using baseband information (such as when a complete transmitter is being tested). RF testing will be covered first.

Most two-tone or multitone testing is performed with the aid of signal generators of some description, at least during the design phase. Modulated carriers from 'real' system sources are usually only applied when satisfactory performance of a prototype with either simulated or CW signals has been achieved. Most laboratory signal generators contain power-levelling diodes electrically close to their RF output port(s) and any signal energy leaking through the power combiner (combining the various signal sources) will cause intermodulation to occur in this diode. This will occur at the output of each signal generator, resulting in a distorted output signal from the power combiner (see Figure 2.21). The level of this distortion will vary with the brand of generator employed, the isolation of the power combiner and the required output level. However as a 'rule of thumb' a distortion level in the vicinity of 50–55 dBc will result from a two-tone test, employing commer-



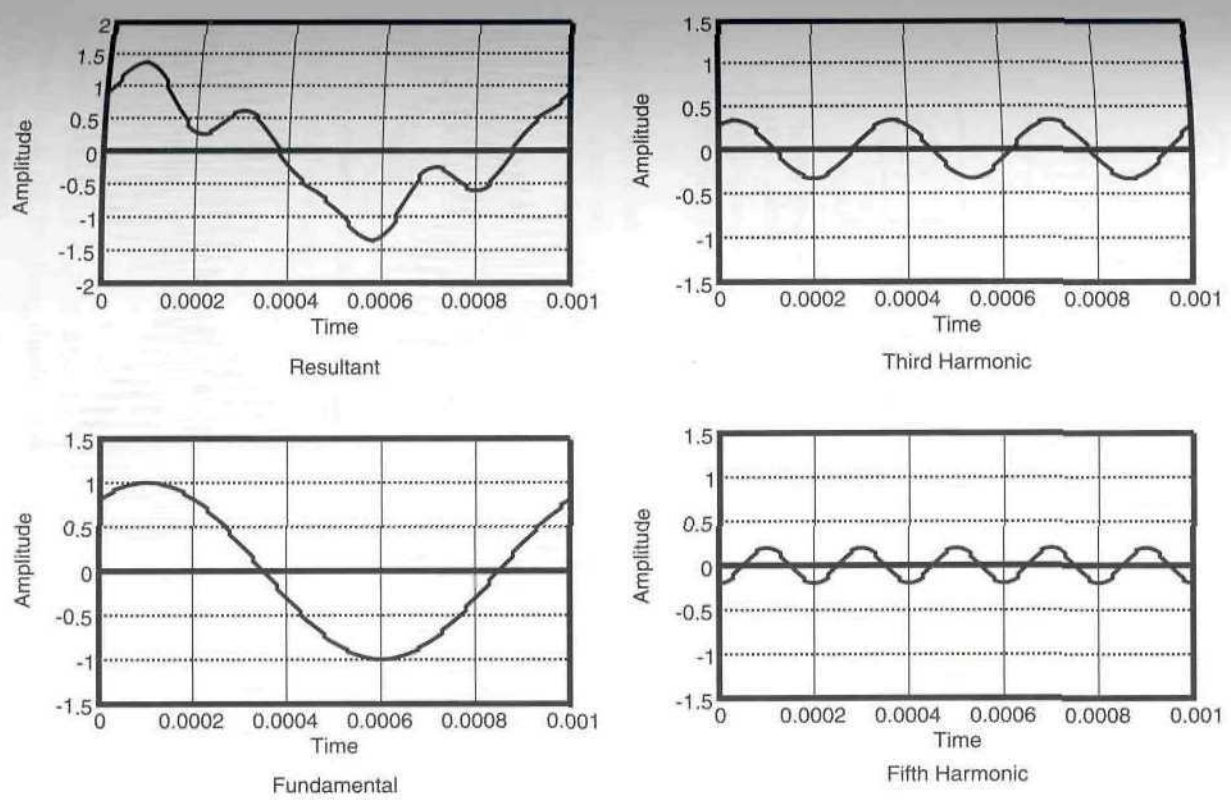


Figure 2.18 Phase distortion of a 1 kHz squarewave. (a) Original signal and (b) Distorted resultant.

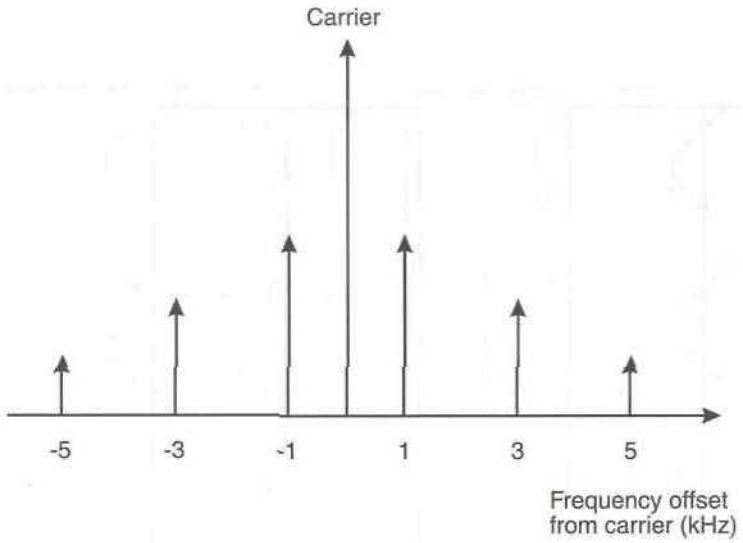


Figure 2.19 Spectrum of a carrier plus 1 kHz squarewave amplitude modulation.

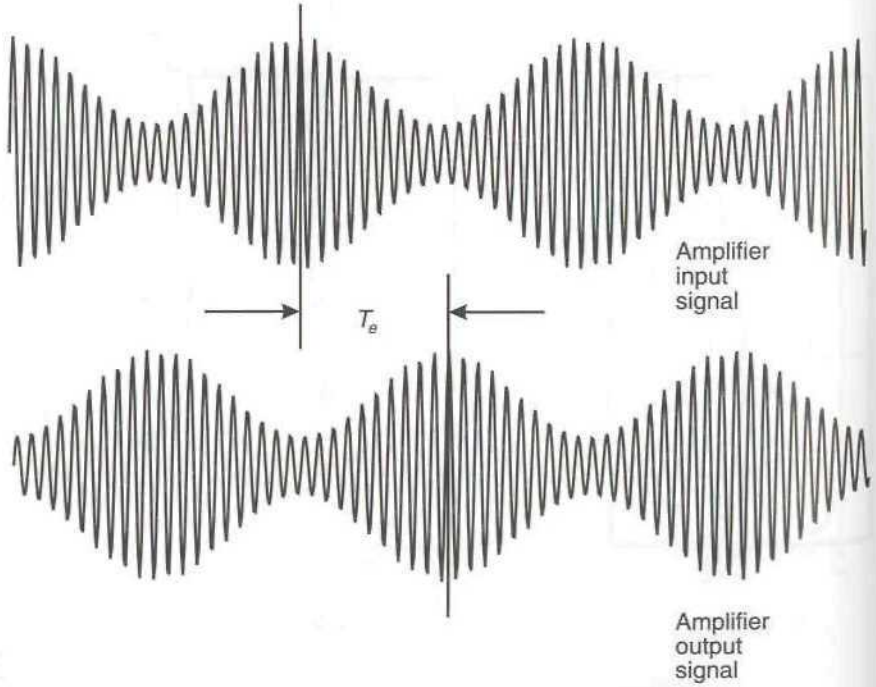


Figure 2.20 Measurement of envelope delay using an oscilloscope.

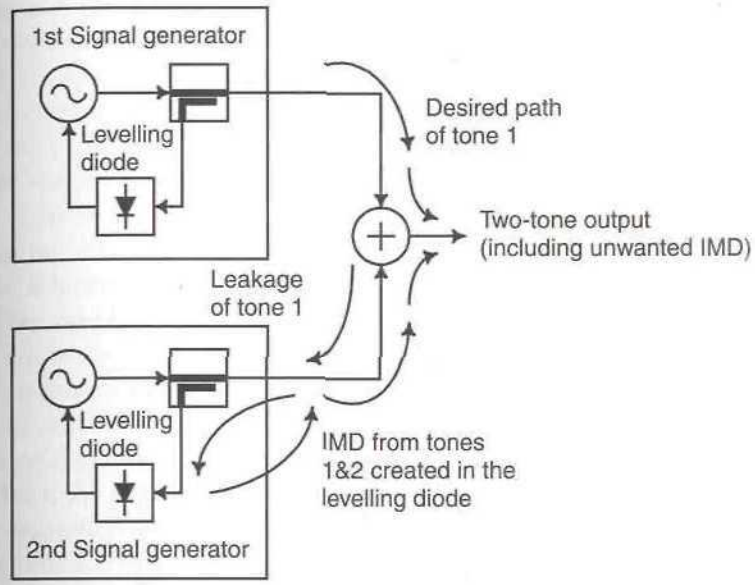


Figure 2.21 Unwanted IMD generation when using commercial signal generators to create a two-tone test (note that a similar process, not shown, occurs in the first signal generator).

cial bench signal generators at a power output of + 10 dBm per tone, using a transformer-based 3 dB hybrid as the combiner. It must be stressed that this figure is only a rough guide and a particular test arrangement may yield better or worse performance than this.

To enable the satisfactory testing of higher performance amplifiers, such as feedforward systems for base station applications, it is necessary to achieve an IMD specification for the test signal approaching 80 dBc. To achieve this level of performance requires a high degree of isolation to be introduced between the signal generators and the power combiner.

A typical test arrangement is illustrated in Figure 2.22. In this arrangement, each signal source is provided with either an attenuator or an isolator in its output path. An attenuation value of at least 10 dB is recommended for 60–70 dBc IMD performance and must be provided externally, as the levelling diode is often placed after the internal attenuators in the generator.

The use of an attenuator will obviously reduce the drive level of the test signal and this is usually undesirable. A better solution is to use one or more isolators in each generator output. In this manner, multitone test

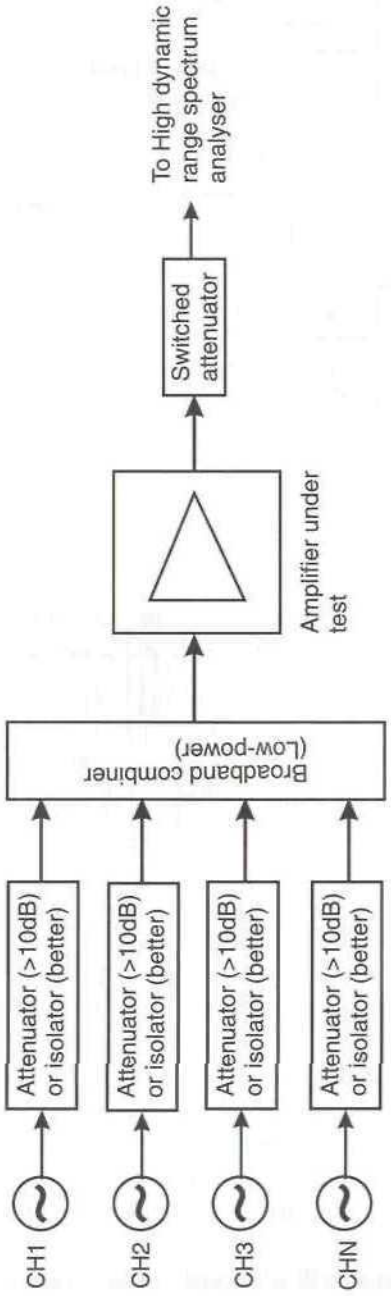


Figure 2.22 Practical realisation of a multitone test at RF.

signals with an IMD performance of 80 dBc or greater may be created, with relatively little loss of drive level (1–2 dB is typical).

High-quality baseband test signals are generally easier to create, but require careful thought if an in-phase and quadrature test signal is required (such as for Cartesian loop and CALLUM transmitters). It is possible to arrange the test tones to appear above, below or either side of the carrier, with the latter usually being the preferred option. In the case of a complete transmitter these test tones can also be used to judge the gain and phase balance achieved between the I and Q channels in the upconversion process, since they may be slightly offset (i.e., non-symmetrical about the carrier) and hence allow the image signals to be seen. The gain and phase balance of the input signals can then be adjusted until the desired image cancellation is achieved and the gain and phase 'error' required to do this, noted. Further details on diagnostic testing of this nature are provided in Chapter 4.

Table 2.2 illustrates the I and Q signals required to provide a two-tone test surrounding the carrier in an SSB-type system. Such a signal is ideal for testing both Cartesian loop and CALLUM transmitters and may also be used in adaptive baseband predistortion systems. The tones should be straightforward sinewaves, with an amplitude appropriate for the transmitter under test. Modern four-channel audio-frequency synthesisers are ideal for this task.

After I/Q upconversion, the two tones will have a frequency spacing of $f_1 + f_2$ and should be of equal level. If the carrier is at a frequency, f_C, the tones will be located at: $(f_C - f_1)$ and $(f_C + f_2)$; the image sidebands will occur at: $(f_C + f_1)$ and $(f_C - f_2)$. Thus, for example, to achieve a tone spacing of 20 kHz, set $f_1 = 9$ kHz and $f_2 = 11$ kHz. The terms δa and $\delta\phi$ indicate the gain and phase error respectively in the upconversion process (note that in the case of transmitters involving baseband feedback, for example, Cartesian loop, the gain and phase error is set in the error path, that

Table 2.2
I and Q baseband signals to create a two-tone test around a carrier in a quadrature-based transmitter

Channel	Frequency	Amplitude	Phase
I	f_1	a	0°
I	f_2	a	180°
Q	f_1	$a + \delta a$	$90° + \delta\phi$
Q	f_2	$a + \delta a$	$90° - \delta\phi$

is, by the downconverter). Both δa and $\delta\phi$ must be varied by equal amounts in the directions indicated by their respective signs. Thus, for example, to correct for a 3° phase error in the upconversion process, the phase of f_1 in the Q channel should be increased by 3° and the phase of f_2 in the Q channel should be reduced by 3°.

2.9 Two-Tone Test With Unequal Tone Powers

The above discussions have all assumed that the two tones used are identical in level, hence providing an infinite (theoretically) envelope variation, although the peak to mean ratio is only 3 dB. If the tone amplitudes are made unequal, then it is possible to test particular parts of the amplifier's transfer characteristic. In particular it is often useful to be able to test the top few dB of the characteristic (up to the 1 dB compression point), as this allows the amplifier's effect on quasi-linear modulation formats to be assessed (e.g., π/4-DQPSK). These formats have envelope variations which do not fall to zero, despite (in many cases) having a typical peak-to-mean ratio of around 3 dB.

Figure 2.23 shows the peak-to-minimum ripple obtained from a two-tone test for a range of values of amplitude difference between the tones.

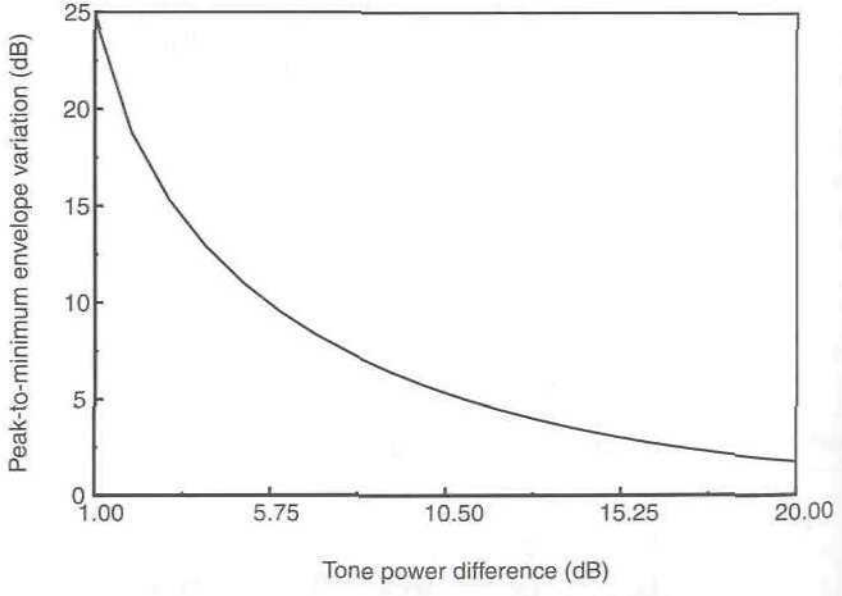


Figure 2.23 Peak-to-minimum envelope variation of a two-tone test with unequal tone powers.

Table 2.3
Peak-to-minimum envelope variation obtained from a two-tone test for a range of values of amplitude difference between the tones

Tone Amplitude Difference (dB)	0	1	2	3	4	5	6	7	8	9
Peak-to-minimum Envelope Variation (dB)	∞	24.8	18.8	15.3	12.9	11.1	9.6	8.3	7.3	6.4
Tone Amplitude Difference (dB)	10	11	12	13	14	15	16	17	18	19
Peak-to-minimum Envelope Variation (dB)	5.7	5.0	4.5	4.0	3.5	3.1	2.8	2.5	2.2	2.0

This data is also reproduced in Table 2.3. Using these it is possible to select an appropriate tone inequality in order to simulate the effect of amplifier nonlinearity for a given modulation format, without the need for advanced test equipment. Note that the results will only be approximately correct, but that they will generally give a more realistic idea of performance than would a conventional, equal-power two-tone test.

A comparison between the (measured) equal power two-tone performance of an amplifier and the equivalent TETRA π/4DQPSK performance is provided in Chapter 4. This shows clearly that a conventional two-tone test is very pessimistic for this application.

2.10 White Noise Testing of Amplifier Linearity

Although the two-tone test is widely used to measure the linearity of amplifier stages it does not accurately simulate the effects of speech signals or of linear data schemes (such as 16-QAM). It is also a poor indicator of linearity for TWTA systems involving a large number of carriers (> 10). A more realistic test for these systems is to use a band-limited white noise signal which has a flat power spectrum across a single channel within the operational bandwidth of the amplifier (or an appropriate band representative of a number of carriers in a multicarrier amplifier). At all frequencies outside of the channel (or band, in the case of a multicarrier system), the signal should be zero.

If this signal forms the input to an amplifier, then any signals in the output spectrum which appear outside of the channel are due to amplifier nonlinearity. In the case of a multicarrier system, the white noise is simulating a large number of carriers of random amplitude and phase and

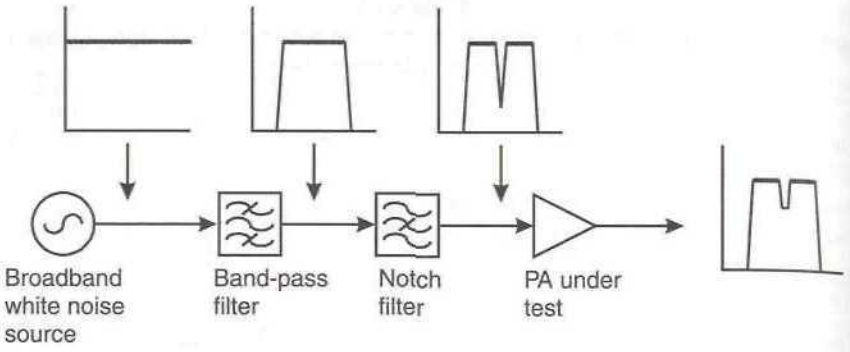


Figure 2.24 White noise testing of amplifier linearity.

a narrow notch filter may be used to leave a 'gap' in the center of the band of interest. This gap will then be 'filled' with any IMD resulting from nonlinearities in the power amplifier, and the ratio of the noise power in this notch to that in the bands either side is the noise power ratio (NPR) for the amplifier.

A block-diagram of this approach is shown in Figure 2.24. The width of the notch should be about 1% or less of the width of the bandpass-filtered noise source and this may most easily be produced at IF, with subsequent (high-linearity) upconversion to the required RF band of interest. The notch depth of the resulting test signal should be at least 10 dB greater than the maximum NPR to be measured [20].

In practice, it may be difficult to produce an input signal of the form described above, particularly for narrow-band systems, and so a carrier, amplitude modulated by a pseudo-random binary sequence (PRBS) of ±1 in level, may be used instead. The resulting signal has zeros at the edges of the channel and at the edges of the adjacent channels throughout the band, assuming that the bit-rate is chosen appropriately. Thus for a channel of bandwidth $2f_b$, the bit rate of the PRBS must be f_b and zeros will occur at $\pm f_b$, $\pm 3f_b$, $\pm 5f_b$, and sequentially. Any energy present in the output of the amplifier at these frequencies is then a measure of its nonlinearity [21].

2.11 Spurious Signals

Some signals appearing at the output of an amplifier bear no obvious relationship to the input signals being amplified. They may appear and disappear at random and may change frequency and level at will. Such

signals are referred to as *spurious products* and consist of parasitic and subharmonic oscillations together with unwanted external interference which is inadequately shielded by the amplifier casing.

Most spurious signal problems are the result of poor RF constructional techniques and inadequate case design or build quality. The problems are usually made more acute by the constant desire to make RF circuitry as small as possible, particularly in the hand-portable radio market. Amplifiers for use in products in this area must be extremely stable and robust.

2.12 Cross-Modulation

In a nonlinear system in which two input signals are being processed, one desired and one undesired, amplitude modulation present on the undesired signal can be imposed upon the desired signal and this can cause potentially severe interference to the wanted signal (see Figure 2.25). This is a particular problem for receiver systems, as amplitude modulation present on a strong unwanted signal can transfer onto a weaker, wanted signal and interfere with, or even mask, the wanted amplitude modulation present on that carrier. Figure 2.25 illustrates this problem for a two-tone modulated 'unwanted' signal and a CW 'wanted' (weaker) signal. Note that third-, fifth-, and seventh-order distortion only is assumed in this example.

This affect can also, however, have a bearing upon multicarrier power-amplifier systems in a power-controlled environment. In such a scenario, it is possible to have a large signal, required to serve a user close to the cell

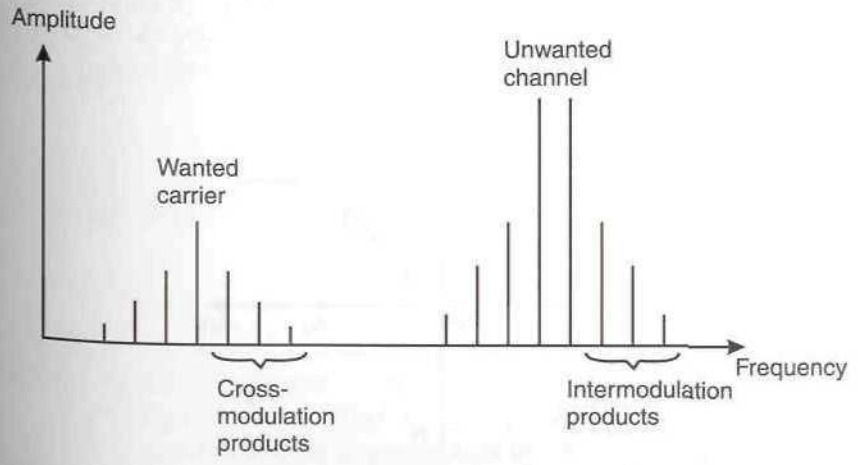


Figure 2.25 Cross-modulation affecting a weak carrier.

boundary, in the presence of a much smaller signal, required to serve a user close to the base station transmitter. Amplitude modulations, such as those present on filtered digital modulation schemes, can transfer from the higher powered signal to the lower powered signal, and thereby degrade the signal vector error of the lower powered signal.

The likely effect of this in a practical system scenario is that the mobile will report a poor signal quality and the base station will therefore increase its output power in order to provide the mobile with a better signal. This mechanism will succeed, at some power level; however, this power level will be higher than that strictly necessary to provide the required service quality (in the absence of nonlinearity in the power amplifier). The result of this will be the potential for additional interference elsewhere in the network, as unnecessary power is being transmitted.

2.13 Modelling of Amplifier Nonlinearities

2.13.1 Ideal Transfer Characteristic

The model of an ideal linear amplifier, with an optimal transfer characteristic, may be represented by an ideal linear limiter, with a transfer function given by:

$$f\{A(t)\} = \begin{cases} A(t) & \forall\, |A(t)| \leq A_0 \\ K & \forall |A(t)| > A_0 \end{cases} \tag{2.48}$$

The form of this (amplitude) characteristic is illustrated in Figure 2.26 for both positive and negative input signal swings. The phase characteristic is assumed to be linear (constant delay) at all input amplitude levels.

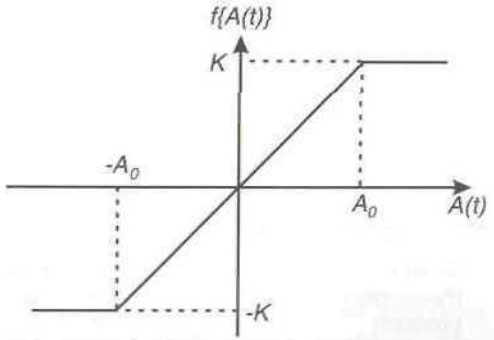


Figure 2.26 Ideal linear limiter as a model of a linear amplifier characteristic.

2.13.2 Memoryless (Instantaneous) Nonlinear Model

As its name suggests this form of model assumes that the power amplifier has no memory effects, that is, that it has no knowledge of past events and hence the present output signal is only a function of the present input signal(s). It is possible to further simplify this type of model by restricting the analysis to signals and distortion contained within the first harmonic zone; such models are then referred to as *bandpass memoryless nonlinear models* and are further simplified since they do not need to model even-order nonlinearities (none result in distortion in the first-harmonic zone).

When applying a memoryless nonlinear model to an RF amplifier, the characteristics of primary interest are the nonlinear gain response (AM/AM characteristic) and amplitude to phase conversion (AM/PM characteristic). These may be modelled in polar (amplitude and phase) (Section 2.13.5) or Cartesian (I and Q) form (Section 2.13.7).

2.13.3 AM–AM and AM–PM Conversion in a Nonlinear Amplifier

Most of the discussions so far in this chapter have concentrated on amplitude nonlinearity, that is the nonlinear relationship between input power and output power present in all practical amplifiers. This is often termed *AM–AM conversion* since it is a conversion between the amplitude modulation present on the input signal(s) and the modified amplitude modulation present on the output signal (the 'modification' being due to the amplitude nonlinearity of the amplifier).

Another effect is, however, also present and that is a conversion from amplitude modulation on the input signal to phase modulation on the output signal; this is known as *AM–PM conversion*. Consider a sinusoidally-modulated input carrier, $P_{in}(t)$, with a modulating signal defined by:

$$M(t) = A_M \cos(\omega_M t) \tag{2.49}$$

For an ideal amplifier, the output phase is given by:

$$\Phi(P_{in}(t)) = K_\phi \tag{2.50}$$

where K_ϕ is a constant. In other words, the output phase remains constant, irrespective of the amplitude (or envelope level) of the input signal.

In the case of a real amplifier, however, amplitude modulation present on the input signal will result in phase modulation of the output signal (i.e., AM–PM conversion), hence (2.50) becomes:

$$\Phi(P_{in}(t)) = K_\phi \cos[\omega_C t + A_M \cos(\omega_M t)]$$
$$= K_\phi \sum_{n=-\infty}^{\infty} J_n(A_M) \cos[(\omega_C + n\omega_M)t] \quad (2.51)$$

where J_n is a Bessel function of order n and ω_C is the angular frequency of the carrier signal. The resulting spectrum is effectively that of a phase-modulated carrier with a sinusoidal modulating signal, hence forming IMD products similar to those resulting from an amplitude nonlinearity.

A real amplifier will, of course, suffer from both forms of nonlinearity to some degree and the output spectrum will consist of a superposition of both effects. This superposition can result in an asymmetry of the IMD products, since the upper and lower IMD products resulting from AM–AM

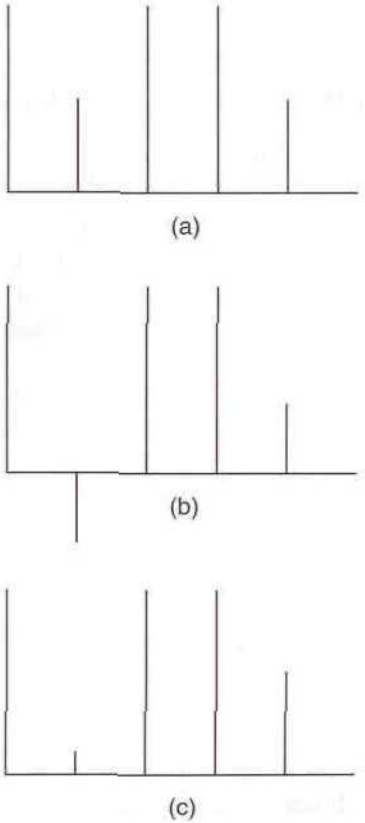


Figure 2.27 IMD resulting from a combination of AM–AM and AM–PM conversion. (a) IMD from AM–AM conversion. (b) IMD from AM–PM conversion, (c) Resultant IMD from combined AM–AM and AM–PM conversion.

conversion are in-phase, whereas a component of those resulting from AM–PM conversion can be 180° out of phase. The phase difference in the IMD products created by the AM–PM conversion process can arise from the mechanisms described in Section 3.13. The effect of the superposition of the AM–AM and AM–PM characteristics is illustrated in Figure 2.27.

2.13.4 Measurement of AM/AM and AM/PM Characteristics

The most convenient method of determining these characteristics from a practical power amplifier is by means of a vector network analyser, although the AM/AM characteristic alone may be measured by a scalar network analyser.

Most modern network analysers contain a 'power sweep' function which permits the source power to be linearly varied by the instrument over a range of 20 dB or more, at a single operating frequency. The amplitude and phase characteristics which result from this measurement are then the approximate AM/AM and AM/PM characteristics of the amplifier under test.

Care must be exercised when making this measurement (as with any other power amplifier measurement) to ensure that the instrument is not damaged by the application of an excessive RF input level.

The measurement is most useful if the (say) 20 dB power variation produced by the network analyser is arranged to just drive the amplifier into compression (by perhaps 1 dB or 2 dB). This will then illustrate what is likely to be the most nonlinear portion of the characteristic and is usually all that is necessary to characterise most amplifiers (the lower portion of the characteristic is usually relatively linear). If, however, it is suspected that significant nonlinearity is contained in the lower portion of the characteristic, then the measurement can be repeated with 20 dB of attenuation inserted before the amplifier under test and the two characteristics combined. Any effects caused by the attenuator itself (e.g., increased delay) will need to be calibrated out. The required data from the two measurements is usually available via the GPIB interface on the analyser, and hence the two characteristics may easily be combined to form the complete model.

Pulsed versions of these measurements provide a more accurate measure of the 1 dB compression point for signals containing a significant amount of AM.

2.13.5 Polar Form of a Memoryless Nonlinear Model

It is possible to treat the AM/AM and AM/PM components of a nonlinearity as separate items and hence produce a model which is effectively a cascade of the two processes. This may be considered at two levels: first, by considering

the effect of the model on a single RF sinewave (CW carrier) and, secondly, by extension of this to a modulated signal. The latter is of considerably greater value, since it will model the distortion of the signal envelope and hence the important in-band effects which are usually of greatest interest.

Consider a CW signal, $C(t)$, at a frequency f_C and with an amplitude A:

$$C(t) = A\cos(2\pi f_C t + \theta) \tag{2.52}$$

The distorted output signal, $D(t)$, of a bandpass memoryless nonlinearity is then:

$$D(t) = f(A)\cos[2\pi f_C t + \theta + g(A)] \tag{2.53}$$

where $f(A)$ describes the AM/AM characteristic of the nonlinearity and $g(A)$ describes its AM/PM conversion.

Work by Minkoff [22,23] has shown, by means of measurement, computer simulation and analytical methods, that (2.52) and (2.53) can be applied to modulated signals as well as to a single CW carrier. Thus:

$$C(t) = A(t)\cos[2\pi f_C t + \theta(t)] \tag{2.54}$$

where $A(t)$ describes the envelope modulations present on the carrier.

The resulting distorted output signal, $D(t)$, is now:

$$D(t) = f\{A(t)\}\cos[2\pi f_C t + \theta(t) + g\{A(t)\}] \tag{2.55}$$

This model can therefore characterise *envelope nonlinearities* and can be used to describe the in-band distortion of a signal containing some form of envelope variation. Such variations are present on a wide variety of signals, including full-carrier AM, SSB, filtered π/4-DQPSK, 16-QAM (and above), all multicarrier signals and many other forms of modulation. A block diagram of this form of envelope nonlinearity is shown in Figure 2.28.

The AM/AM and AM/PM characteristics of a typical class-A amplifier

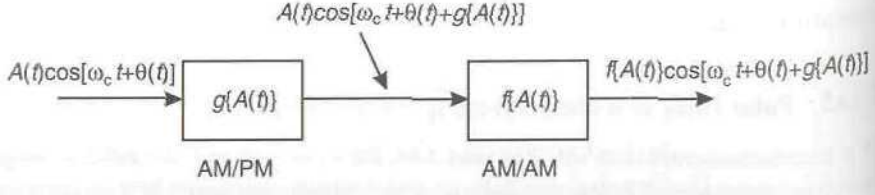
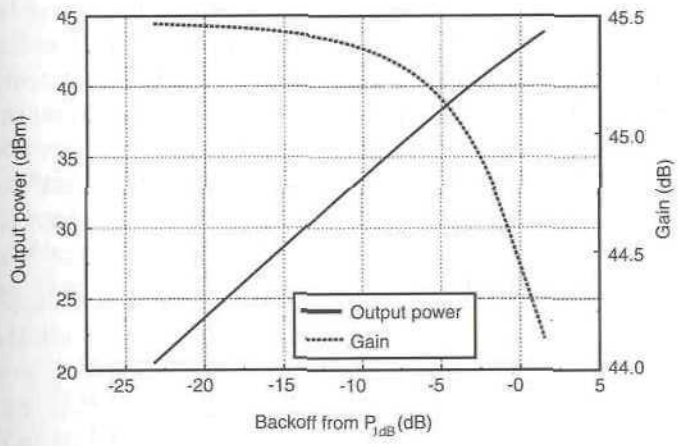
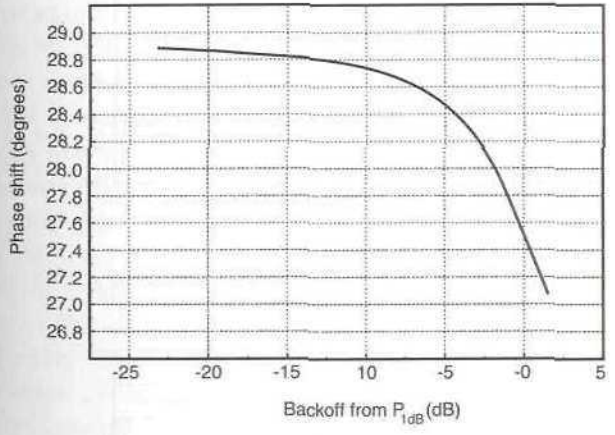


Figure 2.28 Polar form of an envelope nonlinearity for both AM/AM and AM/PM distortion.



(a)



(b)

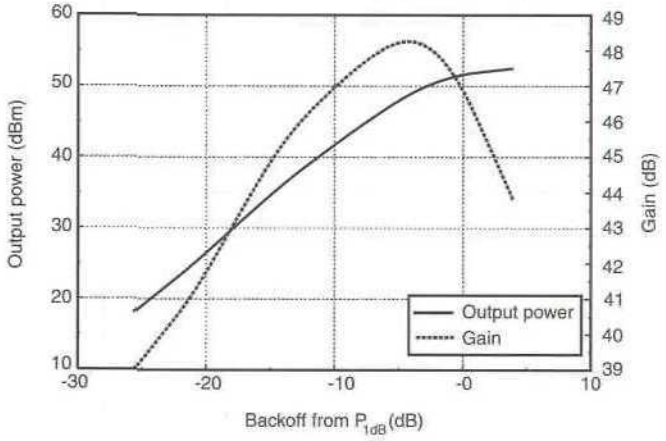
Figure 2.29 AM/AM (a) and AM/PM (b) conversion of a 900 MHz 40W class-A amplifier.

are shown in Figure 2.29 (results obtained from measurements on an actual class-A amplifier at 900 MHz). By eye, this amplifier looks to have an extremely linear characteristic, although its actual third-order IMD performance is only around 27 dBc at its 1 dB compression point.

The AM/AM and AM/PM characteristics of a typical class-C amplifier are shown in Figure 2.30 (results obtained from measurements on an actual class-C amplifier at 400 MHz). This amplifier has a very small amount of bias applied which has the effect of improving amplifier stability at the point

where the devices turn on; it therefore does not exhibit the abrupt 'turn-on' characteristic usually associated with amplifiers operated in class-C.

The overall DC to RF conversion efficiency of the amplifier, even with the addition of the bias, is almost 60%. It is evident therefore that the added bias has very little effect on the power efficiency of the amplifier. The amplifier does, however, exhibit a much steeper transfer characteristic than the class-A stage (roughly a 1.5 dB increase in output signal for every 1 dB increase in input signal over much of its characteristic). Feedback linearisation techniques, for example, the Cartesian loop, must have sufficient



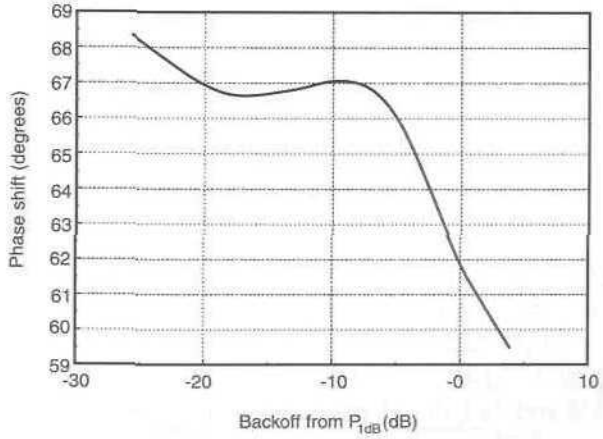


Figure 2.30 AM/AM (a) and AM/PM (b) conversion of a 400 MHz 150W class-C amplifier.

bandwidth to cope with these rapid transitions and it is this factor which often limits the loop gain available from such techniques when applied to a class-C stage. The AM–PM conversion characteristic also has a point of inflection, which is not present on that of the class-A stage.

2.13.6 Effect of AM–AM and AM–PM Conversion on Digital Modulation Formats

Increasing use is being made of digital modulation formats in both voice and data communications. Many of these formats are severely affected by the AM–AM and AM–PM conversion processes which take place in nonlinear amplifiers, irrespective of adjacent-channel considerations. The formats which are affected are those with envelope variations present on the RF signal (i.e., after any relevant filtering) and include two of the most popular schemes currently in use, namely QAM and filtered QPSK.

It is possible to highlight the effects of amplifier nonlinearity on the modulation itself, by examining the constellation and eye diagrams for these two schemes both before and after passing through a highly nonlinear (class-C) power amplifier. The resulting characteristics are shown in Figures 2.31 and 2.32 and the relevant modulation scheme parameters in Table 2.4. The class-C power amplifier had an output power rated at 10W peak and the modulation peaks were arranged to be at this level (i.e., the amplifier was not clipping the signal peaks). The channel center frequency was 435 MHz.

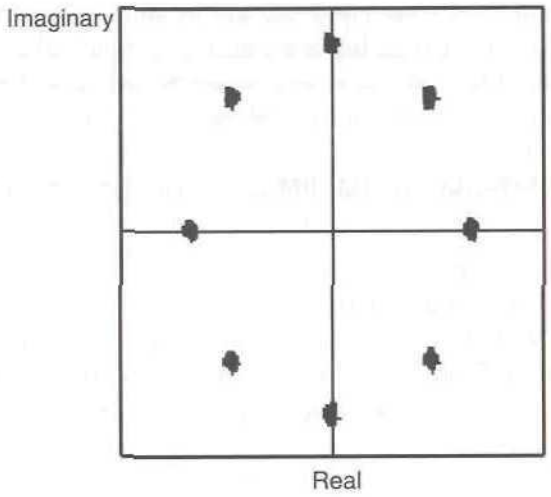
Note that in both cases, the constellation points have become ill-defined; the QAM constellation also betrays the severe AM–PM conversion present in the class-C stage through the obvious rotation of the entire constellation. The eye diagrams show clearly the effect on bit-error performance, as the sampling points (where the eyes should be 'open') are almost completely closed, thus making error-free detection impossible, even in the absence of other noise or interference.

2.13.7 Cartesian Form of a Memoryless Nonlinear Model

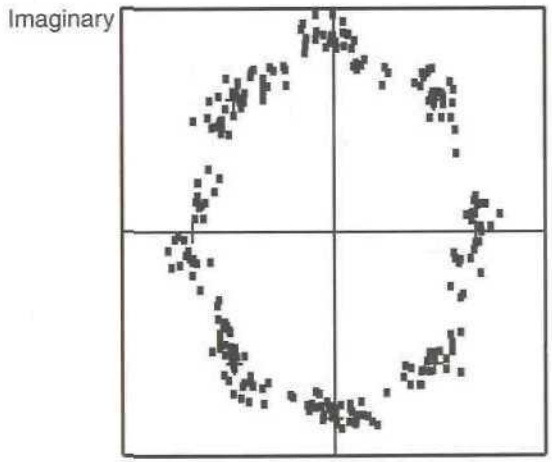
In a similar manner to the above polar nonlinear model, an equivalent Cartesian model can be generated. This model has the advantage that it can be constructed from two nonlinear *amplitude* models, $I\{A(t)\}$ and $Q\{A(t)\}$, thereby avoiding the potential complexity of an AM/PM model.

The distorted output from a bandpass memoryless nonlinearity (from (2.55)) may be expanded to give:

$$D(t) = f\{A(t)\}\cos[g\{A(t)\}]\cos[2\pi f_C t + \theta(t)] - f\{A(t)\}\sin[g\{A(t)\}]\sin[2\pi f_C t + \theta(t)] \quad (2.56)$$

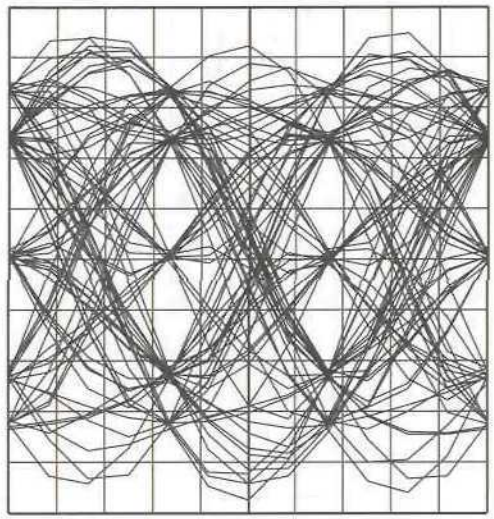


(a)

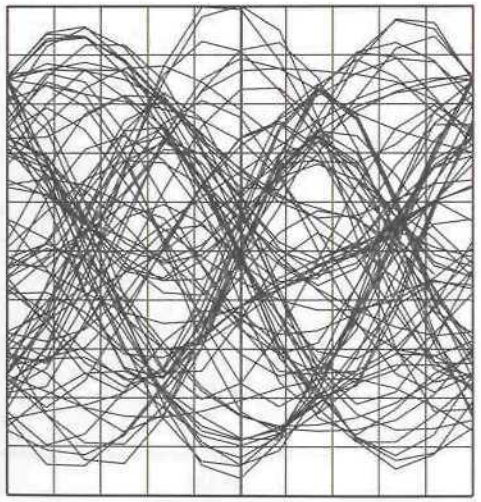


(b)

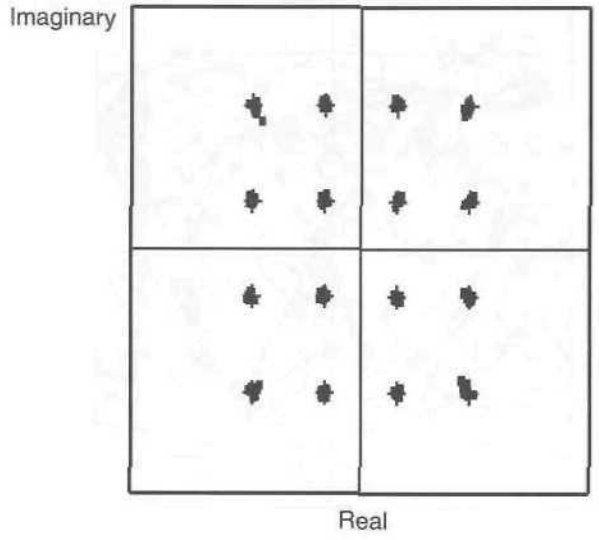
Figure 2.31 The effect of a class-C power amplifier upon a $\pi/4$-DQPSK (TETRA) signal. (a) Input signal constellation. (b) Output signal constellation. (c) Input signal eye diagram. (d) Output signal eye diagram



(c)



(d)



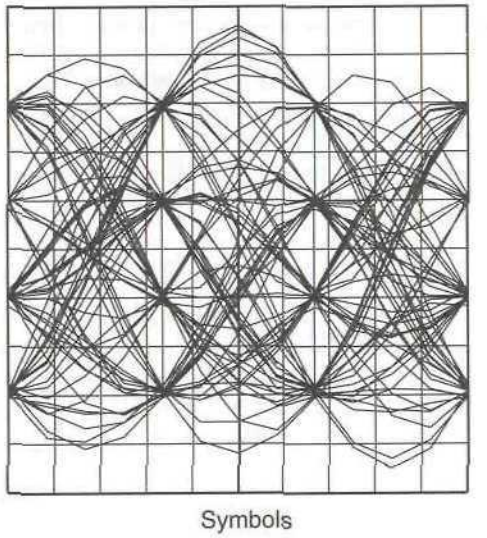


Figure 2.32 *The effect of a class-C power amplifier upon a 16-QAM signal. (a) Input signal* constellation. (b) Output signal constellation. (c) Input signal eye diagram. (d) Output signal eye diagram.

Table 2.4
Digital modulation scheme parameters used to obtain Figures 2.31 and 2.32

Modulation format	Measurement filter and transmitter filter	Reference filter	Symbol rate
π/4-DQPSK (TETRA)	Root-raised cosine $\alpha = 0.35$	Raised cosine $\alpha = 0.35$	18 kbps
16-QAM 4 bits/symbol	Root-raised cosine $\alpha = 0.35$	Raised cosine $\alpha = 0.35$	15 ks/s (60 kbps)

This may be expressed in quadrature components as:

$$D(t) = I\{A(t)\}\cos[2\pi f_C t + \theta(t)] - Q\{A(t)\}\sin[2\pi f_C t + \theta(t)] \quad (2.57)$$

where:

$$I\{A(t)\} = f\{A(t)\}\cos[g\{A(t)\}] \quad (2.58)$$

and

$$Q\{A(t)\} = f\{A(t)\}\sin[g\{A(t)\}] \quad (2.59)$$

and is shown diagrammatically in Figure 2.33.

The Cartesian form of the AM/AM and AM/PM characteristics of a typical class-C stage are shown in Figure 2.34 (for the same amplifier as

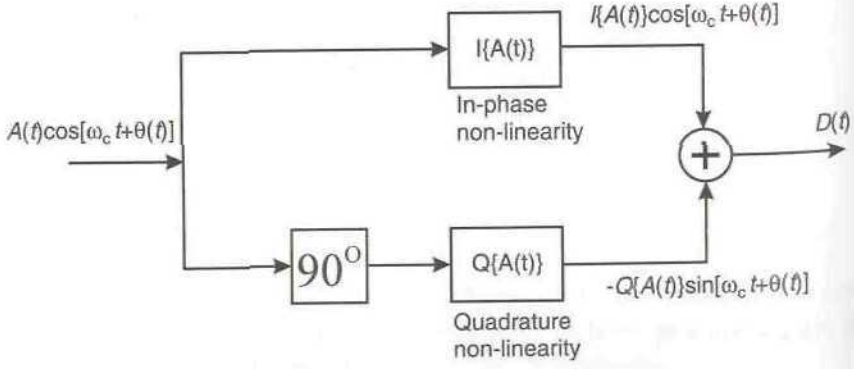


Figure 2.33 Cartesian form of an envelope nonlinearity for both AM/AM and AM/PM distortion.

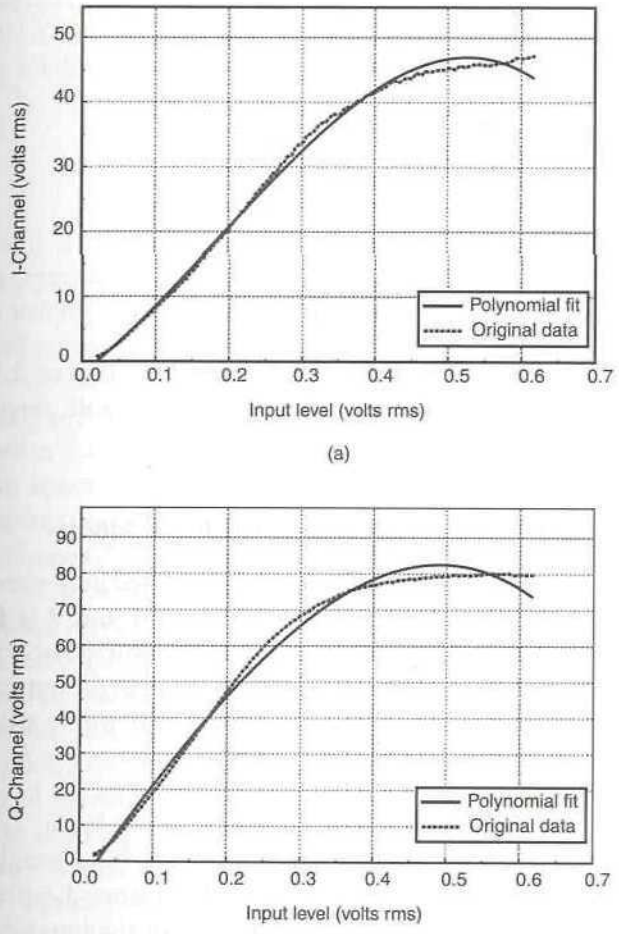


Figure 2.34 Cartesian form of the AM/AM and AM/PM characteristics of a typical class-C amplifier. (a) I-channel, with cubic polynomial fit: $V_{OUT,I} = 86.79V_{in} + 197.45V_{in}^2 - 353.92V_{in}^3$. (b) Q-channel, with cubic polynomial fit: $V_{OUT,Q} = 290.025V_{in} - 67.39V_{in}^2 - 309.70V_{in}^3$.

depicted in Figure 2.30). At first glance, the AM–AM and I-channel characteristics appear very similar and this may be explained as follows.

From (2.56) and (2.57), it can be seen that for low levels of AM/PM conversion, where:

$$\sin[g\{A(t)\}] \approx g\{A(t)\} \quad (2.60)$$

and

$$\cos[g\{A(t)\}] \approx 1 \tag{2.61}$$

then (2.58) and (2.59) may be simplified to become:

$$I\{A(t)\} = f\{A(t)\} \tag{2.62}$$

and

$$Q\{A(t)\} = f\{A(t)\}g\{A(t)\} \tag{2.63}$$

Thus, over most of the AM/PM range depicted in Figure 2.30, these approximations hold true and hence explain the overall shape of the characteristics in Figure 2.34.

2.13.8 Approximate Forms of Memoryless Nonlinear Model

The polar and Cartesian modelling techniques described above require full knowledge (and hence measurement) of the AM/AM and AM/PM characteristics of the nonlinear stage and also often require interpolation between data points, which may prove time consuming. A number of simpler models, involving more analytical functions may be used to improve speed at the expense of model accuracy.

2.13.8.1 Taylor Series

The use of this approach has already been illustrated, for a simple case, in Section 2.2. The concept involves describing the distorted output of the system, $D(t)$, in the form of a Taylor series [24,25] of the input, $v_{in}(t)$:

$$D(t) = \sum_{n=1}^{\infty} a_n v_{in}^n(t) \tag{2.64}$$

where a_n are constants.

This form of model has the advantage that the relative level of each order of distortion is clear from its coefficient a_i and the level of each IMD product may therefore be calculated. It is most useful for systems which contain relatively few orders of distortion, such as TWT amplifiers which have a predominantly third-order characteristic, and hence the series may be truncated after a manageable number of terms without loss of accuracy. Most semiconductor amplifiers, however, generate a very large number of orders of distortion and hence yield a long Taylor series expansion.

A standard Taylor series, as described by (2.64), will only characterise amplitude (AM/AM) distortion. If significant AM/PM distortion exists within the nonlinear device, then a complex power series must be used:

$$D(t) = \sum_{n=1}^{\infty} (x_n + jy_n)v_{in}^n(t) \tag{2.65}$$

where x_n and y_n are constants.

This approach can be shown to be valid, since intermodulation distortion produced by AM/PM conversion is orthogonal to that produced by AM/AM conversion [26,27].

An alternative is the use of two series, one representing the I-characteristic and the other the Q-characteristic. This is effectively a splitting of the terms in (2.65) and results in a model of the form shown in Figure 2.35.

If this approach is used on the class-A and class-C amplifier characteristics presented previously (Figures 2.29 and 2.30), then the characteristics shown in Figures 2.36 and 2.37 result, when terms of up to a seventh-order are used. The resulting models are sufficiently accurate for most purposes and yet are very straightforward to understand and use. Note that u_I and u_Q should be zero (in Figure 2.35), since a properly-functioning amplifier should have an output of zero when no input signal is present.

Note also that in both cases (class-A and class-C), the characteristics and the models extend beyond the 1 dB compression point (see Figures 2.29 and 2.30). In general, this will lead to a poorer model fit than if it were restricted to this range, however, most of the nonlinearity, in the case of the class-C stage, occurs well before the compression point and hence the effect on model accuracy will be relatively small. The models presented do, however, allow 'overload' effects to be modelled and this is useful in assessing the performance of a system under these circumstances.

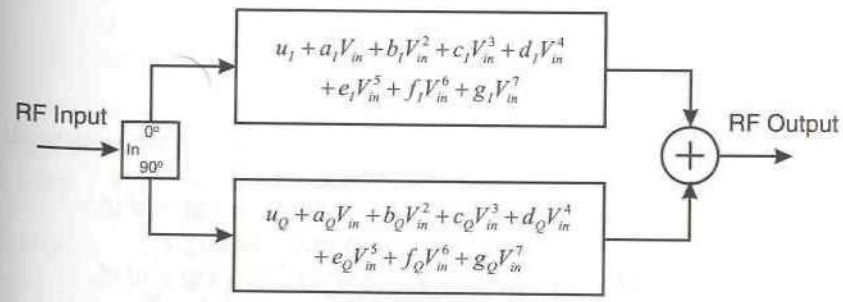


Figure 2.35 Quadrature Taylor series amplifier model.

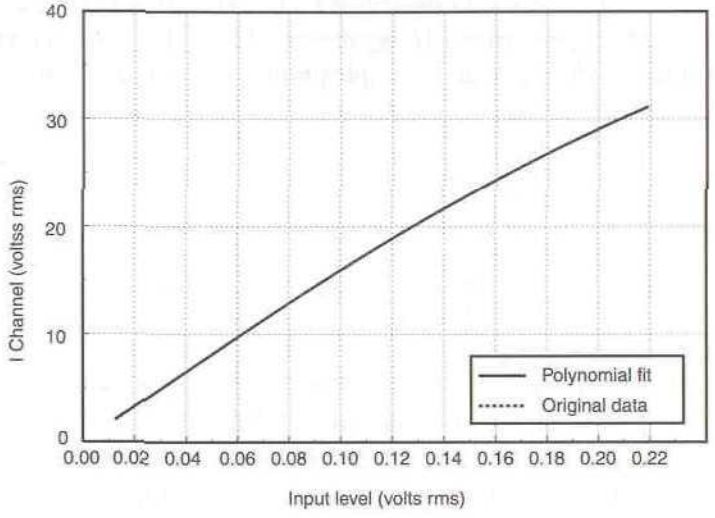


Figure 2.36 I- and Q-channel characteristics ((a) and (b), respectively) and seventh-order polynomial fit for a 900 MHz 40W class-A amplifier. Fit equations: $V_{OUT,I} = 166.50V_{in} - 109.02V_{in}^2 + 1856V_{in}^3 - 25900V_{in}^4 + 1.53 \times 10^5 V_{in}^5 - 4.63 \times 10^5 V_{in}^6 + 5.75 \times 10^5 V_{in}^7$ $V_{OUT,Q} = 91.80V_{in} - 67.35V_{in}^2 + 1124V_{in}^3 - 17600V_{in}^4 + 1.087 \times 10^5 V_{in}^5 - 3.46 \times 10^5 V_{in}^6 + 4.52 \times 10^5 V_{in}^7$.

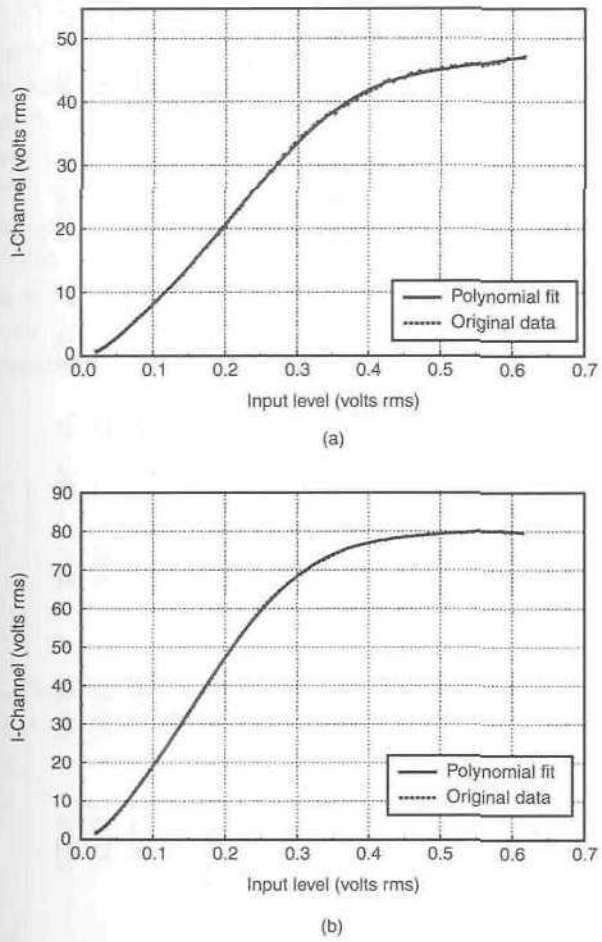


Figure 2.37 I- and Q-channel characteristics ((a) and (b), respectively) and seventh-order polynomial fit for a 400 MHz 150W class-C amplifier. Fit equations: $V_{OUT,I} = 73.02V_{in} + 159.59V_{in}^2 + 106.8V_{in}^3 + 2270V_{in}^4 - 1750V_{in}^5 + 32700V_{in}^6 - 19100V_{in}^7$ $V_{OUT,Q} = 129.78V_{in} + 654.62V_{in}^2 + 3370V_{in}^3 - 30800V_{in}^4 + 73300V_{in}^5 - 72600V_{in}^6 + 25900V_{in}^7$.

A summary of the quadrature PA model characteristics for the class-A and class-C amplifiers considered above is given in Table 2.5, for a model of the form given in (2.66).

$$V_{OUT,I} = a_I V_{in} + b_I V_{in}^2 + c_I V_{in}^3 + d_I V_{in}^4 + e_I V_{in}^5 + f_I V_{in}^6 + g_I V_{in}^7$$
$$V_{OUT,Q} = a_Q V_{in} + b_Q V_{in}^2 + c_Q V_{in}^3 + d_Q V_{in}^4 + e_Q V_{in}^5 + f_Q V_{in}^6 + g_Q V_{in}^7 \quad (2.66)$$

Table 2.5
Summary of polynomial coefficients for the class-A and class-C quadrature power amplifier models

Amplifier class	Model order	Channel	*a*	*b*	*c*	*d*	*e*	*f*	*g*
Class-A	Cubic	I	165.49	−20.72	−395.87	—	—	—	—
		Q	91.32	−13.34	−340.39	—	—	—	—
	Quintic	I	164.17	−4.64	−356.06	−1182	3641	—	—
		Q	90.58	−8.23	−231.77	−1270	3420	—	—
	Seventh-order	I	166.50	−109.02	1856	-2.59×10^4	1.53×10^5	-4.63×10^5	5.75×10^5
		Q	91.80	−67.35	1124	-1.76×10^4	1.087×10^5	-3.46×10^5	4.52×10^5
Class-C	Cubic	I	86.79	197.45	−353.92	—	—	—	—
		Q	290.025	−67.39	−309.70	—	—	—	—
	Quintic	I	59.99	284.42	127.41	-2.26×10^3	2.25×10^3	—	—
		Q	36.32	2220	−8040	1.06×10^4	-4.90×10^3	—	—
	Seventh-order	I	73.02	159.59	106.80	2.27×10^3	-1.75×10^4	3.27×10^4	-1.91×10^4
		Q	129.78	654.62	3370	-3.08×10^4	7.33×10^4	-7.26×10^4	2.59×10^4

2.13.8.2 Saleh Functions

These functions have been suggested as a simple method of approximating a nonlinear characteristic in either polar or Cartesian form [28]. They have subsequently been applied to the problem of predistortion linearisation of TWT amplifiers and the prediction of carrier-to-intermodulation ratios for multicarrier and Gaussian input signals [29].

The basic concept of Saleh functions is to approximate the AM/AM and AM/PM characteristics of a nonlinearity using two formulas. The constants within these formulas are determined by means of a least-squares curve fit to experimentally-determined AM/AM and AM/PM data. For a polar representation, these functions are:

$$f(r) = \frac{\alpha_a r}{[1 + \beta_a r^2]} \tag{2.67}$$

and

$$g(r) = \frac{\alpha_\theta r^2}{[1 + \beta_\theta r^2]} \tag{2.68}$$

where α_a, α_θ, β_a, β_θ are constants and r is the envelope of the input signal.

The corresponding Cartesian representation is:

$$I(r) = \frac{\alpha_I r}{[1 + \beta_I r^2]} \tag{2.69}$$

and

$$Q(r) = \frac{\alpha_Q r^3}{[1 + \beta_Q r^2]^2} \tag{2.70}$$

where α_I, α_Q, β_I, β_Q are constants.

These functions have been shown [29] to provide an accurate, yet simple, model for a bandpass memoryless nonlinearity.

An example of the use of these functions (in polar form) is provided in [30] where the normalised amplitude and phase nonlinearities of a travelling wave tube amplifier are given as:

$$f(r) = \frac{2r}{[1 + r^2]} \tag{2.71}$$

and

$$g(r) = 60^\circ \frac{r^2}{1+r^2} \tag{2.72}$$

However, as the nonlinearity of the device increases, the modeling accuracy of Saleh functions decreases. Hence, they are appropriate for quasi-linear amplifiers, such as class-A and class-AB, but are not appropriate for highly nonlinear amplifier classes, such as class-C.

2.13.9 Bandpass Nonlinear Models Incorporating Memory Effects

The instantaneous nonlinear models described in Section 2.13.8 do not account for memory effects within the amplifier (other than AM/PM conversion) and hence are unable to describe the frequency-dependent behavior of the system. Frequency dependent behavior can often be neglected, since the operating bandwidth of most systems is relatively small with respect to the bandwidth capability of the devices used within them. However, in wideband amplifier designs, this assumption is no longer valid and hence frequency dependent models are required.

The frequency-dependence of the amplifier characteristic causes the AM/AM and AM/PM conversion behavior to alter with frequency. It is possible to approximately model this frequency dependent behavior of a nonlinearity with memory and some of the relevant techniques are described below.

2.13.10 Saleh Model

The approximate equations proposed by Saleh and described in Section 2.13.8.2 can be extended to incorporate memory effects [28,29]. This extension does, however, introduce some significant loss of generality [32] since it restricts the shape of the in-phase and quadrature nonlinearities to being independent of frequency; a scaling factor alone provides the frequency dependence. The general shape of the AM/AM and AM/PM characteristics does not alter radically with frequency (the same also being true of the in-phase and quadrature characteristics), however, this loss of generality will still reduce the accuracy which can be obtained using this model.

The model is extended to incorporate frequency-dependent behavior by altering the coefficients α and β in (2.69) and (2.70) to make them a function of frequency. The resulting equations are:

$$I(r) = \frac{\alpha_I(f) r}{[1+\beta_I(f) r^2]} \tag{2.73}$$

and

$$Q(r) = \frac{\alpha_Q(f) r^3}{[1+\beta_Q(f) r^2]^2} \tag{2.74}$$

where $\alpha_I(f)$, $\alpha_Q(f)$, $\beta_I(f)$, $\beta_Q(f)$ are determined by a number of best-fit approximations at a number of frequencies within the band of interest, resulting in an overall best-fit function in each case. These functions can be decomposed into a cascade of three elements: a filter, a memoryless nonlinearity and a further filter. Finally, an additional block, labeled $\phi_0(f)$, is required to model the small-signal phase (the phase which results as $r \rightarrow 0$). The complete model is shown in Figure 2.38.

2.13.11 Blum and Jeruchim Model

A disadvantage with the models described so far is that the interaction effects of adjacent signals on a signal of interest, have not been assessed. These arise due to the measurement method employed in determining the models, namely that of utilising a single frequency tone and stepping it in both power and frequency in order to generate a series of curves. These curves are then the basis for a curve fitting exercise which yields the parameters for the model.

An improved model has therefore been suggested [32] in which the nonlinear device is 'channel-sounded' by a pseudo-random (PN) sequence, binary phase-shift key (BPSK) modulated onto a carrier at the center of the band of interest (thus forming a direct-sequence spread-spectrum signal). A property of this type of modulated carrier is that a number of spectral lines are generated, with a separation determined by the 'chipping-rate' (clock-rate of the PN generator). These spectral lines act like a multicarrier signal and therefore provide the required adjacent signals, assuming that one spectral line is designated as the tone of interest.

The PN modulated carrier signal is applied to the nonlinear device under test and the resulting output signal passed through a narrow bandpass filter. The output of this filter is a measure of the amplitude of a particular frequency component, with the center frequency being swept across the band of interest to generate a complete frequency characteristic. The input power level is also varied, thereby creating a complete series of curves in a similar manner to that required for the Saleh model above. The PN modulated carrier itself is constant envelope (assuming that it is not filtered

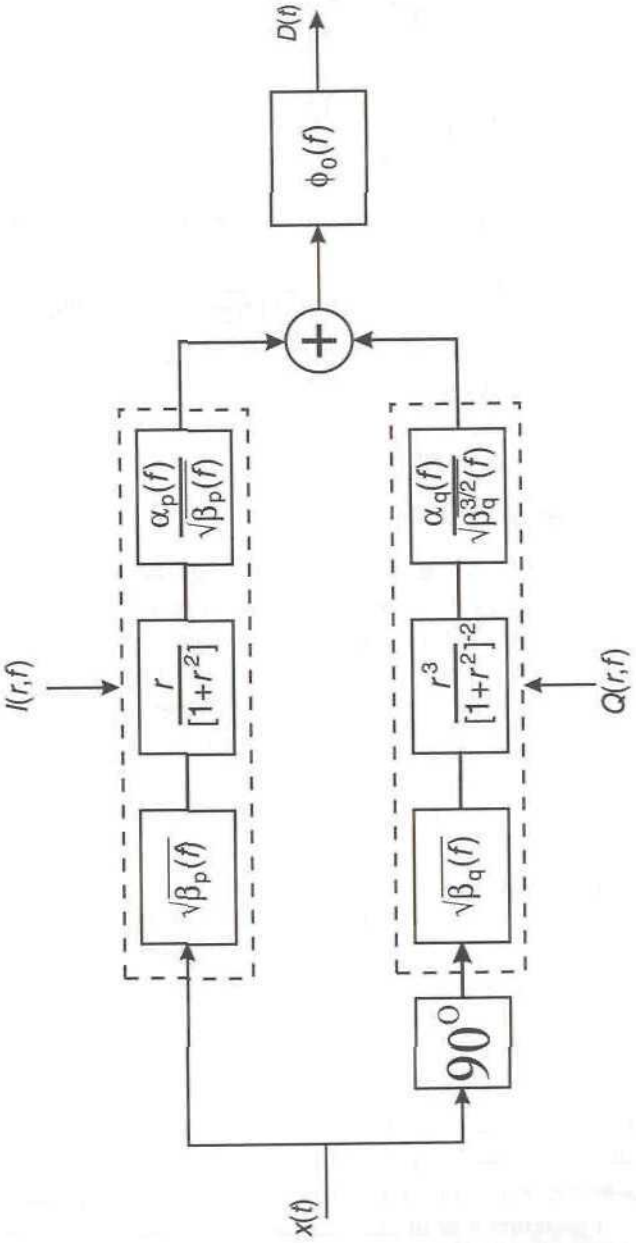


Figure 2.38 Frequency-dependent nonlinear model as proposed by Saleh (after [28]).

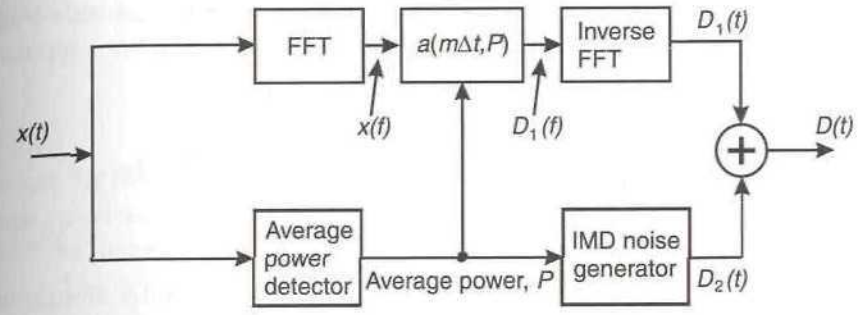


Figure 2.39 Frequency-dependent nonlinear model as proposed by Blum and Jeruchim (after [32]).

before being fed to the input of the device under test) and hence attenuation of its level at the input provides a valid measure of the level dependence of the amplifier's characteristics.

The form of the resulting model is shown in Figure 2.39. The input signal is split, with the upper path feeding a fast-Fourier transform (FFT) to convert the time-domain signal into the frequency domain. The resulting signal is passed through the power-dependent transfer function, with the required average power information supplied from a power measurement algorithm operating on the lower-path split of the input signal. The output of this transfer function is then converted back to the time domain using an inverse FFT to provide the bulk of the modeling process.

It has been acknowledged, however, that the treatment by this model of intermodulation products may be less than ideal and hence an additional block has been added in the lower path. This block takes account of the shortcoming in intermodulation generation by providing a power-dependent intermodulation noise generation function which can then be summed with the output of the inverse FFT. The resulting signal forms the complete model of the nonlinear device.

2.13.12 Volterra Series

A Volterra series [33,34] can be described as a 'Taylor series with memory' [35] and can be used to accurately model an arbitrarily nonlinear system. It does this by describing the distorted output signal, $D(t)$, as an infinite series:

$$D(t) = \sum_{n=0}^{\infty} D_n(t) \tag{2.75}$$

where $D_n(t)$ is the nth order response of the system, formed from an n-fold convolution of the input signal $x(t)$ with the nth-order nonlinear impulse response of the system, $h_n(\tau_1, \ldots, \tau_n)$:

$$D_n(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_n(\tau_1, \ldots, \tau_n) x(t - \tau_1) \ldots x(t - \tau_n) d\tau_1 \ldots d\tau_n \tag{2.76}$$

The functions $h_n(\tau_1, \ldots, \tau_n)$ are known as the nth-order Volterra kernels of the system and the system can thus be completely characterised by a knowledge of these kernels. Thus, for example, h_0 describes the system response to a DC input signal, h_1 the linear response of the system and the higher order kernels, $h_{2\ldots n}$, the higher-order nonlinearities of the system.

The primary advantage of the Volterra series is in its ability to individually describe all orders of distortion present in the system and to allow comparisons to be made between the different components. It does, however, share a number of disadvantages with the Taylor series, for example that of slow convergence. It also has the disadvantage that computation of the higher-order kernels from measured data is difficult.

Equations (2.75) and (2.76) describe the baseband Volterra series nonlinear model; this can be modified to yield a bandpass model by eliminating the DC term and the even-order terms, since these will yield harmonics of the bandpass signal. The resulting equations are complex and the Volterra kernels difficult to compute, making this form of model unattractive in most circumstances.

2.13.13 Generalised Power Series

It is possible to create a frequency-domain power series representation of the output of a nonlinear system (the Taylor and Volterra series being time-domain representations) and the result is a much more efficient series [36,37] which can deal with severe nonlinearities; it can be shown to be related to the Volterra series [38].

The input signal to a nonlinear system may be described as:

$$x(t) = \sum_{n=1}^{N} |x_n| \cos(\omega_n t + \phi_n) \tag{2.77}$$

where $|x_n|$ are the magnitudes of the individual frequency components, ω_n. The distorted output signal from the nonlinear system, $D(t)$, can then be described as a generalised power series:

$$D(t) = A \sum_{i=0}^{\infty} a_i \left\{ \sum_{n=1}^{N} h_n x_n(t - \tau_{n,i}) \right\}^i \tag{2.78}$$

where i is the order of the power series, coefficients a_i and h_n are complex and real respectively, and $\tau_{n,i}$ is a time delay term which depends upon frequency and the order of the power series.

The improved modelling efficiency of this technique over the Volterra series representation stems from its operation in the frequency-domain, although it is unclear how to generate the above coefficients from tabulated measured data. The inclusion of frequency-dependence of the time delay terms and the use of complex coefficients in the model allows a very wide range of nonlinearities to be characterised. This is therefore potentially a powerful modelling method, assuming that the relevant coefficients can be obtained.

References

1. De Carvalho, N. B., and J. C. Pedro, "Compact formulas to relate ACPR and NPR to two-tone IMR and IP3," *Microwave Journal*, December 1999, pp. 70–84.
2. Pedro, J. C., and N. B. De Carvalho, "On the use of multi-tone techniques for assessing RF components' intermodulation distortion," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 47, No. 12, December 1999, pp. 2393–2402.
3. Schetzen, M, *The Volterra and Wiener Theories of Nonlinear Systems*, John Wiley and Sons, New York, 1980.
4. Carvalho, N. B. and J. C. Pedro, "Multi-tone intermodulation distortion performance of 3rd order microwave circuits," *IEEE Symposium on Microwave Theory and Techniques*, Anaheim, USA, Vol. 2, June 1999, pp. 763–766.
5. Boyd, S., "Multitone signals with low crest factor," *IEEE Trans. on Circuits and Systems*, Vol. 33, October 1986, pp. 1018–1022.
6. Bauml, R. W., R. F. H. Fischer and J. B. Huber, "Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping," *IEE Electronics Letters*, Vol. 32, October 1996, pp. 2056–2057.
7. Rudin, W., "Some theorems on fourier coefficients," *Proc. of the American Mathematics Society*, Vol. 10, December 1959, pp. 855–859.
8. Greestein, L. J., and P. J. Fitzgerald, "Phasing multitone signals to minimize peak factors," *IEEE Trans. on Communications*, Vol. 29, July 1981, pp. 1072–1074.
9. Golay, M. J. E., "Complementary series," *IRE Trans. on Information Theory*, Vol. 7, April 1961, pp. 82–87.
10. Popovic, B. M., "Synthesis of power efficient multitone signals with flat amplitude spectrum," *IEEE Trans. on Communications*, Vol. 39, July 1991, pp. 1031–1033.
11. Tellambura, C., "Use of m-sequences for OFDM peak-to-average power ratio reduction," *IEE Electronics Letters*, Vol. 33, July 1997, pp. 1293–1294.

12. Barker, R. H., "Group synchronizing of binary digital systems," *Communication Theory*, 1953, pp. 273–287.
13. Golomb, S. W., and R. A. Scholtz, "Generalized Barker sequences," *IEEE Trans. on Information Theory*, Vol. 4, October 1965, pp. 533–537.
14. Newman, D. J., "An L^1 extrimal problem for polynomials," *Proc. of the American Mathematics Society*, Vol. 16, December 1965, pp. 1287–1290.
15. Schroeder, M. R., "Synthesis of low peak-factor signals and binary sequences with low autocorrelation," *IEEE Trans. on Information Theory*, 1970, pp. 85–89.
16. Jones, A. E., T. A. Wilkinson, and S. K. Barton, "Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission systems," *IEE Electronics Letters*, Vol. 30, December 1994, pp. 2098–2099.
17. Muller, S. H., and J. B. Huber, "OFDM with reduced peak-to-mean power ratio by optimum combination of partial transmit sequences," *IEE Electronics Letters*, Vol. 33, February 1997, pp. 368–369.
18. Choi, B-J, E-L Kuan, and L. Hanzo, "Crest factor study of MC-CDMA and OFDM," *Proc. IEEE Vehicular Technology Conference, Fall '99*, Amsterdam, The Netherlands, Vol. 1, 19–23 September 1999, pp. 233–237.
19. Hardy J. K., *High Frequency Circuit Design*, Chapter 1, Reston Publishing Company, Reston, Virginia, USA, 1979.
20. Katz, A., "TWTA Linearization," *Microwave Journal*, April 1996.
21. Krauss, H. L., C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering*, Chapter 12, John Wiley & Sons, New York, USA, 1980.
22. Minkoff, J. B., "Intermodulation noise in solid-state power amplifiers for wideband signal transmission," *Proc. AIAA 9th Communications Satellite System Conference*, March 1982.
23. Minkoff, J. B., "Wideband operation of nonlinear solid-state power amplifiers—comparisons of calculations and measurements," *Bell System Technical Journal*, Vol. 63, No. 2, February 1984, pp. 231–248.
24. Kuo, Y., "Noise loading analysis of a memoryless nonlinearity characterised by a Taylor series of finite order," *IEEE Trans. on Instrumentation and Measurement*, Vol. IM-22, September 1973, pp. 246–249.
25. Larkin, R., "Multiple signal intermodulation and stability considerations in the use of linear repeaters," *Proc. IEEE 41st Vehicular Technology Conference*, St. Louis, Missouri, USA, May 1991, pp. 747–752.
26. Shimbo, O., "Effects of Intermodulation, AM–PM conversion and additive noise in multicarrier TWT systems," *IEEE Proc.*, Vol. 59, February 1971, pp. 230–238.
27. Sechi, F., "Linearised class-B transistor amplifiers," *IEEE Journal of Solid-State Circuits*, Vol. SC-11, No. 2, 1976, pp. 264–270.
28. Saleh, A. A. M., "Frequency-independent and frequency-dependent nonlinear models of TWT amplifiers," *IEEE Trans. on Communications*, Vol. COM-29, November 1981, pp. 1715–1720.
29. Saleh, A. A. M., "Intermodulation analysis of FDMA satellite systems employing compensated and uncompensated TWT's," *IEEE Trans. on Communications*, Vol. COM-30, No. 5, May 1982, pp. 1233–1242.
30. Saleh, A. A. M., and J. Salz, "Adaptive linearization of power amplifiers in digital radio systems," *The Bell System Technical Journal*, Vol. 62, No. 4, April 1983, pp. 1019–1033.
31. Jeruchim, M., P. Balaban, and K. Shanmugan, *Simulation of Communications Systems*, New York: Plenum Press, 1992, Chapter 2.

32. Blum, R., and M. Jeruchim, "Modeling nonlinear amplifiers for communication simulation," *Proc. IEEE International Conference on Communications*, Boston, USA, June 1989, pp. 1468–1472.
33. Narayanan, S., "Transistor distortion analysis using Volterra series representation," *Bell System Technical Journal*, Vol. 46, May/June 1967, pp. 991–1024.
34. Law, C., and C. Aitchison, "Prediction of wide-band power performance of MESFET distributed amplifiers using the Volterra series representation," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-34, December 1986, pp. 1308–1317.
35. Jeruchim, M., P. Balaban, and K. Shanmugan, *Simulation of Communications Systems*, New York: Plenum Press, 1992, Chapter 2.
36. Steer M., and P. Khan, "An algebraic formula for the output of a system with large-signal, multifrequency excitation," *Proc. of the IEEE*, Vol. 71, January 1983, pp. 177–179.
37. Rhyne, G., M. Steer, and B. Bates, "Frequency-domain nonlinear circuit analysis using generalized power series," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-36, February 1988, pp. 379–387.
38. Steer, M., P. Khan, and R. Tucker, "Relationship between Volterra series and generalised power series," *Proc. of the IEEE*, Vol. 71, December 1983, pp. 1453–1454.

3

RF Power Amplifier Design

3.1 Introduction

Radio frequency power amplifiers (PAs) may be broadly defined by two categories: those which attempt to preserve the original waveshape of the input signal at the output, and those which make no attempt at its preservation. The former category is termed *linear amplifiers* and the latter *constant envelope*, or *nonlinear amplifiers*. The terms 'linear' or 'nonlinear' refer to the shape of the transfer function of the amplifier as outlined in Chapter 2.

Within the two broad categories outlined above, there are a number of sub-divisions, or *classes* of amplifier commonly used. The distinction between the various classes occurs, for example, because of their circuit configurations, operational topologies, linearity, and efficiency.

The three main classes of linear amplifier are A, AB and B, with class-A generally being the most linear and least efficient of the three. Such amplifiers are traditionally most commonly employed in SSB or AM transmitters, where the modulation is transmitted at least partially by means of the amplitude of the RF signal and hence preservation of the signal envelope is important. In such transmitters the nonlinearities which are inevitably present produce distortion in the form of 'splatter' onto adjacent channels as well as distortion within the wanted channel. The linearity performance of these transmitters is therefore an important parameter.

There are a larger number of nonlinear amplifier classes ranging from self-bias schemes to various forms of switching amplifiers. The main classes are designated: C, D, E, F, G, H and S. Class-C amplifiers use a similar circuit

topology to that of class-A amplifiers as the active device is also acting as a current source (for valve or FET circuits), however, the bias arrangement is radically different leading to considerable levels of distortion. Class-D, -E and -S amplifiers utilise the active device as a switch and hence the theoretical maximum efficiency is 100%, assuming that the device has zero switching time, zero on-resistance and infinite off-resistance. Class-F, -G and -H amplifiers are novel configurations which attempt to improve upon the basic efficiency of class-B or -C amplifiers by adopting modified forms of the standard circuit topologies.

The derivations detailing the operation of the various amplifier classes covered in this chapter are based on those of Raab, and further detail can be found in [1].

3.2 Power Semiconductors

The amplifier linearisation techniques discussed in this book are applicable to all forms of amplifier, however, it is anticipated that they will most frequently be used in power amplifier and transmitter design. It is therefore worth examining briefly some of the characteristics of the semiconductor devices which are usually employed in such circuits.

There are two main types of power semiconductor currently employed in high-power RF power amplifiers: bipolar junction transistors (BJTs) and MOSFET devices (in the form of VMOS, TMOS, and LDMOS technologies).

The more recently adopted MOS devices have a number of advantages over the more traditional BJT. They are considerably easier to bias, both when used as linear devices in class-A amplifiers and when negatively biased for class-C operation. The bias circuitry does not need to take account of the temperature of the power device since most MOSFET devices have a negative temperature coefficient and are thus thermally stable. Close thermal coupling is required in BJT bias circuits due to their positive temperature coefficient and consequent tendency toward thermal runaway. This close coupling (usually of a biasing diode to the transistor case) is generally unnecessary for MOS devices.

MOSFET devices are also less susceptible to secondary breakdown effects occurring within the device. Such breakdown is caused by excessive instantaneous power levels, due to the peak voltage-current product, which cause hot-spots within the transistor and subsequent device failure.

One disadvantage, however, is the variability of the threshold voltage, V_T, between different examples of the same device. This can result in biasing

differences between production units and must be taken into account in the amplifier design. In particular, there is much concern over 'V_{GS} drift' in LDMOS devices, with various techniques being suggested to overcome the problem. Essentially the problem amounts to a change in the threshold voltage of 20% or more over the lifetime of the device, with much of this change occurring within the first few hours after first use. Techniques which have been suggested to overcome the problem include:

1. A period of 'burn-in' of a given amplifier during post manufacturing 'test', to allow the bulk of the change to occur prior to final bias setting. This can prove expensive in test time (and hence product cost).
2. Automatic (feedback-based) bias control, usually based upon a measure of the average drain current under particular operating conditions.
3. Open-loop bias trimming based on elapsed time from first switch on. The drift profile of a range of samples of a given device may be used to determine a 'typical' drift characteristic and this can be stored in a look-up table which is accessed based on the elapsed time since the device was first used.
4. Bias control based upon gain monitoring. A change in the bias level will cause an associated change in amplifier gain and this can be monitored by comparing the input and output power levels to the amplifier (or just the output power level, if the input level is known to be fixed and constant).
5. Intelligent monitoring of any linearisation scheme surrounding the amplifier. As the bias level changes, the degree of linearisation required to meet a given specification will also change and a well-designed linearisation scheme will be aware of this. It can therefore re-adjust the bias level to return the degree of correction required to that determined during its design.

Metal semiconductor field effect transistors (MESFETs) are increasingly of interest in the RF power amplifier field. They are widely used in GaAs integrated circuit power amplifiers [2], particularly for handset applications and silicon carbide (SiC) devices are now being fabricated experimentally for use in higher power systems [3]. SiC devices, in particular, have a number of interesting and useful properties with regard to high power RF and microwave amplification:

1. SiC has a wide bandgap of 3.2 eV (compared with 1.1 eV for silicon and 1.4 eV for gallium arsenide). This wide bandgap gives rise to a

very high breakdown electric field, some ten times greater than that of either Si or GaAs.

2. The saturated electron velocity is predicted to be around three times that of GaAs at high electric fields.
3. It has a high drain efficiency (largely due to the high breakdown voltage of ~100V).
4. It also has a good channel current density (300 mA/mm for 0.7 μm gate-length devices).
5. Combining 3 and 4 above yields a large RF power density of 3W/mm.
6. Finally, it has a very high thermal conductivity (roughly three times that of silicon and ten times that of GaAs).

The above properties lead to a predicted ultimate power level for a SiC MESFET being at least five times that available from GaAs. This is clearly a significant benefit and hence SiC may well become a key technology in RF amplifier designs in the near future.

Many RF transistors are designed with their application in mind and are often unsuitable for other applications seemingly within their ratings. Devices intended for class-C operation, for example, may be destroyed by even modest amounts of standing bias intended to induce class-A operation. Selection of devices for a particular application is therefore not as straightforward as is the case with audio-frequency power amplifier designs.

A large signal semiconductor device may be characterised by three regions of operation: *cut-off*, linear or *active* and *saturation*. These definitions will be used in subsequent discussion of the various amplifier classes and therefore will be briefly explained here.

Cut-off refers to the region of operation in which there is insufficient forward bias on the device for conduction to occur. In the case of a silicon NPN BJT, cut-off occurs when the base-emitter voltage is less than about 0.7V. The equivalent situation for a MOSFET occurs when a gate-source voltage of less than the threshold voltage, V_T, is applied. The value of V_T is around 2V to 3V for a typical MOS power FET. In the cut-off region the device may essentially be considered as an open circuit between all terminals.

As the forward bias on the device is increased and eventually exceeds the relevant threshold ($V_{BE} > 0.7$V or $V_{GS} > V_T$), the active region is entered. In the case of a BJT, the base-emitter junction becomes a forward biased diode and the collector-emitter junction becomes a current source whose value is equal to the base current times the current gain of the device, β—a linear relationship. To operate in this region the base current must be

small enough such that the device does not enter saturation, in other words the collector-emitter voltage must be larger than that at saturation, V_{SAT}. A typical value of this saturation voltage is around 0.2V, although it will vary depending upon the device in question.

In the case of a large-signal FET, the active region is represented by a current source between the drain and source terminals of value $i_D = g_m (V_{GS} - V_T)$.

When an FET is driven into saturation, it appears as a pure resistance, R_{on} (to a first approximation). It will enter saturation if a drain voltage of less than $i_D R_{on}$ appears due to the load, whilst the device is operating in the active region.

A BJT will enter saturation when the voltage across the load, V_L, is such that:

$$V_{CE} = V_{sat} \tag{3.1}$$

In this region, the collector-emitter voltage is roughly constant ($=V_{sat}$). The limiting condition for entering this state is that I_B (or V_{BE}) and therefore I_C

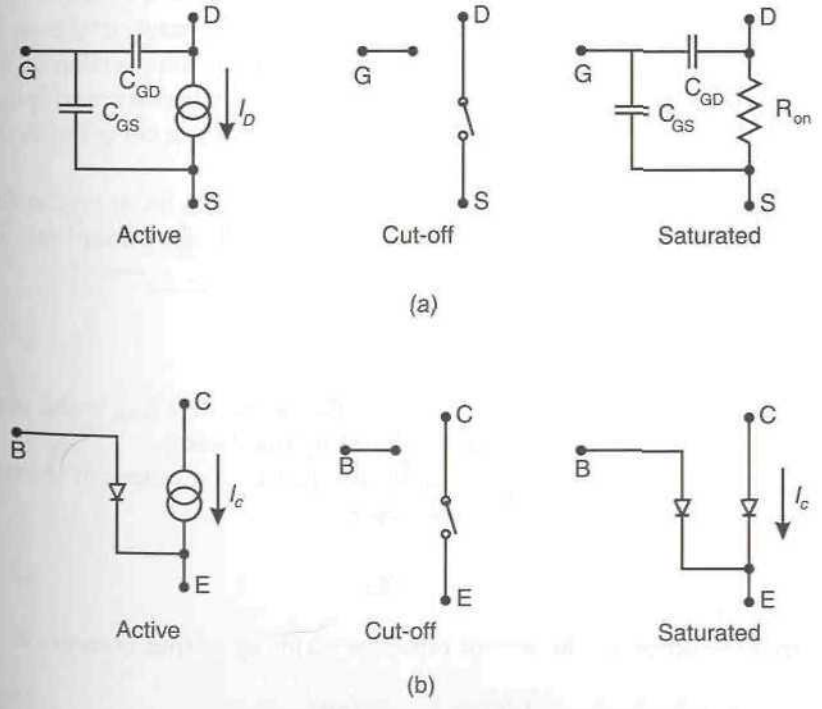


Figure 3.1 FET (a) and BJT (b) simplified device models for the major regions of operation.

are increased to a level such that V_C falls to V_B. In other words the base-collector junction is on the verge of becoming forward biased. Further increases of I_B (or V_{BE}) and therefore I_C, lead to V_C falling very close to V_E and limiting at $V_{CE,sat}$.

The various modes of operation of MOSFET and BJT semiconductor devices are summarised in Figure 3.1.

3.3 Class-A Amplifiers

A class-A amplifier is one in which the operating point and input signal level are chosen such that the output current (collector or drain current) flows at all times. It therefore operates in the linear portion of its characteristic and hence the signal suffers minimum distortion.

The circuit of an RF class-A amplifier is very similar to that of its small signal counterpart. In practice the transistor will have a complex impedance and hence will require matching, however, the matching circuit and its effects will be neglected in the following analysis and the reader is referred to the literature [4] for further information on matching network design.

The basic configuration of a single-ended class-A stage is shown in Figure 3.2. The radio frequency choke (RFC) ensures the suppression of RF currents which would otherwise be drawn from the supply and the coupling capacitor, C_{out}, prevents the standing bias of the device from being fed to the load.

If the transistor is biased to operate in the center of its linear region and a signal voltage superimposed on this bias level, the collector current may be written as:

$$i_C(t) = I_{CQ} - I_{\max} \sin(\omega t) \tag{3.2}$$

where I_{CQ} is the standing bias drawn from the supply and $I_{\max}$ is the peak value of the sinusoidal RF current produced by the device.

The input current from the supply is equal to the quiescent current above due to the action of the choke, hence:

$$I_{DC} = I_{CQ} \tag{3.3}$$

Similarly, the action of the output capacitor yields an output current of:

$$i_o(t) = I_{\max} \sin(\omega t) \tag{3.4}$$

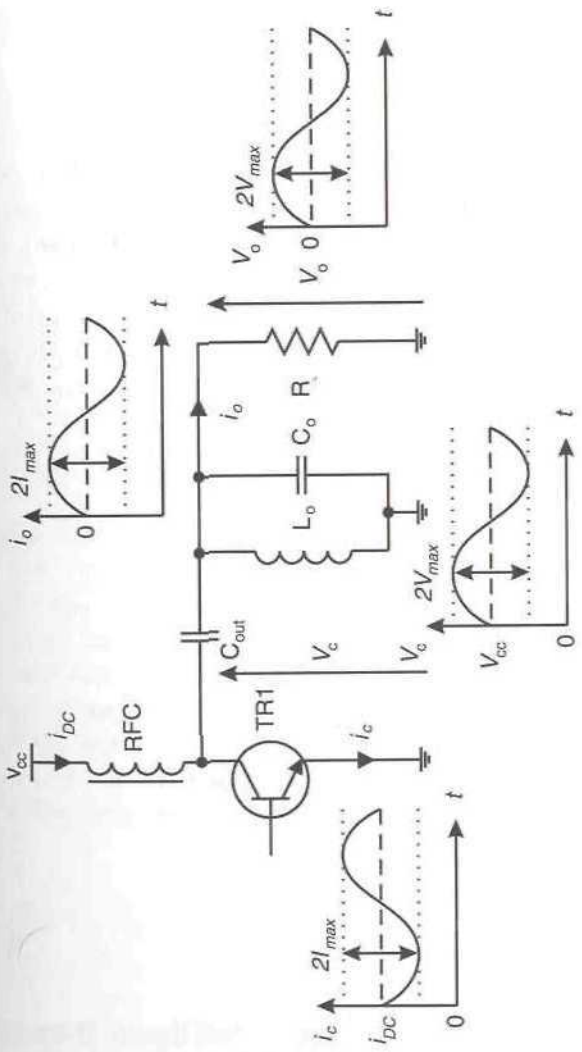


Figure 3.2 Configuration of a single-ended class-A power amplifier showing the various voltage and current waveforms.

If the components are assumed perfect, then the output voltage may be written as:

$$v_o(t) = I_{\max} R_L \sin(\omega t) = V_{\max} \sin(\omega t) \tag{3.5}$$

Finally, the collector voltage consists of a DC component from the supply and an AC component which forms the signal output voltage, $v_o(t)$, hence:

$$v_c(t) = V_{CC} + V_{\max} \sin(\omega t) \tag{3.6}$$

To prevent the device from entering the cut-off region, the collector voltage, $v_c(t)$, must be kept positive. The energy stored in the RFC (inductor) permits the collector voltage to swing higher than the supply in a manner analogous to that of a transformer-coupled class-A audio amplifier. It is thus possible to limit the peak output signal excursion to V_{CC} instead of the usual $V_{CC}/2$ required by an audio class-A amplifier with a resistive collector load. In practice the maximum signal excursion must be less than V_{CC} due to the non-zero saturation voltage of the transistor.

The output current must similarly be positive, giving:

$$I_{DC} = I_{CQ} = \frac{V_{\max}}{R_L} \leq \frac{V_{CC}}{R_L} \tag{3.7}$$

The power from the supply is then:

$$P_S = V_{CC} I_{DC} = \frac{V_{CC}^2}{R_L} \tag{3.8}$$

and the RMS output power supplied to the load is:

$$P_L = \frac{V_{\max}^2}{2R_L} \leq \frac{V_{CC}^2}{2R_L} \tag{3.9}$$

Finally, the overall efficiency is:

$$\eta = \frac{P_L}{P_S} = \frac{V_{\max}^2}{2V_{CC}^2} \leq \frac{1}{2} \tag{3.10}$$

The power dissipated in the active device is the difference between the supply power and the output power; this will be at least 50% of the supply power, in a class-A stage. Careful consideration must therefore be paid to cooling in the design of this type of amplifier.

In a practical class-A amplifier harmonic components will still exist, due to the inherent nonlinearity of the device, and these must usually be removed. This may be done either by careful choice of the output matching network (that is, using a low-pass design) or by the use of a separate output filter. The effect of these harmonic components on the efficiency of the class-A amplifier is considered to be negligible since they will form less than 1/100th of the output power in a typical amplifier.

The effect of the finite saturation voltage of practical devices acts to reduce the efficiency calculated above. The maximum collector voltage swing is reduced to:

$$V_{sw} = V_{CC} - V_{sat} \tag{3.11}$$

Thus, V_{CC} in (3.7) and (3.9) must be replaced by V_{sw} and the overall efficiency is reduced by a factor of V_{sw}/V_{CC}. The maximum efficiency then becomes:

$$\eta_{\max} = \frac{V_{sw}}{2V_{CC}} \tag{3.12}$$

The value of V_{sat} and hence V_{sw} is frequency dependent. At low frequencies ($f < f_T/10$, where f_T is the transition frequency of the device), the device has sufficient time to become fully saturated and V_{sat} will be typically around 0.2V for a silicon device. At higher frequencies, the average saturation voltage will rise and can be as high as a few volts at the maximum operating frequency of the device.

In the case of MOS devices there is an equivalent effect on the saturation, or 'on' resistance. A similar maximum voltage swing exists and is given (for single ended operation) by [1]:

$$V_{sw} = \frac{R_L}{R_L + 2R_{on}} V_{DD} \tag{3.13}$$

3.4 Class-B Amplifiers

A class-B amplifier is one in which the operating point is at one or other extreme of its characteristic, so that the quiescent power is small. The quiescent current or the quiescent voltage of a class-B stage is approximately zero and hence if the excitation is sinusoidal then amplification takes place for only one-half of a cycle.

Class-B operation is significantly more efficient than class-A for use in linear power amplifiers, whilst still providing useful levels of linearity. The standard configuration of a push-pull transformer coupled design is shown in Figure 3.3. This differs from its audio frequency counterpart in the use of two NPN devices, rather than the complementary NPN/PNP configuration usually adopted at lower frequencies. The poor availability of high power PNP devices at higher frequencies has tended to restrict the use of complementary class-B amplifiers to the HF bands.

A single-ended class-B amplifier conducts for 50% of the input cycle (sinusoidal excitation will be assumed) and hence produces significant distortion. The push-pull configuration essentially employs two class-B stages operating in antiphase, with each conducting for a separate half of the input waveform. This antiphase arrangement ensures that whilst one device is conducting, the other is off and consumes no power.

When a device is conducting, it is assumed to be operating in its active region as a perfectly linear current source. The output from each device will be a half sinusoidal current waveform of peak amplitude $I_{\max}$ and the output from both devices will combine in the transformer to produce a sinusoidal output current, $i_o(t)$:

$$i_o(t) = \frac{n_1}{n_2} I_{\max} \sin(\omega t) \tag{3.14}$$

This produces a voltage across the load resistance, R_L, of:

$$v_o(t) = i_o(t) R_L = \frac{n_1}{n_2} I_{\max} R_L \sin(\omega t) = V_{\max} \sin(\omega t) \tag{3.15}$$

The bidirectional action of the transformer means that this load voltage waveform will be transformed back and appear at the collectors of the transistors. Thus:

$$v_{C1}(t) = i_o(t) R_L = V_{CC} + \frac{n_1}{n_2} V_{\max} \sin(\omega t) = V_{CC} + \frac{n_1^2}{n_2^2} I_{\max} R_L \sin(\omega t) \tag{3.16}$$

Examining this more closely reveals two things. First, the peak collector voltage for each of the transistors, $V_{C,\max}$, is given by:

$$V_{C,\max} = \frac{n_1}{n_2} V_{\max} = \frac{n_1^2}{n_2^2} I_{\max} R_L \tag{3.17}$$

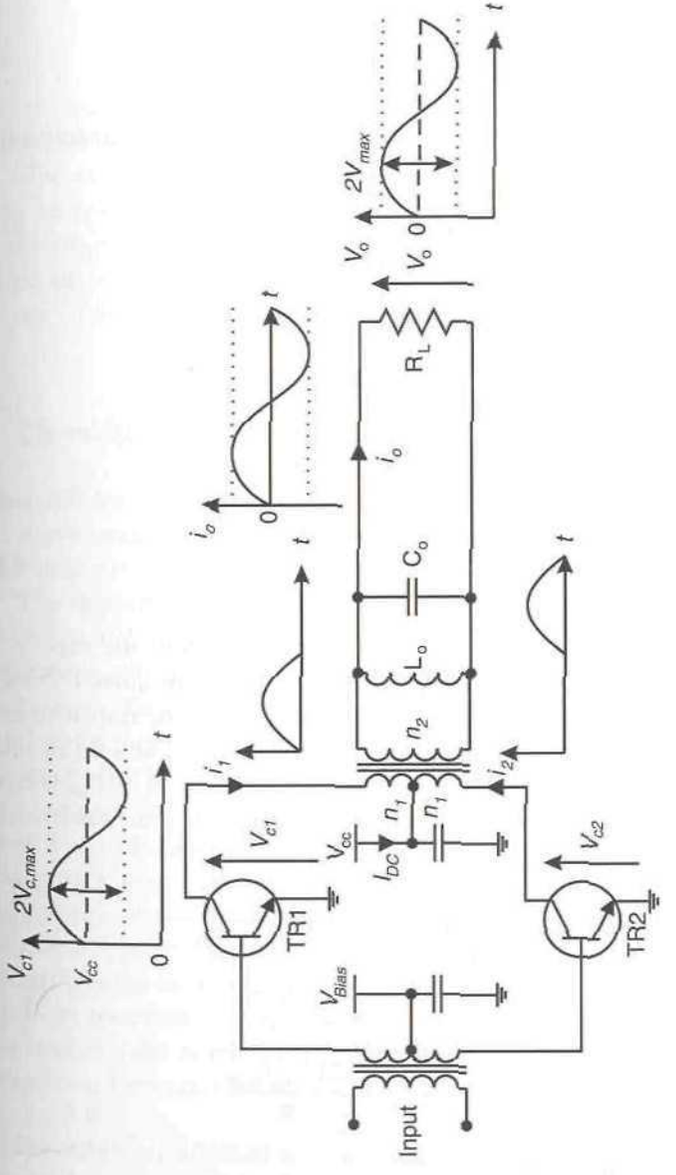


Figure 3.3 Configuration of a push-pull transformer coupled power amplifier, showing the voltage and current waveforms at various points.

and secondly, the transformation of the load resistance, R_L, back to that resistance which is seen by each of the transistors (across one-half of the primary winding), R_p, gives:

$$R_p = \frac{n_1^2}{n_2^2} R_L \tag{3.18}$$

As in the class-A amplifier, the transistor must not go into saturation and therefore $V_{C1} > 0$, hence:

$$\frac{n_1}{n_2} V_{\max} \leq V_{CC} \tag{3.19}$$

The maximum RMS output power is then limited to:

$$P_L = \frac{V_{C,RMS}^2}{R_p} = \frac{n_1^2/n_2^2}{2R_p} V_{\max}^2 \leq \frac{V_{CC}^2}{2R_p} \tag{3.20}$$

Substituting for R_P from (3.18) gives:

$$P_L = \frac{V_{CC}^2}{2(n_1^2/n_2^2)R_L} \tag{3.21}$$

To find the input power, that is, the power drawn from the supply, it is necessary to obtain an expression for the mean value of the current drawn by the transistors. This current is fed to the centre-tap of the transformer and consists of half cycles of current drawn by each device. The sum of these is therefore equivalent to a full-wave rectified sinewave, $i_{1,2}(t) = I_{\max}|\sin(\omega t)|$. Thus the average, or DC component, of the supply current waveform will be:

$$I_{DC} = \frac{1}{2\pi} \int_0^{2\pi} i_{1,2}(t)\, dt = \frac{2I_{\max}}{\pi} \tag{3.22}$$

The supply power is therefore:

$$P_S = V_{CC} I_{DC} = \frac{2V_{CC} I_{\max}}{\pi} \tag{3.23}$$

Finally, the efficiency is then:

$$\eta = \frac{P_L}{P_S} = \frac{V_{CC}^2/2(n_1^2/n_2^2)R_L}{2V_{CC} I_{\max}/\pi} \tag{3.24}$$

Substituting for $V_{CC} = (n_1/n_2)V_{\max}$ from (3.19) and $I_{\max}R_P = (n_1/n_2)V_{\max}$ from (3.17) and (3.18) results in:

$$\eta = \frac{\pi}{4} \tag{3.25}$$

The optimum efficiency, assuming ideal components, is therefore approximately 78.5%.

The total power dissipated in the transistors is $P_S - P_L$, with each taking an equal (half) share.

In practice the idealised efficiency and power dissipation expressions derived above will be degraded by the effects of device saturation and by the presence of residual reactive elements in the load [1].

3.5 Class-AB Amplifiers

A class-AB amplifier is a compromise between the two extremes of class-A and class-B operation. The output signal of this type of amplifier is zero for part, but less than one-half of the input sinusoidal signal.

The distortion added by a class-AB amplifier is consequently greater than that of a class-A stage, but less than that of a class-B stage. Conversely the efficiency will be less than that of a class-B stage and greater than that of a class-A stage. The degree of distortion improvement (or degradation) will depend upon the level of standing bias applied and the consequent level of inefficiency which can be tolerated in a given application. Quite good levels of performance may be achieved with modest levels of standing bias in a push-pull configuration and such designs are consequently popular for higher-power amplifiers at both audio and radio frequencies.

A practical class-B amplifier will exhibit significant crossover distortion due to the imperfect (nonlinear) transition from cut-off to the active mode. This transition is also usually accompanied by an offset voltage, such as is required in overcoming the base-emitter voltage of the devices, and these effects lead to a flat portion being observed in the output waveform during the transition between the active devices. The resulting waveform is shown in Figure 3.4.

The addition of biasing to produce a small quiescent current in the collectors (or drains) of the devices will overcome this problem with only a small loss of efficiency. The improvement in linearity of the amplifier is usually well worth the small sacrifice in efficiency which must be made and is often assumed by many designers when discussing class-B amplifiers. The

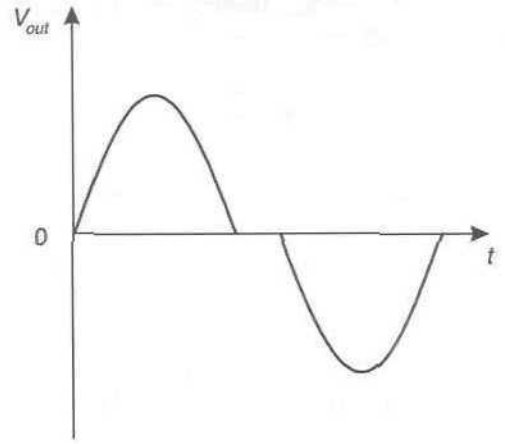


Figure 3.4 Crossover distortion in a class-B amplifier.

addition of the small bias level strictly speaking results in the amplifier being a class-AB design, although this term is often reserved for designs in which a significant amount of bias is applied.

3.6 Class-C Amplifiers

A class-C amplifier is characterised by the operating point being chosen such that the output current (or voltage) is zero for more than one-half of an input sinusoidal signal cycle. This class of amplifier will thus result in significant distortion of the input signal waveshape during the amplification process, thus making it unsuitable for 'linear' amplification applications.

There are two different types of class-C stage: the original or 'classical' class-C PA which was used in valve applications and is now widely used in MOS amplifiers, and the analytically more complex 'class-C mixed mode' configuration utilising a BJT. The former of these will be analysed first, with the principle differences being highlighted later.

3.6.1 Classical Class-C Stage

As in the class-A stage discussed earlier, the original class-C amplifiers involved the active device (valve or FET) operating as a current source driving a load via a tuned filter (to remove harmonics). The configuration and circuit waveforms at various points are shown in Figure 3.5.

The current waveform produced may be modelled in a variety of ways,

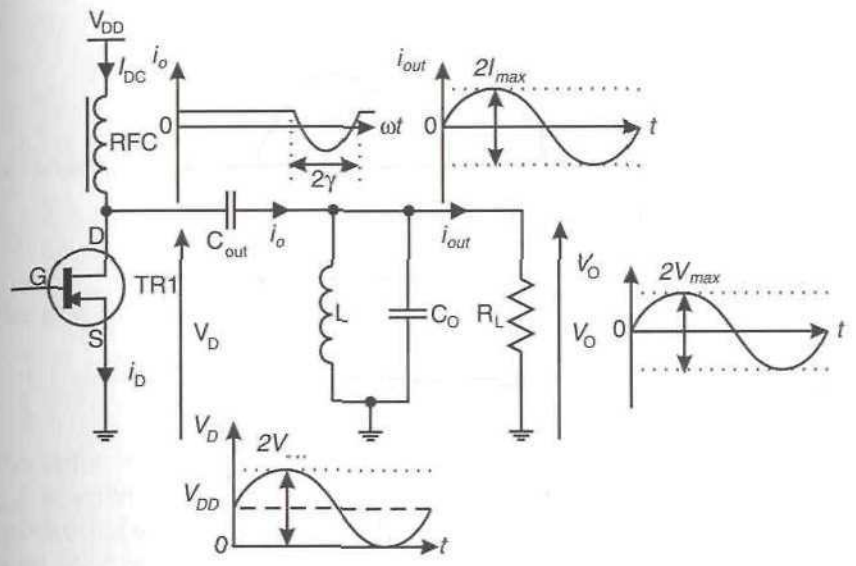


Figure 3.5 Configuration of a 'classical' class-C power amplifier, showing the voltage and current waveforms at various points.

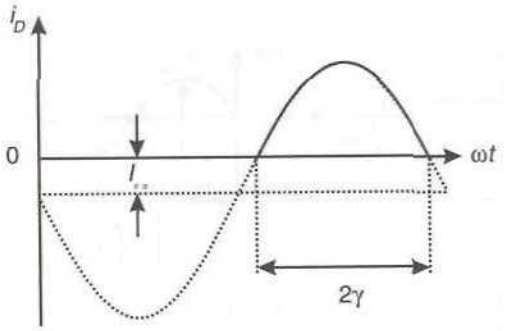
however, the simplest model is that of a biased portion of a sinewave, where the current, I_{DQ} (analogous to the quiescent current in class-A and -B amplifiers), is negative for a class-C PA:

$$i_D(t) = \begin{cases} I_{DQ} - I_{DD}\sin(\omega t) & , \quad I_{DQ} - I_{DD}\sin(\omega t) \geq 0 \\ 0 & , \quad I_{DQ} - I_{DD}\sin(\omega t) \leq 0 \end{cases} \tag{3.26}$$

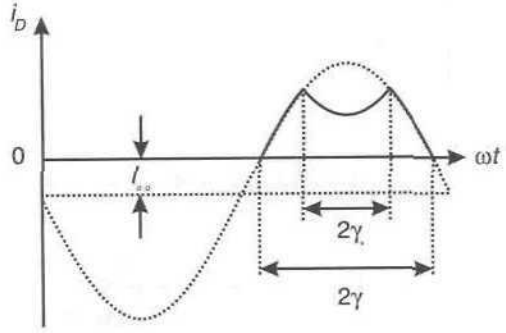
It can thus be envisaged that the greater the (negative) value of I_{DQ}, then the smaller the portion of the input sinewave which is amplified and reaches the output (I_{DD} being positive). This portion of an RF cycle may be specified in terms of a conduction angle γ, where 2γ is the portion of a cycle which the device spends in its active region (see Figure 3.5). The conduction angle is related to the bias and drive currents by:

$$-I_{DQ} = I_{DD}\cos\gamma \tag{3.27}$$

Recalling that the circuit configuration for the class-C stage is identical to that of the class-A and single-ended class-B stages, these may be represented by the above expression when $\gamma = \pi$ (class-A) and $\gamma = \pi/2$ (class-B). Class-C operation is represented by $\gamma < \pi/2$ and the resulting drain current waveform is shown in Figure 3.6.



Normal operation



Saturated operation

Figure 3.6 Drain current waveform for a saturating class-C amplifier.

The average drain current resulting from a given conduction angle is the DC current drawn by the stage from the power supply. This may be found from:

$$I_{DC} = \frac{1}{2\pi}\int_0^{2\pi} i_D(\theta)\, d\theta \tag{3.28}$$

where $\theta = \omega t$. The device is only in its active region (or saturation) during the period specified by the conduction angle, that is, $3\pi/2 + \gamma < \theta < 3\pi/2 - \gamma$. Hence the expression for the DC current may be rewritten as:

$$I_{DC} = \frac{1}{2\pi}\int_{\frac{3\pi}{2}-\gamma}^{\frac{3\pi}{2}+\gamma} i_D(\theta)\, d\theta \tag{3.29}$$

or

$$I_{DC} = \frac{1}{\pi}(\gamma I_{DQ} + I_{DD}\sin\gamma) \tag{3.30}$$

substituting for I_{DQ} from (3.27) gives:

$$I_{DC} = \frac{I_{DD}}{\pi}(\sin\gamma - \gamma\cos\gamma) \tag{3.31}$$

The supply power is then:

$$P_S = V_{DD}I_{DC} = \frac{V_{DD}I_{DD}}{\pi}(\sin\gamma - \gamma\cos\gamma) \tag{3.32}$$

The action of the coupling capacitor C_{out} is to isolate the AC output current, i_{out}, from the DC supply passing through the RFC. This output current will be delivered to the load via the tuned circuit. If this tuned circuit is assumed to be ideal, it will have an infinite impedance to currents at its resonant frequency and a low impedance to all other frequency components. Its action will therefore be to short out all of the harmonic components generated by the stage and to leave the wanted signal component unaffected to pass to the load. This output voltage will be sinusoidal since the harmonic elements have been removed by the filter and will be of the form:

$$V_{out} = V_{\max}\sin(\omega t) \tag{3.33}$$

The output signal voltage, $V_{\max}$, generated across the load is therefore:

$$V_{\max} = -\frac{1}{\pi}\int_0^{2\pi} i_D(\theta)R_L\sin\theta\, d\theta \tag{3.34}$$

where, again, $\theta = \omega t$.

There will only be an output current flowing, and hence an output voltage generated, when the device is operating in its active region. Hence:

$$V_{\max} = -\frac{1}{\pi}\int_{\frac{3\pi}{2}-\gamma}^{\frac{3\pi}{2}+\gamma} i_D(\theta)R_L\sin\theta\, d\theta \tag{3.35}$$

Substituting for $i_D(\theta) = I_{DQ} - I_{DD}\sin\theta$ from (3.26) gives:

$$V_{\max} = \frac{R_L}{2\pi}(4I_{DQ}\sin\gamma + 2I_{DD}\gamma + I_{DD}\sin 2\gamma) \tag{3.36}$$

Finally, substituting for I_{DQ} from (3.27) results in:

$$V_{\max} = \frac{I_{DD} R_L}{2\pi}(2\gamma - \sin 2\gamma) \tag{3.37}$$

The efficiency of a class-C stage is generally highest when the drain voltage swing is equal to the supply voltage, V_{DD}. The efficiency when this is assumed is a function of the conduction angle, 2γ and may be found from:

$$\eta = \frac{P_L}{P_S} \tag{3.38}$$

where:

$$P_L = \frac{V_{\max}^2}{2R_L} = \frac{V_{DD}^2}{2R_L} \tag{3.39}$$

and P_S is given by (3.32). Substituting for $V_{DD} = V_{\max}$ and for $V_{\max}$ from (3.37) gives:

$$\eta = \frac{2\gamma - \sin 2\gamma}{4(\sin\gamma - \gamma\cos\gamma)} \tag{3.40}$$

The relationship between the output signal level, $V_{\max}$, and the drive current I_{DD} (from (3.37)) is nonlinear since the conduction angle, 2γ, is a function of the drive current. Two exceptions to this occur when $\gamma = \pi$ (class-A) and $\gamma = \pi/2$ (class-B). Between these two values, operation is in class-AB (Section 3.5) and for $\gamma < \pi/2$, operation is in class-C. Note that it is consequently possible to operate a single-ended class-B amplifier as a linear stage provided that an L-C tank circuit is used as an output filter in the manner shown in Figure 3.5.

The characteristic of efficiency versus conduction angle is shown in Figure 3.7. From this it is evident that an ideal efficiency of 100% may be obtained by decreasing the conduction angle to a small value. Unfortunately this efficiency is obtained at the expense of requiring somewhat idealised device ratings and a compromise must be sought. From (3.37) it is evident that as $\gamma \to 0$, $V_{\max} \to 0$ for a fixed I_{DD}. Thus, to obtain the same power output, the peak value of the drain current must be increased and may exceed the device rating. The design of a class-C stage must therefore be a compromise between efficiency and device ratings.

As in the case of class-A and -B amplifiers, practical considerations

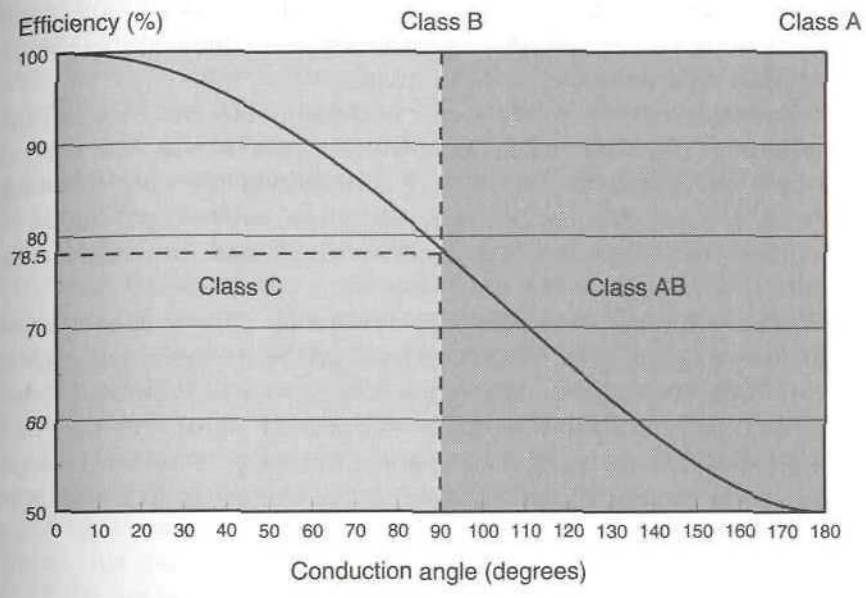


Figure 3.7 Efficiency as a function of conduction angle (γ) for a class-C amplifier.

limit the efficiencies which may be achieved from a class-C stage. The saturation 'on' resistance R_{on} of the FET will limit the maximum voltage swing to V_{sw} as given in (3.13). The practical efficiency may then be obtained by using V_{sw} in place of V_{DD} in all equations except that for input power (from (3.32)).

3.6.2 Saturation of Class-C Stages

Since the operation of a class-C stage is highly nonlinear, it is possible to drive it into saturation without loss of performance. Indeed, the conventional method of driving a class-C stage is to ensure that it achieves saturated operation as this has a number of advantages.

When the active device (an FET) is saturated, the drain current is given by:

$$i_D(t) = \frac{V_{DD} + V_{dm}\sin(\omega t)}{R_{on}} \tag{3.41}$$

where V_{dm} is the drain voltage swing. The driving current and bias current determine the conduction angle, 2γ as previously, however, a saturation angle, γ_s is now introduced which occurs when the voltage drop caused by

the drain current flowing through the drain resistance equals the drain voltage. The tuned output filter is again assumed to remove the harmonic currents, leaving a sinusoidal output current.

Numerical analysis is required to calculate the efficiency of the saturating stage [1], however, the peak efficiency occurs when the stage is just driven into saturation. Thereafter it decreases slightly with increasing drive level. The peak efficiency of the saturating stage is greater than that of the non-saturating stage and this is one of the reasons behind its wide adoption.

The second main advantage of the saturating stage is in its ease of amplitude modulation. The RF voltage produced by the saturating stage is primarily dependent on the drain supply voltage, V_{DD}. Variation of this supply voltage in sympathy with the information signal will therefore amplitude modulate the stage. This important property is used in *Envelope Elimination and Restoration* (EE&R) systems and will be covered in greater detail in Section 3.6.4, and again in Chapter 7.

3.6.3 Class-C Mixed Mode Stage

The use of a bipolar junction transistor as the active device in a 'class-C' stage is considerably different to the situation described above for an FET. Both the circuit configuration (Figure 3.8) and its method of operation bear only a passing resemblance to those described above.

A BJT is more difficult to bias in class-C than an FET as it is less

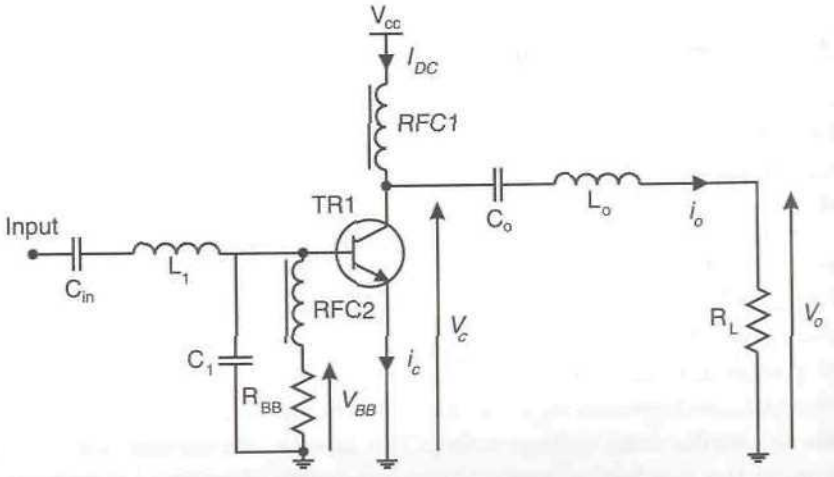


Figure 3.8 Configuration of a class-C mixed-mode amplifier.

straightforward to simply add a negative bias voltage to the sinusoidal input signal. This problem leads to a subsequent difficulty in controlling the duty cycle of operation and hence the collector waveform. The input circuit shown in Figure 3.8 contains an RFC and resistance to ground which provide the bias. The resistance is often unnecessary in practice as the internal base-spreading resistance, r_{bb}, tends to reverse bias the base-emitter junction. This situation is known as *self bias*.

A BJT will have a considerably lower saturation resistance than an FET and also contains a high amount of varactor capacitance between collector and emitter. The combination of these effects leads to a non-sinusoidal collector voltage waveform. The low saturation resistance dominates the parallel output filter and results in a flattening of the collector waveform, whilst the varactor capacitance in parallel with the inductance of the tuned filter does not act as a simple resonant circuit. Harmonics are thus produced in the collector voltage waveform despite the use of a sinusoidal input signal and even when it is of a level such that the transistor is not driven into saturation.

The use of a series tuned circuit in the mixed-mode stage allows the presence of a non-sinusoidal collector voltage waveform (the harmonics are not shorted to ground). The action of the series filter is merely to remove these harmonic components from the final output and prevent them from being delivered to the load.

Analysis of the mixed-mode stage is complex as many of the waveforms appearing throughout the circuit are highly non-sinusoidal. The transistor is required to operate in cut-off, the active region, and saturation by the application of the input drive current. An analytic design is therefore very difficult from first principles and the large-signal impedances usually specified by transistor manufacturers must be used instead for matching purposes. These values must be used with care as they are only valid under the conditions (e.g., carrier frequency, drive level) at which they were measured. Alternative use of the device should therefore be avoided and an appropriate device for the intended application selected instead.

3.6.4 Amplitude Modulation of a Class-C Stage

The use of class-C stages in AM transmitters was briefly discussed in Section 3.6.2. The basic principle is that of introducing modulation, in the form of a variation in the supply voltage, to a saturating class-C stage. The envelope of the resulting RF output will then vary in sympathy with the message signal, resulting in full carrier amplitude modulation.

A convenient method of superimposing the amplitude variations on

the supply voltage is to use a transformer. The secondary of the transformer is connected in series with the supply voltage, whilst the primary is fed from an audio power amplifier. A typical circuit configuration is illustrated in *Figure 3.9 for a BJT stage.*

For 100% modulation to be possible, the collector voltage swing must be capable of varying between 0V and $+2V_{CC}$. The drive capability of the previous (driver) stage must also be large enough to supply the required power for this level of modulation. This requires the drive level to be great enough to maintain saturation and cut-off at the highest effective supply voltage ($2V_{CC}$). If these conditions are met, the collector voltage waveform will be as shown in Figure 3.9, with the transistor being driven between cut-off and saturation. Under these conditions, the collector voltage waveform

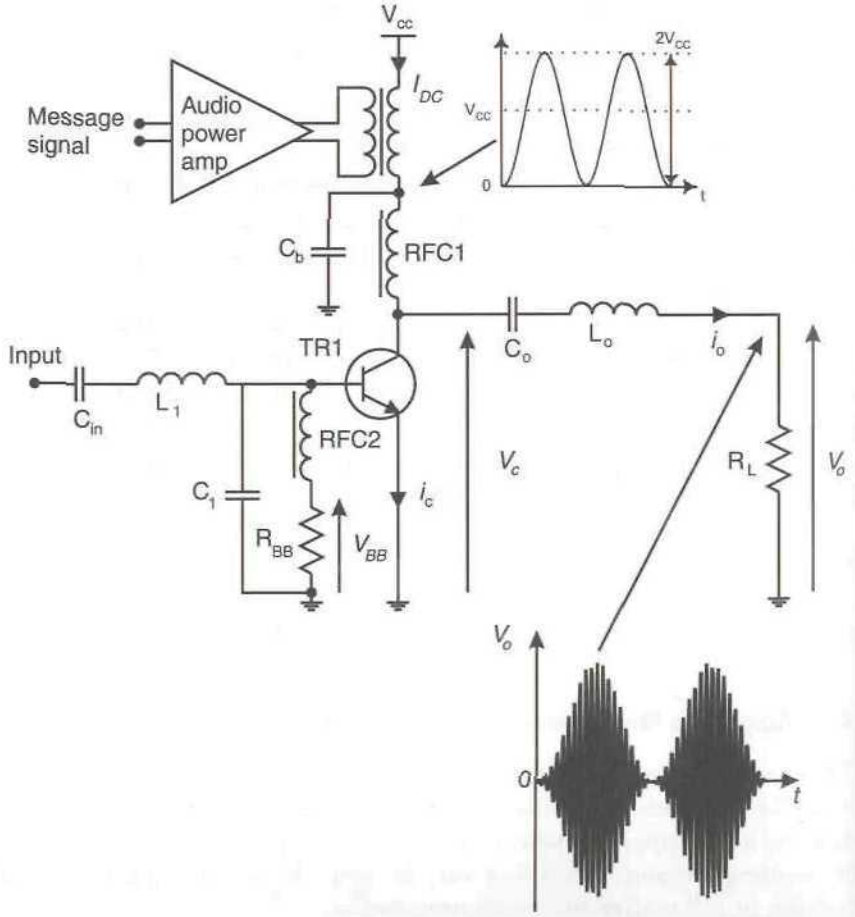


Figure 3.9 Amplitude modulation applied to a class-C PA.

may peak at four times the power supply voltage because of the collector inductor. Care should be taken to ensure that this peak voltage does not exceed the collector-emitter junction breakdown voltage of the RF device.

The capacitor C_b is a bypass capacitor for the carrier frequency. Its presence ensures that the modulation supply voltage appears to come from a very low impedance source and hence it eliminates any effects on the transistor's operation from inductive and capacitive effects in the transformer windings. The function of the output filter has also changed. It now has the additional tasks of removing the harmonic components of the modulation sidebands and the original modulating frequency, as well as its original task of removing the carrier harmonics and DC.

The audio power amplifier, which is used to apply the modulation to the class-C stage, must be capable of supplying half of the DC input power to that stage. It is thus required to be capable of a significant power output in its own right and must also be linear to preserve the fidelity of the message signal. In a high power AM transmitter, therefore, a significant amount of power may be consumed in this stage.

There are a number of problems associated with this type of AM system, most of which result in an increase in the distortion of the message *signal. A typical output waveform from a practical modulated stage is shown* in Figure 3.10. If a sufficient drive level is provided to maintain saturation and cut-off at the peaks, then too much will be available at the troughs and this will prevent the output from falling to zero as is required for 100% modulation. If this drive level is lowered, then a compromise may be reached in which the peaks and troughs are 'missed' by an equivalent amount. This is the situation illustrated in Figure 3.10.

The high drive level has a further disadvantage. The excessive drive causes feedthrough in the troughs of the modulating waveform and prevents it falling to zero. The feedthrough signal is coupled through a reactance which is generally larger than the collector load resistance and this causes the feedthrough signal to be in phase-quadrature with the amplified output signal. The result of this process is an AM to PM conversion (phase modulation of the carrier) which results in additional (unwanted) sidebands, causing the signal to spread into adjacent channels. This signal spreading will, in turn, cause interference to the adjacent channels, an effect which is obviously *undesirable.*

This situation may be improved by partially modulating the driver stage in an attempt to improve the depth of modulation. The drive level will thus more closely resemble that required, in an ideal case, by the final stage. The linearity (fidelity) of the modulation may also be improved by the addition of feedback as shown in Figure 3.11. The transmitter output signal

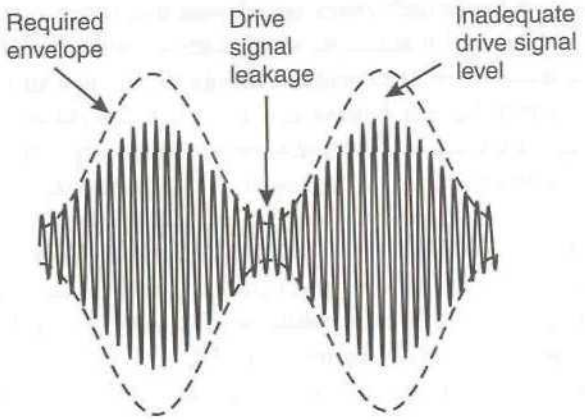


Figure 3.10 Modulated output from a practical AM transmitter.

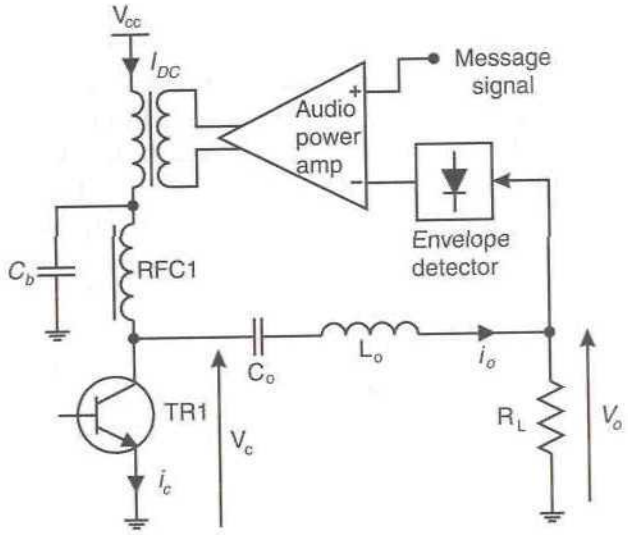


Figure 3.11 Modulation feedback used to improve the linearity of an AM stage.

is demodulated and the resulting audio signal compared with the original modulation. The audio drive signal is then corrected to minimise the difference between the modulation appearing at the output and its intended form. This simple feedback system forms the basis of a number of the amplifier linearisation techniques to be discussed later.

3.7 Class-D Amplifiers

An optimally efficient amplifier would consist of a controlled switch, in which the 'on'-resistance was zero, the 'off'-resistance was infinite and the transition time was zero. The output signal would then consist of the power supply switched at the rate of the input (carrier) signal, with no losses in the switching device. This, somewhat idealised, situation is illustrated in Figure 3.12.

The efficiency of this system would be 100% since there are no losses within the device either due to its resistance or due to a finite switching time. The idea of a perfect RF switch leads to the concept of the class-D amplifier in which this ideal is simulated as far as is possible with practical devices.

It may seem odd to consider such a highly nonlinear class of amplifier in a book on high-linearity amplification, however, such classes of amplifier do have an increasingly important role in the field of linear transmitter design. This role will be detailed in Chapter 7, as they are used in the 'RF synthesis' methods of generating high-power linear signals. Such methods have the potential to realise the highest possible efficiency in a linear power amplifier or transmitter.

There are three main configurations for the class-D stage (the complementary switch and voltage or current transformer coupled designs) and these are based on the class-B designs described earlier. The complementary voltage switching configuration is the most straightforward of these and will be described first.

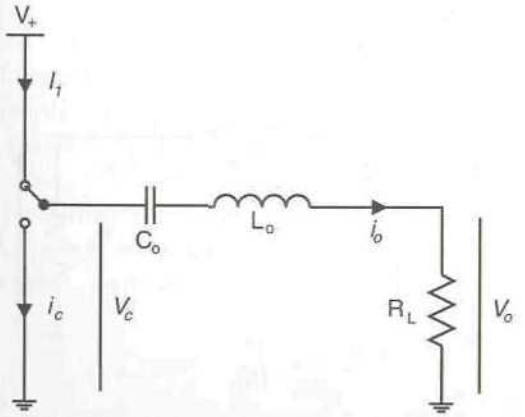


Figure 3.12 Idealised switch used as an amplifier.

3.7.1 Class-D Complementary Voltage Switching Amplifier

A simplified schematic for this design, together with its equivalent circuit, is shown in Figure 3.13. The RF input signal is fed to a splitter transformer which causes TR1 and TR2 to be driven with currents which are 180° out of phase. Thus when TR1 is turned on, TR2 is turned off and vice versa. The result of this arrangement is equivalent to a changeover switch alternating between the supply and ground. The voltage across TR2 is therefore a squarewave alternating between 0V and V_{CC}, ideally with a 50% duty cycle.

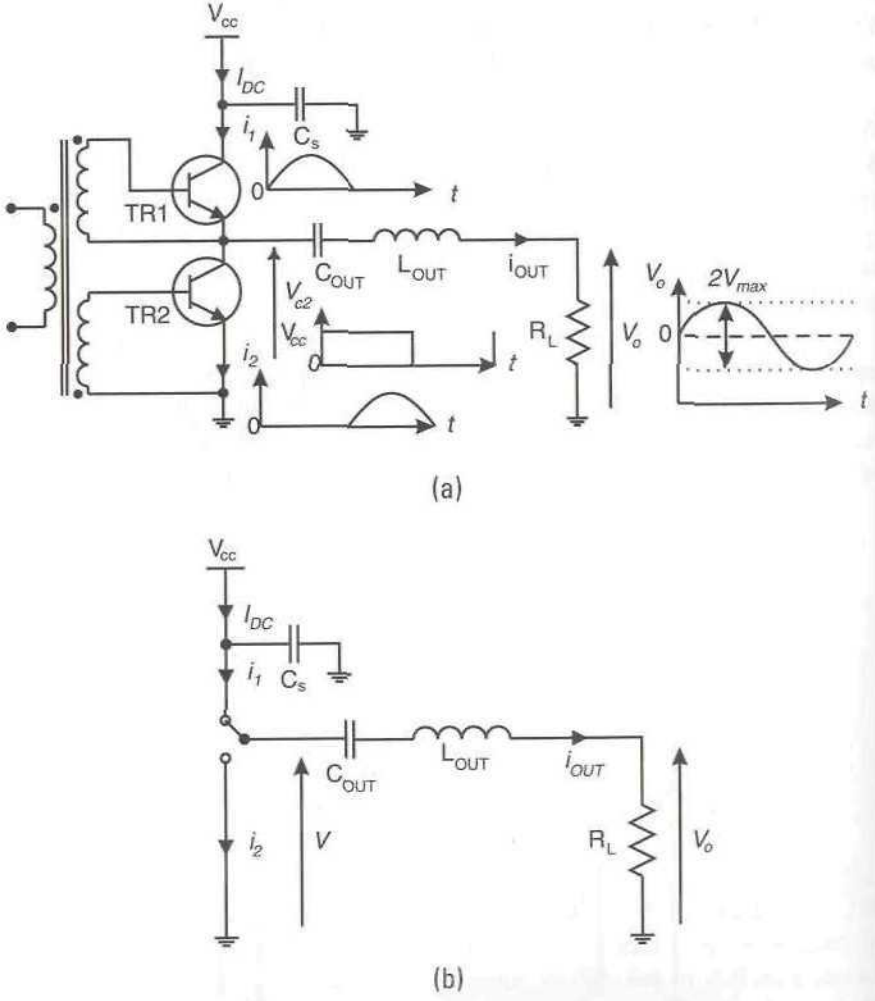


Figure 3.13 Class-D complementary voltage switching amplifier (a) and equivalent circuit (b).

An alternative configuration to that shown in Figure 3.13 would be to use both NPN and PNP devices in a manner akin to that of audio frequency class-B power amplifiers. This is uncommon in RF stages due to the lack of suitable high frequency, high performance PNP devices.

The analysis of a class-D amplifier is very straightforward due to the simple collector voltage waveshape. The voltage across TR2 may be written as:

$$V_{C2} = V_{CC}\left[\frac{1}{2} + \frac{1}{2}b(\omega t)\right] \tag{3.42}$$

where $b(\omega t)$ is given by:

$$b(\omega t) = \begin{cases} +1 & \sin(\omega t) \geq 0 \\ -1 & \sin(\omega t) \leq 0 \end{cases} \tag{3.43}$$

The resulting squarewave collector voltage waveform may be expanded using Fourier analysis from:

$$b(\omega t) = \frac{4}{\pi}\left\{\sin(\omega t) + \frac{1}{3}\sin(3\omega t) + \frac{1}{5}\sin(5\omega t) + \cdots\right\} \tag{3.44}$$

giving

$$V_{C2}(\omega t) = V_{CC}\left\{\frac{1}{2} + \frac{2}{\pi}\sin(\omega t) + \frac{2}{3\pi}\sin(3\omega t) + \frac{2}{5\pi}\sin(5\omega t) + \cdots\right\} \tag{3.45}$$

The output tuned network consisting of L_{out} and C_{out} will be accurately tuned to the fundamental frequency of the squarewave collector voltage. It will thus have a negligible impedance at that frequency and a high impedance to the harmonic frequencies. The action of the capacitor will also remove the DC component from the output waveform and it will thus appear as a sinusoid at the fundamental frequency of the squarewave with a zero DC level. Since the output is assumed to consist exclusively of the fundamental frequency component, we have:

$$i_{out}(\omega t) = \frac{2V_{CC}\sin(\omega t)}{\pi R_L} \tag{3.46}$$

The positive half of this sinusoidal current will be supplied by TR1 and the

negative half by TR2, with the peak value of the resulting complete sinusoid being:

$$\hat{i}_{out} = \frac{2V_{CC}}{\pi R_L} \tag{3.47}$$

The peak output voltage is therefore:

$$V_{\max} = \frac{2V_{CC}}{\pi} \tag{3.48}$$

and hence the output power is:

$$P_{out} = \frac{V_{\max}^2}{2R_L} = \frac{2V_{CC}^2}{\pi^2 R_L} \tag{3.49}$$

The DC input current can be found from (3.47) since the output current flows from the supply through TR1:

$$I_{DC} = \frac{\hat{i}_{out}}{\pi} = \frac{2V_{CC}}{\pi^2 R_L} \tag{3.50}$$

Finally, the power drawn from the supply is:

$$P_S = V_{CC} I_{DC} = \frac{2V_{CC}^2}{\pi^2 R_L} \tag{3.51}$$

The theoretical efficiency is therefore:

$$\eta = \frac{P_{out}}{P_S} = 1 \tag{3.52}$$

that is, 100%. This result is perhaps obvious as the active devices are operating (theoretically at least) as perfect switches and hence have no voltage across them when conducting and no current through them when off. They therefore dissipate no power and hence all must be supplied to the load.

3.7.2 Class-D Transformer-Coupled Voltage Switching Amplifier

This is the first of two transformer-coupled class-D configurations both of which are similar in structure to the transformer-coupled class-B design described in Section 3.4. The circuit configuration, together with typical waveforms, is shown in Figure 3.14.

As in the complementary class-D amplifier, the input transformer results in the transistors turning on and off alternately. The supply voltage, V_{CC}, is thus placed across alternate halves of the primary

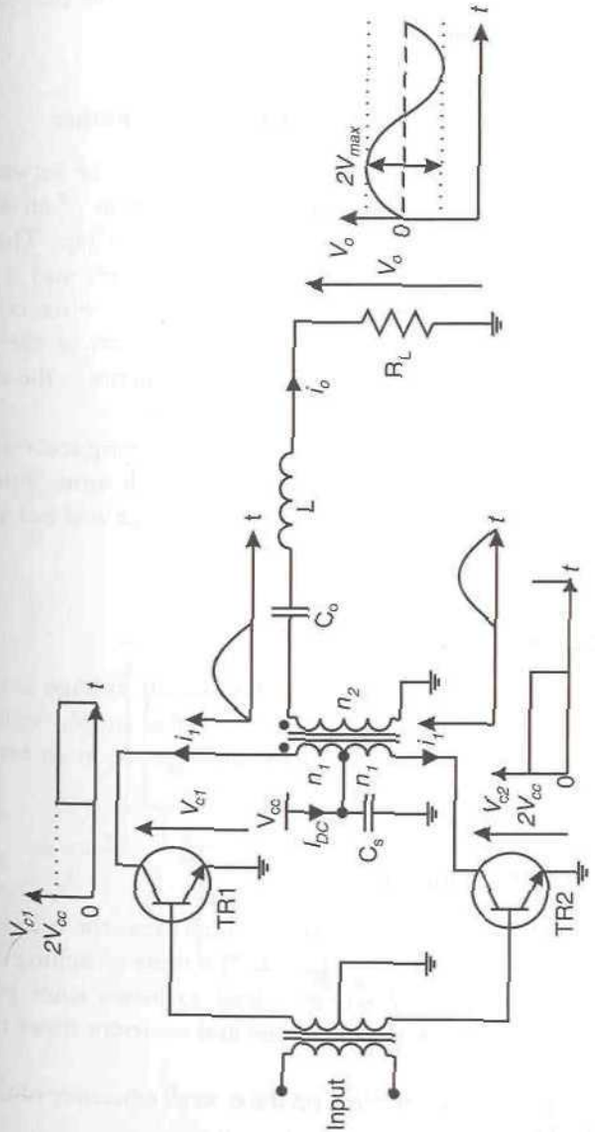


Figure 3.14 Class-D transformer-coupled voltage switching amplifier and circuit waveforms.

winding, resulting in positive and negative transformed versions of V_{CC} across the secondary. The resulting secondary waveform is again a squarewave and the collector voltages are also squarewaves switching between 0V and $+2V_{CC}$.

The efficiency may again be shown to be 100%, with the output filter (L_{out}, C_{out}) performing the same function.

3.7.3 Class-D Transformer-Coupled Current Switching Amplifier

Referring to Figure 3.15 it can be seen that the only difference between the current and voltage switching configurations is the addition of an RFC in series with the supply line feeding the transformer center-tap. The RFC forces the circuit to draw a constant current from the supply and it is this current which is switched by the active devices. Thus it is now the collector current waveform which is squarewave in nature and each of the corresponding circuit waveforms are the duals of their equivalents in the voltage switching configuration.

The principle advantage of the current switching configuration is that of device switching speed. The active devices will switch more quickly in this configuration if they are operating as current switches and not voltage switches as in the previous configuration.

3.7.4 Amplitude Modulation of a Class-D PA

In an ideal class-D PA, the amplitude of the RF output voltage is directly proportional to the supply voltage. Variation of the supply voltage in response to an applied message signal will therefore result in an extremely linear amplitude modulation of the RF carrier.

3.7.5 Practical Class-D Amplifiers

There are a number of practical considerations which cause the behavior of a class-D amplifier to differ from that anticipated. The finite switching times of the devices, together with stray reactances, lead to losses since power is consumed whilst a finite voltage appears across and a current flows through the devices.

The effect of a finite transition time on the overall efficiency of the stage can be shown to be [1]:

$$\eta_{real} = \eta_{ideal} \frac{\sin \phi_s}{\phi_s} \tag{3.53}$$

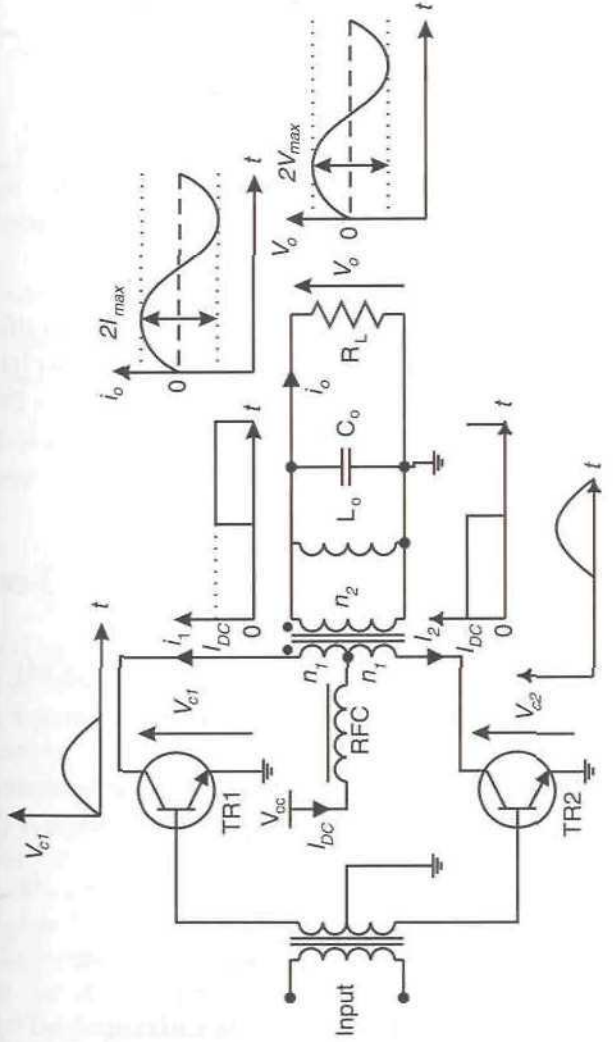


Figure 3.15 Class-D transformer-coupled current switching amplifier and circuit waveforms.

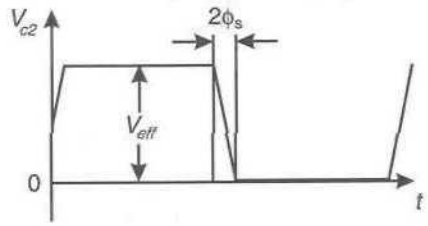


Figure 3.16 Finite switching time of the devices in a class-D PA.

where ϕ_s is the angular portion of the RF cycle taken for a device to switch as shown in Figure 3.16. This expression assumes the somewhat idealised case of a linear transition between states.

As has been mentioned previously, the devices will exhibit a non-ideal behavior when saturated and this will again reduce the overall efficiency. In the case of a BJT, the supply voltage, V_{CC}, should be replaced by an effective voltage, V_{eff}, in all expressions except that for input power. The overall efficiency is then reduced from the ideal of 100% to:

$$\eta_{real} = \frac{V_{eff}}{V_{CC}} \tag{3.54}$$

where:

$$V_{eff} = V_{CC} - V_{sat} \quad \text{(transformer coupled)} \tag{3.55}$$

$$V_{eff} = V_{CC} - 2V_{sat} \quad \text{(complementary)} \tag{3.56}$$

The 'on'-resistance of an FET may be dealt with in a similar manner to that of the class-B amplifier discussed earlier. The effective voltage, V_{eff}, is given by:

$$V_{eff} = V_{DD}\frac{R_L}{R_{on} + R_L} \tag{3.57}$$

Protection of the active devices is required when driving a load which has a reactive element at the fundamental frequency. It is therefore a wise precaution in most practical designs involving bipolar devices. Device protection is required in order to provide a path for the reverse currents which result from a phase shift of the output current relative to the collector current caused by the stray reactance.

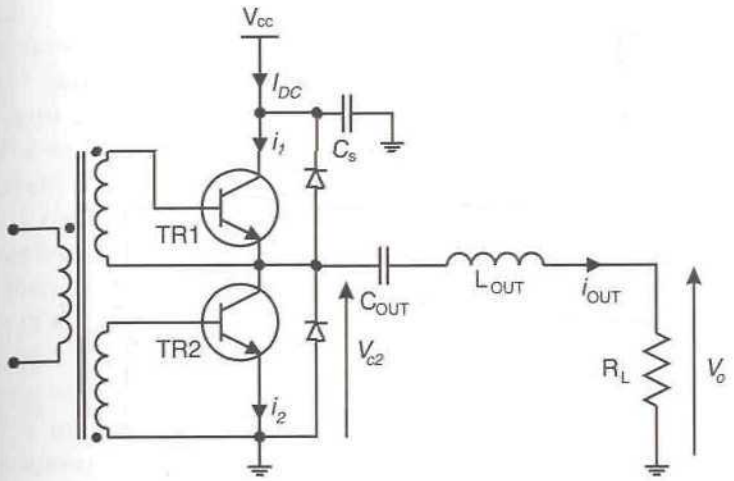


Figure 3.17 Diode protection applied to a complementary class-D amplifier with a reactive load.

A reverse direction path is provided by a diode connected across each of the active devices as shown in Figure 3.17.

3.8 Class-E Amplifiers

The class-E amplifier is a single-ended switching configuration with a passive load network as shown in Figure 3.18. The simplest form of load network consists of a series tuned L-C circuit and a shunt capacitance across the collector-emitter junction (the emitter being grounded). One of the principal advantages of this configuration is immediately apparent in that this shunt capacitance may be composed of the inherent junction capacitance of the transistor plus some additional capacitance to ensure correct circuit operation. Thus the transistor's inherent capacitance is no longer a source of power loss but becomes an essential part of the circuit's operation.

In the operation of the class-E stage, the transistor is assumed to operate as an ideal switch with zero 'on'-resistance and infinite 'off'-resistance. The equivalent circuit can then be drawn as in Figure 3.19, with the two shunt capacitances combined to form a single equivalent capacitance, C.

The derivation of the voltage and current waveforms of a class-E PA is complex since neither waveform may be simply determined for the entire RF cycle. Equations have been presented in the literature [5,6] to describe the

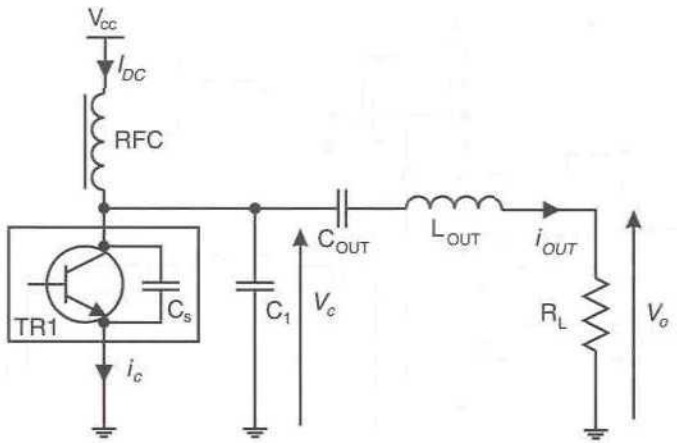


Figure 3.18 Class-E amplifier circuit configuration.

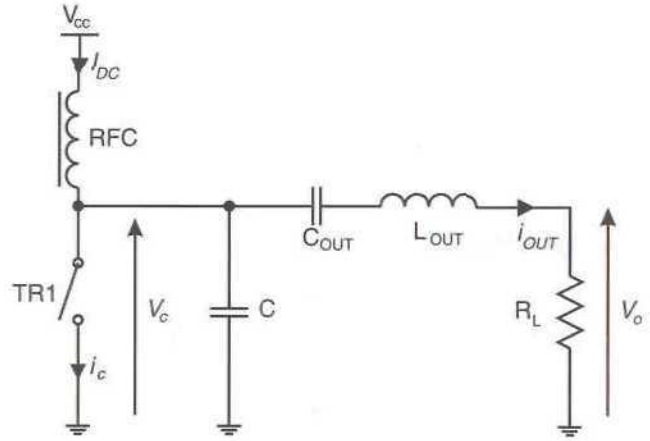


Figure 3.19 Equivalent circuit of a class-E PA.

operation of the class-E amplifier and the reader is directed to these for further details. The idealised efficiency of a class-E stage is 100% although practical considerations of circuit Q, saturation voltage and switching time reduce this in practice.

3.9 Class-F Amplifiers

A class-F amplifier has a load network which is resonant at one or more harmonic frequencies in addition to its resonance at the fundamental

frequency. The addition of harmonics, at the correct level, to the fundamental causes a flattening of the collector voltage waveform. The collector voltage waveform thus begins to approximate a squarewave with a consequent improvement in both efficiency and output power [7].

The configuration of a class-F stage in which only the third harmonic is added is shown in Figure 3.20. The transistor acts as a current source and the resulting collector current waveform is a half-sinewave as was the case in a single-ended class-B configuration. The action of the fundamental-frequency tuned circuit composed of L_1 and C_1 is to produce a sinusoidal output voltage. The third-harmonic resonator (L_3, C_3) permits the collector voltage waveform to have a third harmonic component, which in turn results in the flattening of the collector voltage waveform described above.

The amplitude and phase relationship of the third harmonic relative to the fundamental component is important to ensure that the correct waveshape is produced at the collector. This is usually accomplished by the correct tuning of both tuned circuits and by utilising some device saturation.

An improved design replaces the harmonic resonant circuit with a quarter-wave transmission-line, which simulates the equivalent of an infinite number of resonators. The parallel-tuned output circuit is a short to ground for all of the harmonic frequencies, however, the quarter-wave line transforms this into short circuits at each of the even harmonic frequencies and open circuits at each of the odd harmonics. The resulting fundamental and sum of odd harmonics leads to a squarewave collector voltage waveform and hence to the device operating as a switch. The efficiency of the stage is thus further improved.

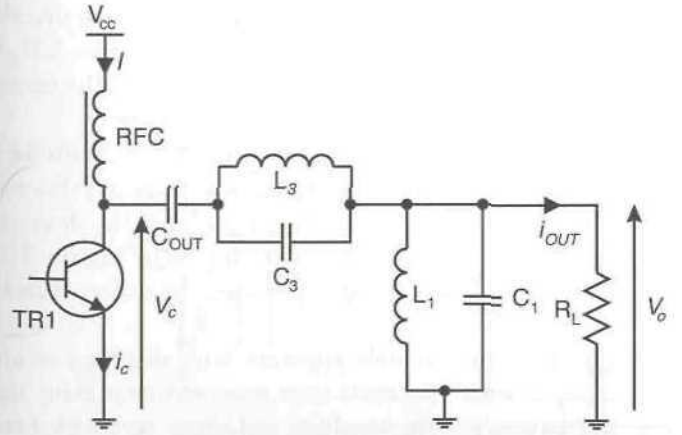


Figure 3.20 Configuration of a class-F power amplifier.

3.10 Class-G and -H Amplifiers

Both of these amplifier classes are principally limited to audio frequency applications due to the wide bandwidths required for the active components.

Class-G amplification requires the use of more than one supply voltage and at least two pairs of active devices (for two different supply voltages). The first pair of devices is connected to the lower supply voltage (which is, say, half of the higher supply voltage) and operate as current sources when the output is of a low level. As the output level increases, these devices are cut off and the devices connected to the higher voltage supply become active and operate as current sources. The efficiency improvement comes when the lower level signals are being amplified as the efficiency at the lower supply voltage will be twice that at the higher supply. The result is an overall improvement in efficiency for signals which do not continuously require the higher voltage supply.

The class-H configuration consists of a highly efficient amplifier supplying the collector voltage level of the main amplifying stage in order to ensure that it remains just above that of the output signal. The result of this is that the collector dissipation in the current-source configured main amplifier is greatly reduced and the efficiency may be made close to 100%. *This type of system is covered in more detail in Chapter 8.*

3.11 Class-S Amplifiers

Pulse-width modulation (PWM) is used in many applications requiring high efficiency energy conversion or control. The basic principle involves the creation of a rectangular waveform of variable duty-cycle such that the mean level varies to form the desired waveshape, as shown in Figure 3.21, where a triangular waveform is compared with the input waveform, the result of this comparison forming the PWM waveform.

Class-S amplification utilises a PWM input waveform, with the desired output waveshape contained in its mean value, and amplifies this switching waveform efficiently before low-pass filtering it to leave the desired output waveform. A typical circuit configuration is shown in Figure 3.22, with reverse-current protection diodes inserted between the collector and emitter of each device.

Two difficulties are immediately apparent with this type of amplifier. Firstly, the frequency at which the transistors must switch is many times that of the final output frequency of the amplifier and hence very wide bandwidth devices are required. This currently limits such techniques to low RF

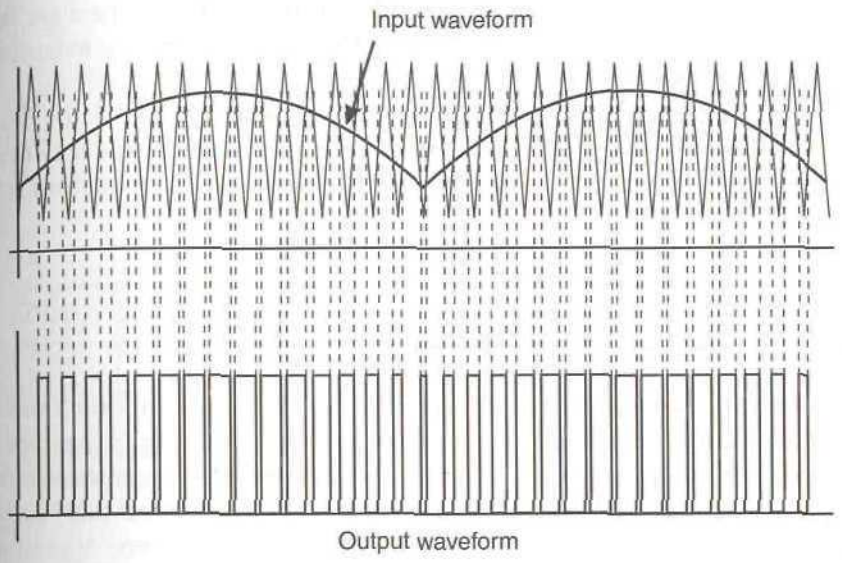


Figure 3.21 Creation of a PWM waveform.

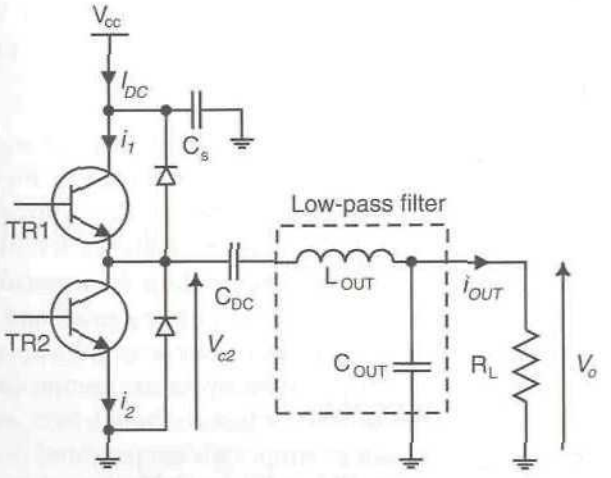


Figure 3.22 Class-S amplifier configuration.

frequencies at best. Secondly, TR1 is a PNP device, and since there are few radio frequency PNP devices available, this again limits the maximum frequency of operation of these amplifiers.

A single-ended Class-S stage may be utilised as a modulator since operation here is only required at audio frequencies. Such a stage may form the basis of the modulator in an Envelope Elimination and Restoration type transmitter.

3.12 Biasing for Linear Operation

The use of class-A and class-AB amplifiers for linear power amplification relies on the use of a standing bias current, applied to the base (in the case of a bipolar transistor) in order to bias the RF device into partial or full conduction. This bias current must remain constant, despite the varying envelope of the input signal to the amplifier, which will cause significant variations in the collector current (and hence base current) required by the device.

When considering low power RF amplification, any of the conventional resistive biasing techniques may be used (e.g., potentiometer biasing), however, such techniques are not appropriate for medium- and high-power amplifiers due to their requirement for a low-impedance, relatively high-current, voltage source. A number of more appropriate techniques are described in this section.

3.12.1 Diode Biasing

Perhaps the simplest technique for RF power amplifier biasing is the use of a clamping diode [8–11] and a typical circuit is shown in Figure 3.23. The clamping diode provides a low-impedance voltage source for the bias voltage of the transistor and hence requires a higher standing current to flow through it than the bias current required by the transistor. This diode must therefore be an appropriately-rated power device and must also be physically connected to the power transistor or heat sink in order to provide temperature compensation. For perfect temperature compensation, both the clamping diode and the RF power transistor should have similar thermal characteristics (changes in voltage drop with temperature).

The base-emitter resistor (R1 in Figure 3.23) is used to slightly lower the actual base voltage below the forward voltage of the clamping diode, D1. The bias current (diode forward current) is determined by the DC resistance

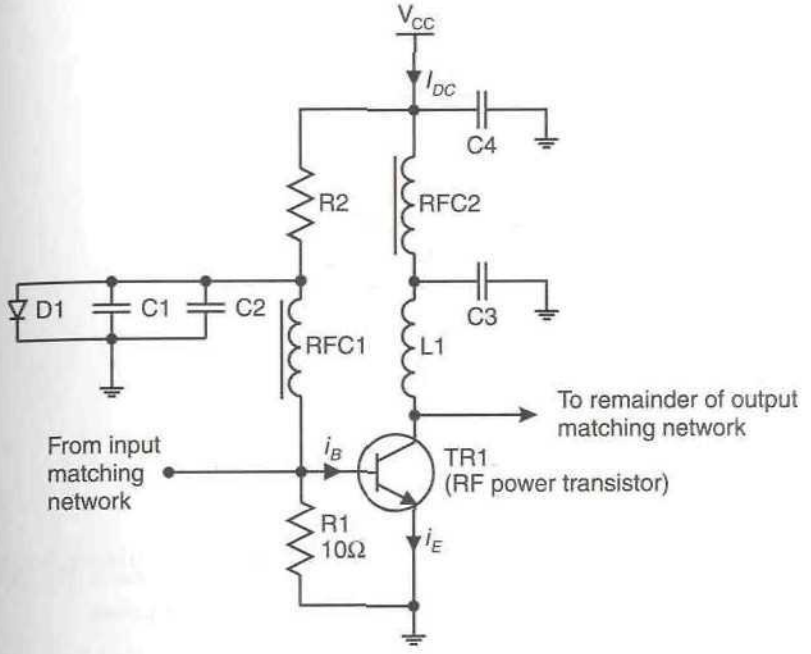


Figure 3.23 Clamping diode bias scheme for an RF power transistor.

of the RF choke, RFC1, and the value of R2. This dropping resistor is required to dissipate quite a high power and hence must be adequately rated. This power is also wasted and hence further reduces the efficiency of the stage. The power dissipated in this resistor will be:

$$P_{R2} = (V_{CC} - V_b)I_{b,\max} \tag{3.58}$$

3.12.2 Amplified Diode Biasing

The clamping diode bias scheme of Figure 3.23 can be improved [8,11] by the addition of an emitter follower to amplify the diode current, and hence reduce the required current through the diode by a factor equal to the current gain of the transistor employed. This arrangement employs two diodes in series, since one is required to compensate for the base-emitter voltage drop of the biasing transistor, TR1. The circuit arrangement for this scheme is shown in Figure 3.24.

Since the required diode current is now relatively small (in most applications), signal diodes may be employed and a suitable AF power

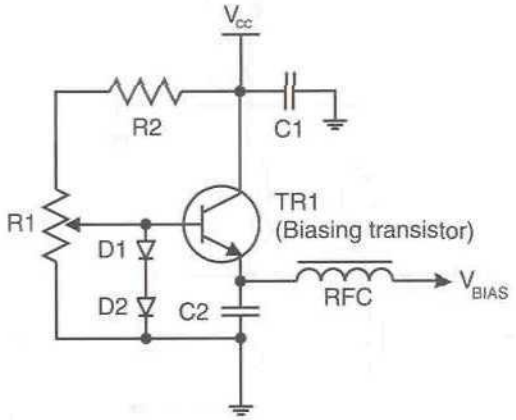


Figure 3.24 Amplified diode bias scheme for an RF power transistor.

transistor (or Darlington pair) employed for TR1. The current flowing through D1 and D2 is given by:

$$I_{D1,D2} = \frac{I_{BIAS}}{h_{FE}} \tag{3.59}$$

where I_{BIAS} is the required bias current for the RF power transistor and h_{FE} is the DC current gain of the biasing transistor, TR1.

R1 and R2 set the bias idle current and the radio-frequency choke (RFC) and capacitor, C1, prevent the RF signal at the base of the RF power transistor from reaching the bias transistor. This could cause high-frequency instability of the bias circuit and would also severely affect the matching of the RF stage.

Temperature compensation may be provided in a similar manner to that used above, namely to mount one of the diodes as close as possible to the RF power device, either on the heatsink, or better still, on the rear of the power transistor case. The other diode and the biasing transistor, TR1, should remain at ambient temperature, although TR1 may well require mounting on a heatsink in many applications.

3.12.3 Low Source-Impedance Biasing

An improved, but still relatively simple, active biasing arrangement is shown in Figure 3.25 [8,12]. In this arrangement, TR2 is a high-power audio-

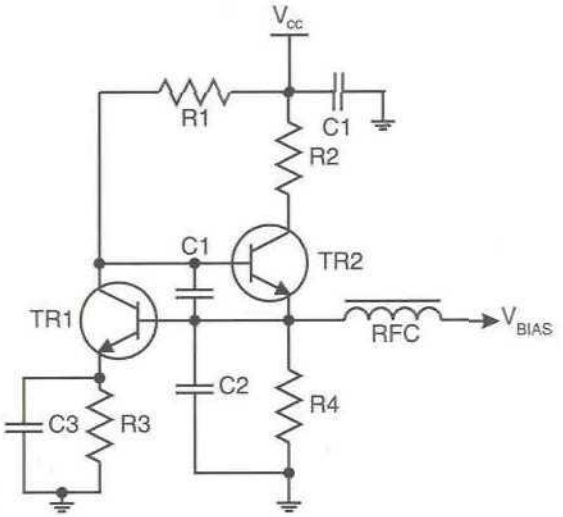


Figure 3.25 A two-transistor biasing scheme for an RF power transistor, featuring a low source impedance.

frequency, general purpose device which must be rated sufficiently to supply the bias current—its power dissipation will typically be in the order of a few watts and hence it will usually need to be mounted on a heatsink. Care must be taken with heatsink mounting to ensure that the transistor case (or tab) is isolated from the ground. TR1, on the other hand, should be capable of being mounted on the main heatsink, as close as possible to the RF transistor, in order to perform temperature compensation (although again it will need to be isolated from ground). It need only supply the drive current for TR2 and hence can be any general purpose, low power device. The only requirement is that its forward base-emitter voltage drop, at the required (low) drive current, must be lower than that of the RF power transistor at its operating bias level.

Having chosen the required semiconductor devices, it only remains to choose the various resistor values. The base-emitter forward voltage drop of TR1, plus the voltage drop across R3 sets the output bias voltage, V_{BIAS}. R3 may therefore be chosen accordingly, or alternatively a suitable variable resistance may be used to allow the bias level to be set up during the testing phase.

R1 provides the base drive current for TR2 and should be chosen accordingly; the required current may be determined from the specified bias

current for the RF device and the h_{FE} of TR2:

$$R_1 \leq [V_{CC} - (V_{BIAS} + V_{BE})] \Big/ \frac{I_{BIAS,RF}}{h_{FE,TR2}} \tag{3.60}$$

where $I_{BIAS,RF}$ is the required RF transistor bias current. R2 limits the maximum bias current which may be obtained from the circuit and should be adequately rated to handle the necessary power dissipation. The value of R2 may be found from:

$$R_2 = \frac{V_{CE} - V_{CE,SAT}}{I_{BIAS,RF}} \tag{3.61}$$

Finally, the capacitors C1–C3 may be necessary to suppress high-frequency oscillations, depending upon the design and layout of the circuit.

3.12.4 Voltage Regulator Biasing

It is also possible to utilise IC voltage regulators as constant-current sources for RF transistor biasing. Care must be taken with such devices, as many have a minimum output voltage well in excess of 0.7V, but these may still be used in a suitable circuit configuration [8].

3.12.5 Biasing of FET Devices

Biasing of FET devices is usually very much simpler than that described above for BJT devices. In many cases, it is sufficient to provide a bias voltage directly on to the gate through a suitable impedance. This impedance must be capable of minimising the amount of RF feeding into the bias supply, this is important for reasons which will be outlined in Section 3.13. It must also present a suitable impedance to prevent the device from going unstable. This is particularly important for GaAs FET devices, which usually require a low resistance from their bias supplies (sometimes none at all).

A typical network, therefore, is shown in Figure 3.26. This consists of a voltage supply (e.g., from a variable voltage regulator, as shown), fed via a bias resistor whose primary function is to aid in the decoupling of the gate from the bias supply, and finally through a transmission line which is a quarter wave long at the center-frequency of interest. It is this latter element which provides the primary RF isolation between the gate and the bias supply. Decoupling is often provided at both the ends of the transmission line and also at the bias resistance (as shown).

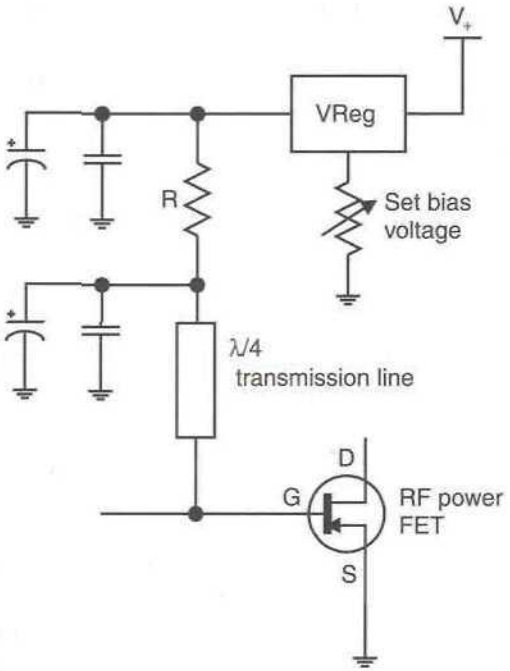


Figure 3.26 Simple FET biasing technique.

3.13 Sources of Inequality for IM Products

3.13.1 Introduction

Testing of a practical RF amplifier design over a range of tone-spacings will usually result in some differences emerging in the levels of the upper and lower intermodulation distortion products of a given order or a change in the level of both, equally, with changes in tone spacing. The origin of these differences is not obvious from classical AM–AM and AM–PM amplifier models and no single factor can explain the effect in all cases. This section presents a number (but probably not all) of the possibilities relating to the design of an individual amplifier stage or cascade of stages.

3.13.2 Supply Modulation Effects

The variation in the current drawn from the power supply for a linear RF power amplifier ranges from almost zero (theoretically, actually zero) for class-A amplifiers, to the full current capability of the supply, for class-B amplifiers (with a rectified sinewave waveshape, in the case of a two-tone

test). This large current variation should be isolated from the power supply circuitry (which usually has a very poor RF and envelope-frequency response) by means of the decoupling network. This network should be mounted as close as possible to the required point and should consist of a range of values and technologies to ensure that all envelope frequencies of interest are adequately decoupled.

Imperfections in the decoupling network will result in a supply voltage variation in sympathy with the envelope current drawn from the supply. This in turn will result in an unwanted envelope modulation being applied to the (already modulated) output signal and is usually more prevalent at particular envelope frequencies, where, for example, a tantalum capacitor is beginning to become inductive and a parallel ceramic capacitor has yet to reach a sufficiently low impedance. Given the reactive nature of the decoupling network, it is very likely that this envelope modulation will be applied at a different phase angle to the wanted modulation. This in turn will result in the AM–PM phase angle being non-zero and hence in an asymmetry in the resulting IM products [13,14]—see also Section 2.13.3.

The use of a range of capacitors, each of which is effective in a sub-range of the required envelope frequencies, is widely applied to solve this problem. A drawback of this technique, however, is that as a low-frequency capacitor becomes inductive, it can form a resonant circuit with a smaller value (high-frequency) capacitor. This in turn can result in a large asymmetry at the resonant frequency. This effect is analysed in [14] and the use of a broadband emitter follower as the supply mechanism is suggested as a better solution than a range of decoupling capacitors.

3.13.3 Bias Modulation Effects

The device biasing circuitry can also suffer from a similar problem to that outlined above, although not generally due to cycling currents being drawn from the supply. The function of the bias network is to supply an appropriate (and constant) voltage or current to the gate or base of the device. So far as the RF signals *or modulation* are concerned, it should appear to be a high impedance (or possibly a very low impedance in the case of modulation signals)—it should certainly not allow such signals to pass to any part of the power supply and be reflected back to the input of the active device.

Signals at the modulation frequency (or harmonics thereof) are generated by nonlinearities in the base or gate of the active device and these are often reflected, rather than absorbed by the matching network. Indeed this property is sometimes used as a predistortion mechanism (see

Chapter 6). These modulation frequency signals can then impinge upon the bias network and this must absorb them (ideally) and certainly not reflect phase-modified or further distorted signals (from the power supply circuitry) back to the gate or base of the device. These signals will combine with the wanted input signals and cause a modification of the IM products; this may well include asymmetry.

3.13.4 Interstage Reflections

As mentioned above, the input to an active device can (and usually does) generate distortion. This distortion can be passed by the matching network (perhaps arbitrarily modified in amplitude or phase) to the output of the preceding stage. This in turn will modify, again arbitrarily in amplitude and/or phase (and possibly further distort) the reflected signal and will reflect it back to the originating input. Clearly this additional input signal can have an effect on the resulting output distortion of the stage and one effect is that of *inequality in the IMD products.*

Solutions to this problem include: changes in the matching network design to reduce or eliminate these signals (e.g., arranging the network to have a low impedance at the harmonics and/or the envelope frequencies) or including interstage isolation (e.g., an isolator or pad).

References

1. Krauss, H. L., C. W. Bostian and F. H. Raab, *Solid State Radio Engineering*, Chapter 12, John Wiley & Sons, New York, USA, 1980.
2. Curtice, W. R., "A MESFET model for use in the design of GaAs integrated circuits," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 28, No. 5, May 1980, pp. 448–456.
3. Allen, S. T., R. A. Sadler, T. S. Alcorn, J. W. Palmour, and C. H. Carter, "Silicon carbide MESFETs for High-Power S-Band Applications," *IEEE International Microwave Symposium 1997*, Denver, Colorado, USA, Vol. 1, June 1997, pp. 57–60.
4. Bowick, C., *RF Circuit Design*, Howard Sams & Co., Indianapolis, USA.
5. Raab, F. H., "Idealized operation of the Class-E tuned power amplifier," *IEEE Trans. on Circuits and Systems*, Vol. CAS-24, December 1977, pp. 725–735.
6. Raab, F. H., "Effects of circuit variations on the Class-E tuned power amplifier," *IEEE Journal of Solid State Circuits*, Vol. SC-13, April 1978, pp. 239–247.
7. Raab, F. H., "Class-F power amplifiers with maximally flat waveforms," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 45, No. 11, November 1997, pp. 2007–2012.
8. Dye, N., and H. Granberg, *Radio Frequency Transistors—Principles and Practical Applications*, Chapter 4, Butterworth Heinemann, MA 02180, USA, 1993.
9. *RF Device Data*, DL110, Rev 4, Vol. II, Motorola Inc. Semiconductor Products Sector, Phoenix, Arizona.

10. *The Acrian Handbook*, Acrian Power Solutions, 490 Race Street, San Jose, California, 1987, pp. 622–674.
11. Application Note AN-779, Motorola Semiconductor Sector, Phoenix, Arizona.
12. Koppen, M. J., *Electronic Applications Laboratory Report*, ECO7308, Philips Components, Discrete Semiconductor Group, 1974.
13. Cripps, S. C., *RF Power Amplifiers for Wireless Communications*, Norwood, MA: Artech House, 1999, Chapter 7.
14. Sechi, F. N., "Linearized class-B transistor amplifiers," *IEEE Journal of Solid State Circuits*, Vol. SC-11, No. 2, April 1976, pp. 264–270.

4

Feedback Linearisation Techniques

4.1 Introduction

Perhaps the simplest and most obvious method of reducing amplifier distortion is by some form of feedback. Since its invention and subsequent patenting by H. S. Black [1], some years after the long-neglected feedforward technique, feedback has been applied almost universally to error correction of all forms. It is extensively used for plant and process control, but has found perhaps its largest domestic market in audio amplifiers.

At audio frequencies, device gain is usually inexpensive to accomplish, and stability is relatively easy to achieve due to the comparatively small bandwidths required (20 kHz or so). The inherent disadvantages of feedback, in terms of gain-bandwidth product limitations and the sacrifice of gain for linearity, are not a problem. The low frequencies involved also mean that the time-delay through a single stage is relatively small by comparison to the time taken for a complete cycle of the desired signal—this again makes the problem of achieving amplifier stability considerably easier.

When considering RF amplification, the problems of using feedback become more pronounced and the overall design must be considered much more carefully. The bandwidths involved are often larger than at audio frequencies, the cycle-times much smaller and the forward-path gain more expensive to achieve—all of these factors must now be considered. The final factor to be considered is that the linearity required in an RF system is often much greater than that required in an audio system. A typical audio system

specification of 0.1% total harmonic distortion at full output would equate to a nonlinearity which would produce intermodulation products of perhaps 30 dB or 40 dB down for a two-tone test. Many RF systems require IMD products to be 70 dB or more below the wanted signals—a much more stringent requirement.

This chapter looks at the various feedback techniques which have been applied to the field of RF power amplifier linearisation. Broadly speaking these can be broken down into four categories: RF feedback, envelope feedback, polar feedback, and Cartesian feedback, all of which are covered in detail, along with their derivatives. The envelope, polar and Cartesian feedback systems should, more accurately, be referred to as *transmitter linearisation techniques* since each design is a complete transmitter rather than simply a linear amplifier. This distinction is important and needs to be drawn for many of the other techniques described in this book.

4.2 Feedback Theory

Consider the generalised feedback system shown in Figure 4.1. Contained in the forward path is the system requiring control, which for the purposes of this text is an RF amplifier (or upconverter and RF amplifier). The feedback path may contain a number of signal processing elements, however in the simplest case, a (potential) divider will provide *linear feedback*.

The comparator is the component which derives an error signal from the input and fed-back output signals. It is most often a voltage subtracter, but this is not universally so. For example, in a phase-locked loop it is the difference in phase between the input and output signals which is of importance, and hence a phase subtracter, or *phase detector* is used as the

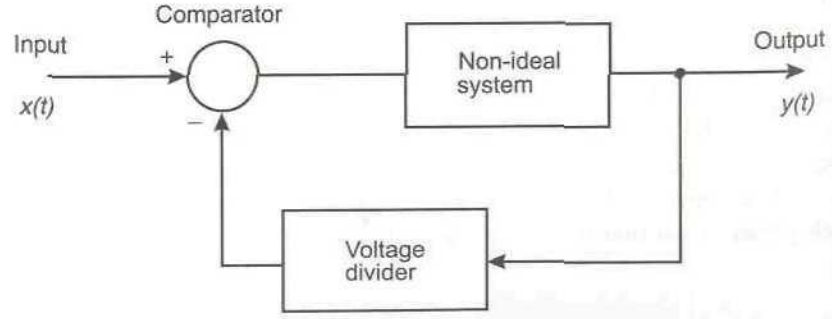


Figure 4.1 General feedback system

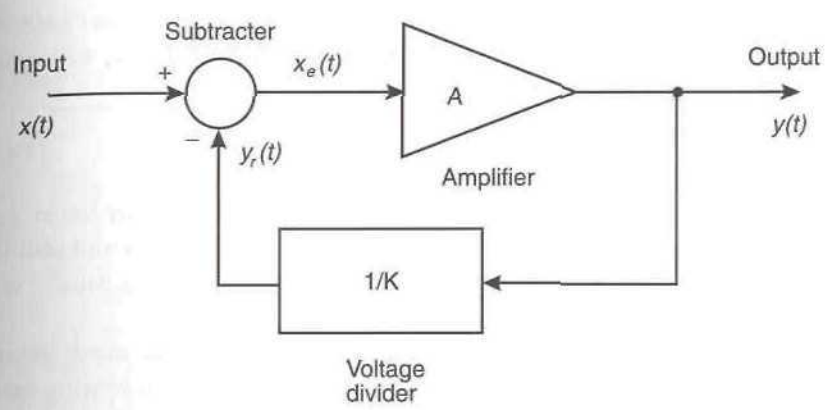


Figure 4.2 Feedback applied around an amplifier system.

comparator [2]. In the case of feedback linearisation of amplifiers the comparator will take the form of a voltage or current subtracter and this will be assumed here.

Figure 4.2 illustrates the use of negative feedback around an imperfect amplifier. The forward path is now an imperfect amplifier of gain, A, and the feedback path is a voltage divider of gain, $1/K$. The comparator is a voltage subtracter of unity gain.

A system input signal, $x(t)$, will result in an error signal, $x_e(t)$ appearing at the input to the amplifier. This signal will be amplified to yield a system output signal, $y(t)$, given by:

$$y(t) = Ax_e(t) \tag{4.1}$$

This output signal is processed by the voltage divider, resulting in the reference signal for the subtracter, $y_r(t)$, where:

$$y_r(t) = \frac{y(t)}{K} \tag{4.2}$$

The error signal, $x_e(t)$, is now defined by:

$$x_e(t) = x(t) - y_r(t) \tag{4.3}$$

Finally, combining (4.1), (4.2) and (4.3) gives:

$$y(t) = \frac{KA}{K + A}x(t) \tag{4.4}$$

Assuming, for the moment, that the amplifier gain is much greater than the voltage division ratio, that is, $A >> K$, then $K + A \approx A$ and, to a good approximation,

$$y(t) = Kx(t) \tag{4.5}$$

The resulting system, therefore, although having a considerably lower gain than the original amplifier, has gained in terms of the reliability and stability of that gain, since the division ratio of the voltage divider (K) is usually very stable with, for example, temperature.

The requirement for the main amplifier gain, A, to be much greater than the division ratio of the potential divider has a significant bearing on an RF system. RF gain, particularly at higher power levels, is expensive to achieve, due to both the relatively low gain and high costs of RF power devices. Only small amounts of feedback can therefore be applied to individual power stages, with the resulting effect on distortion being proportionately less (Section 4.2.1). Applying feedback around the whole of a multistage RF amplifier would partially alleviate this problem, as the gain could be sacrificed in the lower power stages, where is it easier and cheaper to recover. However, such far reaching feedback may well result in instability, due to the significant delay through the forward path of the system relative to the period of a single cycle of the input waveform at an RF operating frequency.

4.2.1 Feedback Distortion Reduction

In the above discussion, the effect of the distortion added by the amplifier was not considered explicitly. The effect of noise and distortion may be analysed by the addition of a summing junction in the forward path into which the noise and distortion ($d(t)$) are added. This situation is illustrated in Figure 4.3.

The analysis proceeds in a similar manner to that above, giving:

$$y(t) = Ax_e(t) + d(t) \tag{4.6}$$

and

$$y_r(t) = \frac{y(t)}{K} \tag{4.7}$$

The error signal, $x_e(t)$, is similarly defined by:

$$x_e(t) = x(t) - y_r(t) \tag{4.8}$$

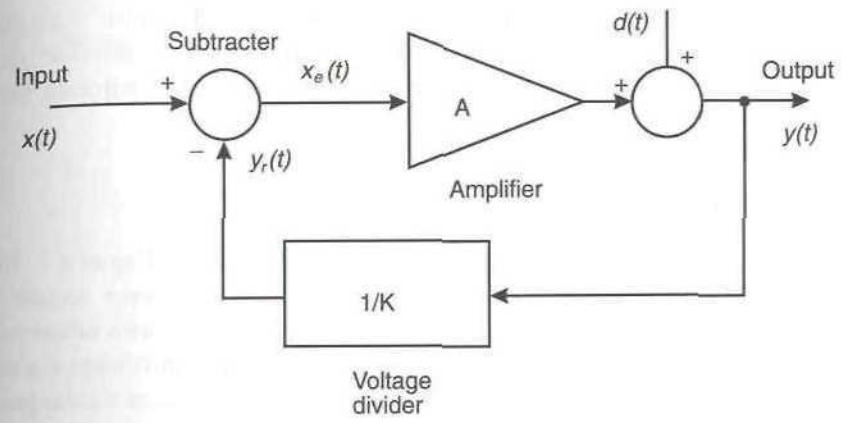


Figure 4.3 Feedback applied around an amplifier system with distortion.

Combining (4.6), (4.7), and (4.8), gives:

$$y(t) = \frac{K(Ax(t) + d(t))}{K + A} \tag{4.9}$$

Assuming, for the moment, that the amplifier gain is much greater than the voltage division ratio, that is, $A >> K$, then $K + A \approx A$ and therefore:

$$y(t) = Kx(t) + \frac{Kd(t)}{A} \tag{4.10}$$

The distortion produced by the main amplifier is therefore reduced by a factor K/A by the inclusion of the feedback loop. Thus, for example, if $A = 10K$, then the distortion level in the main amplifier will be improved by 20 dB, with the corresponding sacrifice of 20 dB of gain.

4.3 RF Feedback

The two simplest and most common forms of *passive* RF feedback are similar to their audio frequency counterparts, namely *series feedback* and *shunt feedback*. Both are concerned with linearising an individual stage, rather than a complete multistage amplifier and both are extremely simple to implement.

They may be applied to linearise class-A, class-AB or class-B stages and hence will not result in a high efficiency amplifier. However, the improvement in distortion performance is stable and predictable and the relative simplicity of the two methods makes them popular for high-reliability applications.

They are most often applied to correct for 'linear distortion' (i.e., gain and phase ripple) in broadband amplifiers and also to provide gain stabilisation for a design, where a particular level of gain is required for a large number of production units, for example.

4.3.1 Series Feedback

The configuration of a series feedback amplifier is shown in Figure 4.4. In a normal amplifier stage (without feedback) the emitter resistor would be shunted by a capacitor with negligible reactance at all frequencies of interest. This capacitor is absent from the circuit of Figure 4.4 and hence an AC component is permitted in the emitter voltage. This variation also appears on the base, since V_{be} remains approximately constant with changes in device current, hence providing negative feedback due to the inverting behavior of the stage.

The action of the feedback may be described mathematically as follows. The input signal, v_{in}, is almost exactly reproduced at the emitter, since the base-emitter voltage, V_{be}, remains essentially constant. Thus, the emitter current may be deduced:

$$i_e \approx \frac{v_{in}}{R_E} \tag{4.11}$$

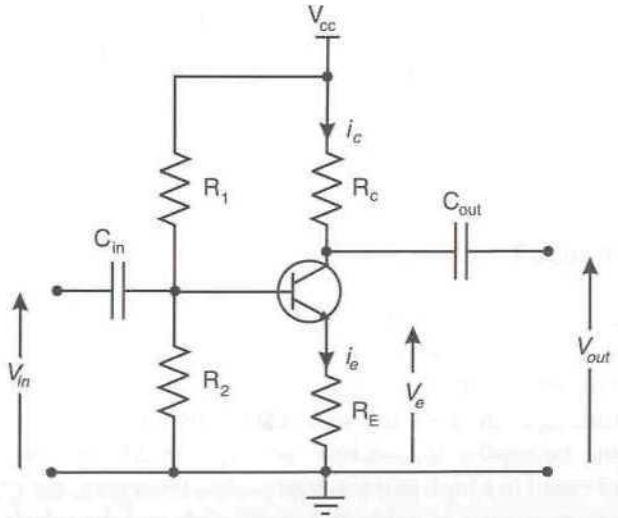


Figure 4.4 Series feedback applied to an RF amplifier stage.

The collector current and emitter current are approximately equal, that is, $i_c \approx i_e$, hence:

$$i_e \approx \frac{v_{in}}{R_E} \tag{4.12}$$

or

$$\frac{1}{R_E} \approx \frac{i_c}{v_{in}} \tag{4.13}$$

The series feedback amplifier may therefore be regarded as a transconductance amplifier of gain, $g_T = 1/R_E$. The gain is therefore specified by a resistance which is substantially constant, for example, with temperature, bias level, time, and, more importantly, is also linear. By contrast, similar analysis on a standard common emitter amplifier yields a transconductance gain of $1/r_e$, where r_e is given by:

$$r_e = \frac{kT}{qI_E} \tag{4.14}$$

where k is Boltzmann's constant, q is the electronic charge, T is the device temperature (in Kelvin) and I_E is DC value of the emitter current. This resistance is not only poorly defined with temperature, it is also nonlinear.

The voltage gain of the stage may similarly be derived since:

$$v_{out} = -i_c R_C \tag{4.15}$$

The stage voltage gain is therefore:

$$A_v = \frac{v_{out}}{v_{in}} \approx -\frac{R_C}{R_E} \tag{4.16}$$

The equivalent circuit of the series feedback amplifier is shown in Figure 4.5 and from this it is possible to derive the input (driving) impedance of the circuit. Analysis of the circuit gives:

$$\begin{aligned} v_{in} &= i_b r_{be} + (i_b + \beta i_b) R_E \\ &= i_b (r_{be} + (1+\beta) R_E) \end{aligned} \tag{4.17}$$

The input impedance is therefore:

$$r_{in} = \frac{v_{in}}{i_b} = r_{be} + (1+\beta) R_E \tag{4.18}$$

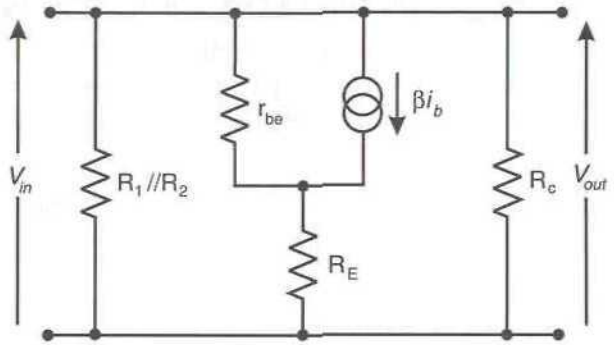


Figure 4.5 Equivalent circuit of a series feedback amplifier.

which is considerably higher than that achieved without feedback. Design of the input matching network and drive circuitry may therefore be greatly simplified.

4.3.1.1 Practical Considerations

Many RF power transistors include some emitter resistance as part of their structure, in order to ensure an even division of current within the device. This emitter or *ballast* resistance results in an inherent (small) level of series feedback in many PAs.

Source feedback in an FET stage, although common in small signal stages, is not generally practical for high power designs. The large values of gate voltage result in a consequent large voltage across the source resistance and hence an excessive power dissipation in that resistance.

Finally, it is important to note that series feedback will have no effect if the amplifier stage is being current (rather than voltage) driven. Current drive is frequently used in PAs due to the inherently more linear relationship between base and collector currents, rather than between base voltage and collector current (exponential relationship).

As a result of these considerations a second feedback configuration is more often used, namely shunt, or collector, feedback.

4.3.2 Shunt Feedback

Shunt feedback could be viewed as the dual of series feedback, in that current drive is now required for the feedback to operate and no linearity improvement will be experienced if voltage drive is attempted instead. The following discussion will assume that current drive is provided and hence no

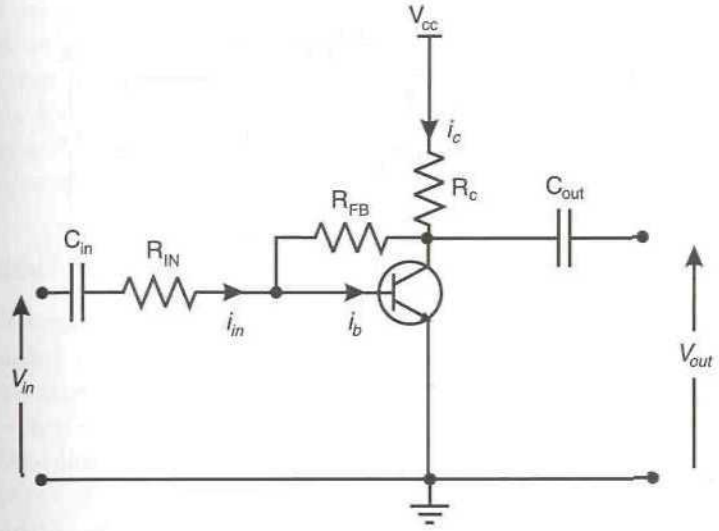


Figure 4.6 Shunt feedback applied to an RF amplifier stage.

input series resistor is required in the circuit configuration shown in Figure 4.6 (i.e., $R_{IN} = 0$).

It will be assumed that the current gain, β, of the transistor is large, hence making the base current negligibly small, and also that the base-emitter voltage remains constant with applied signal (i.e., $v_{be} = 0$).

The signal input current (i_{in}) therefore flows through the feedback resistor creating a voltage drop across it equal to $-i_{in}R_{FB}$, and, since $v_{be} = 0$:

$$v_{out} = -i_{in}R_{FB} \tag{4.19}$$

Therefore the transresistance ($r_T = v_{out} / i_{in}$) is:

$$r_T = \frac{v_{out}}{i_{in}} = -R_{FB} \tag{4.20}$$

The input current/output voltage relationship is therefore linear and again independent of, for example, temperature effects.

Analysis of the equivalent circuit shown in Figure 4.7 yields the small signal current gain:

$$A_i = \frac{i_{out}}{i_{in}} = -\frac{R_{FB}}{R_C} \tag{4.21}$$

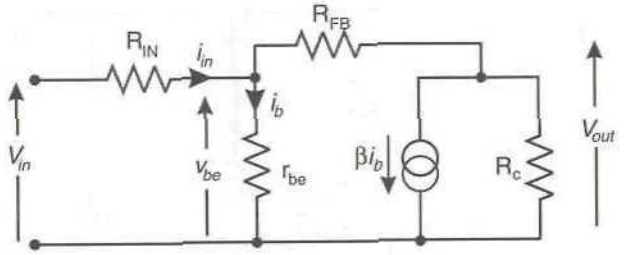


Figure 4.7 Equivalent circuit of a shunt feedback amplifier.

and the input resistance, r_{in}:

$$r_{in} = \frac{R_C + R_{FB}}{1 + \dfrac{R_{FB} + (1+\beta)R_C}{r_{be}}} \tag{4.22}$$

which is generally low in comparison to that of the series feedback stage.

4.3.2.1 Practical Considerations

The practical implementation of shunt feedback at RF is most easily achieved using the network shown in Figure 4.8. The blocking capacitor, C_{FB}, should have negligible reactance at the frequency of interest and the inductor, L_{FB}, may be added to compensate for a fall in device gain at high frequencies. The effect of phase-shifts due to the inductor, transistor and associated leads must be considered when evaluating circuit stability.

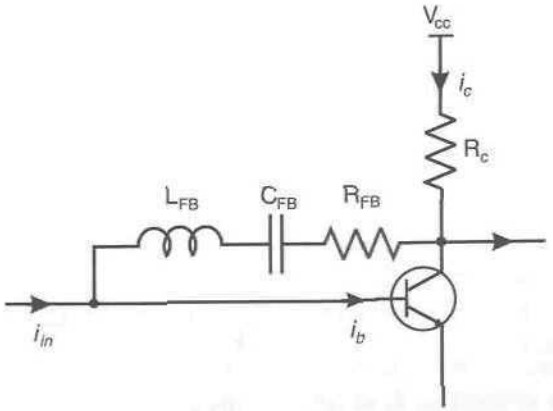


Figure 4.8 Practical implementation of a shunt feedback amplifier.

As was mentioned above, current drive must be used for this form of feedback to operate. This is a distinct advantage as the basic linearity of the current transfer characteristic (and hence current-voltage transfer characteristic) of a transistor is much better than that of its voltage-current transfer characteristic. Hence less feedback will be required to yield a given level of absolute performance.

4.3.3 Other Configurations

A tap or separate winding on the collector load transformer, in a transformer-coupled amplifier, may be used to couple a portion of the amplifier output current back to the base. In the case of a BJT this will linearise the current transfer characteristic; with an FET, the voltage transfer characteristic will be linearised.

Passive feedback is often used merely to improve the broadband characteristics of an RF amplifier, rather than to improve its IMD or harmonic distortion properties. It may be used, for example, to compensate for a poor impedance match at certain frequencies within the desired passband or to introduce a degree of gain levelling in a broadband amplifier. These advantages are, however, achieved at a price (as in all feedback systems) since gain is lost due to the action of the feedback and efficiency is also sacrificed due to power dissipation in the feedback network [3]. The presence of feedback also lowers the effective input impedance of the device.

The goal of RF feedback for gain levelling (particularly at lower RF frequencies, e.g., HF) is usually to effect a gain reduction at low frequencies, where the gain is generally high, in order to balance the lowering of gain with increased frequency, occurring due to device roll-off. The feedback network is therefore designed with a low-pass response, thereby providing greater feedback at lower frequencies.

Figure 4.9 illustrates three transformer-based configurations for implementing RF feedback. In Figure 4.9(a), collector to base feedback is provided by the transformer T1, with the RF input forming a tap to the primary of T1. The impedance 'seen' by the feedback voltage is set by T1 and hence its level and the consequent power dissipation in the feedback network are also set by this transformer. The capacitor, C, is purely for DC blocking and should be chosen so as not to have a significant effect in the band of interest.

In Figure 4.9(b), feedback is achieved by means of a third turn on the output transformer, T2, hence allowing a less complex (and lower impedance) input transformer to be used. The impedance of the feedback source is now no longer that of the collector and hence may be lowered as desired.

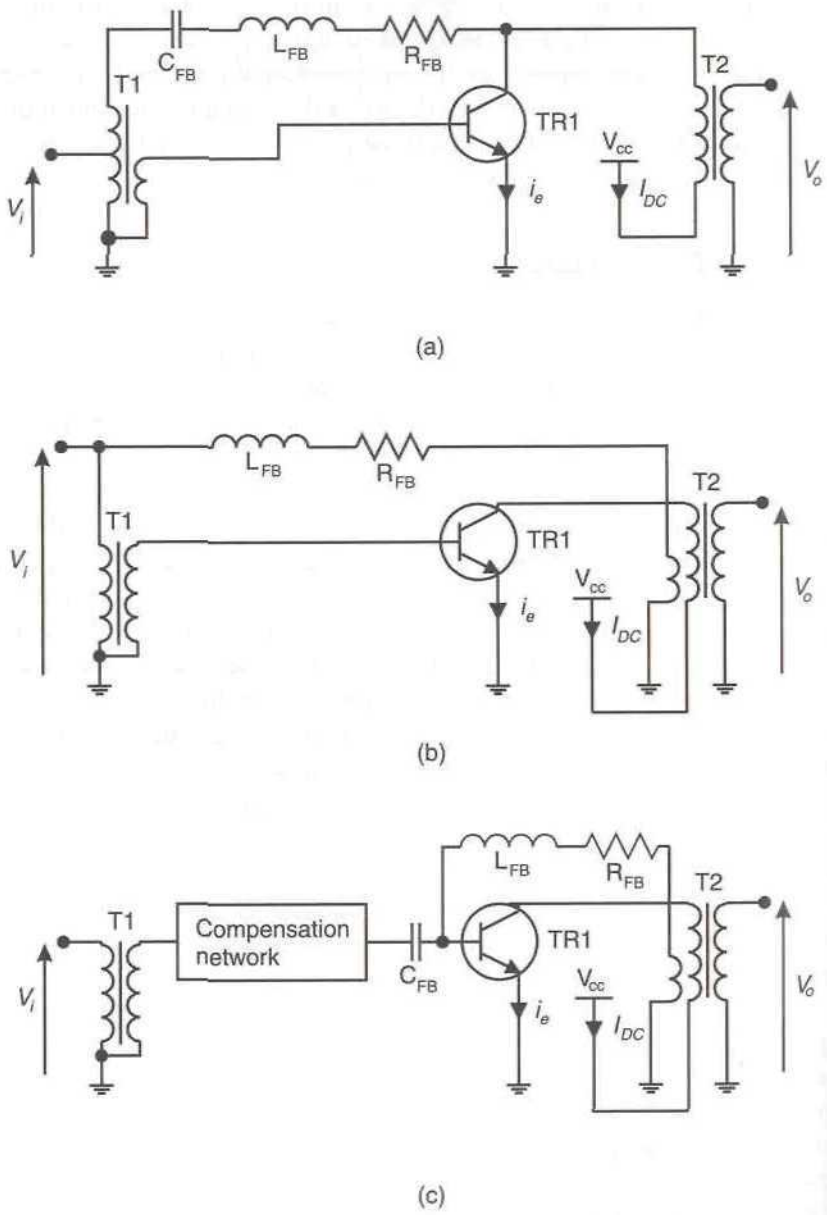


Figure 4.9 Negative feedback applied around a BJT, utilising input and output transformers (after [3]).

Finally, Figure 4.9(c) shows the feedback connected directly to the base (a low impedance point) and hence a low impedance feedback source is required. This again can be provided from an additional winding on the output transformer, T2, as shown. This is the preferred arrangement for bipolar devices since the base impedance is well defined, and hence the level of feedback may be determined purely by the turns ratio on T2.

4.3.4 Example System

A practical example of the use of passive feedback in a multistage RF amplifier was presented in a paper by Mitchell [4]. The application was for a medium-power submarine cable repeater amplifier for use in a 10,800 carrier frequency division multiplex system of up to 500 nautical miles in length. A summary of the final specification is given in Table 4.1.

The design of this amplifier, although not recent, illustrates many of the techniques and features inherent in feedback RF amplifiers. Negative feedback was originally chosen for this design as it can be used to set and control the gain response and terminal impedances at the input and output. A principal advantage of negative feedback over most of the other linearisation techniques is that of simplicity. The resultant number of active devices required is small and hence the reliability is likely to be

Table 4.1
Summary of specifications for the high-band submarine cable repeater amplifier

Parameter	Achieved specification	Notes
Frequency range:		
F_{min}	79 MHz	
F_{max}	135 MHz	
Output power	25 dBm	
Gain:		
At F_{min}	26 dB	
At F_{max}	35 dB	
Third-order IMD	−102 dBm	Measured using a two-tone test
Noise figure	5 dB	For complete amplifier
Input return loss	>20 dB	
Output return loss	>20 dB	

good. This is a particularly attractive feature for the submarine cable application considered here since the design life had to be greater than 25 years, with repair being virtually impossible.

A disadvantage of feedback amplifiers is also illustrated in that in order to ensure loop stability, the open loop transfer function must have a bandwidth of approximately an order of magnitude greater than the highest operational frequency. This requirement is a significant problem at higher frequencies as the devices required tend to have a poorer performance and lower power capability than their lower frequency counterparts. They also tend to be very expensive.

4.3.4.1 Amplifier Design

To achieve the required intermodulation specification, a loop gain of 15 dB was required in the feedback loop, hence requiring an open loop gain of 50 dB in order to achieve the required overall gain of 35 dB (at 135 MHz). The circuit diagram is shown in Figure 4.10, with the circuit consisting of three common emitter stages, plus input and output impedance transformation.

The design illustrates three different types of feedback, namely: series, shunt, and overall feedback around the complete amplifier. Each stage has feedback applied in an alternate series-shunt-series arrangement for the three transistors and this is intended to stabilise the amplifier performance against changes in the active device characteristics over its lifetime. The shunt feedback applied to the second stage is frequency selective, and this provides the necessary shaping of the loop gain response to restrict the amount of in-band feedback, and provide large amounts of out-of-band feedback, hence realising the necessary bandpass frequency response. This illustrates another advantage of feedback in this application, in that an almost arbitrary modification to the frequency response can be achieved.

It can be deduced from the previous sections on series and shunt feedback that the alternate application of these techniques throughout the amplifier will lead to a considerable mismatch between stages. The interaction between stages is therefore small and this will help to improve the overall stability of the design.

The overall feedback around the complete amplifier is employed to set the terminal impedances to 88Ω (for optimum noise figure of the active devices) and to tailor the overall passband frequency response to the approximately $\sqrt{f}$ response required to compensate for the response of the cable. The feedback components employed are a tapped transformer at input and output together with the current feedback resistors, R1 and R2; the damped resonant circuit at the bottom of the diagram provides the

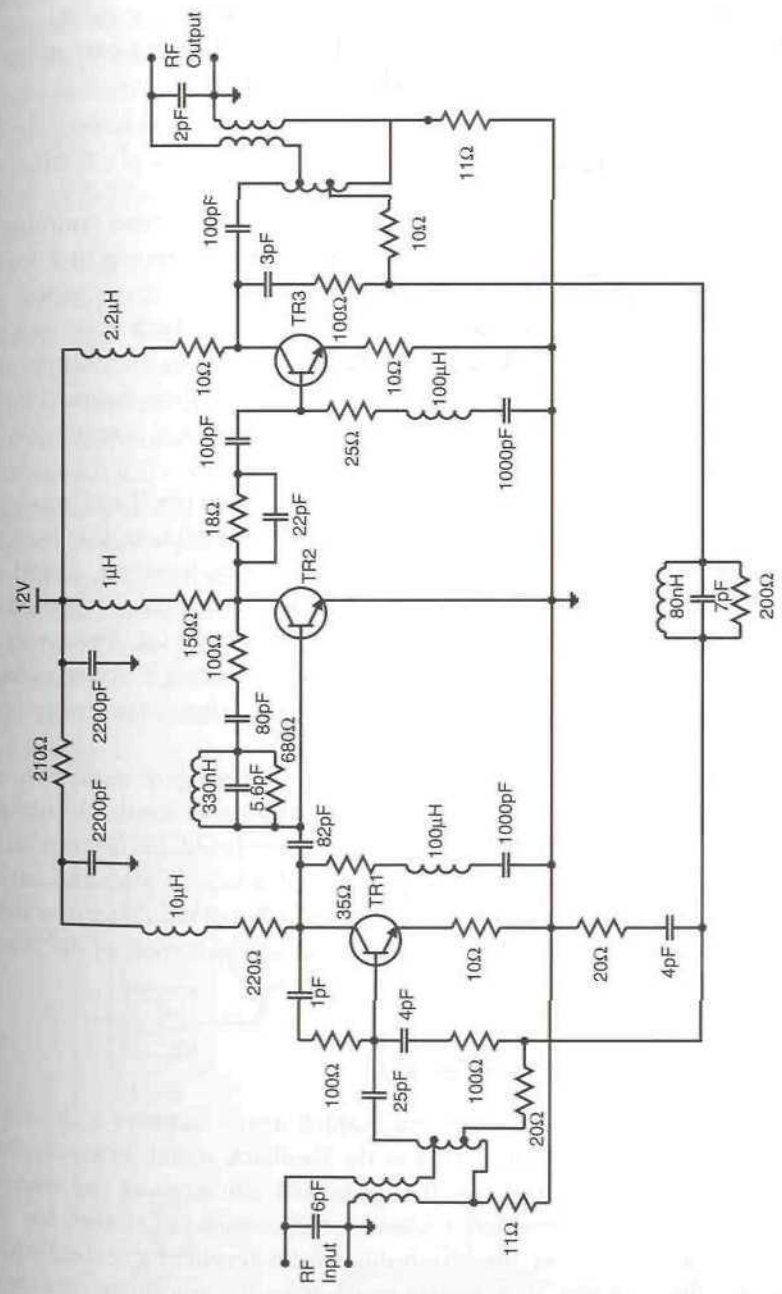


Figure 4.10 Circuit diagram of a 135 MHz feedback amplifier for submarine cable applications (omitting base-biasing components). TR1 and 2 are both Avantek AT4641, and TR3 is HP type 35853E (from [4] © IEE 1979).

frequency shaping. The 88Ω impedances are then transformed by the input and output transformers to the required 50Ω impedance of the cable.

The achieved final loop gain of around 17 dB resulted in a corresponding 17 dB reduction of the third-order IMD products.

4.3.5 Active RF Feedback

If the 'voltage divider' in Figures 4.1–4.3 is replaced by an active (amplifier) stage, gain is provided in the feedback path and hence less of the main amplifier power needs to be dissipated in the feedback components. In addition, this 'auxiliary amplifier' (the amplifier in the feedback path) can be designed to be relatively nonlinear and hence introduce its own distortion into the feedback path. This distorted feedback signal is then summed with the input signal and, by suitable control of its gain and phase, a reduction in the main amplifier distortion may be achieved.

Perez *et al.* describe a system based on this concept [5–7] and report a gain reduction of only 3 dB for a corresponding reduction in third-order IMD of 12 dB. The level of distortion cancellation was, however, found to be dependent upon signal level, with relatively poor performance reported at signal levels other than that on which the system was set up. However, if only a single operating point is required (i.e., power control is not required) then it may be possible to achieve satisfactory performance from this technique.

The improved performance, in terms of a minimal gain reduction, of this active feedback technique over conventional passive feedback, occurs primarily because of the increased distortion introduced by the auxiliary amplifier. This (additional) distortion, when added back into the input signal, results in a proportionately greater cancellation of the distortion than of the (wanted) linear signal. Therefore a smaller gain reduction of the linear signal results.

4.3.6 Difference-Frequency Feedback

This is a little used feedback technique, which again employs a modified form of the amplifier's own distortion as the feedback signal, in an attempt not to significantly lower the amplifier gain and efficiency at the wanted frequency. In this case, apparent third-order distortion is created by the feeding-back of some of the distortion due to the amplifier's second-order nonlinearity; the amplifier then creates third-order intermodulation distortion from this fed-back second-order distortion. Careful adjustment of the feedback components can therefore allow this 'false' third-order IMD to

appear and act like conventionally fed-back third-order IMD, and hence linearise the amplifier.

A second-order nonlinearity, when faced with a standard two-tone test, will produce components at the difference frequency, $f_1 - f_2$, and components centered around the second harmonic of the channel frequency. It is usually the low (difference) frequency components which are chosen to be used in the feedback path, as these lead to a simpler practical implementation.

A block diagram of this approach is shown in Figure 4.11. The diplexer at the output is required to separate the second-order distortion (in the DC zone) from the fundamental frequency (wanted) signal, thereby only allowing feedback of the second-order distortion component. Feedback of the fundamental component as well would, of course, result in a lowering of the amplifier's gain at the wanted frequency. The diplexer at the input allows the recombination of the (modified) second-order distortion with the input signal. Diplexing is necessary due to the filtering action of the input and output matching networks which would otherwise completely remove the low-frequency second-order distortion.

Since the signals to be split and combined by the diplexers are of such widely differing frequencies, it is a simple matter to make use of RF chokes to handle the second-order distortion at baseband and inject it directly into the transistor base (for example); similarly at the output (e.g., at the collector).

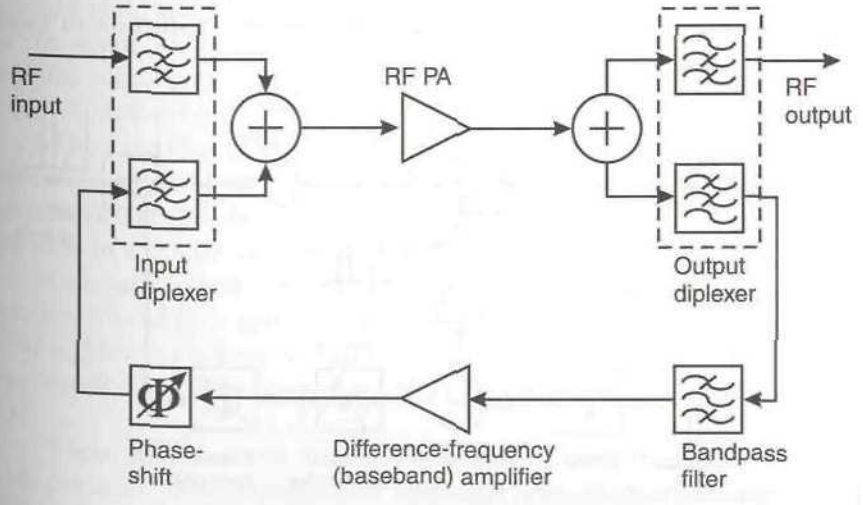


Figure 4.11 Difference-frequency feedback applied around an RF amplifier.

The bandpass filter in the feedback path prevents both DC and unwanted high-frequency components from entering the feedback amplifier.

Use of this technique on a 10 GHz amplifier is reported in the literature [8]. The resulting improvement in third-order IMD performance was 12 dB when using a two-tone test with 10 MHz tone spacing.

4.3.7 Distortion Feedback

This is another attempt to overcome the shortcomings of conventional feedback when applied to RF amplifiers, in this case by cancelling the fundamental frequency components from the fed-back signal in a manner akin to that used in a feedforward system (see Chapter 5). Thus only the nonlinear distortion components remain in the fed-back signal, hence providing an improvement in nonlinear distortion performance of the overall amplifier. However, like the difference-frequency feedback technique outlined above, no improvement in linear distortion (gain or phase performance with frequency) is achieved.

A block diagram of the distortion feedback approach is shown in Figure 4.12. It suffers from many of the inherent problems of the feedforward technique, since its performance is dependent to some extent on the quality of a cancellation process. The gain and phase controllers should, ideally, be automatically controlled to compensate for temperature drift, component aging and changes in the gain or phase response of the amplifier(s) with signal loading.

Only modest improvements in linearity have been achieved to date by using this technique. Gajda and Douville [9] report a reduction in third-

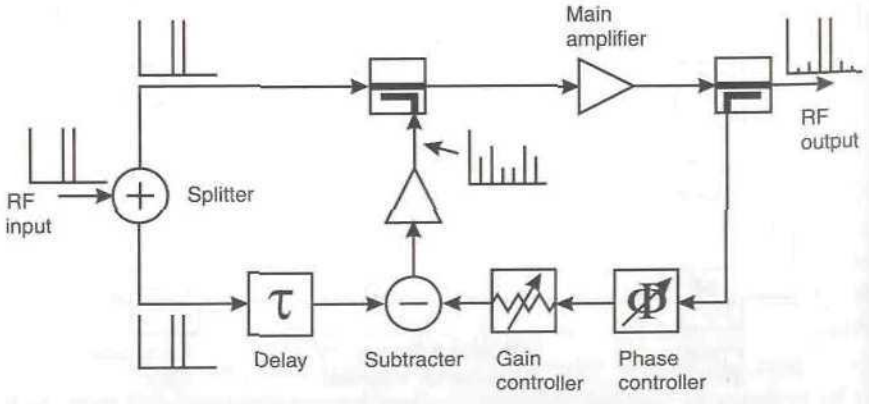


Figure 4.12 A distortion feedback amplifier.

order IMD level of 6 dB when applying the technique to an amplifier at 300 MHz with a 10 MHz bandwidth. Significantly better performance would be necessary to justify the complexity involved in any sort of control technique to be used in conjunction with this type of amplifier.

4.3.8 RF Feedback Employing Cartesian Compensation

The compensation filtering required in RF feedback systems can be provided in a number of ways. Conventional designs have employed cavity resonators in the forward path [10]; however, these have many obvious drawbacks. An alternative technique involves the use of a Cartesian loop (Section 4.6); however, with the input signal supplied at RF prior to the quadrature demodulator (that is, subtraction to form the error signal performed at that point). This technique therefore utilises Cartesian compensation in the forward path and this has the advantage of allowing simple baseband RC filtering to be used in place of the cavity resonators mentioned above.

The use of Cartesian compensation has the disadvantage, over conventional RF feedback, of requiring an on-frequency local oscillator signal to drive the down- and upconversion processes. In addition, noise present in the down- and upconverters will only be compensated within the loop bandwidth, hence wideband noise can be generated which would not be present in a conventional RF feedback system.

A block diagram of the Cartesian compensated RF feedback system is shown in Figure 4.13. Both the demodulator and modulator are in the forward path, and hence nonlinearities in either should be compensated for by the action of the feedback. In addition, IQ vector errors in either the modulator or demodulator should be less critical than for the Cartesian loop system (Section 4.6).

The I and Q compensation circuits may be fabricated as single-pole RC filters with a suitable cut-off frequency (e.g., 500 kHz for a 1 MHz RF output spectrum). The RF phase-shift is necessary to compensate for delays around the loop in a similar manner to that of the Cartesian loop.

Feedback stability criteria dictate the degree of IMD cancellation which can be achieved for a given bandwidth. As an example, a system with a loop delay of 36ns can achieve an IMD improvement of around 18 dB over an RF bandwidth of 400 kHz, decreasing to 5 dB over an RF bandwidth of 4 MHz [11].

These improvements are somewhat more modest than can be achieved with Cartesian loop; however, the technique does allow an RF input signal to be used without additional downconversion.

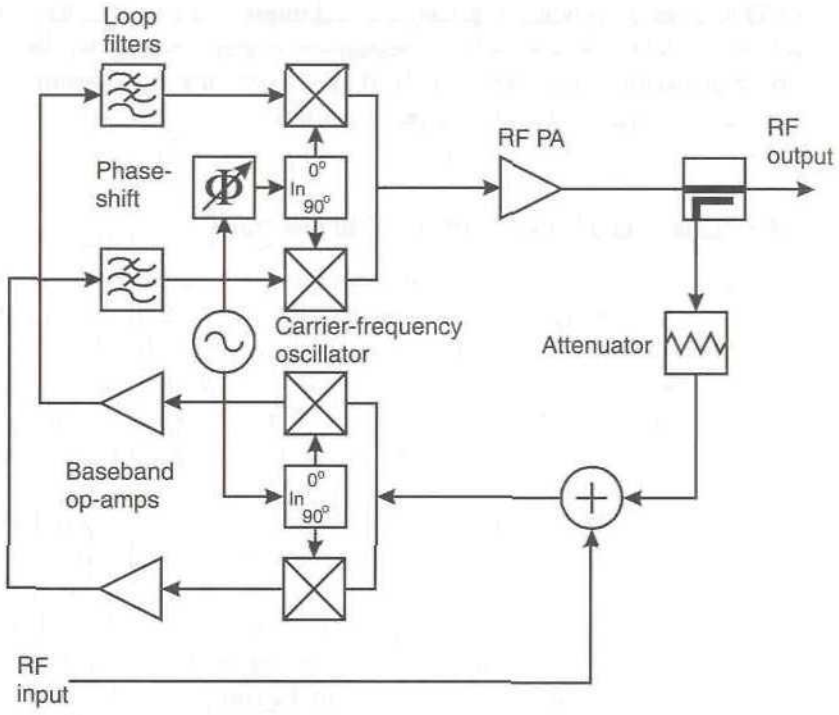


Figure 4.13 RF Feedback system employing Cartesian compensation.

4.3.9 Conventional Quasi-Linear Transmitters With Feedback Power Control

Many quasi-linear transmitters (i.e., those requiring only modest degrees of linearity) do not use any form of linearisation for spectral purity reasons, but do employ feedback to linearise the power ramping and power control functions. A typical arrangement is shown in Figure 4.14.

The I and Q input signals (from the baseband DSP or similar) are upconverted by a quadrature upconverter (usually a single IC), before being fed to the RF transmit chain. This transmit chain consists of an attenuator, for power control and ramping, followed by a pre-driver and power amplifier. In some systems, the power amplifier has sufficient gain to eliminate the pre-driver, and in others, ramping and power control are realised via a gain control pin on the power amplifier IC.

The characteristics of the variable attenuator (or the gain control pin) are generally far from linear, and hence some form of linearisation is required for the power control and ramping functions. This is provided by the output

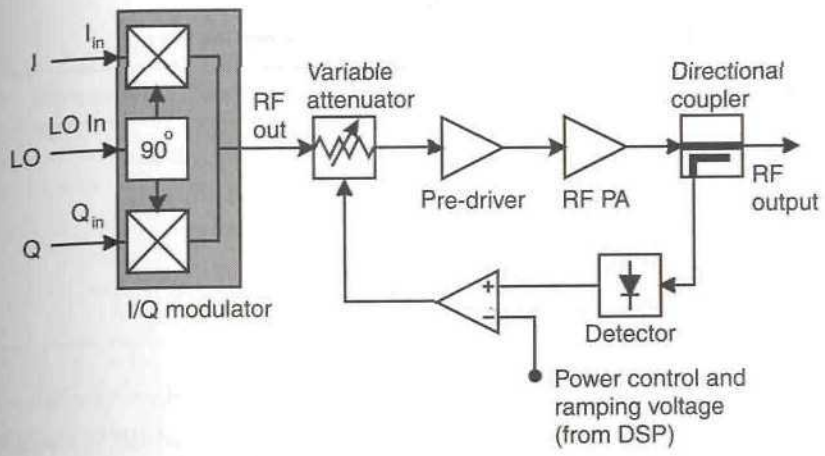


Figure 4.14 Conventional quasi-linear transmitter architecture.

coupler, detector, and feedback mechanism, which together ensure that the power control setting and ramp profile provided by the DSP are closely followed by the transmitter output signal.

This arrangement has a number of disadvantages, most notably, the need for temperature control or careful temperature characterisation of the detector diode. In some systems this must be performed for each transmitter during the production process. This is obviously costly and undesirable.

The linearity requirements of quasi-linear systems are sufficient that conventional class-C power amplifiers cannot be employed in the architecture of Figure 4.14. Thus some form of linear PA or linearisation system employed around a class-C PA is required. Virtually all quasi-linear transmitters currently available employ the former of these solutions.

The adjacent channel requirements of DAMPS (American digital cellular) and PDC (Japanese digital cellular) are modest, but nevertheless important. PDC, for example, has the specifications shown in Table 4.2. It can be noted that the emissions in the first adjacent channel are not specified, and hence the third-order intermodulation produced by the RF power amplifier is not of significant importance. Similarly, the specification does not become more stringent until the fourth adjacent channel, by which time 11th-order IMD is the dominant factor (based on a 21 kHz modulation bandwidth).

It is therefore possible to meet the PDC specification by employing a power amplifier with significant IMD, far worse than could be achieved with, say, a linear class-A or -AB biased PA, operating with the modulation peaks at its 1 dB compression point. The conventional solution to the

Table 4.2
PDC (Japanese Digital Cellular) performance requirements

Parameter	Value
Modulation format	π/4-DQPSK
Modulation bandwidth	21 kHz
Channel bandwidth	25 kHz
Max. power at ±50 kHz (second adjacent channel)	−45 dBc
Max. power at ±100 kHz (fourth adjacent channel)	−60 dBc

handset PA challenge, namely that of achieving the adjacent channel specifications of Table 4.2 whilst consuming the minimum battery power, is generally to employ a linear class of amplifier IC, operating with the modulation peaks well beyond the 1 dB compression point and generally into saturation. This will obviously result in significant intermodulation and spectral spreading, however, the relatively relaxed emission requirements of quasi-linear systems allow this design method to be employed without penalty.

4.4 Modulation Feedback

The use of envelope (or modulation) feedback is a logical extension to the basic notions of feedback in RF amplifiers. By returning the feedback problem to an essentially audio frequency environment, many of the stability problems, although not eliminated altogether, are considerably alleviated. The basic concept is quite straightforward, and is similar to that discussed for AM linearity improvement in class-C stages (Chapter 3). Figure 4.15 shows a general schematic of an envelope feedback system. Its operation is effected by the comparison of a detected version of the output modulation with the intended form of that modulation appearing at the input; the error signal thus generated forming a 'predistorted' version of the input signal. Subsequent amplification of the error signal will therefore generate the intended signal at the output of the transmitter. Note that this linearisation scheme operates on a complete transmitter, since the input signal is now the required modulation and the output generated is an RF signal containing that modulation.

The most basic form of envelope feedback employs non-coherent (envelope) detection and therefore cannot compensate for phase distortions within the amplifier. This simple technique is thus most often employed in

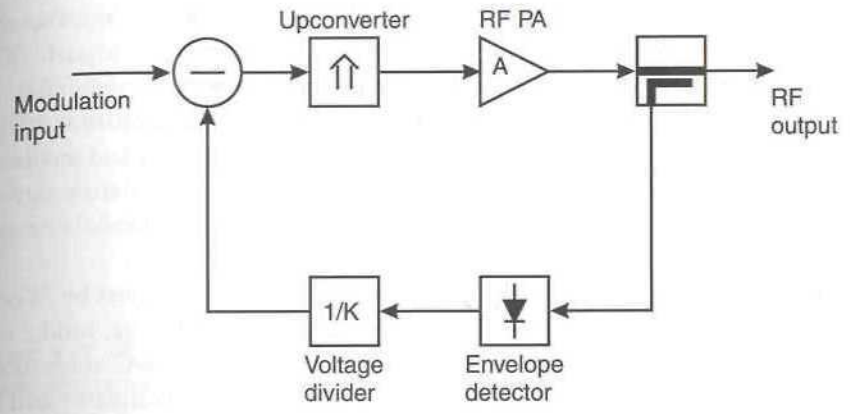


Figure 4.15 Schematic of a transmitter employing envelope feedback.

full-carrier AM transmitters, where detection will ultimately be performed by an envelope detector, and phase information is not required. More advanced derivatives of the basic technique preserve the phase information lost in the basic system by utilising polar or Cartesian signal formats. Both of these will be described in detail later in this chapter.

The feedback bandwidth required of an envelope feedback transmitter will depend upon the modulation format(s) employed in the transmitter. Some signal types can result in a very large feedback bandwidth being required, despite the modulation bandwidth (at RF) being more limited. A good example of this is a two-tone test, which has an envelope of the form of a full-wave rectified sinewave (see Chapter 2). This waveform has a cusp as the envelope goes through its minimum and hence potentially a very wide bandwidth. This will consequently require a large envelope feedback bandwidth for complete linearisation.

An early example of envelope feedback was described [12] and patented [13] by Wood and Arthanayake (although it is similar to an earlier system by Bruene [14]). They principally developed the technique to overcome the problems of low stage gain and large carrier phase-shift in high frequency power amplifiers. The low stage gains available, particularly in higher power output stages, mean that feedback applied around a single stage is inappropriate. The problems of carrier phase-shift in multistage feedback amplifiers, on the other hand, lead to the stability problems alluded to earlier. Envelope feedback, as has already been indicated, can overcome these problems and enable efficient class-C amplifiers to be employed for AM transmission.

The configuration applied in this system is a little different from the complete transmitter described in Section 4.1, in that it is purely an amplifier

and not a complete transmitter. Additional complexity is employed at the input in order to detect the modulation on the incoming signal. The modulator is then only required to correct for the distortion and not to perform the full modulation as well. This technique overcomes some of the difficulties, which are encountered with the direct modulation and feedback method described in Chapter 3, in producing high-power modulating signals to modulate the supply of a PA stage; both the original AM modulator and the error signal modulator can operate at a low power level.

Figure 4.16 shows the basis of the configuration employed by Wood and Arthanayake. Modulation is employed before the final stage, unlike the direct modulation case shown in Chapter 3, and the class-C stage then amplifies the resultant signal. Both the nonlinearity of the modulator and of the final stage are compensated for by the feedback, and this is demonstrated below.

The open loop response of the system (neglecting the input and output coupler losses) is:

$$V_o = a_1 a_2 V_i \tag{4.23}$$

where a_1 is the modulator stage gain and a_2 is the power amplifier stage gain. If the output coupler through-path loss is now included, the open-loop gain of the complete amplifier becomes:

$$A_o = \frac{V_o(1-\beta)}{V_i} = a_1 a_2 (1-\beta) \tag{4.24}$$

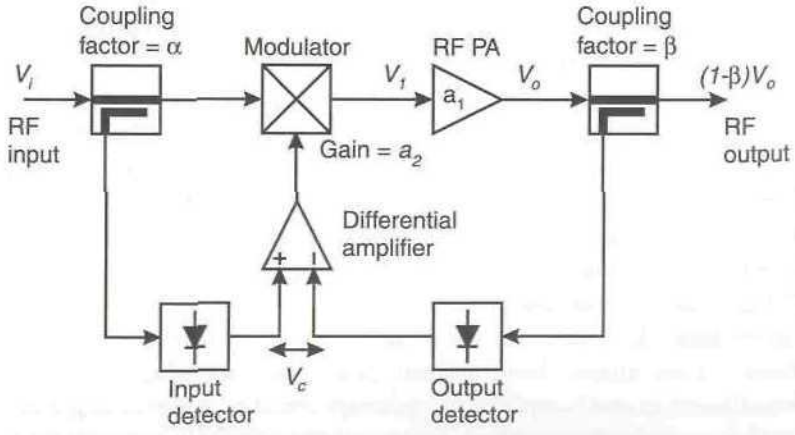


Figure 4.16 Envelope feedback linearised amplifier.

Thus, as would be expected, the open loop gain is a function of the nonlinear gains, a_1 and a_2.

The incorporation of envelope feedback results in the following input–output relationship:

$$V_o = a_2 V_1 \tag{4.25}$$

and

$$V_1 = (a_1 + \gamma\eta V_e) V_i \tag{4.26}$$

where $\gamma = da_1/dV_e$ is the modulator sensitivity and η is the efficiency of the diode detectors (assumed identical for both).

Since,

$$V_e = \alpha V_i - \beta V_o \tag{4.27}$$

then:

$$V_o = a_2 V_i [a_1 + \gamma\eta(\alpha V_i - \beta V_o)] \tag{4.28}$$

Rearranging this gives:

$$V_O = a_2 V_i \left[\frac{a_1 + \alpha\gamma\eta V_i}{1 + a_2\beta\gamma\eta V_i}\right] \tag{4.29}$$

The closed-loop gain is therefore:

$$A_C = a_2(1-\beta)\left[\frac{a_1 + \alpha\gamma\eta V_i}{1 + a_2\beta\gamma\eta V_i}\right] \tag{4.30}$$

As $\gamma \to \infty$, the above expression simplifies to:

$$A_c = \frac{\alpha(1-\beta)}{\beta} \tag{4.31}$$

which is independent of γ (which may be nonlinear) and the nonlinear gain terms a_1 and a_2.

It is thus evident that a large value of modulator sensitivity will lead to small amounts of amplitude distortion at the amplifier output. The main limitation on the achievable linearity is that of the diode detectors, particularly at low feedback signal levels. Any nonlinearity in these detectors effectively appears as nonlinearity in the α and β terms in (4.31) above. This nonlinearity can be minimised by utilising the detectors in their quasi-linear regions, that is, utilising a high drive level.

The inherently greater linearity of FETs, together with their good efficiency, has led to their consideration for use in AM transmitters. Petrovic and Gosling [15] describe a simple AM transmitter, employing gate modulation and envelope feedback, for VHF/UHF applications. The schematic of this transmitter is given in Figure 4.17 and can be seen to be very simple. Feedback is again effected by simple envelope detection, although here the modulation function is also performed in this single-stage.

The advantage of gate modulation over drain (or collector) modulation is obvious. The modulating signal may be applied at a low signal level, in this case by a standard operational amplifier, and the inherent distortion of the stage may be removed by feedback. The alternative requirement for high-power audio stages is thus removed, and this leads to a greater overall efficiency. The efficiency of this 10 Watt PEP (peak envelope power) stage is quoted at 49% for a modulation depth of 96%. The raw CW efficiency of the device when operating in class-C is quoted at 56%; the complete AM

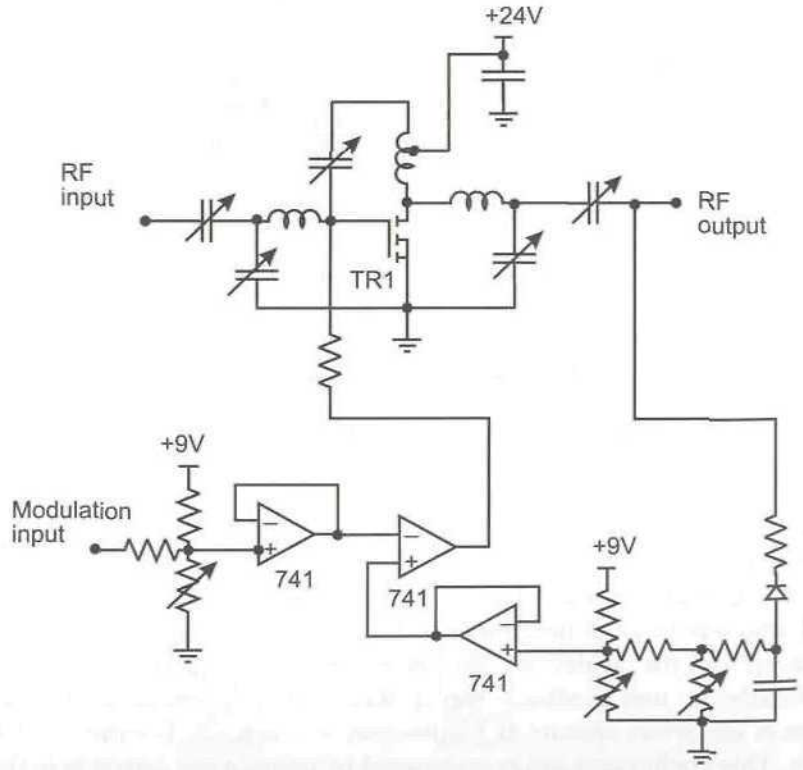


Figure 4.17 Gate-modulation VMOS transmitter (from [15] © IEE 1978).

transmitter is thus operating close to the maximum efficiency available from the device.

An added advantage of this technique is that of output power stabilisation. The transmitter output is virtually unaffected by slight variations in tuning, load, RF drive, or power supply voltage.

The simple non-coherent forms of envelope feedback described above laid the foundations for the consequent development of coherent envelope feedback systems which are able to compensate for amplitude and phase nonlinearities in a transmitter. The first of these schemes was proposed by Petrovic and is known as the *polar loop*.

4.5 Polar-Loop Transmitter

The polar-loop transmitter [16] is an extension of two previous linear amplifier schemes, namely: envelope feedback and envelope elimination and restoration (see Chapter 7). It overcomes some of their principal disadvantages and results in an extremely linear transmitter architecture.

A schematic of the transmitter is shown in Figure 4.18 (note again that this is a transmitter, rather than an amplifier, linearisation technique). The radio frequency section of the transmitter is extremely simple and consists of a voltage-controlled oscillator (VCO), operating at the final output frequency, and an RF amplifier stage (or chain). The final stage of the RF amplifier chain forms the amplitude modulator, for correction of the distortions introduced in previous stages (and its own). This aspect of operation is similar to that described for envelope feedback (Section 4.4).

The modulated output of the transmitter is attenuated, and converted to a convenient IF by the synthesizer and frequency converter (mixer). The resulting IF signal is then resolved into phase and amplitude (polar form) using a limiter and demodulator respectively. The demodulator utilises the output of the limiter to detect the unlimited input signal, hence performing a form of coherent amplitude detection. The output of the demodulator is compared with a similarly detected version of the input signal and the resulting error signal feeds the final stage modulator. This is therefore a basic envelope feedback system. The phase information (limiter output) is similarly compared with the phase of the input signal and the resulting error signal controls the VCO. The phase-control portion of the system is therefore a simple phase-locked loop. Thus, both the amplitude and phase of the transmitter's output signal can be carefully controlled within independent feedback loops.

The analysis of the system may be broken into two parts, both of which are already well understood. The amplitude correction loop is analytically

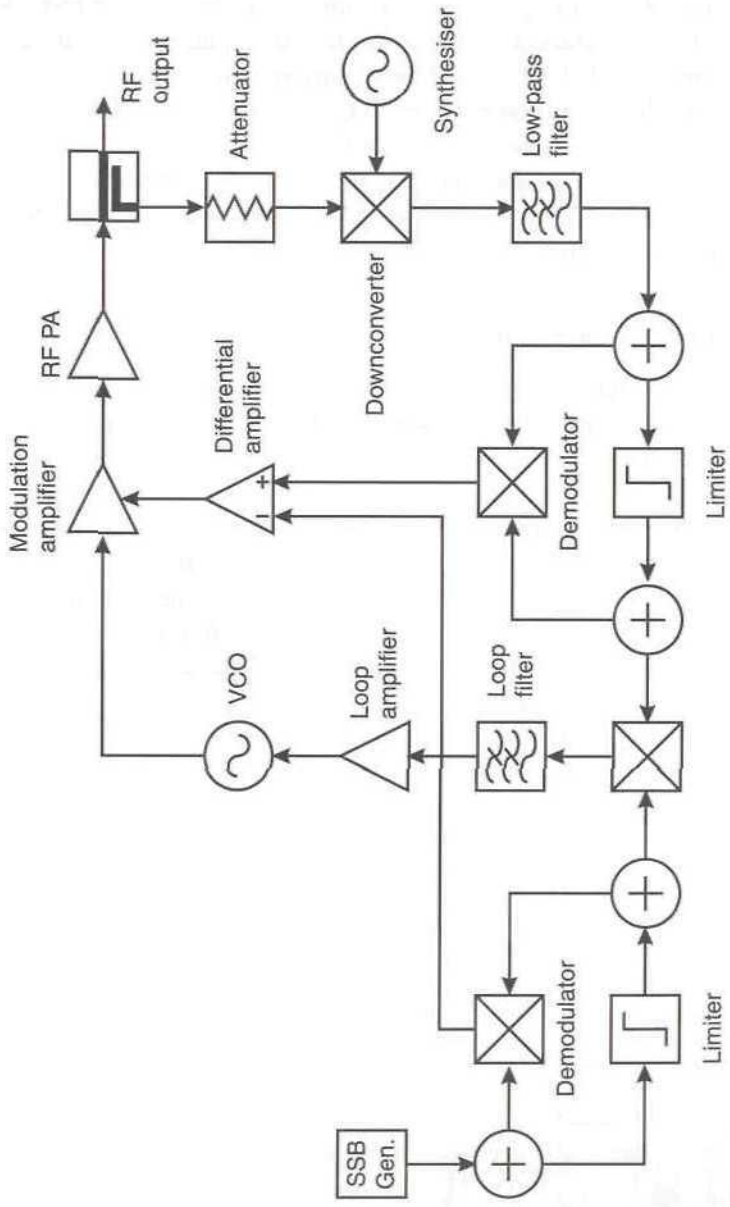


Figure 4.18 Polar-loop transmitter (from [16] © IEE 1979).

similar to the envelope feedback system described in Section 4.4, and its analysis proceeds in a similar manner. The phase correction loop may be analysed by phase-locked loop techniques [2] and suitable choices made for the loop filter, VCO constant, and loop amplifier. When considering two-tone test signals, the sharp phase discontinuities inherent in that type of signal mean that a first-order loop is, ideally, required.

These phase discontinuities result in a potentially wide feedback bandwidth being required for the phase-feedback loop. For similar reasons to those discussed above with respect to envelope feedback, the envelope feedback loop bandwidth can also be large. These high feedback bandwidth requirements are a major limiting factor in the performance of the polar loop technique.

Results for the basic polar-loop transmitter indicate that a two-tone third-order intermodulation performance of better than −50 dBc is possible, with relatively inexpensive components, over a typical SSB channel bandwidth.

The polar-loop transmitter has many advantages over earlier, incoherent, feedback linearisation schemes. Some of these advantages may be summarised as follows:

1. Since the VCO and the RF modulator (which can appear prior to the final amplifier stage) are included within the feedback loop their linearity performance is not critical. Low-cost modulators of virtually any type may therefore be utilised without compromising the final system performance.
2. High-efficiency class-C amplifiers may be utilised in the RF chain, creating a power-efficient linear transmitter.
3. The RF portion of the transmitter is very simple, as it contains only a VCO and high-efficiency power amplification all operating at the final output frequency.
4. No upconversion is employed in the RF chain and hence image-reject filtering is not required.
5. The use of gate modulation is permissible in the final stage, despite its inherently poor linearity, and hence a low-power modulating (differential) amplifier may be used.
6. The use of feedback means that the transmitter is insensitive to, for example, tuning, component aging, and supply voltage variations.
7. The technique is applicable to other linear modulation schemes, for example, AM, suppressed-carrier AM, as well as the various forms of SSB.

4.5.1 Improvements to the Polar Loop Technique

A slightly modified form of the basic polar loop was later described by Petrovic and Smith [17] and is shown in Figure 4.19. In this case the feedback modulation is applied earlier in the RF amplifier chain, to a low-level modulation amplifier. Two filters have also been added, one to remove transmitter harmonics at the output and a low-pass filter at IF to remove image frequencies before the limiter. The third-order IMD performance of this configuration is quoted as better than −60 dB relative to PEP for a two-tone test. This equates to the third-order products being at a level of −54 dB relative to the tones (−54 dBc).

A further modification of the polar loop technique has been applied to effect a complete HF transmitter [17]. The schematic of this configuration is shown in Figure 4.20.

In this case the feedback loop bandwidths have been considerably reduced, from around 200 kHz for the narrow-band (3 kHz) SSB transmitter described above to 50 kHz for the HF transmitter. The bandwidth reduction is necessary in order to ensure stability of the HF transmitter configuration. The feedback loops now only correct for the distortion introduced by the RF amplifiers alone, however, adequate performance at HF can still be achieved.

The transmitter shown in Figure 4.20 was capable of 100W PEP with intermodulation products some −66 dB relative to PEP for a two-tone test.

The restriction of the polar-loop transmitter to a limited range of linear modulation schemes (due primarily to bandwidth restrictions in the phase and envelope feedback paths) led to further development of the ideas behind it. As a result a general, feedback optimised linear transmitter configuration was conceived, which is capable of transmitting any form of linear modulation contained within an appropriate bandwidth—the *Cartesian loop*.

4.6 Cartesian-Loop Transmitter

The Cartesian loop technique [17,18] was first proposed by Petrovic in 1983 as a superior form of modulation feedback transmitter. It was primarily designed for SSB transmission, but has since been applied to many other linear and quasi-linear modulation schemes.

A block diagram of the basic Cartesian-loop transmitter is shown in Figure 4.21. The operational principle is similar to the polar loop described above, however, the baseband signal information is now processed in Cartesian (I and Q) form. The modulating signal is split into quadrature components by the broadband phase-shifting network, and these are fed into

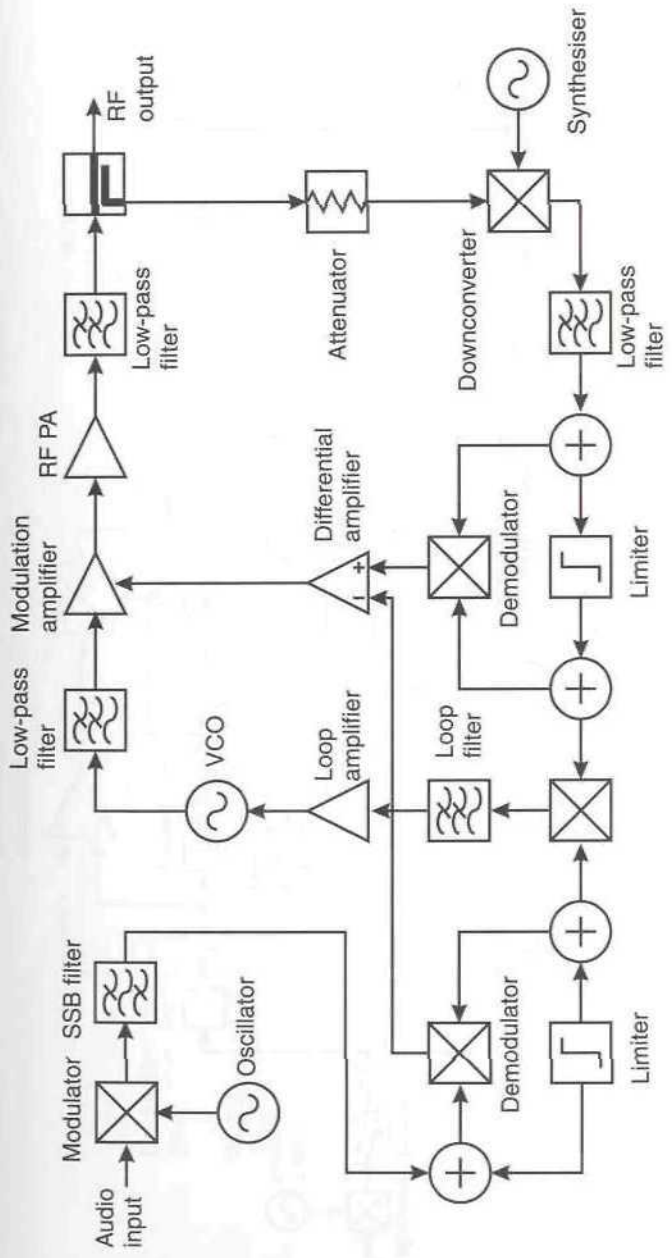


Figure 4.19 Modified polar-loop transmitter (from [17] © IEE 1984).

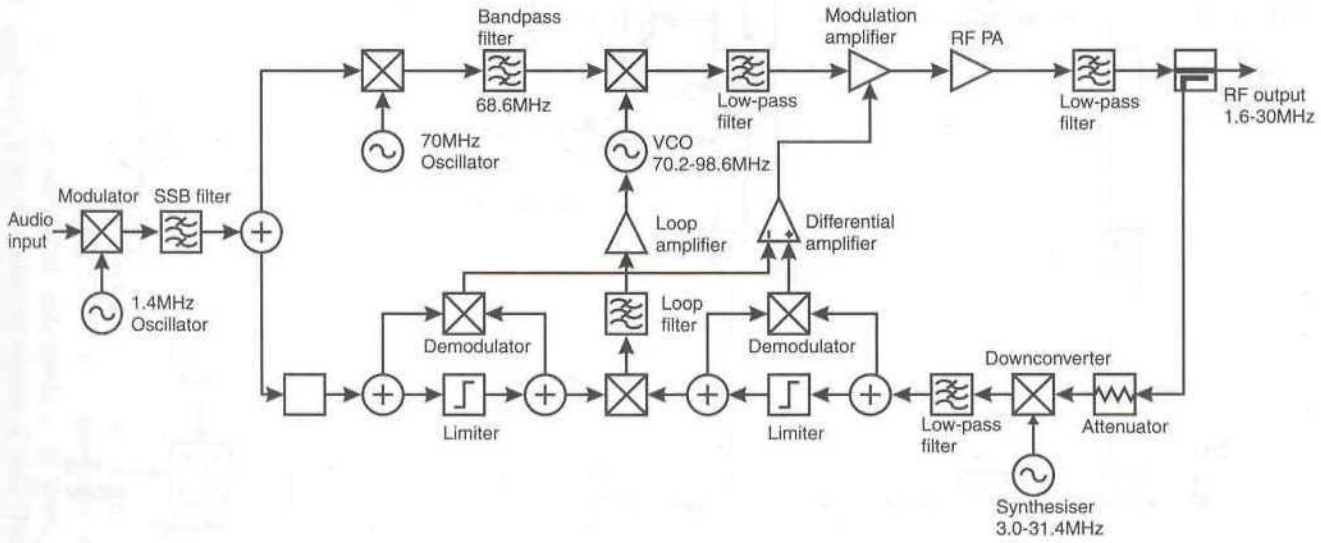


Figure 4.20 HF polar-loop transmitter (from [17] © IEE 1984).

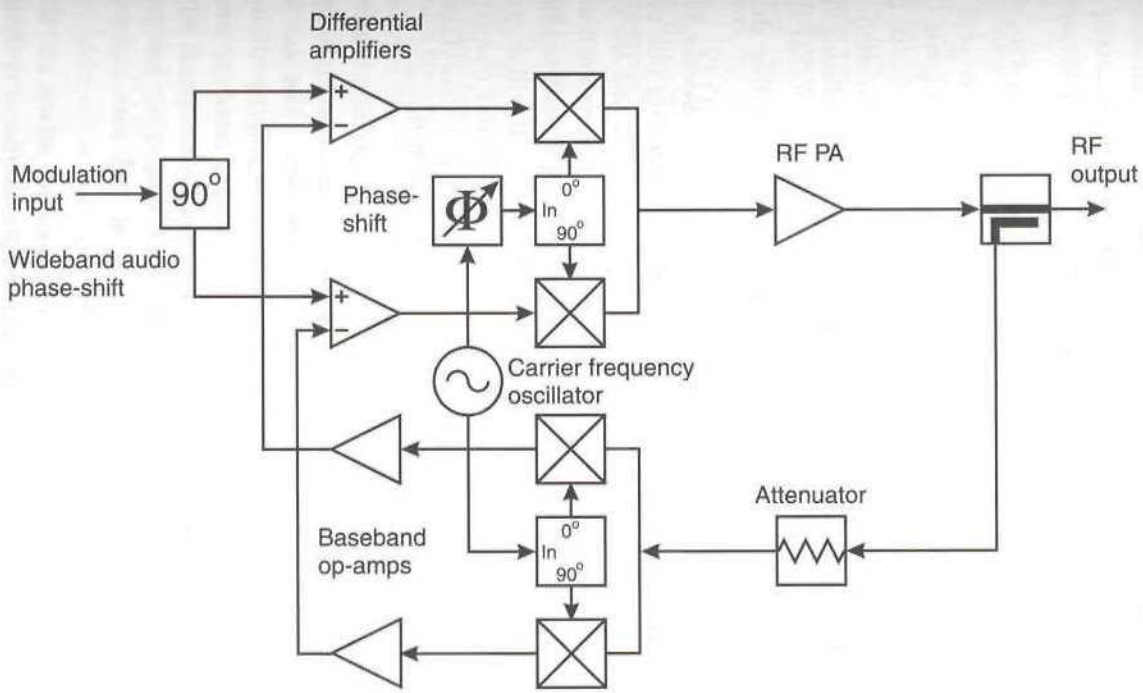


Figure 4.21 Cartesian-loop transmitter.

differential amplifiers which form the subtraction process in order to generate the error signals. The outputs of the differential amplifiers are upconverted to RF utilising a quadrature RF oscillator. The resultant RF signals from the two paths are then combined to produce the complex RF output signal. This low-power RF signal is then amplified by the nonlinear power amplifier before feeding the antenna.

The output from the RF amplifier is sampled by a directional coupler and attenuated to a suitable level to feed the downconversion mixers. These mixers are supplied with exactly the same local oscillator signals (appropriately phase-shifted) as were the upconversion mixers and hence the up- and downconversion processes are coherent. The downconverted output signal forms the feedback path to the differential amplifiers, closing the two loops. The orthogonal nature of the feedback system means that the two feedback paths operate completely independently, thus ensuring that both AM–AM and AM–PM characteristics are linearised.

A phase shift is required between the up- and downconversion processes and is provided in the local oscillator path feeding one or other of these components. The phase shift is adjusted to ensure that the up- and downconversion processes are correctly synchronised, despite the finite time delay of the RF power amplifier (and any IF processing, if an IF is employed within the loop). Any error in the setting of this phase shift will degrade the loop phase margin and this issue will be covered later in this chapter.

The operation of the Cartesian loop has a number of advantages over that of the polar loop and these may be summarised as follows:

1. Removal of the need for a dynamic PLL/VCO. The requirement for a fast PLL to track the rapid phase-changes which can occur in some linear signals is no longer a problem. The consequent tracking and phase-error problems are also eliminated. In addition, the PLL arrangement in the polar loop can have problems tracking or locking at low envelope levels, such as occur when the IQ vector passes through zero. This will result when using signals such as a two-tone test, an SSB signal and with some digital modulation formats (e.g., 16-QAM).
2. The modulation process is reduced to a simple mixer, and the need for a separate modulator at the final output frequency is eliminated.
3. Simplicity of implementation.
4. Applicable to any modulation scheme.
5. A standard hardware configuration results, which allows a flexible approach to the choice of modulation scheme.

6. Significant reduction (and equalisation) of the two feedback loop bandwidth requirements. This is a key benefit of the Cartesian loop technique over both envelope feedback and the polar loop. It is this benefit which largely explains the popularity of the Cartesian loop technique and the relative unpopularity of polar loop.

Petrovic described the application of the Cartesian loop technique to a practical HF transmitter operating from 1.6 MHz to 30 MHz [17]. A schematic of this transmitter is shown in Figure 4.22; the overall transmitter being rated at 10W PEP.

The modulation and combining process in this transmitter takes place at an intermediate frequency of 45 MHz; the phase-shifting network for the carrier frequency oscillator is thus only required to operate on a single, fixed frequency, which greatly simplifies its design. The IF signal is downconverted to the required carrier frequency by a variable frequency oscillator (VFO) operating between 46.6 MHz and 75 MHz. The additional filtering throughout the loop is employed to ensure that the correct mixer products are selected at each stage.

4.6.1 Phase-Shift Network Design

The carrier frequency phase-shift network can be fabricated from a quadrature hybrid based network, as shown in Figure 4.23. A number of other techniques are also applicable, such as the use of high-pass or low-pass filters, quadrature coils, or microstrip techniques (at higher frequencies).

The broadband phase-shift network for the audio frequency modulation inputs is somewhat more difficult to realise in analog hardware. The method suggested by Petrovic (Figure 4.24) is based on generating an SSB signal at 10.7 MHz, using the filter method, and then demodulating it with quadrature carriers. This method does ensure that quadrature is maintained to a high degree of accuracy over the entire audio band although it is complex and expensive in hardware. A particular disadvantage is the requirement for an SSB filter, as these tend to be bulky and expensive.

An alternative method is to use a DSP device to generate the modulation signals. Such devices allow high accuracy quadrature signals to be generated at audio frequencies by means of a Hilbert transform filter and this was the method employed by Bateman *et al.* [19], which is described below.

The feedback loop bandwidth of the HF transmitter was set to 100 kHz as a compromise between stability and correction bandwidth. It is therefore unable to remove harmonic distortion products, and an output filter is

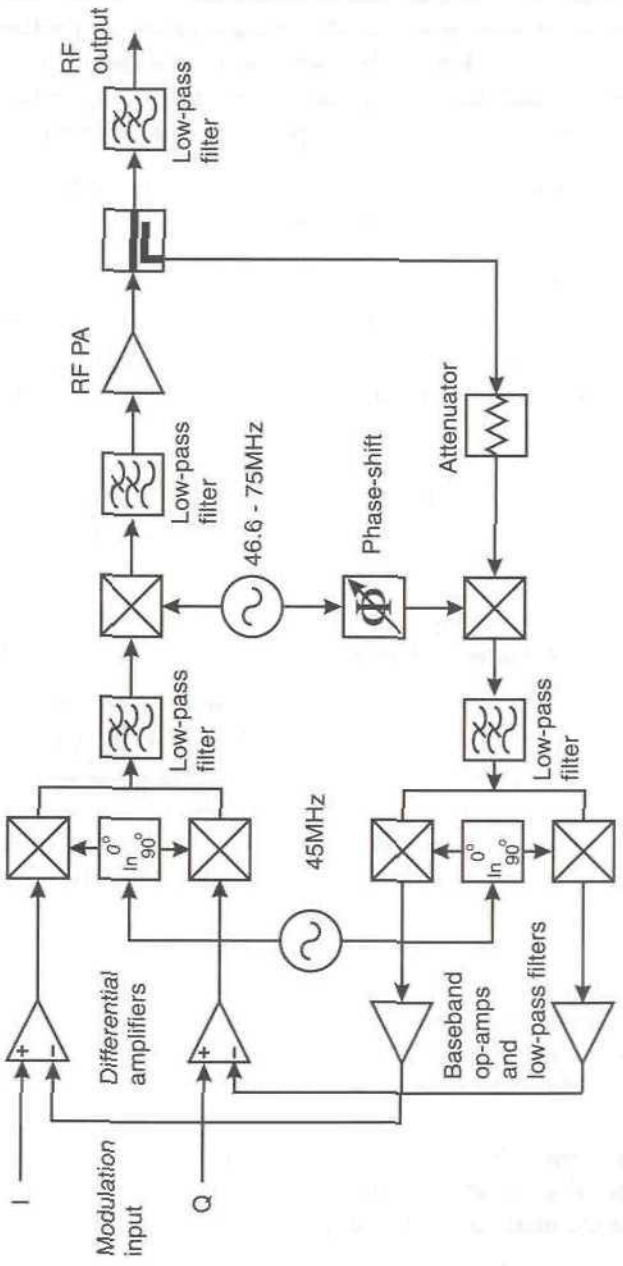


Figure 4.22 HF Cartesian, loop transmitter (from [17] © IEE 1984).

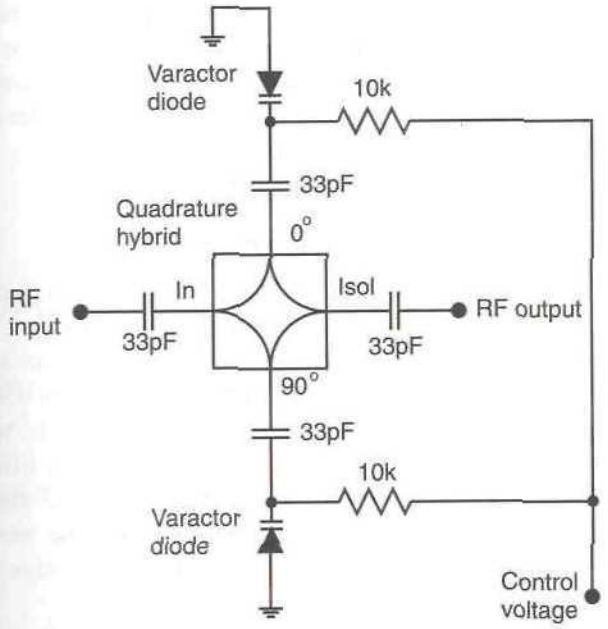


Figure 4.23 Carrier phase-shift network (coupling capacitor values shown are for 900 MHz operation).

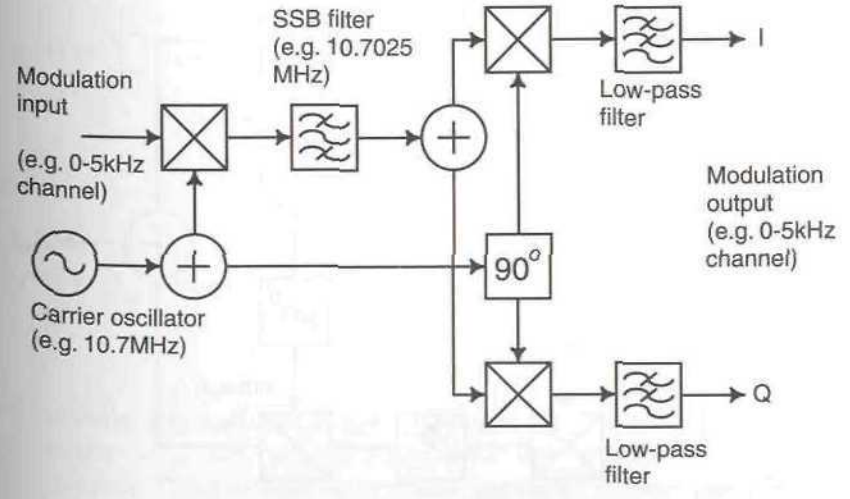


Figure 4.24 *Modulation phase-shift network (from [17] © IEE 1984).*

required. The reported suppression of the in-band products (intermodulation distortion) is very good. The third-order products from a two-tone test were some 65 dB down on the tones, which is just over 70 dB down on PEP; some 33 dB of suppression is therefore obtainable with the design, over a typical SSB channel bandwidth (3 to 5 kHz).

4.6.2 Weaver Method SSB and the Cartesian Loop

The Weaver method (or *third method*) of SSB generation was first proposed in 1956 [20] as an additional method of generating SSB without the requirement for a narrow-band crystal filter. The phasing method of SSB generation suffers from the fact that the remaining unsuppressed image band appears adjacent to the wanted frequency band. Although, in theory, this image should not exist, imperfect system components mean that it cannot entirely be eliminated. This is a particular problem in a mobile radio environment, as the adjacent channel performance must be very good—hence the need for highly linear amplification and an alternative method of SSB generation.

The principal advantage of the Weaver method is that the image channel falls within the band of the wanted channel and hence the suppression specification is greatly relaxed. A Weaver method SSB generator is shown in Figure 4.25. It is a direct conversion architecture and allows many of the system components to be implemented in a DSP device.

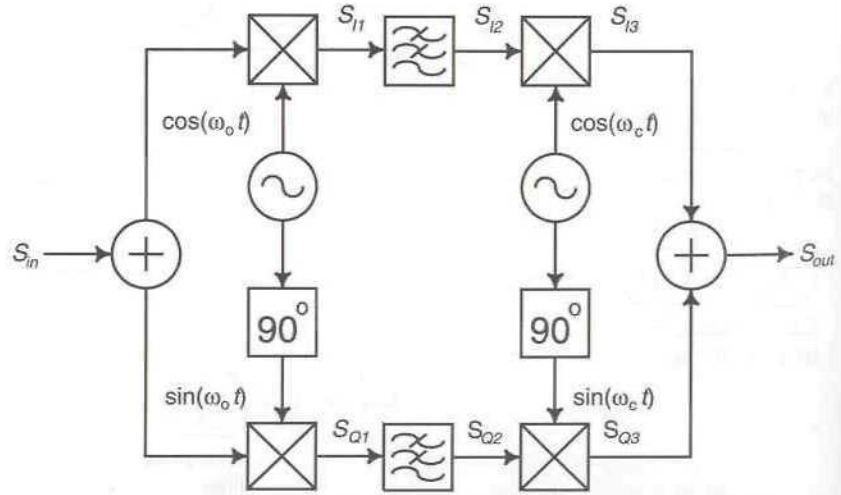


Figure 4.25 Weaver method SSB generator.

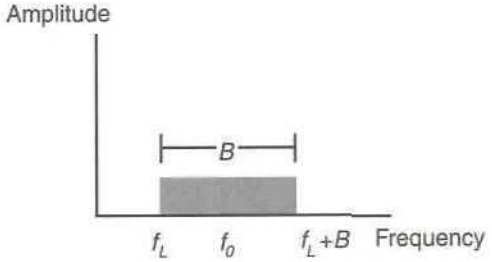


Figure 4.26 Baseband input signal spectrum.

The operation of a Weaver generator may be described, with reference to Figure 4.25, as follows: the audio input signal is restricted to a bandwidth, B, with a band centre frequency, f_0, and a lower limit, f_L, as shown in Figure 4.26. The input band may be considered as a summation of sinusoids:

$$s_{in}(t) = \sum_{n=1}^{N} E_n \cos(\omega_n t + \phi_n) \tag{4.32}$$

The audio input signal is mixed with a quadrature oscillator operating at half of the required modulation bandwidth. Two audio-frequency quadrature paths are thus formed where the audio band in each has been 'folded on top of itself', occupying half of the original bandwidth. The resulting signals are:

$$s_{I1}(t) = \sum_{n=1}^{N} E_n \cos[(\omega_n - \omega_0)t + \phi_n)] + \sum_{n=1}^{N} E_n \cos[(\omega_n + \omega_0)t + \phi_n)] \tag{4.33}$$

and

$$s_{Q1}(t) = \sum_{n=1}^{N} -E_n \sin[(\omega_n - \omega_0)t + \phi_n)] + \sum_{n=1}^{N} E_n \sin[(\omega_n + \omega_0)t + \phi_n)] \tag{4.34}$$

The resulting spectrum appears as shown in Figure 4.27; note the gap between the top of the required audio band and the bottom of the mixer products band. This provides a convenient region for the low-pass filter roll-off and will be 600 Hz wide for a 300 Hz to 3.4 kHz audio input spectrum.

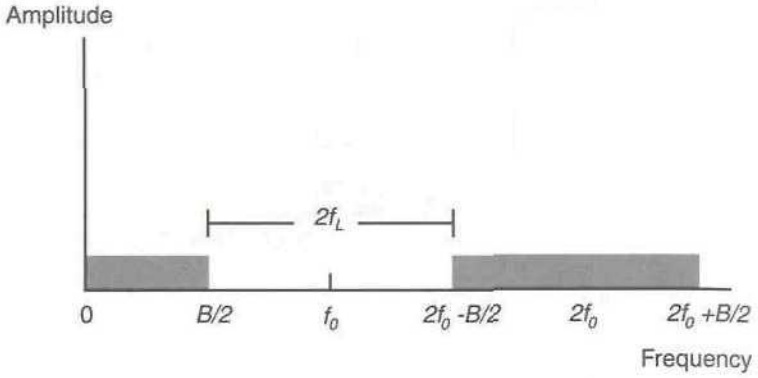


Figure 4.27 Signal spectrum at the output of the first balanced modulators.

The resulting, filtered, signals will be:

$$s_{I2}(t) = \sum_{n=1}^{N} E_n \cos[(\omega_n - \omega_0)t + \phi_n)] \tag{4.35}$$

and

$$s_{Q2}(t) = \sum_{n=1}^{N} -E_n \sin[(\omega_n - \omega_0)t + \phi_n)] \tag{4.36}$$

Each path is then upconverted to the final channel frequency by a quadrature local oscillator operating at the center of the channel, that is, not at what would be the carrier frequency in a conventional filter-based SSB system. The resulting RF output signals are therefore:

$$s_{I3}(t) = \sum_{n=1}^{N} \frac{E_n}{2} \cos[(\omega_c + \omega_n - \omega_0)t + \phi_n)] + \sum_{n=1}^{N} \frac{E_n}{2} \cos[(\omega_c - \omega_n + \omega_0)t - \phi_n)] \tag{4.37}$$

and

$$s_{Q3}(t) = \sum_{n=1}^{N} \frac{E_n}{2} \cos[(\omega_c + \omega_n - \omega_0)t + \phi_n)] - \sum_{n=1}^{N} \frac{E_n}{2} \cos[(\omega_c - \omega_n + \omega_0)t - \phi_n)] \tag{4.38}$$

The two paths are then summed to produce an SSB channel in which the image from the final upconversion process appears in-band and the suppression of which is mainly governed by the quadrature accuracy of the oscillator and the leakages involved in the RF summing junction:

$$s_{out} = s_{I3} + s_{Q3} \tag{4.39}$$

Hence

$$s_{out} = \sum_{n=1}^{N} E_n \cos[(\omega_c + \omega_n - \omega_0)t + \phi_n] \tag{4.40}$$

The two quadrature audio frequency paths, created by use of the Weaver method, lend themselves rather neatly to application in a Cartesian-loop transmitter [21]. The schematic of such a system is shown in Figure 4.28 [22] where differential amplifiers and a Cartesian feedback loop have been added to the basic Weaver transmitter.

The Weaver–Cartesian-loop transmitter has proved very successful in implementing a TTIB (transparent tone-in-band) based SSB transmitter [22]. The format of a TTIB signal is shown in Figure 4.29 and it can be seen that the pilot tone in the center of the band can be made to coincide with the upconversion oscillator position in the Weaver upconversion process. The requirement to suppress the feedthrough and leakage of this oscillator to a high degree is thus removed, as acceptable levels of performance may be obtained with more realistic levels of feedthrough suppression.

The audio components of the system, to the left of the differential amplifiers in Figure 4.28, may be implemented in a DSP device and thus audio quadrature can be maintained to a very high degree of accuracy. The RF power amplifier may be a class-C device and is more often composed of a cascade of stages, the final one of which will be class-C. The technique is equally applicable to class-A and class-AB stages, and a consequent improvement in performance will be obtained; however, this will be at the expense of size and efficiency.

It is useful, at this point, to draw the distinction between the *linearising* and *operational* bandwidths of a linear transmitter. The transmitter is capable of performing linearisation over a certain channel bandwidth, which is determined by the bandwidth of the feedback loop. This is termed the *linearising bandwidth* and will obviously depend on the gain employed in the feedback loop, as well as its bandwidth. A practical limit on this bandwidth is in the region of a few hundreds of kHz (for high levels of loop gain and a standard RF power module), with typical systems employing channel

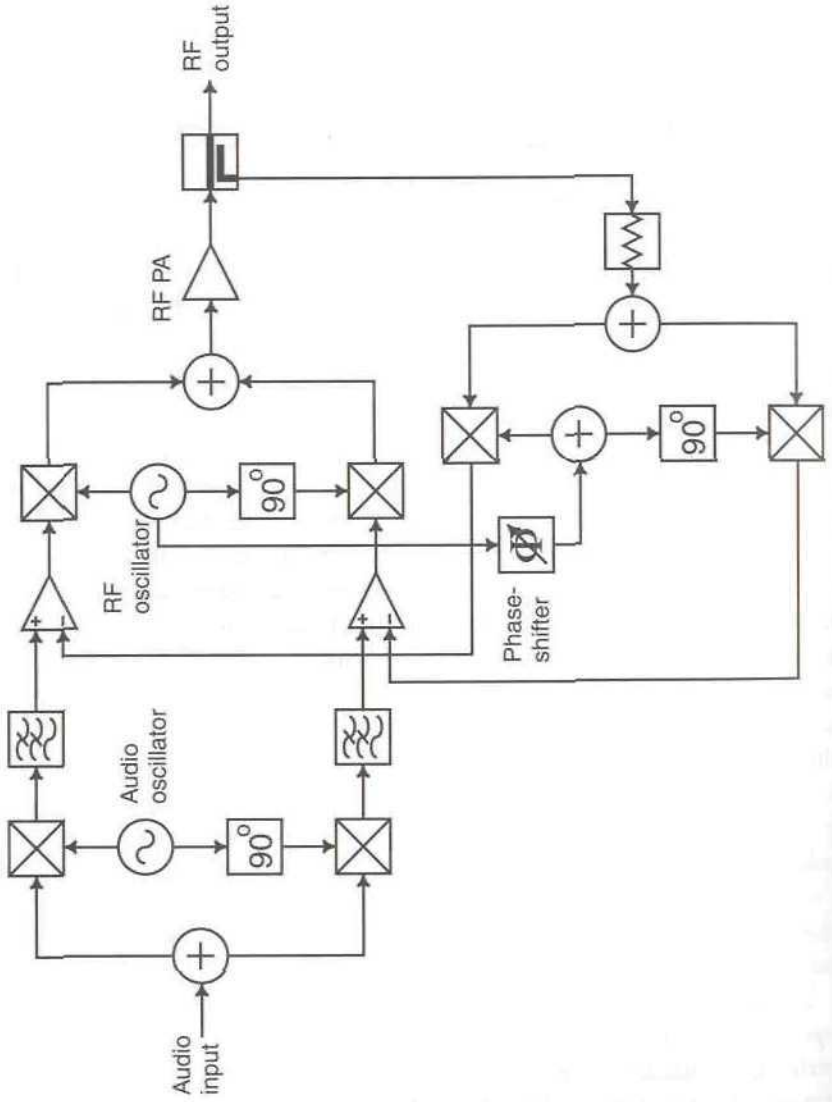


Figure 4.28 Cartesian loop techniques applied to a Weaver transmitter.

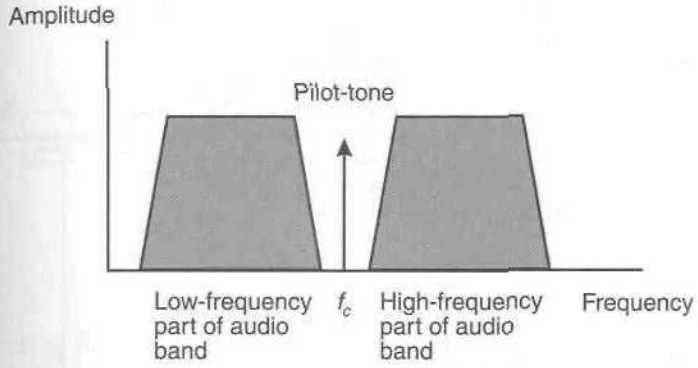


Figure 4.29 Spectrum of a TTIB channel.

bandwidths of 30 kHz (= 6 × 5 kHz SSB channels or 1 × DAMPS channels) or 5 kHz (for a single TTIB/SSB channel). Higher linearisation bandwidths are possible with very low delay power amplifiers, such as integrated circuit PAs found in handsets or broadband MMIC amplifiers, which can have delays of less than 1 ns.

The *operational bandwidth* is defined by the circuit components, that is, the bandwidth of the power amplifier chain, the quadrature bandwidth of the local oscillators, and the phase-shift network. It is the bandwidth within which the linearising bandwidth can appear whilst still maintaining acceptable performance; this may be several tens of megahertz for a typical design. For example, a typical transmitter operating in a mobile radio system utilising SSB modulation would have a channel bandwidth of 5 kHz (and a consequent linearisation bandwidth of perhaps 3–5 times this figure) and an operational bandwidth of 30 MHz, that is, the 5 kHz linear channel could appear anywhere in the 30 MHz spectrum allocation and (importantly) could be reallocated simply be re-programming the channel synthesiser.

4.6.3 Analysis of a Cartesian-Loop Transmitter

Consider the Cartesian loop schematic shown in Figure 4.30. The loop filters and high-gain baseband amplifiers have now been included, along with the required symbology.

The open-loop output signal due to an I-channel input signal, $S_I(s)$, may be written as:

$$S_{out}(s) = F_1(s)M_1(s)G_A(s)S_I(s) + D(s) \qquad (4.41)$$

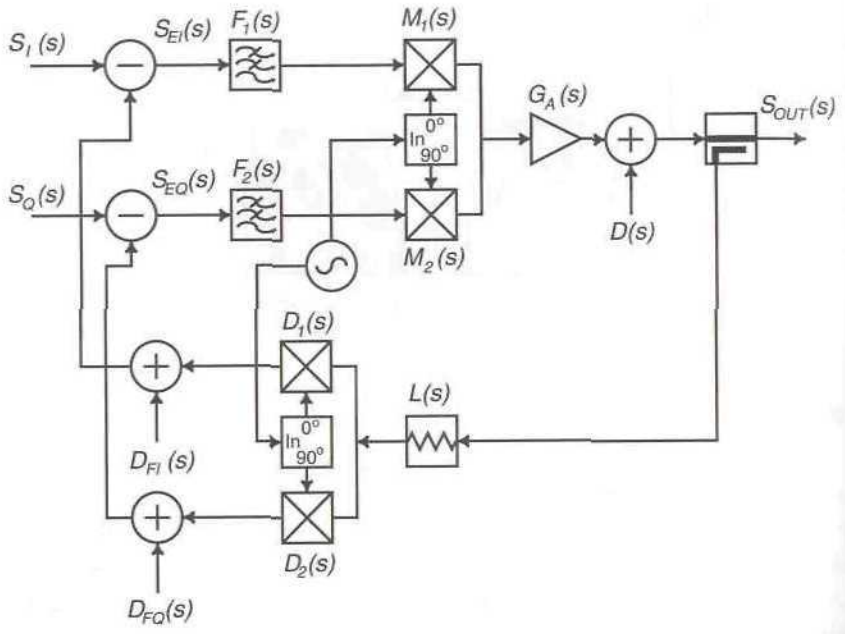


Figure 4.30 Analysis diagram of a Cartesian-loop transmitter.

where $D(s)$ is a term representing the distortion added by the power amplifier and the other terms are the various component gains, as indicated in the diagram. The situation for a Q-channel input signal is completely analogous and will be considered at the end.

When the loop is closed, the above expression becomes:

$$S_{out}(s) = F_1(s)M_1(s)G_A(s)S_{EI}(s) + D(s) \tag{4.42}$$

where S_{EI} is the I-channel error signal term, and it, in turn, is given by:

$$S_{EI}(s) = S_I(s) - [L(s)D_1(s)S_{out}(s) + D_{FI}(s)] \tag{4.43}$$

The term, $D_{FI}(s)$, represents the distortion present in the feedback loop and this is assumed to be added at the output of the downconversion mixer. Experience has shown that the distortion in these components is the most significant (of the feedback components) in determining the overall performance.

Combining (4.42) and (4.43) gives the closed-loop output signal for the

transmitter with an I-channel input signal.

$$\begin{aligned} S_{out}(s) = {} & \frac{F_1(s)M_1(s)G_A(s)S_I(s)}{1 + F_1(s)M_1(s)G_A(s)L(s)D_1(s)} \\ & - \frac{F_1(s)M_1(s)G_A(s)D_{FI}(s)}{1 + F_1(s)M_1(s)G_A(s)L(s)D_1(s)} \\ & + \frac{D(s)}{1 + F_1(s)M_1(s)G_A(s)L(s)D_1(s)} \end{aligned} \tag{4.44}$$

Substituting for:

$$A_I = F_1(s)M_1(s)G_A(s) \tag{4.45}$$

and

$$\beta_I = L(s)D_1(s) \tag{4.46}$$

and assuming that the loop gain, $A_I\beta_I >> 1$, the closed loop output signal becomes:

$$S_{out}(s) \approx \frac{S_I(s)}{\beta_I} + \frac{D(s)}{A_I\beta_I} - \frac{D_{FI}(s)}{\beta_I} \tag{4.47}$$

The forward path distortion, $D(s)$, which predominantly consists of the distortion introduced by the RF power amplifier, may therefore be suppressed to an arbitrarily large extent, commensurate with the required gain and phase margins. It is important to note that the same is not true of the distortion introduced in the feedback path, generally by the nonlinearities in the downconversion mixers. These mixers must therefore be carefully chosen to have a high third-order intercept point, and must also be driven at a suitably low level.

Equation (4.47) is similar to the expression derived by Johansson and Mattsson [23,24], with the exception of a different assumed position for the added distortion in the feedback loop. They also note that a similar situation exists for noise added within the loop, and note that low-noise components must be utilised in the feedback path in order to maintain an acceptably low-noise system (see Section 4.7).

4.6.4 Loop-Filter Design

The two loop filters, one each in the I and Q baseband channels, are usually low-pass designs and are a compromise between loop stability and correction (or linearisation) bandwidth. The optimum design of these filters requires their bandwidth to be just sufficiently wide to permit all of the channel information to pass unhindered, with a roll-off just gradual enough to ensure that the IMD products are suppressed to a uniformly low level. There is, for example, little point in making the filter bandwidth wide enough to ensure the ninth-order products receive the same 30 dB suppression (say) as the third-order products, since the ninth-order products are likely to be considerably smaller to start with.

Consider a Cartesian-loop transmitter, with negligible distortion in the feedback path. The output signal (from (4.44)) is then:

$$S_{out}(s) = \frac{F_1(s)M_1(s)G_A(s)S_I(s)}{1 + F_1(s)M_1(s)G_A(s)L(s)D_1(s)} + \frac{D(s)}{1 + F_1(s)M_1(s)G_A(s)L(s)D_1(s)} \tag{4.48}$$

If it is assumed that all of the loop components, with the exception of the loop filter, are broadband, then this equation becomes:

$$S_{out}(s) = \frac{M_1 G_A S_I(s) F_1(s)}{1 + M_1 G_A L D_1 F_1(s)} + \frac{D(s)}{1 + M_1 G_A L D_1 F_1(s)} \tag{4.49}$$

The first term in the above expression indicates that any input signal within the bandwidth of the loop filter will appear amplified at the output. Channel filtering is therefore better performed on the input signals and not by the loop filters, as these will be selected for optimum distortion suppression and not for optimum channel shaping.

Concentrating on the second term of (4.49), it is evident that distortion suppression is critically dependent upon the loop filter bandwidth and roll-off characteristic. The residual distortion appearing at the output is therefore:

$$D_{out}(s) = \frac{D(s)}{1 + M_1 G_A L D_1 F_1(s)} \tag{4.50}$$

This equation allows prediction, to a first approximation, of the IMD suppression achievable in a given transmitter with a given loop filter characteristic.

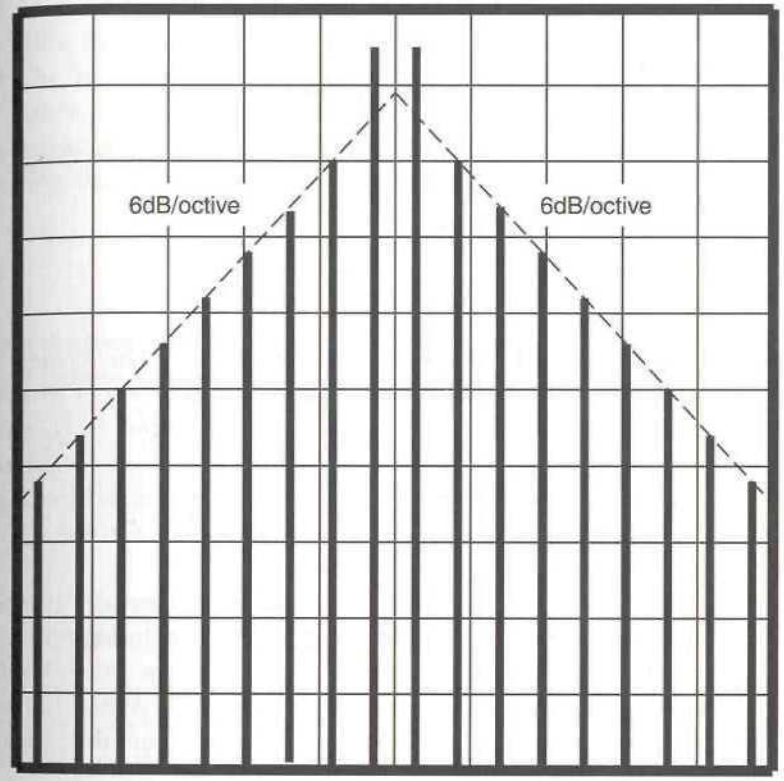


Figure 4.31 Approximate (idealised) form of output spectrum from a class-C amplifier; vertical scale 10 dB/division. Note that the 'octave' referred to is relative to the envelope frequency of the two-tone test, not to the RF centre frequency.

As an example, consider a two-tone test, with the tone spacing set to the desired channel bandwidth, $\Delta\omega$, applied to a class-C amplifier. The resultant, uncorrected, output spectrum would resemble that shown in Figure 4.31.

The IMD sidebands will, for the present, be assumed symmetrical and can be approximated by a single-pole roll-off (6 dB/octave of audio bandwidth from the channel centre) as shown. Note that this is purely arbitrary, and is chosen for the purposes of this illustration; a given, practical, class-C stage may exhibit a very different characteristic.

The frequency characteristic of the distortion (when downconverted to baseband) may therefore be approximated by a single-pole low-pass filter:

$$D(j\omega) = \frac{D_{max}}{1 + j\omega(\Delta\omega/2)} \tag{4.51}$$

Hence:

$$D_{out}(j\omega) = \frac{D_{\max}}{M_1 G_A L D_1 F_1(j\omega)\{1 + j\omega(\Delta\omega/2)\}} \quad (4.52)$$

If $F_1(j\omega)$, therefore, is a single-pole low-pass filter, with $CR = \Delta\omega/2$, then the resulting output distortion is:

$$D_{out}(j\omega) = \frac{D_{\max}}{M_1 G_A L D_1} \quad (4.53)$$

which is a constant level, independent of frequency. The system gain components may therefore be set to the required values in order to achieve the target specification. The filter applied above may therefore be considered to be an optimum filter for the particular amplifier distortion characteristic assumed at the outset: the loop bandwidth has been minimised for a given degree of correction and the roll-off has been optimised for the amplifier distortion characteristic.

All of the arguments used above may be applied equally to the Q-channel, yielding similar results and drawing similar conclusions.

In the case of an asymmetric IMD characteristic, the filter should be matched to the side having the greater IMD power, as it is this side which will determine the minimum level of performance for the linearised amplifier.

Equation (4.50) may also be used to calculate the level of suppression which can be expected at a given frequency away from the channel center frequency, if the optimum filter-type cannot be used or is physically unrealisable.

In the case of a first-order loop, where:

$$F_1(j\omega) = \frac{\omega_n}{\omega_n + j\omega} \quad (4.54)$$

the distortion attenuation factor at relatively low frequencies ($\omega << AB\omega_n$), where the loop gain is much greater than unity, is given by:

$$D_{atten}(\omega) \approx \frac{AB\omega_n}{\sqrt{\omega_n^2 + \omega^2}} \quad (4.55)$$

where A is the composite gain in the forward path and B is the composite gain in the feedback path. The attenuation level can be maximised by maximising the gain bandwidth product, $AB\omega_n$, and minimising the filter

3 dB bandwidth, ω_n. The maximum value of gain bandwidth product for a given time delay may be found from (4.81) in Section 4.6.5.

The maximum attenuation level is achieved when the loop gain approaches infinity (i.e., the loop filter bandwidth approaches zero). The value of this maximum attenuation is frequency dependent and can be found by setting the loop bandwidth, ω_n, to zero in (4.55):

$$D_{atten,\max}(\omega) \approx \frac{AB\omega_n}{\omega} \quad (4.56)$$

Setting the loop filter bandwidth, ω_n, to zero implies that the optimum loop filter is therefore an integrator, with the resultant loop being first-order.

Note that the above discussion concentrates on the optimum form of loop filter from the perspective of distortion reduction and not that of loop stability. It is usually loop stability limits which dictate the loop filter design and hence the criteria outlined in Section 4.6.5 usually prevail in a typical design.

4.6.5 Cartesian Loop Stability Analysis

The above arguments have assumed that the desired level of loop gain is achievable with the required loop filter characteristic, and no account has been taken of delays around the loop. In practice, the ideal loop filter characteristic may well not be achievable with a suitable loop gain to provide optimum suppression. For this reason, the majority of commercially-produced Cartesian loop designs utilise a first-order loop, designed primarily around loop stability criteria.

It is instructive, therefore, to examine the characteristics of a first-order loop in detail and particularly methods of determining the maximum achievable loop gain whilst still maintaining stability. This in turn will determine the maximum level of IMD suppression which can be achieved for a particular loop.

For the purposes of analysis, a first-order Cartesian loop may be approximated by the components shown in Figure 4.32. Only a single feedback channel is shown, with both channels assumed to be identical. In practice, the loop delay, τ, will be distributed throughout the loop, with the main amplifier providing a significant proportion. Any RF filtering present (e.g., in the upconversion path) will also contribute a significant delay. For this reason, it is preferable to place RF harmonic filtering after the loop, rather than immediately following the main amplifier, although local oscillator harmonics can result in unwanted downconversion of PA

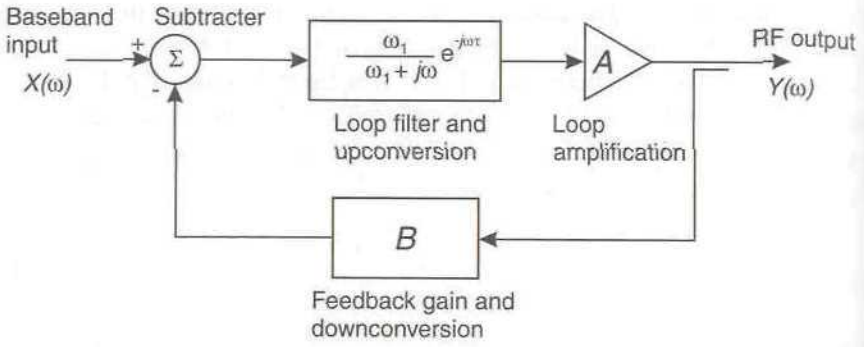


Figure 4.32 Equivalent circuit of a first-order Cartesian loop (one channel only).

harmonics with a consequently detrimental effect on performance. This also explains why it is undesirable to incorporate additional upconversion within the loop (i.e., employ an IF within the loop), since the inevitable filtering associated with this process will add significantly to the loop delay.

The up- and downconversion processes are assumed to be transparent for the purposes of analysis, as they will have no practical effect on the stability. Any delay introduced by these processes is absorbed in the overall delay term, τ.

The subtraction process occurs in the loop differential amplifiers, which are assumed to have infinite gain (and therefore to be perfect). The first-order loop 'filter' will typically be configured as an integrator which will usually have a relatively small timeconstant with respect to the channel bandwidth. The loop amplifier, A, is assumed to incorporate all of the gain in the loop following the subtraction process, up to the output sampling coupler. The feedback gain, B, includes the coupling loss in the sampling coupler, the loss in the downconversion attenuators and the gain of the baseband pre-amplification.

The forward path transfer function is therefore:

$$F(\omega) = \frac{A\omega_1}{\omega_1 + j\omega} e^{-j\omega\tau} \tag{4.57}$$

where ω_1 is the angular frequency of the open-loop 3 dB point. This is determined by the integrator timeconstant:

$$\omega_1 = \frac{2\pi}{R_I C_I} \tag{4.58}$$

where R_I and C_I are the integrator resistor and capacitor values respectively.

The open-loop transfer function is then:

$$G(\omega) = \frac{AB\omega_1}{\omega_1 + j\omega} e^{-j\omega\tau} \tag{4.59}$$

The complete transfer function for the loop is now:

$$H(\omega) = \frac{Y(\omega)}{X(\omega)} = \frac{A\omega_1 e^{-j\omega\tau}}{\omega_1 + j\omega + AB\omega_1 e^{-j\omega\tau}} \tag{4.60}$$

The stability of the loop may be analysed by means of a Nyquist diagram, with the gain and phase margins deduced by the same means. It is important to note, however, that the various gains and attenuation factors in the loop must be expressed in a consistent manner. Thus, for example, the gains and losses in the RF parts of the system should be expressed in voltage terms (rather than the more common power format) in order to maintain consistency with the baseband voltage gains in that part of the loop.

As an example, consider the loop configuration shown in Figure 4.33. This is a simplified Cartesian loop, with I and Q inputs at baseband and some

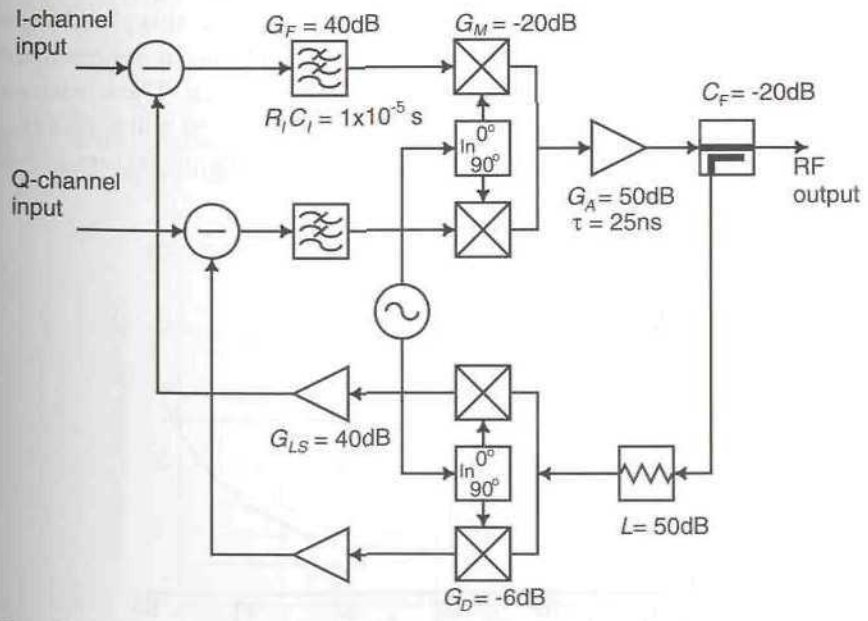


Figure 4.33 Example of a Cartesian loop system incorporating some typical values of gain in the loop. The filters and up- or downconverters are assumed to have 0 dB gain.

typical gain and attenuation values shown for the various parts of the loop (the modulator 'loss' assumes an active upconverter and is based on the input and output *voltages*). The loop time delay is 25 ns and the loop dominant pole, $\omega_1 = 628$ krad/s. This is a reasonable figure for a system channel bandwidth in the vicinity of 25 kHz, depending upon the IMD characteristics of the RF power amplifier. Using these values yields a loop (voltage) gain of 49.6, which, ideally, will yield an IMD power suppression in the region of 34 dB. Thus, if the RF power amplifier had an uncompensated two-tone IMD level of 26 dBc, the complete linearised transmitter should achieve an IMD level of 60 dBc (a better performance would be achieved with π/4-DQPSK, as a better starting point would be achieved—see Figure 4.41). This assumes, of course, that no significant distortion is introduced in the downconversion chain, prior to the differential amplifier. It is often the performance of these components (particularly the I/Q downconverter itself) which limits the overall performance of the transmitter.

If these values are incorporated into (4.59) (giving $A = 3200$ and $B = 0.016$), the resulting Nyquist diagram is that shown in Figure 4.34. It can be seen from this diagram that the loop is stable, but by a relatively small margin; the Nyquist locus does not encircle the (−1,0) point.

The frequency response characteristic shown in Figure 4.35 also indicates that the loop is close to instability. The peak in this characteristic at around 9 MHz will appear as a pair of noise sidebands at approximately 9 MHz either side of the carrier in the output RF spectrum. These sidebands are a tell-tale sign that the loop is operating at its limits and will probably go unstable given the slightest provocation. Such provoking factors include

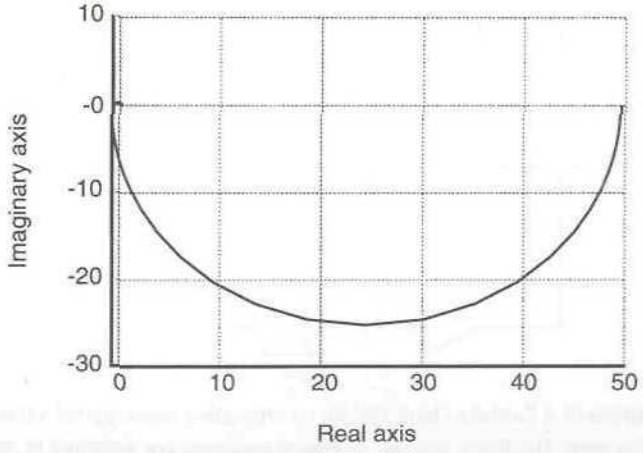


Figure 4.34 Nyquist diagram for the example Cartesian loop.

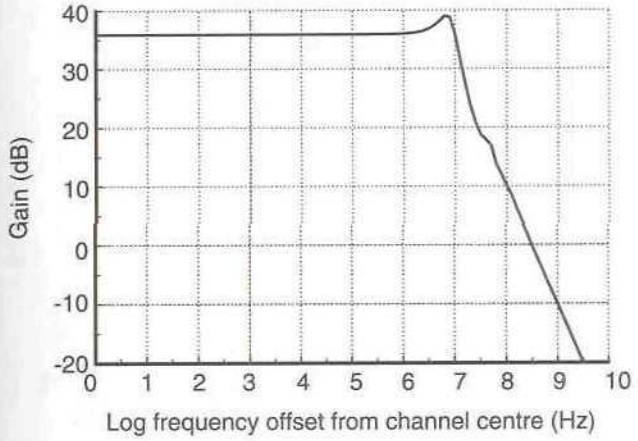


Figure 4.35 Frequency response of the example Cartesian loop.

temperature variations (usually resulting in loop gain variations), poor VSWR antenna loading and alterations in loop phase due to changes in power level (e.g., due to the presence of power control in the system).

The intermodulation suppression characteristic is given by:

$$I(\omega) = \frac{1}{1 + AB\omega_1 e^{-j\omega\tau}} \tag{4.61}$$

A plot of this characteristic for the example given above is shown in Figure 4.36. This shows the IMD cancellation of which the loop is capable for a

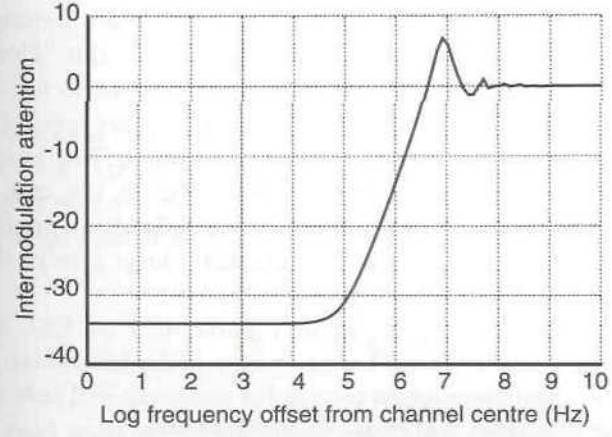


Figure 4.36 Intermodulation suppression characteristic for the example Cartesian loop.

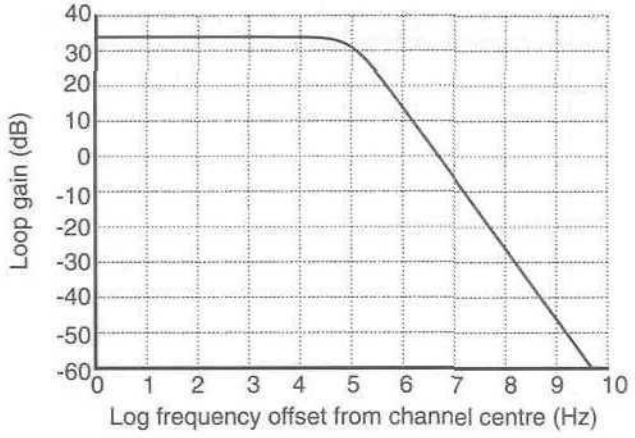


Figure 4.37 Loop gain characteristic of the example Cartesian loop.

range of frequency offset values from the channel frequency (both above and below). Note the 'peak' in the characteristic around 9 MHz; this indicates that a Cartesian loop will not only produce no benefit at this frequency offset, but will actually enhance (slightly) any intermodulation present at this frequency offset. Fortunately, most amplifiers intended for a 25 kHz channel bandwidth (for example) will not produce a significant level of intermodulation at this large offset.

Finally, a plot of the loop gain characteristic for the example loop is given in Figure 4.37. Note the difference between the open- (Figure 4.37) and closed-loop (Figure 4.35) bandwidths (approximately two orders of magnitude).

Increasing the power gain of the PA by a 10 dB, whilst leaving all other parameters unchanged, will result in the loop going unstable. This situation is illustrated by Figure 4.38. The Nyquist locus now encircles the (−1,0) point (just) and hence the loop will oscillate at a frequency approximating to that which brings the locus closest to the (−1,0) point. This latter point is only true when the loop is marginally unstable; in the case of gross instability, the frequency of oscillation is more difficult to predict (and is usually only of academic interest). The (unstable) loop gain in this case is 44 dB.

The example system presented in Figures 4.34 to 4.37 is a useful illustration of the behavior of the Cartesian loop under ideal conditions. It is generally a good approximation to real life for relatively well behaved power amplifiers, such as class-A and class-AB designs (since these have predominantly AM/AM nonlinear characteristics) [25]. The situation is greatly

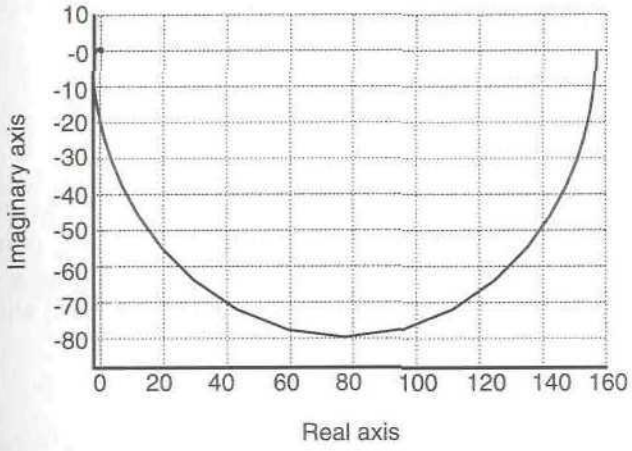


Figure 4.38 Nyquist diagram showing the effect of increasing the PA power gain by 10 dB for the example Cartesian loop.

complicated if a class-C amplifier is incorporated in the loop, and the above analysis will yield a very optimistic value of the achievable loop gain (and hence of the achievable IMD suppression). This situation is outlined further below.

4.6.5.1 Gain and Phase Margins

Using the above expression for the open-loop transfer function of a first-order system (4.59), the gain and phase margins may be deduced from:

$$GM = \frac{1}{|G(\omega_\pi)|} \tag{4.62}$$

$$PM = \angle G(\omega_{GU}) + \pi \tag{4.63}$$

where ω_π is the angular frequency at which the loop phase becomes π radians (the phase crossover frequency) and ω_{GU} is the angular frequency at which the loop gain reduces to unity.

The magnitude and phase of the open-loop transfer function are:

$$|G(\omega)| = \frac{AB\omega_1}{\sqrt{\omega_1^2 + \omega^2}} \tag{4.64}$$

$$\angle G(\omega) = -\left[\tan^{-1}\left(\frac{\omega}{\omega_1}\right) + \omega\tau\right] \tag{4.65}$$

The phase crossover frequency may now be found by setting $\angle G(\omega) = \pi$ in the above equation, giving:

$$\tan(\omega\tau) = -\left(\frac{\omega}{\omega_1}\right) \tag{4.66}$$

The tangent function above is periodic, with asymptotes occurring at $\omega = \pi/2\tau$. It is therefore necessary to ascertain the first crossing point of this function with the straight line $-\omega/\omega_1$. Since this line has a negative slope and passes through the origin, the first intercept point with $\tan(\omega\tau)$ will occur in the range:

$$\frac{\pi}{2\tau} < \omega < \frac{\pi}{\tau} \tag{4.67}$$

Equation (4.66) may be solved by means of a Taylor series expansion of $\tan(\omega\tau)$ about $\pi/2$:

$$\tan(\omega\tau) \approx -\frac{2}{2\omega\tau - \pi} + \frac{2\omega\tau - \pi}{6} + O\left[\omega\tau - \frac{\pi}{2}\right]^3 \tag{4.68}$$

Combining (4.66) and (4.68) and simplifying gives:

$$\omega^2\left(12\tau + 4\omega_1\tau^2\right) - \omega(6\pi + 4\pi\omega_1\tau) + \left(\pi^2 - 12\right)\omega_1 \approx 0 \tag{4.69}$$

For small $\omega_1\tau$, this has solutions:

$$\omega_\pi \approx \frac{\pi}{2\tau} \tag{4.70}$$

or

$$\omega_\pi \approx \frac{\pi\omega_1\tau}{2\tau(\omega_1\tau + 3)} \tag{4.71}$$

The latter solution may be neglected as it falls outside of the range defined by (4.67), and hence the phase crossover frequency is defined by (4.70).

The gain margin for the loop may now be found from (4.62), (4.64) and (4.70), giving:

$$GM \approx \frac{\sqrt{\omega_1^2 + \pi^2/4\tau^2}}{AB\omega_1} \tag{4.72}$$

For small values of $\omega_1\tau$ such that $\omega_1\tau << \pi/2$, the above equation simplifies to:

$$GM \approx \frac{\pi}{2AB\omega_1\tau} \tag{4.73}$$

For the example system detailed in the previous section (Figure 4.33), this calculates to 6.1 dB. It will be less than 0 dB for an unstable system.

Finally, since the gain margin of a stable loop must be greater than unity, the maximum possible value of stable loop gain is given by:

$$(AB)_{\max} \approx \frac{\pi}{2\omega_1\tau} \tag{4.74}$$

The phase margin may be found by setting the magnitude of the gain (4.64) to unity and solving for ω_{GU}:

$$\omega_{GU} = \omega_1\sqrt{A^2B^2 - 1} \tag{4.75}$$

This may be simplified at all realistic levels of loop gain (such that $AB >> 1$) to give:

$$\omega_{GU} = AB\omega_1 \tag{4.76}$$

The loop phase at this frequency is then (from (4.65)):

$$\angle G(\omega_{GU}) = -\left[\tan^{-1}\left(\sqrt{A^2B^2 - 1}\right) + \omega_1\tau\sqrt{A^2B^2 - 1}\right] \tag{4.77}$$

or

$$\angle G(\omega_{GU}) \approx -\left[\frac{\pi}{2} + AB\omega_1\tau\right] \qquad \text{For } AB >> 1 \tag{4.78}$$

The phase margin of a high-gain first-order loop is therefore (from (4.63)):

$$PM = \frac{\pi}{2} - AB\omega_1\tau \tag{4.79}$$

For the example system detailed in Section 4.6.5 (Figure 4.33), this calculates to 45.3°. It will be negative for an unstable system.

Since the phase margin of a stable loop must be positive, the loop gain must fulfil the condition:

$$AB < \frac{\pi}{2\omega_1\tau} \tag{4.80}$$

By examining (4.74) and (4.80), the condition for loop stability may be stated as:

$$AB\omega_1\tau < \frac{\pi}{2} \tag{4.81}$$

As an example, consider the system described previously. Using (4.81) above yields a maximum value of loop gain, AB, of 100 times. This is consistent with Figure 4.34, which indicated that the loop was just stable with a loop gain of 49.6.

It is important to note that the above analysis is only valid for a *monotonic* amplifier AM/PM characteristic. This characteristic must also be *unipolar* (i.e., must not pass through zero), although this condition may be ensured over a range of frequencies by the use of a variable phase-shift between the up- and downconversion local oscillator paths. Most class-A and class-AB amplifiers posses a monotonic AM/PM characteristic and hence the above equations may be used to calculate the maximum achievable value of stable loop gain. Most class-C amplifiers do not posses a monotonic AM/PM characteristic as they usually display a 'peak' at the point where the active device turns on (see Chapter 2 for an example). The achievable gain from a Cartesian loop employing a class-C amplifier is therefore usually much less than that which would be calculated using (4.81).

Analysis of a Cartesian loop has been performed in a similar manner to that described above by Boloorian [26] and Briffa [27].

4.6.5.2 Alternative Loop Filter Designs

The above analysis assumes that the loop filter is a simple integrator and hence the resulting loop is first-order. It is also possible to construct alternative forms of loop filter and higher order loops. One example is outlined in [28], where the benefits of a phase-lag compensation method are highlighted. It is shown that a factor of 2 increase in the linearisation bandwidth may be achieved, although the technique is not suitable for transmitters where the first signs of instability (with increasing loop gain) occur at frequencies close to the carrier.

The circuit diagram of the phase-lag compensator is shown in Figure 4.39 and this is cascaded with the existing loop integrator. The frequency response of the circuit is given by:

$$H(j\omega) = \frac{1 + j\omega T}{1 + j\omega\alpha T} \tag{4.82}$$

where ω is the angular frequency in rad/sec and α and T are defined below.

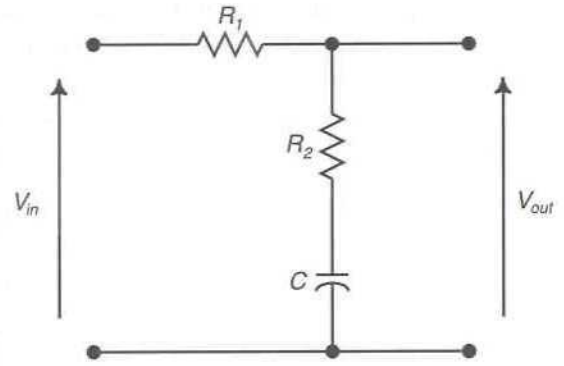


Figure 4.39 Circuit diagram of the phase-lag compensator.

The lower break frequency of the response is $1/\alpha T$ and the higher is $1/T$, where:

$$T = R_2C_2$$
$$\alpha = 1 + \frac{R_1}{R_2} \tag{4.83}$$

Above the higher break frequency, the magnitude response falls to $1/\alpha$. The higher break frequency of the compensator is chosen to be smaller than the frequency separation from the carrier at which the onset of loop instability becomes evident. The value of α may then be chosen to provide an appropriate level of attenuation to prevent the $(-1,0)$ point being reached.

The values of T and α must be carefully chosen to ensure that the phase-rotation does not result in instability close to the carrier, due to the rotation of the original Nyquist plot toward the $(-1,0)$ point.

4.6.5.3 Gain and Phase Margins When Using Highly Nonlinear Amplifiers

The preceding analysis is valid for amplifiers with a quasi-linear transfer characteristic (e.g., class-A and class-AB); however, as the degree of nonlinearity increases, the available loop gain whilst maintaining loop stability decreases from that predicted above. The levels of loop gain predicted by these methods therefore represent the upper bound of achievable loop gain and the degree of loop gain 'back-off' required in a practical system will depend upon the amplifier nonlinearity. A class-C amplifier, for example, will achieve stability at a loop gain significantly below that predicted from this analysis.

A typical class-C amplifier will exhibit a significant gain variation with power level as the transistor goes from cut-off through its linear region into saturation. This gain variation may be many tens of dB and will obviously have a significant effect on the overall loop gain of a Cartesian loop based around it. It is therefore necessary to design the loop bearing in mind this maximum level of gain (or alternatively to incorporate at a relevant point in the loop a component with an approximately inverse gain characteristic).

In the absence of additional correction, it can be shown [29] that the gain margin of the system must be sufficient to accommodate the expected additional gain due to the gain change with power level of the RF amplifier. Thus, for example, a system with a gain margin of 10 dB should not exceed a level of 10 dB of additional gain due to the gain variation of the RF amplifier. This is equally true of gain variations with temperature, although these are obviously slower acting and may be compensated by other means. In practice, a margin must be added to this requirement in order to prevent excessive 'peaking' in the closed-loop transfer characteristic, resulting in noise sidebands on one or both sides of the wanted channel (usually a number of megahertz away).

In addition, the AM/PM conversion in a highly nonlinear amplifier is very significant and will degrade the phase margin of the system. This again must be taken into account when calculating the available loop gain for a particular system, based on the derivations presented above. Thus, for example, a system with a phase margin of 60° will tolerate up to 60° of phase change due to the changing amplitude level of the input signal (i.e., due to AM/PM conversion), assuming that the loop phase control is perfectly adjusted. In this situation the loop phase control would require precise adjustment for each channel in the transmitter's operational bandwidth, and any other small variations (e.g., in amplifier gain) would render the system unstable. An additional phase margin should therefore be allowed.

It is evident from the above arguments that the AM/AM and AM/PM characteristics of a highly nonlinear amplifier directly degrade the available gain and phase margins of a Cartesian loop formed around it. It is therefore necessary to reduce the loop gain from that predicted using (4.62) to (4.81), often substantially [25]. The Cartesian loop can still, however, provide useful levels of IMD improvement with such amplifiers, and designs have been successfully produced giving −60 dBc IMD products from a class-C amplifier at 900 MHz. Careful selection of the class-C device and design of the amplifier can have a very significant effect on the ultimate performance which can be achieved from a Cartesian linearised system.

4.6.6 Cartesian Loop Applied to a Software Radio

The Cartesian-loop technique has been applied to many linear modulation schemes, including 16-QAM [30,31], Transparent Tone-in-Band (TTIB) [32] and in a generalised transceiver architecture, which is able to transmit any modulation scheme (within a given bandwidth) for which the signals may be generated in I and Q format (from a DSP) [33]—a *software radio*.

The generality of a Cartesian loop linear transmitter architecture is such that it can easily be made modulation independent, and hence can transmit any modulation scheme within its operational bandwidth. A radio may therefore be envisaged in which the baseband processing takes place within a digital signal processor (DSP), and I and Q baseband signals from this feed the transmitter. The modulation scheme in the transmitter is determined purely by software and may be reconfigured at will simply by downloading new software.

The receiver architecture may similarly be designed by employing direct-conversion techniques, with demodulation being performed in the DSP software. The receiver is therefore also flexible and may similarly be reprogrammed to accept a different, or additional, modulation scheme. A block diagram of this type of transceiver configuration is given in Figure 4.40, with a number of practical systems having been built by Wireless Systems International [33] and others.

This type of system has many applications, particularly with the proliferation of recent research into more spectrally efficient modulation schemes. Existing PMR schemes, for example, may employ this type of technology in any new mobiles purchased for their system. Thus, they will still be compatible with the original modulation system (e.g., FM) but are capable of offering additional or alternative schemes which may then gradually take over as obsolete equipment is scrapped. A smooth transition from one type of system to another is therefore possible, making such changes affordable to organisations with a large existing investment in mobile radio technology.

In addition, Cartesian loop has also been suggested as a linearisation technique for the TETRA (Trans European Trunked Radio) system as it employs a filtered π/4-DQPSK modulation scheme, with its attendant amplitude variations. Since the amplitude variations of this and other π/4-DQPSK schemes (such as D-AMPS in the United States) do not produce a full envelope variation, but a much smaller 'amplitude ripple', this leads to the question of what degree of linearity is required of the transmitter.

In the case of the TETRA standard [34], the basic system parameters are given in Table 4.3. A comparison between the basic two-tone IMD

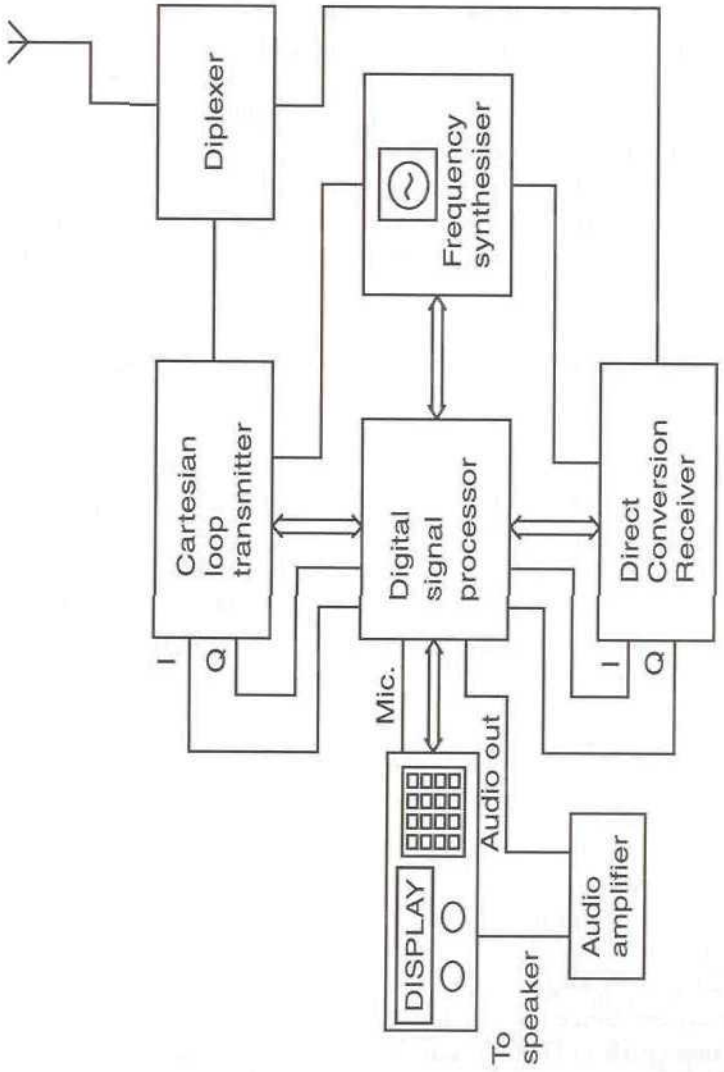


Figure 4.40 Flexible linear transceiver architecture.

Table 4.3
System parameters for TETRA

Parameter	Value
Initial Frequency Band	380–400 MHz
Modulation scheme	π/4-DQPSK
Transmit modulation filter	Root-raised cosine, $\alpha = 0.35$
Channel spacing	25 kHz (or 6.25 kHz)
Multiple access scheme	TDMA
Number of timeslots per frame	4
Timeslot duration	14.1667 msec
Raw data rate	36 kbits/sec (18 kbaud)
Gross data rate per slot	7.2 kbps
Adjacent channel coupled power limit	−60 dBc at 25 kHz
Ultimate IMD limit	−70 dBc at 50 and 75 kHz

performance of a linearised power amplifier and its adjacent channel power, when amplifying the TETRA form of π/4-DQPSK, is given in [35] and is reproduced in Figure 4.41.

Using Figure 4.41 it is possible to approximately predict the required

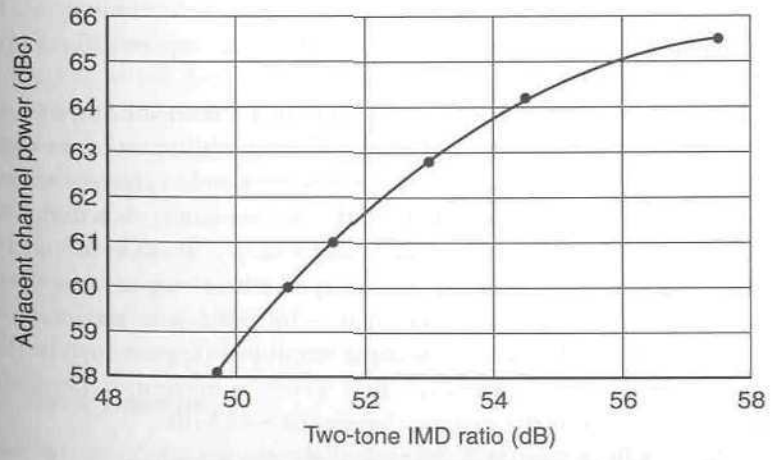


Figure 4.41 Experimental comparison between the adjacent channel power of a linear transmitter employing TETRA π/4-DQPSK modulation and its raw two-tone intermodulation ratio (from [35] © IEE 1994).

transmitter linearity in order to meet the TETRA specification, based on the more familiar two-tone IMD performance of an amplifier. It is also necessary to be able to predict the peak power requirement of the amplifier, based on the desired mean output power specification. This can be achieved using Table 2.1, given that the α value of the root-raised cosine filter is known for a particular system (not only TETRA).

4.7 Noise Performance of a Cartesian Loop

One problem with the CFL transmitter is that its output noise performance is no longer dominated by that of the RF power module employed within it (as is the case in many more conventional transmitter architectures). The use of significant degrees of attenuation, followed by a high level of gain, within the loop results in a noise performance markedly poorer than that of a conventional transmitter, at a range of frequency offsets from the channel center. There are a number of trade-offs which are available to the designer of a CFL transmitter to aid in the optimisation of the output noise performance. This section presents an overview of the noise performance of a Cartesian-loop transmitter and highlights the design methods which may be employed in order to optimise its noise performance.

The noise performance of a Cartesian loop is of particular concern with regard to the noise transmitted in the receive band of the radio system, as this must be kept to a minimum to prevent the blocking of weak wanted signals. Such blocking can occur either due to the mobile's own transmitter in a full-duplex system, or due to a nearby transmitter in any system (full-duplex, half-duplex, or simplex).

Take a full-duplex system as an example. If the transmit output power is 20W (+43 dBm) and the transmit output filter (usually part of the duplex filter) provides an attenuation of 60 dB at the nearest end of the receive band, then the noise must be sufficiently low (in the receive band) such that it does not mask a signal at the receiver's minimum sensitivity. This signal may be in the vicinity of −120 dBm, meaning that the transmitter output noise floor in the receive channel must be better than −103 dBc *plus* the cochannel protection ratio of the modulation scheme employed (approximately 10 dB for many current digital schemes). This gives a maximum permissible transmitter noise floor in the receive channel of −113 dBc.

Conventional transmitter architectures do not usually have significant difficulty meeting this level of noise performance, as they are based around a power amplifier stage with a noise figure in the vicinity of 10 dB. Even if attenuation is employed prior to the amplifier, for the purposes of power

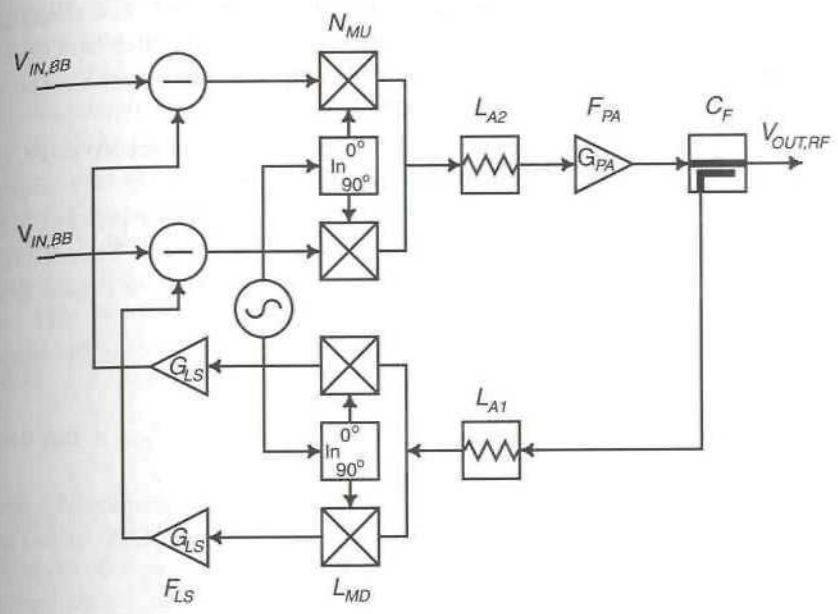


Figure 4.42 Cartesian loop-transmitter showing the system noise parameters (all expressed in dB).

control, a significant margin still exists in the noise floor and hence it is usually of little concern.

In the case of a Cartesian loop-transmitter, however, this is no longer true. The Cartesian loop configuration employs a significant degree of attenuation in the feedback path (L_{A1} in Figure 4.42) and a high degree of gain in the forward path; both of these factors have the potential to introduce noise into the system. The situation is further complicated by the noise removing properties of the feedback itself, making it more difficult to predict, directly, the noise performance of the complete transmitter.

A full analysis of this problem is provided in the literature [36]; however, for most practical purposes a simplified treatment is sufficient and such is provided below.

4.7.1 Noise Analysis

The Cartesian loop may be split into forward and feedback paths and the noise factor of each calculated (see Figure 4.42). The noise within and outside of the loop bandwidth may then be calculated from these, as the feedback path noise figure dominates the noise within the loop bandwidth

and that in the forward path dominates the noise further out. The transition between these two regions is based on the closed-loop transfer function of the loop (see Section 4.6.5 and equation (4.60)), with the overall response being of the form shown in Figure 4.43.

The forward path consists of the loop filter, the I/Q upconverter, the forward path attenuator (often incorporated as part of the power control circuitry) and the RF power amplifier chain. The output noise power of this forward path is usually dominated by the output noise floor of the upconverter, hence, to a first approximation, the output noise power from the forward path is given by (in dBm/Hz):

$$N_{FW} = -174 + L_{A2} + F_{PA} + G_{PA} \qquad \text{dBm/Hz} \qquad (4.84)$$

where L_{A2} is the loss in the forward path attenuator(s), F_{PA} is the noise figure of the power amplifier and G_{PA} is its gain (all in dB).

Note that (4.84) assumes that a passive upconverter is employed within the loop; if an active upconverter is employed instead, then (4.85) below can be used:

$$N_{FW} = N_{MU} - L_{A2} + G_{PA} \qquad \text{dBm/Hz} \qquad (4.85)$$

where N_{MU} is the noise output power of the upconverter in dBm/Hz; this is

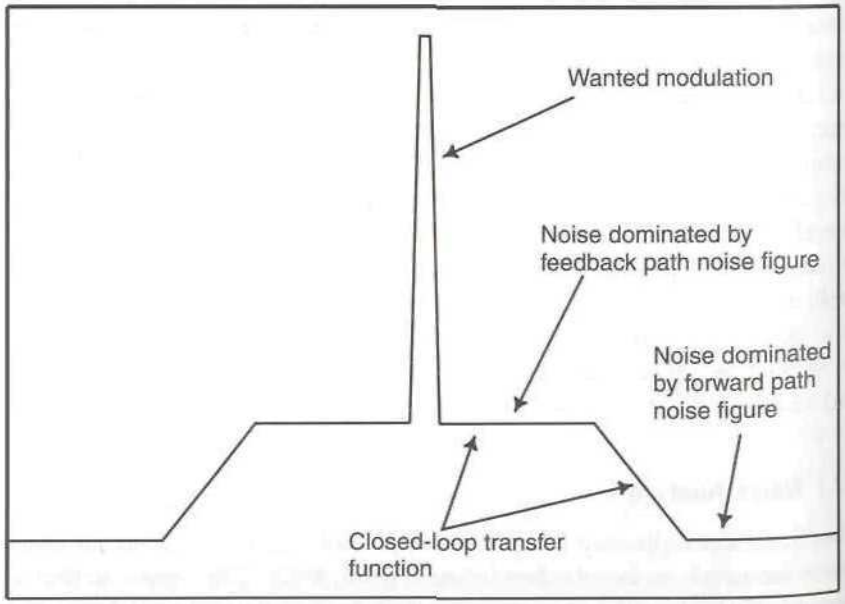


Figure 4.43 Form of Cartesian loop output noise characteristic.

normally specified on most manufacturers' data sheets. This equation assumes that the noise figure of the PA is negligible with respect to the noise output from the upconverter and that the attenuation of the forward path attenuator is not sufficiently large so as to reduce the upconverter noise to that of the thermal noise floor (i.e. $N_{MU} - L_{A2} >> -174$). Both of these assumptions are normally valid for a Cartesian loop operating at full output power. Typical figures are: $N_{MU} = -155$ dBm/Hz (at 0 dBm output power), $L_{A2} = 2$ dB, $G_{PA} = +42$ dB, giving 10W of output power capability (assuming that the PA output stage is appropriately rated) and a noise floor of -115 dBm/Hz. In the case of TETRA (18 kHz equivalent modulation bandwidth), for example, this translates to:

$$\begin{aligned} N_{PR} &= P_{out} - (N_{FW} + 10\log(BW)) \\ &= +40 - (-115 + 42.6) \\ &= 112.4 \text{ dBc} \end{aligned} \qquad (4.86)$$

which would meet the current TETRA specification for noise at an offset of greater than 5 MHz from the wanted channel [34].

The feedback path consists of the output coupler, the feedback attenuator(s), the I/Q demodulator and the baseband level-set op-amp(s) (so called because they are primarily used to set the baseband input level required for the loop to give its rated output power). The transmitter output noise power resulting from this feedback path is given by:

$$N_{FB} = -174 + C_F + L_{A1} + L_{MD} + F_{LS} \qquad \text{dBm/Hz} \qquad (4.87)$$

where C_F is the coupling factor of the output coupler, L_{A1} is the loss in the feedback path attenuator(s), L_{MD} is the loss in the demodulator and F_{LS} is the noise figure of the level-set op-amps (all in dB). Equation (4.87) does, of course, assume that the loop is capable of supplying this noise power (i.e., that it is operating with sensible loop parameters and a PA with sufficient output capability) and also that a passive demodulator is employed.

If an active demodulator is used instead, L_{MD} should be replaced by the cascaded noise figure of the demodulator and baseband op-amps.

Again, taking a typical example, in this case with a passive demodulator:

$$\begin{aligned} N_{FB} &= -174 + C_F + L_{A1} + L_{MD} + F_{LS} \\ &= -174 + 20 + 45 + 9 + 12 \\ &= -88 \text{ dBm/Hz} \end{aligned} \qquad (4.88)$$

For the 10W TETRA PA considered above, this represents a level of −85.4 dBc, which would be adequate to meet the TETRA specification at a frequency offset from the wanted channel of between 100 kHz and 500 kHz, but not between 500 kHz and 5 MHz (beyond this, the loop gain should have rolled-off sufficiently for the forward path to dominate). An improvement in noise figure of around 5 dB (e.g., in the level-set op-amps) would be required to meet the TETRA specification at all points.

Note that it is not always straightforward to accurately measure some of the downconversion chain losses for (4.87). A simple method of estimation is as follows:

$$L_{DC} = -20\log\left[\frac{V_{out,BB}}{V_{in,RF}}\right] + G_{LS} \tag{4.89}$$

where $L_{DC} = C_F + L_{A1} + L_{MD}$ is the loss in the downconversion elements, G_{LS} is the gain in the 'level-set' op-amps, in dB (which can be easily and accurately determined from the resistor values), $V_{out,BB}$ is the baseband output voltage from the level-set op-amps and $V_{in,RF}$ is the RF input voltage to the output coupler (i.e., the RF output voltage from the PA). L_{DC} can then be substituted for $C_F + L_{A1} + L_{MD}$ in (4.87).

The noise figure of the op-amp stages may be found from:

$$N_{NI} = 10\log\left[\frac{e_n^2 + i_n^2(R_S + R_P)^2 + 4kT(R_S + R_P)}{4kTR_S}\right] \text{ dB} \tag{4.90}$$

for a non-inverting configuration and:

$$N_I = 10\log\left[\frac{e_n^2 + R_1^2\left(i_n^2 + 4kT/R_2\right) + 4kTR_1}{4kTR_1}\right] \text{ dB} \tag{4.91}$$

for an inverting configuration (note that R_1 includes the source resistance in this case). The various symbols are defined in Figure 4.44 and R_P is given by:

$$R_P = \frac{R_1 R_2}{R_1 + R_2} \tag{4.92}$$

Equations (4.90) to (4.92) are valid for voltage feedback op-amps; similar equations also exist for current feedback op-amps [37]. Note that matching considerations in a practical loop may well significantly increase the noise figures obtainable in practice from these elements and (4.90) to

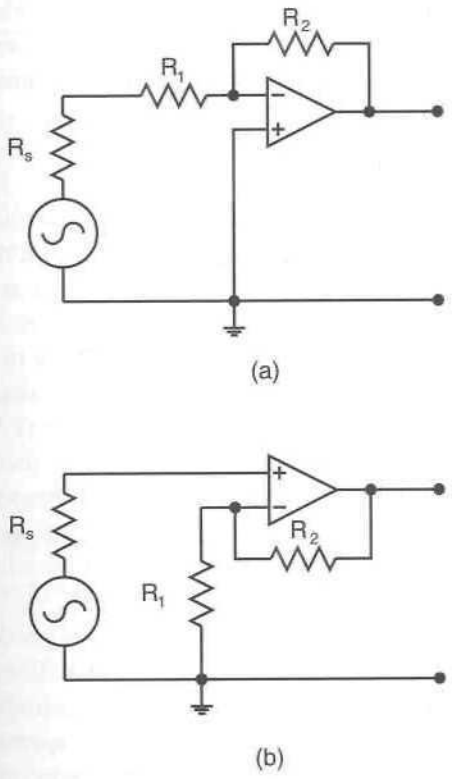


Figure 4.44 Inverting (a) and non-inverting (b) op-amp configurations.

(4.92) should therefore be seen as providing a minimum rather than an easily achievable value.

Finally, note that the output noise power within the loop bandwidth, given by (4.87), assumes that the loop is operating without 'peaking' in the closed-loop transfer characteristic, caused by operating the loop close to its stability limit. The resulting noise sidebands may well break a given noise specification, despite calculations indicating that the specification can be met. In such circumstances, the only available course of action is usually to reduce loop gain.

4.8 Practical Considerations With the Cartesian-Loop Transmitter

There are a number of practical issues which must be considered when constructing a Cartesian-loop transmitter, and a number of design compro-

mises must be taken into account. Many of the problems originally associated with the loop have now been solved and a number of alternative solutions are presented here for the more obvious shortcomings.

4.8.1 Local Oscillator Performance

A problem with the direct conversion architecture transmitter (described above) is that it requires a 'clean' (i.e., low distortion) local oscillator at the output frequency. Local oscillator distortion appears either as discrete spurii within the operating band, or as 'phase noise' immediately adjacent to the wanted frequency. This phase noise, or residual frequency modulation, will ultimately limit the adjacent channel performance of the transmitter, and is a factor in determining transmitter error-vector magnitude (EVM).

A further concern is that the modulated output from the transmitter, or indeed from the antenna, will leak back into the VCO tuned circuits, and thus cause spurious modulation of the VCO. The action of the feedback loop would not eliminate this modulation and hence it would appear as a frequency modulation of the final output signal.

There are a number of alternative solutions to these problems. The first is to use an intermediate frequency for the feedback loop. The local oscillator is not then required to work at the final output frequency and its specification may be relaxed. This is, however, at the expense of additional complexity elsewhere in the loop, and unwanted loop delay. A second solution is to use a multiloop synthesiser, so that there is no VCO running at the RF output frequency. The multiloop design also enables a lower phase noise output to be achieved.

A third option is to mix an off-frequency synthesised (tunable) local oscillator with a fixed oscillator to provide the final output frequency. Neither oscillator is now operating at the transmitter output frequency, hence minimising the chances of unwanted VCO modulation.

4.8.2 Matching of the Quadrature Signal Paths

Another important factor in the design of the transmitter is in the control of the gain and phase matching between the I and Q channels, and the removal of DC offsets from these signals. Any mismatch in the I and Q signals will result in an 'image' signal which is incompletely suppressed, and cochannel with the wanted signal. Any residual DC offset will create a local oscillator component, appearing at the center of the channel. An advantage of incorporating an IF within the loop is that the downconversion mixers are operating at a fixed frequency, where any gain, phase, or DC offsets will

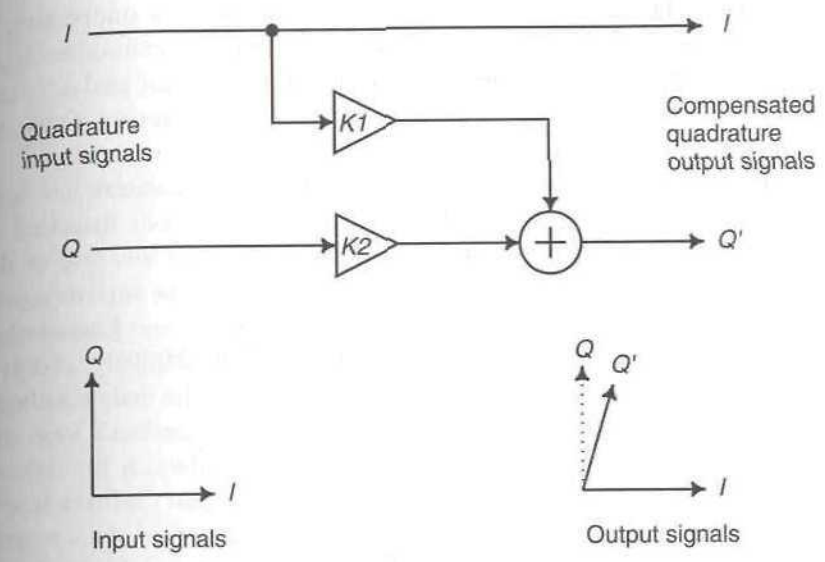


Figure 4.45 Gain and phase compensation in a direct conversion transmitter.

remain constant. In the direct conversion method, the mixers have to operate over the same frequency range as the coverage band of the transmitter, and so compensation for the mixers' frequency-dependent characteristics may be necessary. Note that it is only the mixers in the downconverter whose characteristics affect the final transmitted output: the effects of the upconversion mixers will be removed by the two feedback loops. Hence lower performance mixers may be used in the upconverter, without affecting the transmitted signal.

Gain and phase compensation in the direct conversion transmitter may in fact be achieved by modifying the I & Q baseband signals supplied from the DSP. The operation of this process is shown in Figure 4.45. A small fraction of the I channel signal is added to the Q channel output, and by alteration of the variables K1 and K2, any amount of gain and phase mismatch may be accommodated. This approach requires prior knowledge of the actual gain and phase mismatches, due to the mixers, at every channel frequency.

In practice, the mixers' performance changes only gradually with frequency, and so it is possible to have values of K1 and K2 which are valid for a whole range of frequencies, say a span of 5 MHz. In this way, a 20 MHz operating band could be covered in four sub-bands, each with their own settings for K1 and K2. The channel code for the frequency synthesiser could be used to index these values, and select the ones appropriate for the

channel in use. Hence, the original calibration need only be undertaken at four spot frequencies, in the center of each sub-band. The calibration itself may be performed by observing the output on a spectrum analyser, and repetitively adjusting K1 and K2 for optimum image suppression. Once set, at manufacture, these values would be stored in nonvolatile memory.

An alternative system for adjusting quadrature modulators has been suggested by Faulkner *et al.* [38]. This scheme involves diode detection of the combined output signal from the I & Q modulators and use of the resulting signal as a feedback error signal to a controller. The authors report the successful application of the technique to a predistortion linearisation system. The principal advantage of this system is that of simplicity, as only a single error signal is required to perform the correction. Its main disadvantage for this application is that it involves an additional feedback loop and hence may suffer from the attendant stability and bandwidth limitations, particularly in the form of interactions with the main signal feedback loops. The accuracy provided by this system is also greater than is strictly necessary for the Cartesian loop application, with fixed sub-band weightings being more than adequate in this case.

4.8.3 DC Offsets on the Quadrature Signals

The DC offsets, which result in an LO component in the transmitted signal, arise from a combination of two effects, and two mechanisms are used to reduce them.

The major cause is an accumulation of DC offsets in the downconversion mixers, the two stages of preamplification, and the differential amplifiers. These offsets are all removed by a DC nulling mechanism, as shown in Figure 4.46. This uses a pair of high-gain feedback loops (one for each of the quadrature signals) around the whole of the baseband processing chains. A DC nulling voltage is applied to each path which cancels out all the other offsets. This is accomplished by a low-offset amplifier, and a sample-and-hold amplifier. The nulling voltage is maintained by the hold capacitor, and hence the nulling process must be repeated regularly, say every few minutes of transmit time or at the beginning of each transmission, to combat drifts. The nulling process is performed by rapidly switching-off the RF output, setting the modulation input to zero, and switching both sample-and-hold devices to sample mode. After this DC feedback loop has settled, hold mode is resumed, and the modulation inputs and RF output are restored. The whole process can be performed within 100 microseconds, to minimise interruption to the channel traffic.

Note that if the process of setting the I and Q baseband inputs to zero is

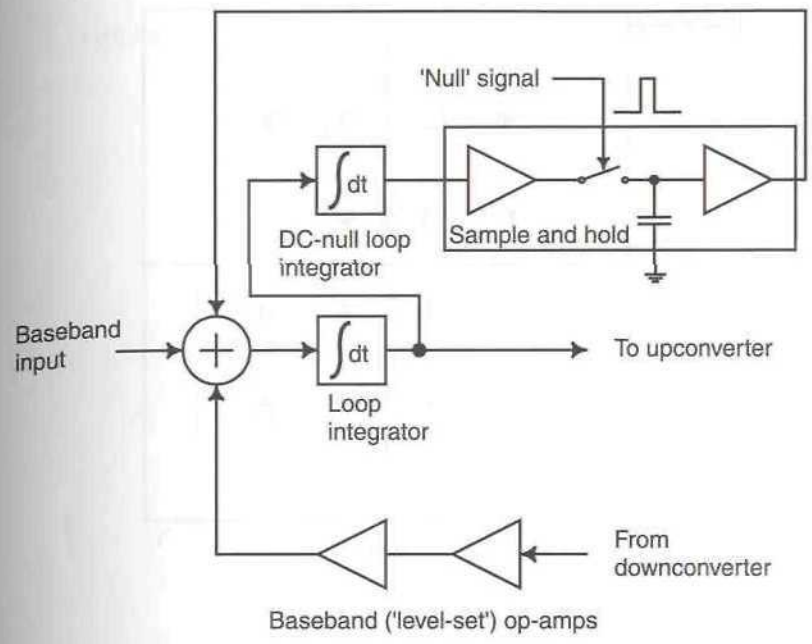


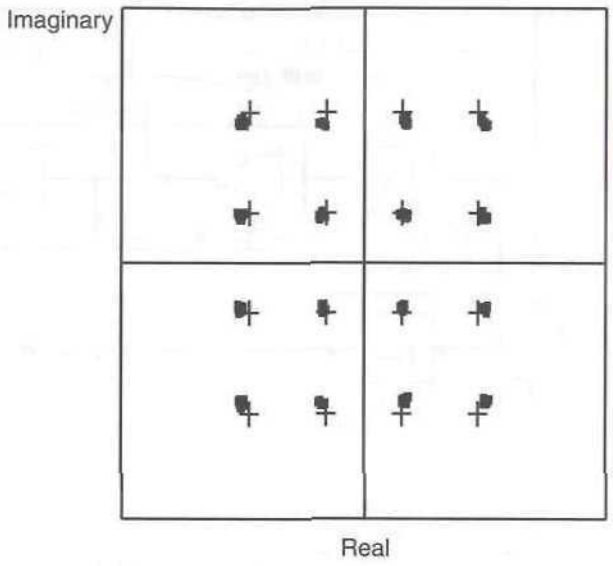
Figure 4.46 DC nulling circuit (I channel only shown; Q channel is identical).

performed within the software of the DSP device providing the baseband signals (i.e., by feeding 'zero' to the DACs), then the DC null mechanism will also compensate for offsets in the DACs themselves and, for example, any op-amp stages following them. Low-offset devices are therefore unnecessary in this part of the transmitter.

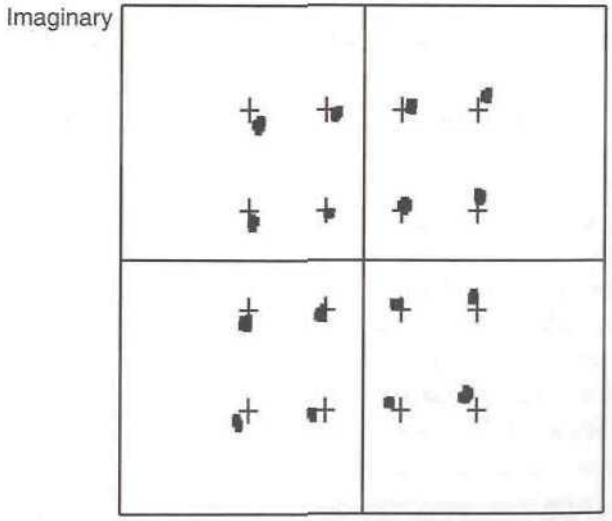
The use of more modern, digital sample-and-hold circuits will reduce the frequency with which the nulling voltage needs to be updated, as they eliminate the capacitor and its associated leakage. The update can therefore be performed upon initialisation of the transmitter and during the hand-off process (say) for a cellular system.

4.8.4 Effect of I/Q Errors and Carrier Leakage on Digital Modulation Formats

The effect of various impairments, introduced by a practical Cartesian loop, are illustrated in Figure 4.47 for a 16-QAM signal (4 bits/symbol). In Figure 4.47(a) an amplitude error of 0.3 dB is introduced and it can be seen that the constellation is effectively 'constrained' in one dimension (imaginary) and expanded in the other (real) and thereby misses the ideal constellation points

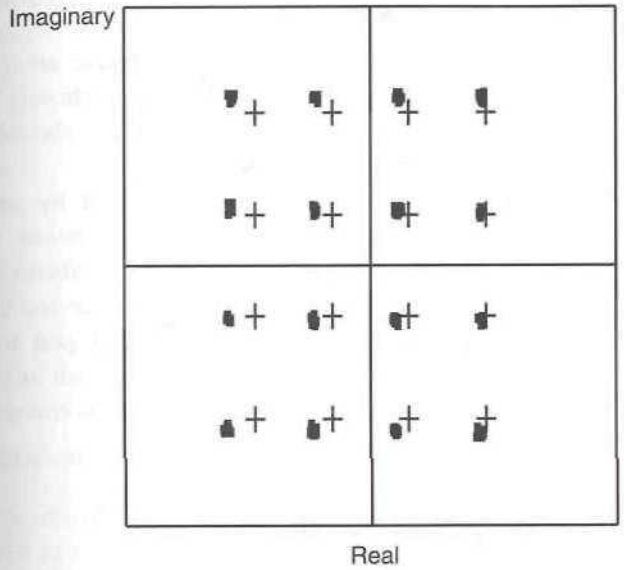


(a)

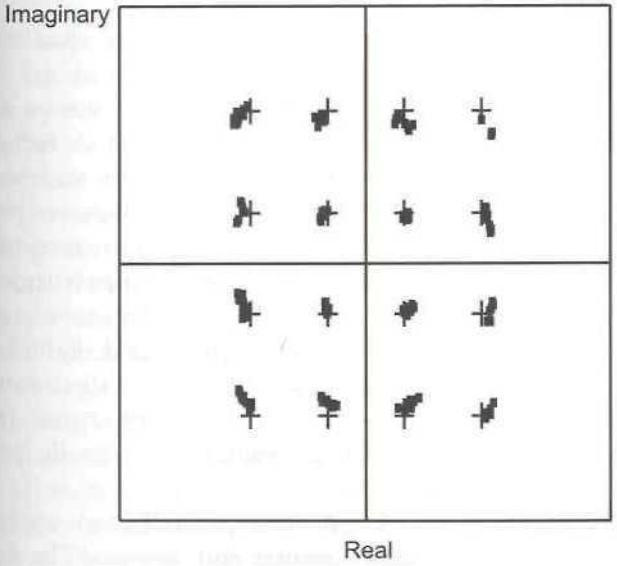


(b)

Figure 4.47 Effect of I/Q imbalance and carrier leakage (DC offset) on a 16-QAM signal. QAM parameters: 60kbps, root-raised cosine filter, $\alpha = 0.35$.
(a) 0.3 dB amplitude error. (b) 10° phase error. (c) 10% carrier leakage.
(d) 0.3 dB amplitude error and 3° phase error.



(c)



(d)

in both I and Q. The error is most obvious on the extremes of the constellation.

Figure 4.47(b) shows the 'skewing' effect of a phase error on the constellation. A deliberately high phase error has been chosen for this illustration, to emphasise the effect, and such a gross error should not be encountered in a correctly-designed practical loop.

Figure 4.47(c) illustrates the fixed offset introduced by unwanted carrier leakage, caused either by a poor DC null performance or by a poorly-performing downconverter (gross DC drift of the baseband op-amps could also be a cause, but this is unlikely with modern IC devices).

Finally, Figure 4.47(d) combines typical values of I/Q gain and phase imbalance (0.3 dB and 3°, respectively) for a practical Cartesian loop. These levels of error would correspond to an image level approximately 30 dB below the wanted signal in, for example, an SSB system.

4.8.4.1 SVE in a Cartesian-Loop Transmitter

All of the above problems will contribute to a degradation of the 'signal vector error' (SVE) (also known as 'error vector magnitude', EVM) of the digital modulation.

Error vector magnitude is now a commonly-quoted specification for both transmitter and receiver performance evaluation (e.g., TETRA, DAMPS), in preference to gain or phase imbalance or image rejection. Typical figures are in the range of 5% to 15% for most mobile radio systems (e.g., TETRA [39]) and various test instruments now incorporate the measurement of this parameter as a standard feature. There are potentially a large number of factors in the design of a transmitter (or receiver) which can contribute to the final measured value, however, in practice, in a well designed system most are usually negligible. Examples include problems with the receiver detection process (normally performed digitally), transmitter nonlinearities (the effect of these can still be significant where linearisation is not employed), synthesiser frequency errors (normally tracked out), and errors in modulation generation (normally performed digitally with very little resultant error).

There are two sources of SVE which are generally non-negligible in a system design and affect both the transmitter and receiver. The first is the gain and phase imbalance present in the transmitter (and the corresponding quadrature demodulator in the receiver [40]), assuming that both are performed by analogue means. In the case of a Cartesian loop, these errors result from imperfect matching between the demodulator mixers (or analogue multipliers) and an imperfect 90° split in the local oscillator

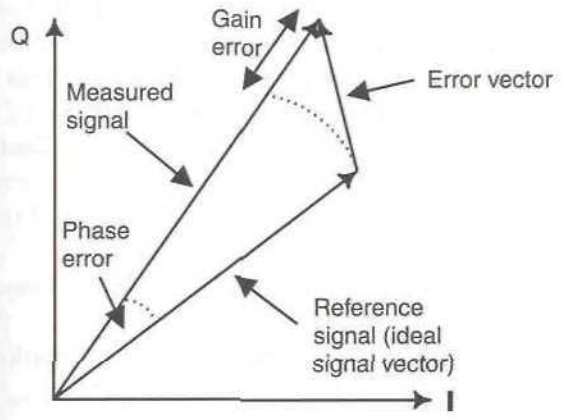


Figure 4.48 Illustration of signal vector error in the I/Q plane.

path. Together these result in a distortion of the vector in the I/Q plane, as illustrated in Figure 4.48.

Although these errors may be corrected digitally [41], this would typically involve a degree of individual testing during production and this is generally undesirable. Figures presented in [40] indicate that a phase error of 1° can be achieved as a 'typical' specification for an IC implementation, although currently available commercial devices usually specify around 0.3 dB of amplitude error and 3° phase error.

The effect of quadrature modulator and demodulator errors on the adaptive predistortion method of RF power amplifier linearisation has been studied in some detail [42]. This study indicates the detrimental effect of such errors on predistorter performance and also highlights some methods for overcoming them. An earlier study highlights the detrimental effect of quadrature errors on the spectral characteristics of a power amplifier [43]; quadrature errors are therefore clearly an undesirable effect in many areas (not only in a Cartesian loop).

The second major contributor to SVE is phase noise present on the local oscillator. This results in a random rotation of the signal vector around the I/Q plane, with a mean error determined by the synthesiser characteristics and the characteristics of the detection and tracking filtering present in the receiver. Phase noise is present on all signal sources and although it is technologically possible to reduce it to a degree whereby it would have a negligible effect on EVM (over and above that generated by the I/Q modulation and demodulation process), this is not usually economic in mobile radio designs. It must therefore be incorporated into any study on EVM in a particular design.

The traditional method of studying the effects on EVM of phase noise, modulator, and demodulator errors is by means of a block-level simulation of the complete transmitter or receiver system (or both). This is a relatively complex and costly process and relies on the availability of sophisticated simulation tools of sufficient accuracy and of operators skilled in the use of these tools and interpretation of the subsequent results. Care must be taken to ensure that sampling rates are chosen appropriately and that blocks are being used as intended and not beyond their capabilities.

The effect of phase noise on a received carrier, as a result of mobile propagation effects, has been studied in detail [44]: however, this analysis is long and complex and does not lend itself easily to adaptation as a design tool for the system designer of a transmitter or receiver.

4.8.4.2 SVE Calculation for a Cartesian Loop

With reference to Figure 4.48, the magnitude of the signal error vector may be determined using the cosine rule as:

$$E_V = \sqrt{[R^2 + M^2] - 2RM\cos(\phi_e)} \tag{4.93}$$

where R is the magnitude of the reference ('ideal') vector, M is the magnitude of the measured ('actual') vector, and ϕ_e is the phase error between them. The measured vector magnitude, M, is composed of the reference vector magnitude, R, plus a component resulting from the gain error present in the system, G_e:

$$M = R + G_e \tag{4.94}$$

If the reference vector magnitude is set to unity, then the resultant error vector magnitude (in %), E_{VM}, may be found from:

$$E_{VM} = 100\sqrt{[1 + M^2] - 2M\cos(\phi_e)} \tag{4.95}$$

The EVM may be plotted as a family of curves over a range of values for the gain and phase errors. Two examples are shown in Figures 4.49 and 4.50, with Figure 4.49 representing a general overview for a wide range of errors and Figure 4.50 showing a detailed view of the range of errors generally encountered in most commercial quadrature modulator and demodulator subsystems, whether integrated circuit or hybrid based.

A typical specification, for example, is a gain error of 0.3 dB and a phase error of 3°; using Figure 4.50, this results in a peak error vector

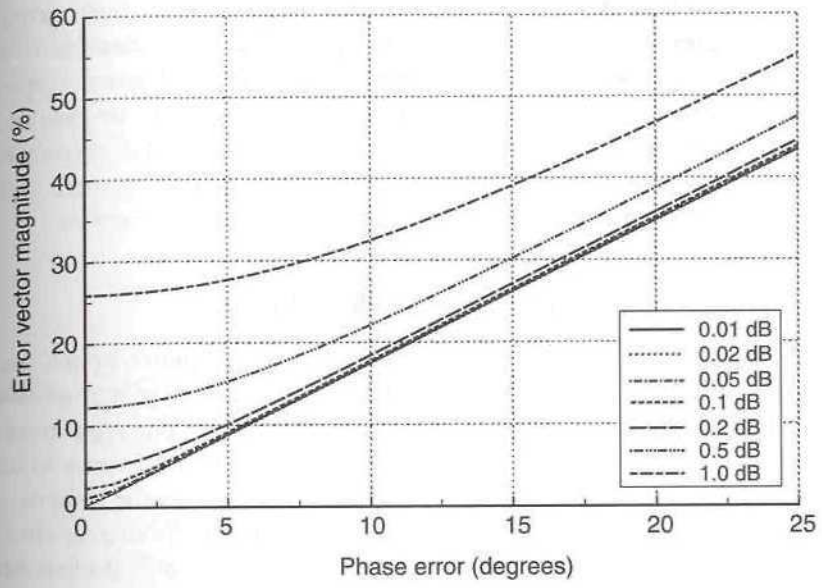


Figure 4.49 Error vector magnitude for a wide range of gain and phase errors.

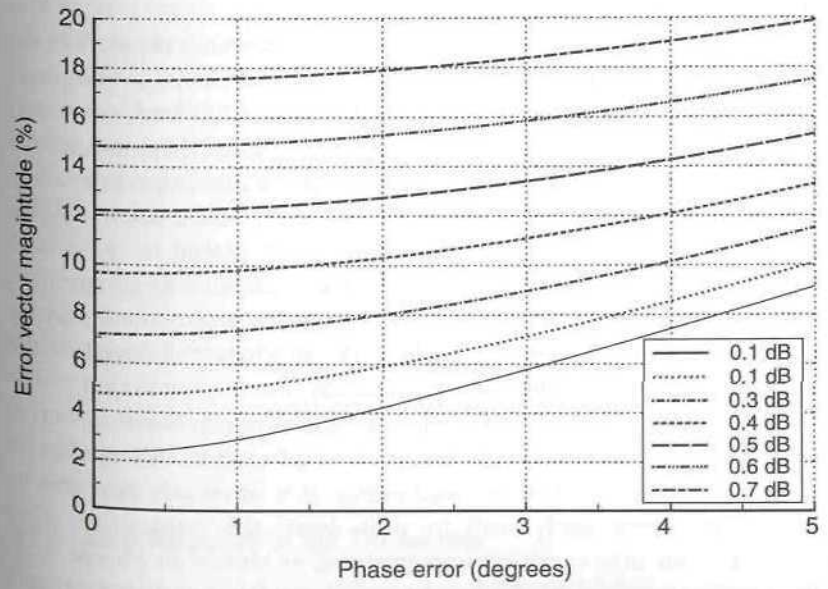


Figure 4.50 Error vector magnitude for a range of gain and phase errors typically found in commercial quadrature modulators and demodulators.

magnitude of 9%. A given component is unlikely to be at both extremes simultaneously and hence may well be acceptable (on a statistical basis) with this specification, despite the relatively high EVM figure. A more acceptable figure for many systems would be around 6% and this can be guaranteed with, for example, gain and phase error values of 0.2 dB and 2° respectively. Note that the r.m.s. values typically specified in mobile radio systems will be much lower than these figures for the above gain and phase errors.

4.8.5 Practical Considerations of Loop Stability

Whilst it is possible to reduce the distortion in an amplifier by the use of feedback, one question always arises with the use of a high-gain loop: that of stability. Obviously in a high-power radio transmitter radiating from an antenna, the dangers of potential instability, in terms of interference to other users, are considerable, and so great care must be taken in the design.

The usual stability criteria apply to the feedback loop: namely that the gain must be less than unity when the phase shift reaches 180°. If a first-order loop with very high gain is used, then the transition below unity gain may be at a frequency of several megahertz, and over this range, delay in the amplifier chain will produce a significant variation in phase through the amplifier. For this reason, a phase compensation is required as the RF operating ('carrier') frequency is changed.

This compensation is achieved by means of a phase shift network in the local oscillator drive to the upconverter (as described earlier). The phase shift could alternatively be placed in the downconverter LO feed, or even in the amplifier chain, but this would result in either increased delay around the loop, or less predictable offsets and mismatches in the downconverter. Any parasitic amplitude change with phase shift would also cause a variation in output power from the loop, if the phase control is placed in the downconverter LO path; this is avoided by placing it in the upconverter path. In a similar manner to the I/Q matching described earlier, this control may be implemented as a single setting for each of the sub-bands, indexed to the synthesiser programming code (channel number).

If a gross change in the phase through the amplifier were to occur, caused for instance by a component failure or a bad antenna mismatch, it is possible that wideband oscillation could occur. It is obviously desirable to detect this state immediately, and to shut down the transmitter. This instability detection may easily be implemented, as shown in Figure 4.51, by high-pass filtering one of the quadrature downconverted signals, and feeding the output to an envelope detector and comparator. The comparator output could then feed a transmit inhibit function, and disable the

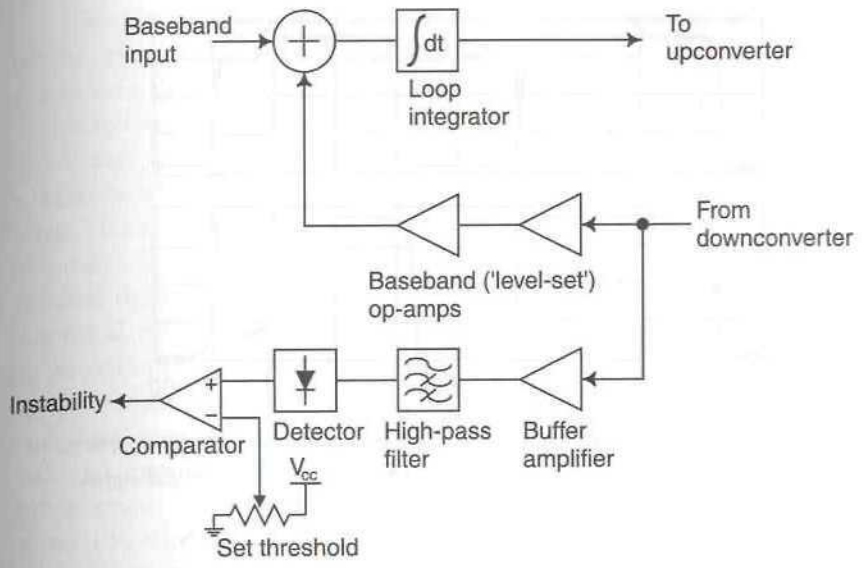


Figure 4.51 Basic instability detector.

transmitter. Note that the required downconverted signal could be taken from anywhere along the 'level-set' op-amp chain and not only at the downconverter output (as shown in Figure 4.51), assuming that the op-amps are sufficiently broadband (> 20 MHz).

This mechanism will detect any wideband oscillation state, but will not detect the condition occurring when the loop phase is around 180° in error. In this case, positive feedback occurs at DC, resulting in a CW output at the LO frequency. This particular state can be detected by checking for limiting DC outputs in the downconverted signal, although conventional wideband oscillation is much more likely in practice, with DC positive feedback usually only being encountered during design debugging. The only mechanism which would generally create this problem is a component failure in the loop phase control mechanism; it is highly unlikely that such a failure would occur *and* result in exactly the correct phase relationship for this problem to be encountered. Most designs do not, therefore, include any mechanism for detecting this condition.

4.8.6 Step Response of the Transmitter

Any transmitter which is used in a rapid-keyed TDMA system, such as the US D-AMPS, or European TETRA systems, must exhibit a rapid step response. That is to say the RF output must rise from zero to the correct

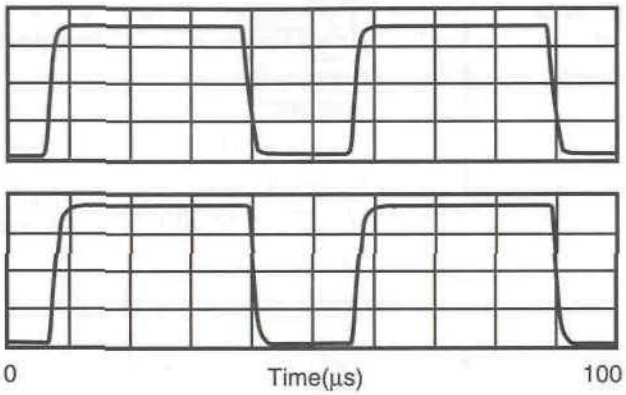


Figure 4.52 Typical step response of a first-order Cartesian-loop transmitter; upper trace: baseband input signal envelope, lower trace: detected RF output signal envelope.

level, without significant delay or overshoot, otherwise the first part of the transmitted message may be distorted.

Cartesian feedback transmitters utilising a first-order loop exhibit excellent step responses. Figure 4.52 shows the response to an off-on-off pulse of 33 microseconds duration. The top trace shows this pulse as applied to the I channel input to the transmitter. The lower trace is obtained from the video output of a spectrum analyser on zero span, that is, the time-domain response at the output frequency. The delay between the traces, in the order of 650 nanoseconds, results from the demodulation process in the analyser. It can be clearly seen that even with a pulse rise time of in the order of 2 microseconds, there is no noticeable overshoot in the response.

This indicates that the transmitter is able to faithfully reproduce a rapidly rising pulse applied to the modulation inputs. Thus, when an appropriately-shaped pulse is generated by the DSP, the RF output has the same envelope shape. Higher order loops must be carefully designed, in respect of their damping factors, in order that significant delays or overshoots are not created.

4.8.7 Power Control in a Cartesian-Loop Transmitter

Power control in transmitters for land-mobile radio systems is desirable, in order to reduce the level of cochannel interference to users in other areas. Two methods for implementing power control in the Cartesian transmitter have been devised, one of which requires no additional hardware, the other is more complex, but provides a more power-efficient system.

Since the RF output is a linearly amplified and frequency-shifted replica of the complex baseband signal, power control may be achieved by controlling the amplitude of this baseband signal. This is achieved simply by multiplying each of the I & Q signals from the DSP by the same value (the square root of the power ratio decrease required). This method has the advantage of simplicity of implementation, requiring no additional hardware. There are, however, disadvantages, especially if the power control has to cover a large range, say in excess of 10 dB. This arises partially from the reduced dynamic range of the modulation signal, since only part of the numerical range is used in the output digital-to-analog converters. Thus quantisation noise and DC offsets become more significant, and the quality of the transmitted signal is reduced. Table 4.4 shows this power control operating over a range of 20 dB, and indicates how the local oscillator leakage component becomes a more significant proportion of the signal. An enhancement would be the use of variable-gain amplifiers from the DAC outputs to the modulation inputs, thus allowing the dynamic range of the signals to be preserved, at the expense of extra hardware. This would still, however, not solve the LO leakage problem.

One other result of implementing power control exclusively at baseband, is a reduced power conversion efficiency of the transmitter. Table 4.5 indicates how the efficiency decreases as the output power level is reduced for a typical class-AB power module. For this reason, another method of power control is sometimes desirable: one entailing RF amplifier switching. To achieve a reduction in output power equal to the gain of the final amplifier stage, this stage may be switched out of circuit, simultaneously with an equal amount of attenuation in the downconverter, as shown in

Table 4.4
Baseband power control of a Cartesian-loop transmitter (measured)

Power level (dB relative)	Peak output (Watts)	Modulation amplitude (normalised)	LO suppression (dB below pk. o/p)
0	40.8	1.00	−46
−3	20.4	0.71	−43
−6	10.3	0.50	−40
−10	4.2	0.32	−36
−16	1.0	0.16	−30
−20	0.4	0.10	−26

Table 4.5
Efficiency of a typical class-AB Cartesian-loop transmitter with power control implemented exclusively at baseband.

Power level (dB relative)	Mean output power (Watts)	Efficiency (%)
0	20.4	25.2
−3	10.2	16.8
−6	5.15	10.8
−10	2.1	5.7
−16	0.5	1.8

Figure 4.53. The phase compensation must also be adjusted, to maintain stability, or an equivalent time delay switched-in in place of the amplifier (as shown). Thus the last stage of the amplifier chain will always be working near to optimum efficiency, and the dynamic range of the signal will be preserved.

In a given design, it may be desirable to use a combination of both methods for fine and coarse control.

A further (more common) alternative is shown in Figure 4.54; in this case, matching attenuators are switched in the forward and feedback paths, thus providing power control, whilst maintaining loop gain. At full power, the forward path attenuators will be switched-out of circuit, whilst those in the feedback path will all be switched-in. To reduce power, an appropriate attenuator in the feedback path is switched-out (i.e., an n dB attenuator to provide an n dB reduction in output power). In order to compensate for the n dB increase in loop gain which will result from this, an n dB attenuator must then be switched-in in the forward path. The attenuators need not match in terms of implementation technology or size; however, if the delay around the loop alters as a result of switching the attenuators then either an equivalent time delay must be switched in instead (as shown) or the loop phase control must be altered to compensate. There are thus two alternatives:

1. Both forward and feedback path attenuators are implemented in an identical manner (e.g., PIN-diode switched 'Π' or 'T' network matched attenuators). Since all of the attenuators will have an identical delay, they will automatically compensate for each other when switched. As a result, no additional compensation by means of the phase-shifter is necessary.

2. The forward path attenuator is implemented in a low-power technology (e.g., in an integrated circuit, since its intercept point is much less critical to the IMD performance of the overall design) and the feedback path attenuator is implemented discretely (to achieve a high intercept point). In this case, the forward and feedback path delays will not be equalised when a pair of attenuators are switched and hence either an appropriate delay must be switched-in instead, or the loop phase control must be updated accordingly (e.g., by means of a look-up table housed in the transmitter baseband DSP).

The attenuator switching mechanism described above has a number of advantages:

1. It overcomes the DC offset (carrier leakage) and D/A converter dynamic range problems inherent in the baseband power control mechanism.
2. It can provide finer power control steps than the amplifier switching method (but not as fine as those of the baseband method).
3. It removes the need for the high-power (and potentially lossy) RF changeover switches inherent in the forward path of the amplifier switching method.

It is not, however, as power-efficient as the amplifier switching technique (in general) and cannot provide the very fine (< 1 dB) and highly-accurate steps available from the baseband technique. It is common, therefore, to use a combination of techniques in practice, e.g., baseband control for fine steps (1 dB say) and attenuation switching for coarse steps (10 dB say).

4.8.8 Effect of External Signals on a Cartesian-loop Transmitter

When connected to an antenna, rather than a screened, matched load, a transmitter output will be subjected to external signals, which arise from the multiplicity of other users of the same radio service (e.g., cellular), all of whom are transmitting potentially similar power levels and many of whom may be in close proximity to the transmitter in question. This is true for handsets and mobiles which can clearly be placed closely together in normal operation, but is also true of most base stations, as many antennas, for different or similar services, are often placed on the same mast. In all cases, the transmit frequencies of the interfering signals may be only a few kilohertz away from the user's own channel, or they may be many megahertz away, it is therefore important to consider both cases. These

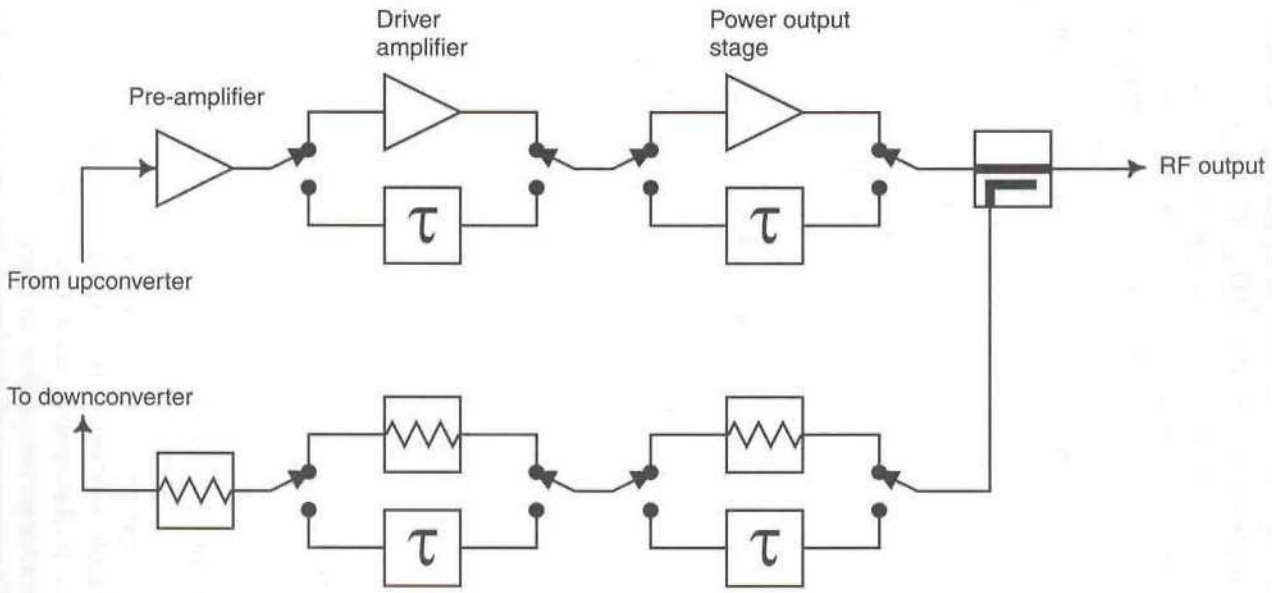


Figure 4.53 Coarse power control of a Cartesian-loop transmitter by power-stage removal.

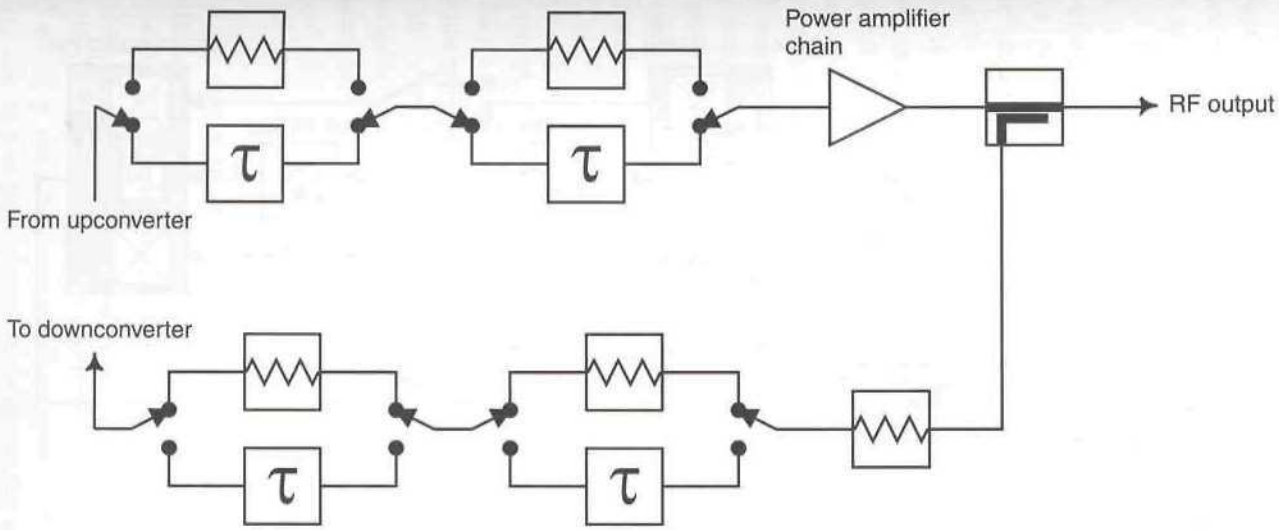


Figure 4.54 Power control of a Cartesian-loop transmitter by attenuator switching; the switch positions shown would provide maximum output power from the transmitter.

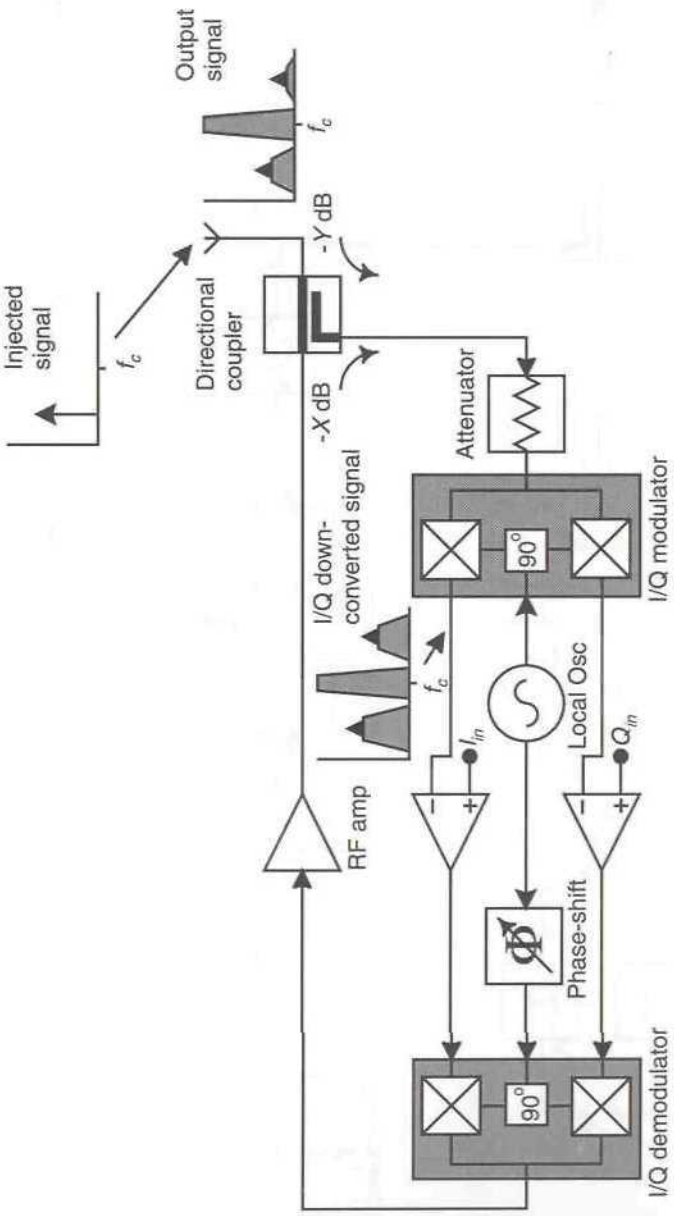


Figure 4.55 Effect of an externally-injected carrier on a Cartesian-loop transmitter.

external signals must not significantly affect the operation of the loop, and in particular must not result in significant distortion from the transmitter output. Equally they must not result in instability of the transmitter, or cause it to generate a significant level of signal vector error (SVE).

In the case of a transmitter employing Cartesian feedback correction, the action of these interfering signals is much more complicated than for an open-loop transmitter; the problems therefore require careful analysis. A number of different mechanisms come into play and the relative effect of these on the output spectrum depends heavily on the transmitter design and the performance of some of the subsystem components (e.g., the I/Q modulator and demodulator). A purely practical analysis of the problem is difficult, as a number of the mechanisms result in coincident signals in the output spectrum; distinguishing between them is extremely difficult in most practical situations (i.e., with real, rather than perfect system components).

The standard method of assessing the susceptibility of a transmitter to an external signal is by means of a carrier injected into the output of the transmitter during operation. The level and spacing of this carrier from the wanted channel vary, depending upon the particular system specification being considered. Typical figures are an injected level of 30 to 50 dBc (below the wanted transmitter output power) and separations varying from a few channels to 1 MHz (or sometimes more). It is therefore important to consider the effect of both close-to-carrier signals and widely-separated signals on the performance of the loop (the mechanisms of interference being different in each case).

The principal effect of an externally-injected signal may be outlined as follows (with reference to Figure 4.55). Suppose that an unwanted carrier $U(t)$ is injected at a frequency separation, $\Delta\omega$, from the wanted channel center frequency:

$$U(t) = C\cos[(\omega_C + \Delta\omega)t + \phi(t)] \tag{4.96}$$

where ω_C is the center frequency of the transmit channel. This signal will have two effects on the loop: first, it will create intermodulation products in the output of the RF power amplifier and these, plus a reflected version of the carrier (due to imperfect matching of the PA output to the antenna) will couple into the feedback path, via the output coupler, in the normal way. The second effect is that the carrier will reverse-couple into the feedback path directly, due to the imperfect directivity of the output coupler. The coupler directivity will be degraded from its nominal (perfectly-matched) value by the relatively poor match of the power amplifier output and antenna

input, in a practical system. The resulting directivity could be only 5 dB in *many practical systems.*

The two versions of the injected carrier (from the PA reflection and from the reverse coupling) will sum vectorially to create a single vector, with a phase and amplitude determined by the path difference, output coupler directivity and PA output mismatch:

$$U_{fb}(t) = C_{fb}\cos\left[(\omega_C + \Delta\omega)t + \phi_{fb}(t)\right] \tag{4.97}$$

that is, a single tone with an amplitude and phase determined by the result of the vector summation process. In addition, intermodulation between the injected carrier and the wanted (modulated) signal, in the output of the PA, will introduce modulation around the injected carrier. This will exhibit similar spectral characteristics to the wanted modulation, but will occupy a greater bandwidth (due to the spectral spreading inherent in the IMD generation process).

The original (wanted) output signal will be unaffected by the injection of the carrier, so long as it is of a sufficiently low level not to significantly affect the peak power at which the transmitter is operating. Clearly, a very large signal could force the PA into saturation and this would then affect the intermodulation performance in and around the wanted channel.

As an example, consider that the wanted channel consists of a two-tone test with equal tone amplitudes, given by (4.98).

$$W(t) = M\cos[\omega_1 t + \phi_1] + M\cos[\omega_2 t + \phi_2] \tag{4.98}$$

The intermodulation processes occurring in the output of the power amplifier will be complex; however, they may be modelled, to a first approximation, by a cubic characteristic. This is sufficient to illustrate the basic mechanisms by which distortion is transferred to the other signals in the output spectrum. The amplifier characteristic may therefore be described as:

$$P(\alpha(t)) = A\alpha(t) + B\alpha^3(t) \tag{4.99}$$

It is not obvious that the value of B for the normally amplified signal will be the same as for the reverse-injected signal (which is not amplified by the amplifier); however, for the purposes of this illustration, it will be assumed that it is.

If it is assumed, for the present, that the loop is not operational, the two basic mechanisms by which signals appear in the downconverter chain may

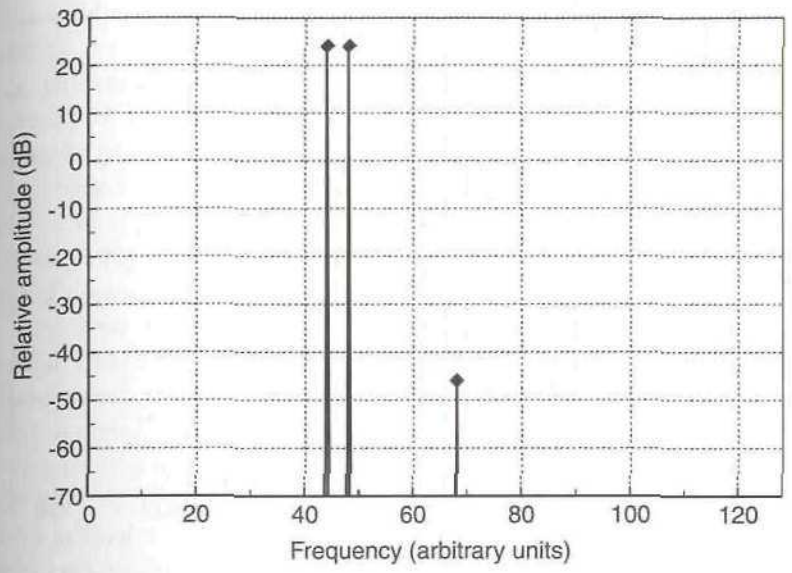


Figure 4.56 Feedback path RF signal assuming imperfect output coupler directivity only.

be illustrated separately. The first is due to imperfect coupler directivity in the output coupler, and this is illustrated in Figure 4.56 for an output coupling factor of 20 dB, an assumed coupler directivity of 20 dB, and an injected carrier level of −30 dBc.

The second mechanism is due to intermodulation occurring in the amplifier and affects both the wanted channel and the injected interferer. This is illustrated in Figure 4.57 for the above coupler and carrier values and the following signal and amplifier parameters: $A = 100$, $B = -3.5$, and $M = 1$.

If these two mechanisms are combined, it is clear that the second dominates with the values chosen above, that is, amplifier IMD is the most significant effect. This situation is illustrated in Figure 4.58.

The action of the loop will be to view the injected carrier as a form of 'distortion', as it does not appear in the original input signals. Similarly, the distortion generated by the injected carrier intermodulating in the output of the PA, will be viewed by the loop as a further form of distortion. It will therefore attempt to correct for these 'distortion' mechanisms in the same way as for any other form of distortion, so long as they appear within the loop bandwidth. This therefore creates two possible cases: injected signals appearing *within the loop correction bandwidth and signals appearing* outside the correction bandwidth. These two cases will be dealt with separately below.

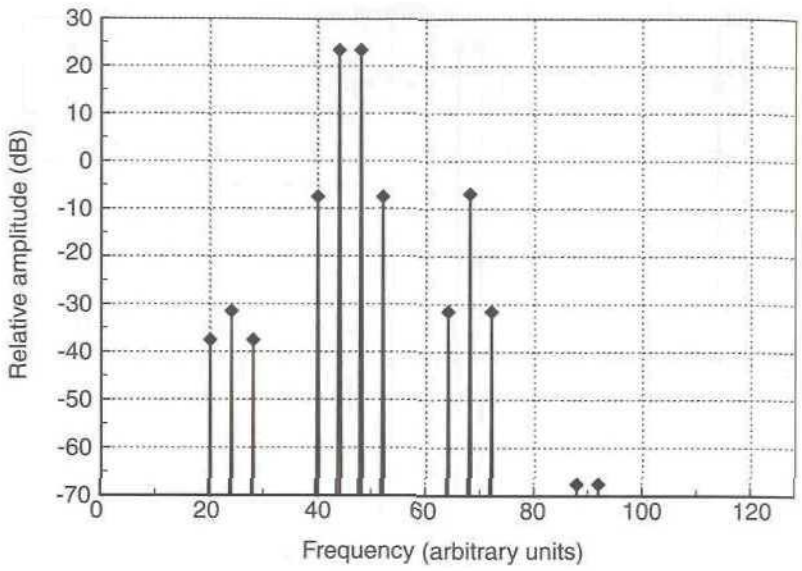


Figure 4.57 Feedback path RF signal caused by amplifier IMD only.

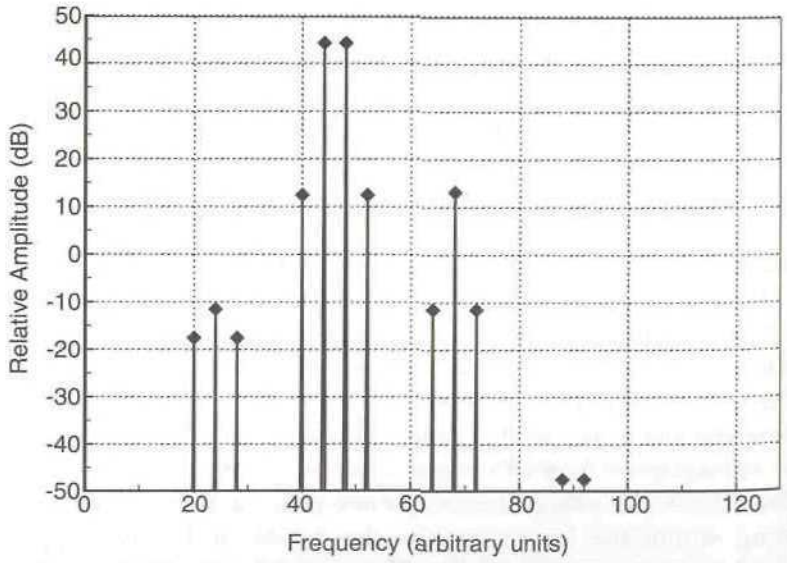


Figure 4.58 Feedback path RF signal caused by both mechanisms.

In addition to the mechanisms described above, the process of generating an SSB correction signal, in order to cancel out the injected carrier, will result in an image of this correction signal also being generated. The level of this image will be determined by the basic loop image suppression performance, which, in turn, is dominated by the downconverter amplitude and phase balance. If this mechanism is added to the situation illustrated in Figure 4.58, the relative level of the image signal can be seen (a 30 dB image suppression is assumed in this simulation). This is illustrated by the filled square markers in Figure 4.59.

The final output spectrum may now be deduced by assuming a particular level of loop gain and hence IMD removal. This will not affect the image level, since it has already been assumed that this is determined by the downconverter performance, which would only be true in the presence of loop gain (the upconverter would otherwise determine this parameter). If a loop gain of, say, 20 dB is assumed, then each of the unfilled diamond markers in Figure 4.59 would be reduced by 20 dB (with the exception of the original two tones). This results in a general spectral improvement of 20 dB, with the important exception of the image signal from the injected carrier (to the left of the wanted channel in Figure 4.59). The distortion surrounding this signal will be reduced, but the CW signal in the center will only be reduced to the level of the filled square marker; this image level then becomes dominant.

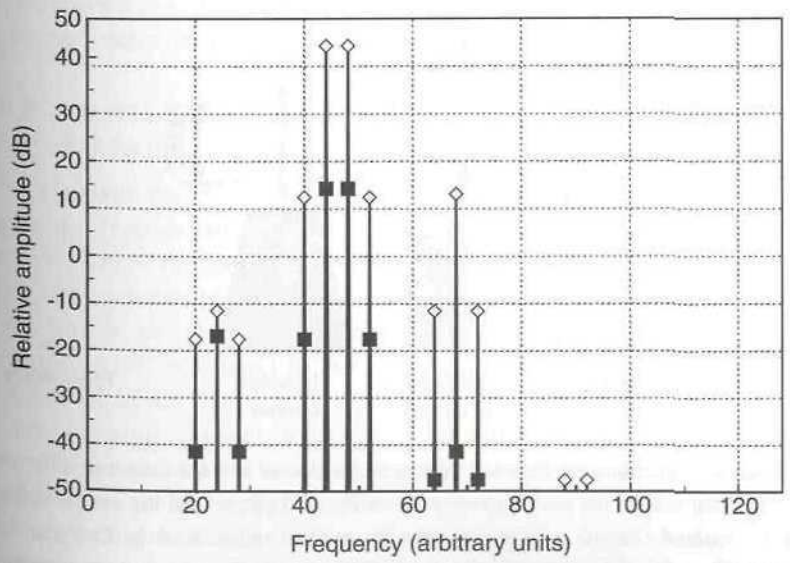


Figure 4.59 Feedback path RF spectrum with a superimposed image signal.

4.8.8.1 Injected Signals Appearing Within the Loop Correction Bandwidth

When the injected carrier appears at a spacing from the wanted channel center frequency such that it can be accommodated within the loop correction bandwidth, the loop will attempt to remove this unwanted 'distortion' as outlined above. The loop will do this by generating an additional signal complementary to the injected signal *in I and Q components*. This additional signal is, of course, 'generated' by the normal error signal creation process within the loop.

When this error signal is upconverted, it will yield two sidebands, one designed to cancel the unwanted injected carrier, and the other caused by the imperfect image rejection of the upconverter. This second sideband (and any error in the first) will in turn be corrected by the action of the I and Q feedback, since the imperfections occur in the forward path of the loop. This correction will, however, itself be imperfect, due to the finite image rejection (I/Q balance) of the downconverter used in the error path. The result is therefore a near-perfect correction of the injected carrier at the output of the loop (to a degree determined by the loop gain), but an imperfect suppression of the resulting image signal created within the loop itself (caused by I/Q imbalance in the downconverter).

The resulting spectrum is therefore of the form shown in Figure 4.60. In this figure, the injected carrier is supplied to the transmitter output via a

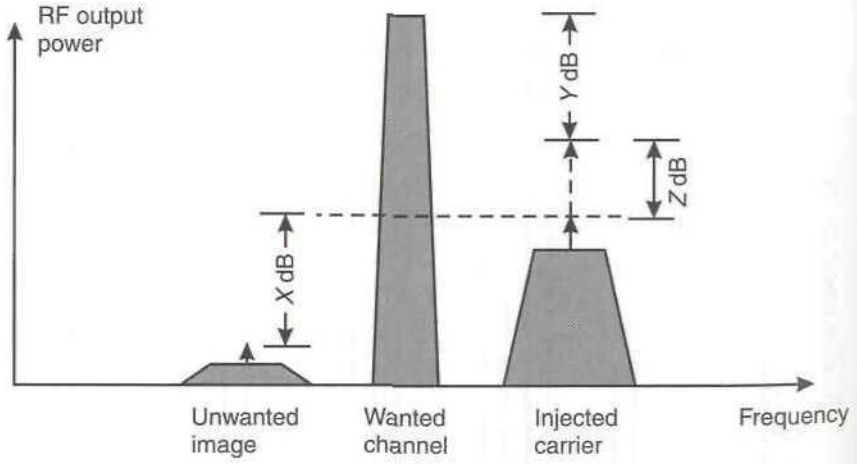


Figure 4.60 Form of spectrum obtained by injecting a carrier into the output of a Cartesian loop, where the spacing between the injected carrier and the centre of the wanted channel is within the loop correction bandwidth of the Cartesian loop. Note that the injected carrier is the CW signal in the centre of the resulting modulation, which surrounds it.

directional coupler (or similar), at a level Y dB below the peak output of the transmitter. This signal, when viewed at the output of the transmitter, will appear at a level Z dB below its injected level, where Z is determined by either:

1. The directivity of the injection coupler *in circuit*. Note that this directivity may well be much poorer than specified by the coupler manufacturer, as the coupler terminating impedances are unlikely to be a good match.

or

2. The loop gain of the Cartesian loop.

If, for example, the loop gain of the Cartesian loop is low and the coupler directivity is good, then the injected carrier level, appearing in the output spectrum, will be determined by the loop gain alone (i.e., Z will equal the loop gain). If the converse is true, then the level is determined purely by the injection coupler directivity.

The suppression of the unwanted image, $(X+Z)$ dB, is determined primarily by the I/Q balance of the downconverter. This will typically (for a narrowband, high-gain loop) be of a similar order to the loop gain and hence result in similar image and injected carrier levels in the output spectrum.

Note that any intermodulation distortion generated in the output of the power amplifier, as a result of the injected carrier, will be removed by the normal action of the loop to a degree directly determined by the loop gain. In the case of a low-gain loop, this IMD may still be significant, but it should not be noticeable in a high-gain design.

4.8.8.2 Injected Signals Appearing Outside of the Loop Correction Bandwidth

In the opposite case to that considered above, the mechanism by which the output spectrum is corrupted is slightly different. Once a signal is injected (well) outside of the loop correction bandwidth, the loop no longer has any ability to generate and supply an appropriate correction signal. The spectral corruption is therefore determined primarily by the degree of reverse intermodulation generated in the PA output (as this is no longer able to be corrected by the loop).

As a secondary effect, an image signal will also be generated, as the loop gain, outside of the loop bandwidth, will not necessarily be sufficiently low for this effect to be negligible. Thus, for example, a loop with a transfer function incorporating a gain of 40 dB within its loop bandwidth and a predominantly first-order roll-off, will have a gain of 30 dB at 10 times its 3 dB frequency-offset from carrier. Therefore, 30 dB of gain could still exist

10 MHz (say) from the carrier frequency. An image signal, generated in the upconverter, from a reverse-injected carrier at this offset, will therefore only exhibit an additional 10 dB of attenuation over that determined by the upconverter I/Q balance specification. The loop will have no gain at this offset to enable the downconverter to correct this image, and hence the image level appearing at the output of the system will be 10 dB below the I/Q modulator image specification at this offset. This could well be of a significant level in many systems.

The amount of reverse IMD generated will be determined by the PA design itself, and will often be significant. It can be recognised as IMD, in the case where the transmitter is supplying a modulated wanted signal, as the IMD will be of a form determined by the modulated channel. For example, if the transmitter is supplying a π/4-DQPSK signal, then the IMD product will exhibit spectral characteristics of a form determined by, but not the same as, the original π/4-DQPSK signal.

The resulting form of the output spectrum is shown in Figure 4.61, with the symbols having the same meanings as discussed above with reference to Figure 4.60. The variables X_1 and X_2 will be determined by the characteristics of the loop under consideration; X_1 will either be set by the image level or the reverse IMD level as described above, and X_2 will be determined by the reverse IMD level.

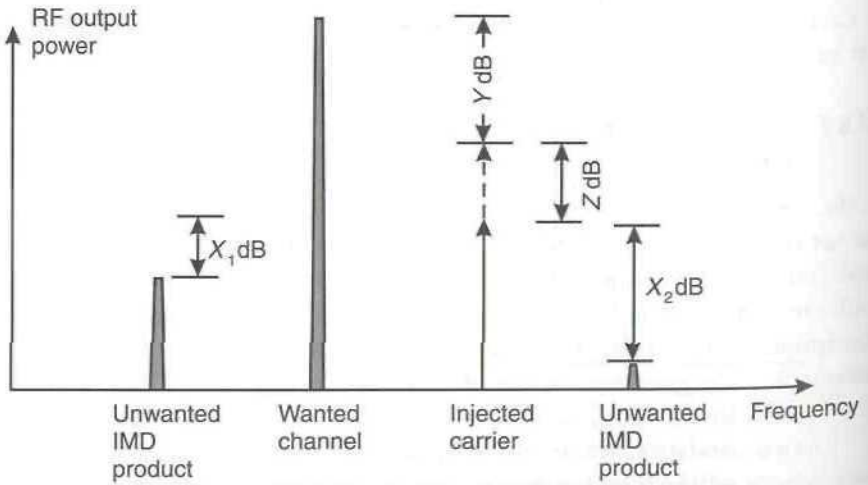


Figure 4.61 Form of spectrum obtained by injecting a carrier into the output of a Cartesian loop, where the spacing between the injected carrier and the centre of the wanted channel is greater than the loop correction bandwidth of the Cartesian loop.

The main difference between the two cases lies in the greater potential for wider spectral corruption where the injected carrier lies well outside of the loop correction bandwidth. This arises from the fact that many IMD products could be created where the PA has a poor reverse-IMD response, rather than the suppressed IMD and image signals which are created in the previous case. In practice, the fact that the injected carrier, in the case of most standard tests, does not appear at the same level as the wanted output signal (i.e., Y is usually around 30 dB, and not 0 dB), is of great benefit. This arises from the fact that the PA output is effectively seeing an 'unequal two-tone test' and hence one IMD product will be significantly larger than its mirror image. This is illustrated by X_2 being much larger than X_1 in Figure 4.61. The smaller IMD product is often small enough to be negligible, leading to, at a first glance, similar looking spectra for the two cases of injection within and outside of the loop bandwidth.

4.8.8.3 Practical Illustration

As an illustration of the mechanisms described above, some practical results are presented from a hardware Cartesian loop transmitter. The output spectrum produced by the loop in response to a two-tone test is illustrated

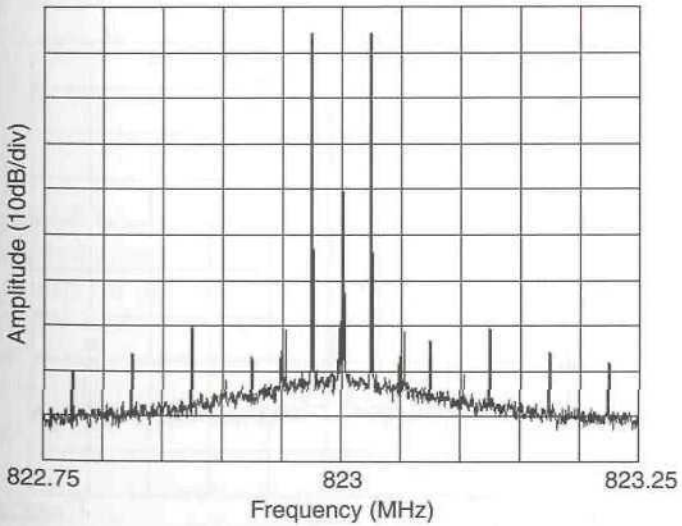


Figure 4.62 Close-in RF output spectrum obtained from a practical 800 MHz loop when injecting a carrier into the output, where the spacing between the injected carrier and the centre of the wanted channel is within the loop correction bandwidth of the Cartesian loop.

in Figure 4.62. This spectrum remains unchanged irrespective of whether or not a carrier is injected into the output of the loop; the basic IMD specification is around 65 dBc with respect to the level of the tones, at 6W PEP output power.

A wider view of this spectrum is illustrated in Figure 4.63; here the injected carrier (at 824 MHz) and its associated distortion can clearly be seen. In addition, the image signal and its associated distortion are clearly visible at 826 MHz. Note the increased spectral spreading of both of these unwanted components relative to the spreading of the original channel; this may be a significant factor in some systems.

The spacing between the injected carrier and the wanted channel is within the loop correction bandwidth in this case, but the spacing is still large (1 MHz) and hence the degree of correction obtained will be small (around 4 dB). This spacing was chosen as it is small enough to allow the individual spectral lines to be clearly distinguished, but large enough that the IMD suppression obtained is small (hence allowing all potential components of the output spectrum to be clearly seen).

It is possible to predict the level of the injected carrier and image signals appearing in the output spectrum as follows.

In the case where the separation between the wanted channel and the injected interferer is within the loop IMD correction bandwidth, the level of

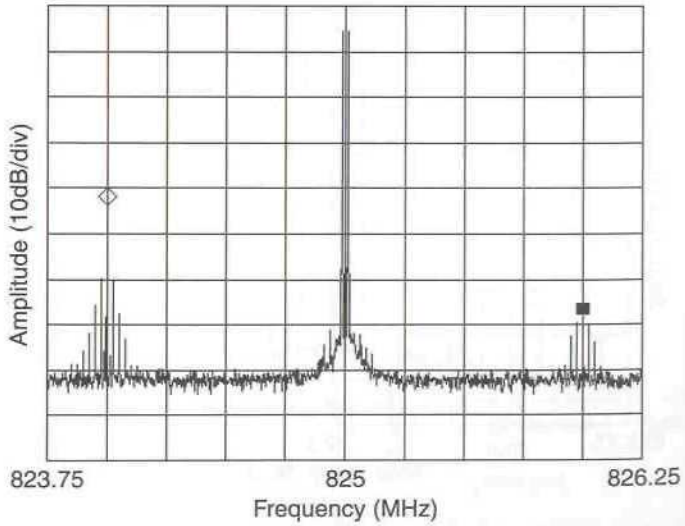


Figure 4.63 RF Output spectrum obtained from a practical 800 MHz loop by injecting a carrier into the output, where the spacing between the injected carrier and the centre of the wanted channel is within the loop correction bandwidth.

the dominant third-order IMD product generated by the injected carrier and the wanted channel (taking account of the loop correction on this product) is given by:

$$I = P_{TX} - Y - D - G_L \tag{4.100}$$

where: P_{TX} is the transmitter output power (wanted channel, dBm), Y is the injected tone level, taking account of coupler directivity and PA output effects (dBc), D is the level of the resulting IMD (dB relative to the injected tone level), and G_L is the loop gain (dB). The above assumes that the downconverter linearity is not the limiting factor (which it may be in some cases, although it too is seeing an unequal and hence less demanding two-tone test). Note that a second third-order IMD product will also be generated, however, this will be significantly lower in level than I above and hence can usually be neglected. The quantity 'D' may be estimated from an open-loop measurement on the power amplifier alone, with an appropriate, unequal, two-tone test to represent the wanted channel and the injected interferer. Note that the wanted channel should be running at an appropriate peak power level (i.e., full output for the transmitter) in order to get a realistic idea of the IMD level. The addition of the injected tone will almost certainly drive the PA into a very nonlinear part of its characteristic and hence give a relatively high level of IMD.

In this case, for example, at a 1 MHz offset for the injected tone:

$$\begin{aligned} D &= +37 - 39 - 13 - 9 \\ &= -24 \quad \text{dBm} \end{aligned} \tag{4.101}$$

This corresponds to a level of −61 dBc (with respect to the wanted channel), which is approximately the level shown in Figure 4.63.

Given that the other third-order IMD product is significantly smaller than that calculated above, the injection side of the carrier is therefore dominated by the residual injected carrier level (taking account of coupler directivity and PA output effects) and its associated IMD. The level of this carrier is given by:

$$C = \begin{cases} P_{TX} - Y - G_L & \text{well within the loop IMD correction bandwidth} \\ P_{TX} - Y & \text{well outside the loop IMD correction bandwidth} \end{cases} \tag{4.102}$$

provided that the in-circuit directivity of the injection coupler is not the dominant effect (which it is for Figure 4.63). In this case, the loop gain, G_L,

can have no corrective effect, as the 'problem' is occurring outside of the loop. The situation both within and outside the loop bandwidth then becomes the same and is set by the injection level, Y, at approximately 39 dBc (for Figure 4.63). An example of the loop providing correction is shown in Figure 4.65; here the injection coupler in-circuit directivity is sufficient to allow cancellation of the injected tone to be seen.

In this case, (4.102) may alternatively be defined in terms of the loop IMD correction characteristic, $G_L(f)$:

$$C = P_{TX} - Y - G_L(f) \tag{4.103}$$

In this manner, the roll-off between the frequency offsets at which the loop is providing significant correction and those at which little or none is possible, may be described.

In the case of the unwanted image signal, this appears at a low level on the opposite side of the wanted channel to the injected carrier. The level of this signal is given by:

$$E = P_{TX} - Y - S_{IR} \tag{4.104}$$

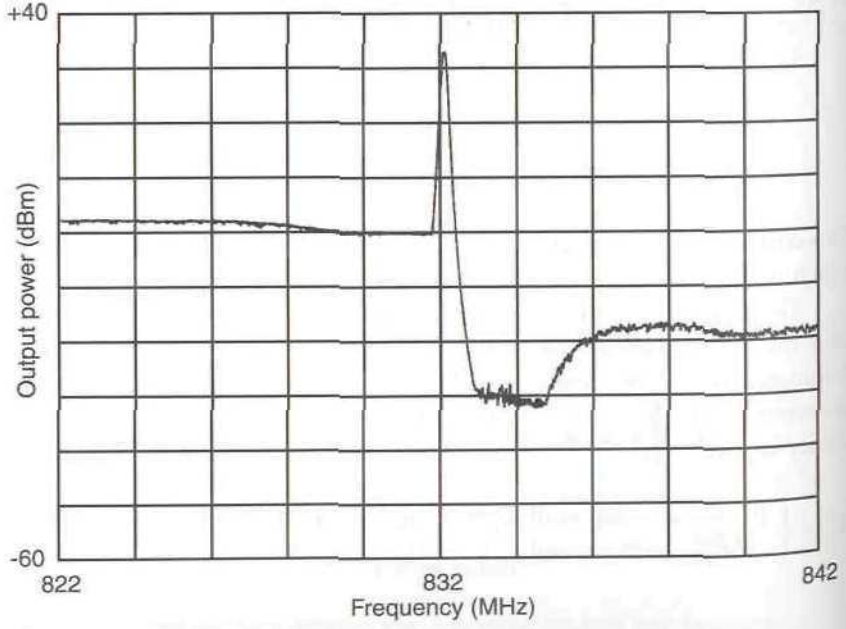


Figure 4.64 Swept reverse-injected carrier response with Cartesian feedback disconnected and the carrier injected below the channel centre frequency.

where S_{IR} is the suppression caused by the image rejection capabilities of the downconverter.

In this case:

$$\begin{aligned} E &= +37 - 39 - 43 \\ &= -45 \quad \text{dBm} \end{aligned} \tag{4.105}$$

This corresponds to a level of 82 dBc (with respect to the wanted channel), which is below the level of the IMD signal in Figure 4.63 (the peak power of the transmitter being +37 dBm).

Note that the injection coupler directivity will be affected by the input and output match it sees and will be very poor, even for a moderately well-matched system. The coupler used in deriving the results shown in Figure 4.63 had a directivity of over 25 dB when used with good 50 ohm terminating impedances, but this value reduced to around 5 dB or less (depending upon frequency) with the PA output impedance substituted.

The previous results illustrate what occurs at one particular frequency offset for the injected carrier. It is also interesting to examine the effect over a range of frequency offsets and to compare the effect of having Cartesian feedback present to that of the power amplifier alone. The former case is illustrated in Figure 4.65 and the latter in Figure 4.64.

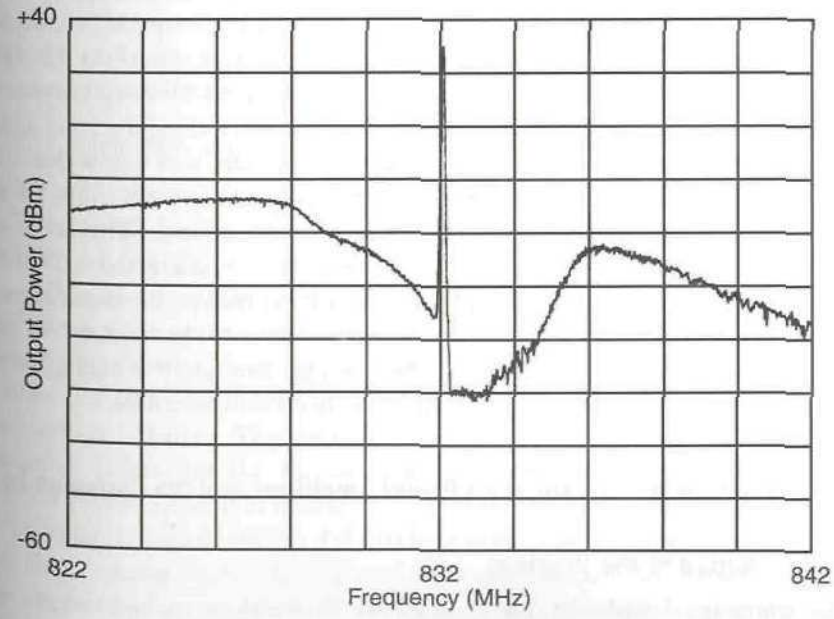


Figure 4.65 Swept reverse-injected carrier response of the complete Cartesian loop with the carrier injected below the channel centre frequency.

In Figure 4.64, it is clear that the output of the PA provides a relatively constant reflection of the injected carrier (or a constant degradation of the injection coupler directivity), as the injected carrier level is almost constant over the 20 MHz frequency range of this plot. The effect of this injected carrier could also be expected to be constant, as no feedback is applied in this case; however, it can be seen to improve at a separation from the carrier of 3 MHz or less. This can be explained by examining the effect of broadband modulation on the PA used in the experiments, as the wider separations between the injected carrier and the channel center frequency effectively create a two-tone test with unequal tone levels and a high envelope frequency. Since this PA was not designed for broadband modulation formats, its decoupling is not adequate at these envelope frequencies and hence the IMD performance of the PA degrades with increasing envelope frequency, giving the 'filter-like' response seen in Figure 4.64. A better design of PA (from this perspective) would yield a flat response, at a level similar to that seen close-in in this plot.

Figure 4.65 shows the swept response to the same type and level of injected carrier as that of Figure 4.64. The resulting characteristic is now somewhat different, however, as the action of the loop dominates in most areas. Taking the trace to the left of the channel center frequency, this begins at a level similar to that of Figure 4.64, increases slightly and then reduces sharply as the channel is approached. This mirrors the loop IMD cancellation response, in which IMD is actually enhanced at large offsets (in this case, around 4 MHz from carrier), before reducing sharply as the loop correction increases. This clearly shows the ability of the loop to reduce the level of the injected interferer appearing at the output of the system, well below that seen without feedback correction. Similarly, the lower trace to the right of the channel exhibits a similar (but more pronounced) response. This is now a genuine image or distortion signal and consequently is corrected within the loop correction bandwidth (in this case to a level below the measurement noise floor of this plot). It also experiences a similar 'enhancement' to that seen for the injected carrier, at an offset of approximately 4 MHz, before reducing to the same level (on the RHS) as that of Figure 4.64.

4.8.9 The Problem of Saturating Power Amplifiers and the Cartesian Loop

4.8.9.1 Nature of the Problem

The Cartesian loop is a linear feedback correction system, where the feedback is performed in quadrature components at baseband. By this mechanism, it is capable of correcting for both AM/AM and AM/PM conversion effects in the RF power amplifier, a feat it achieves by means of a complex predistorted error signal which is upconverted to feed the power amplifier chain. Thus, for the loop to operate correctly, the distortion effects in the forward path (upconverter, pre-driver and power amplifier) must be capable of being corrected by a predistorting signal.

When considering a forward path component being driven into saturation, the only method of correcting for the distortion is to 'replace' the lost power, thereby 'completing' the cut-off waveform. This is the function effectively performed by some alternative linearisation techniques, such as feedforward. In feedforward, the additional power can be supplied by the error amplifier, however, in the case of the Cartesian loop, no addition of output power can be achieved and hence the saturated waveform cannot be corrected.

In some systems with only a modest linearity requirement (e.g., DAMPS and PDC) the distortion created by an open-loop saturating power amplifier is not necessarily sufficient to break the mask. Thus, the Cartesian loop does not need to linearise this part of the characteristic in order to fulfil the required specification in these systems. Unfortunately, however, the loop is not 'intelligent' and hence cannot distinguish between the parts of an amplifier characteristic which could benefit from linearisation, and those parts where it is unnecessary; it will attempt to linearise both.

If no steps are taken to resolve this problem, the use of linearising feedback will appear to be detrimental, since the performance of the linearised transmitter will be worse than its open-loop counterpart.

4.8.9.2 Illustration of the Problem

Figure 4.66 shows the response of a commercially-available IC power amplifier, used in a number of handset designs, which fulfils the PDC specification (see Table 4.2).

When this amplifier is included in a Cartesian loop and operated at the same mean power level, the result is as shown in Figure 4.67. It is evident from this figure that the required specification is far from being met at both 50 kHz and 100 kHz offsets, despite the amplifier being included within a linearisation scheme. The fact that the amplifier is being driven into saturation is causing the Cartesian loop to produce a correction signal which is very wideband in nature, thereby degrading the output spectrum. If the amplifier is backed-off (by 1.3 dB) such that saturation no longer occurs, then the response shown in Figure 4.68 is generated. This shows an IMD improvement of some 18 dB (at 50 kHz offset), which is far greater than would be expected from the back-off alone, thus indicating that the loop is satisfactorily linearising the PA.

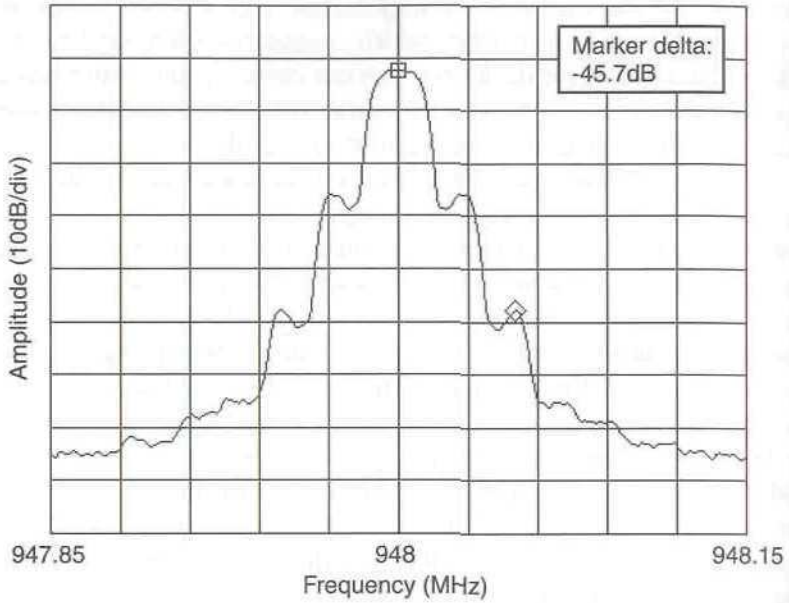


Figure 4.66 Response of the IC PA when operating at full rated output power (+30.5 dBm) without Cartesian feedback.

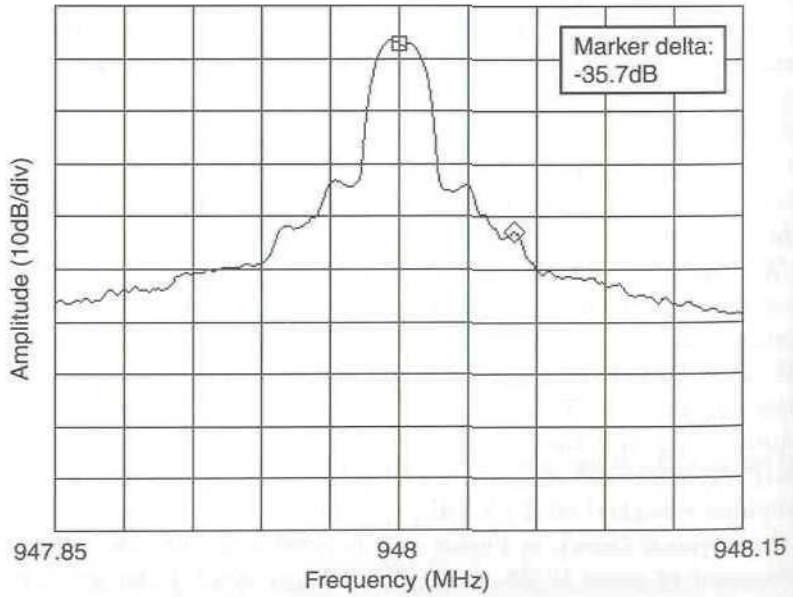


Figure 4.67 Closed-loop response of the Cartesian loop when the IC PA is operated at full rated power (+30.5 dBm).

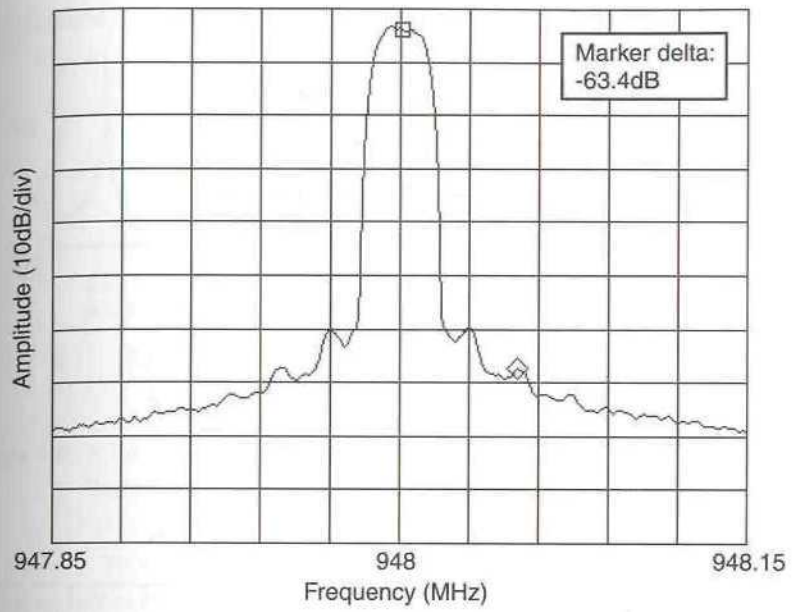


Figure 4.68 Closed-loop response of the Cartesian loop when the IC PA is operated at 1.3 dB mean power backoff (equivalent to 2 dB of baseband input backoff for both I and Q channels).

It is possible to illustrate the mechanism by which the loop generates significant additional distortion by means of an idealised simulation. Consider the ideal amplifier characteristic shown in Figure 4.69. This will generate a clipped RF envelope of the form shown in Figure 4.70, for a sinusoidal envelope variation. The loop error signal generated by a correctly operating Cartesian loop will then be of the form shown in Figure 4.71. The large peaks generated by the loop are intended to correct for the nonlinearity occurring at $|v| > V_C$ in Figure 4.69. The effect of these peaks on the error signal spectrum is clearly illustrated in Figure 4.72, with the spectrum of the clipped envelope (unlinearised) also being shown for reference. It is clear that the spectrum of the clipped sinewave, whilst heavily distorted, is not as severely affected as that of the loop error signal. The loop will therefore clearly generate a much poorer output spectrum than that of the clipping amplifier alone.

4.8.9.3 Baseband Clipping as a Potential Solution

It is clear that the unlinearised power amplifier, as shown in Figure 4.65, is capable of meeting the required specification, despite being driven into

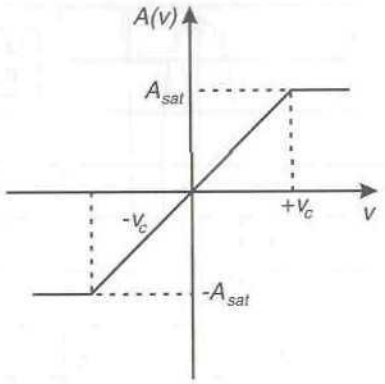


Figure 4.69 RF power amplifier characteristic.

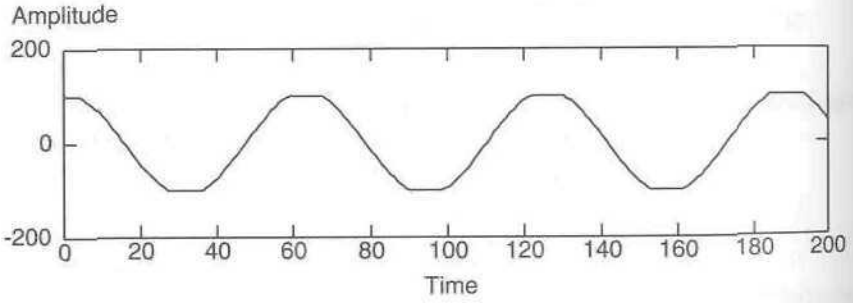


Figure 4.70 Clipped amplifier output signal (open-loop) – arbitrary units.

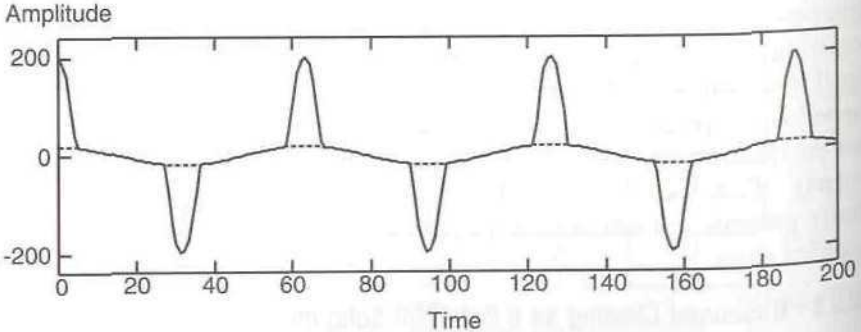


Figure 4.71 Resultant error signal (solid line), with clipped amplifier output signal (dashed line) for reference – arbitrary units.

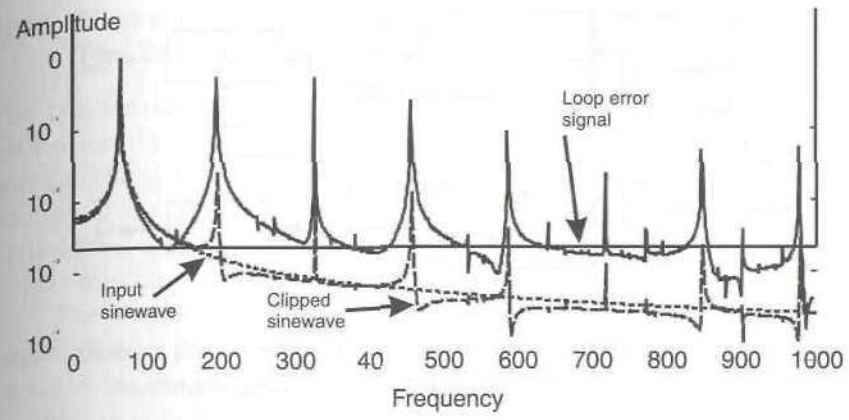


Figure 4.72 FFT-generated signal spectra (arbitrary units). Dotted line: Input sinewave, Dashed line: Clipped sinewave (output of PA), Solid line: Loop error signal.

saturation. Thus clipping of the waveform peaks of the RF signal, *per se*, is not a problem in this type of quasi-linear system, it is only the exaggeration of this limiting action which occurs when the feedback loop overloads, which results in excessive spectral spreading.

It is therefore logical to conclude that predistortion of the baseband input signals, in the form of clipping of these signals, is a good potential solution to the problem. If the input signals are clipped to the extent that they still meet the required emission mask (at baseband), then the linearly upconverted and amplified signals from the Cartesian loop should also meet the required mask. The signal clipping will eliminate the need for the loop to overdrive the PA in an attempt to generate additional corrective power at the saturation level, a point at which it is, by definition, incapable of doing so.

The ideal form of clipping would create a circular, rather than a square constellation, and this can be achieved by constraining (i.e., clipping) the magnitude of the vector in the I-Q plane. This form of clipping would be difficult to achieve accurately by conventional analog means, and would involve a significant amount of analog signal processing. A much simpler solution is to recognise the fact that DSP methods are commonly used to generate the I-Q baseband signals, and would almost certainly be used in the proposed flexible architecture type of terminals (see Section 4.6.6). Clipping may therefore be performed in the digital domain, prior to D/A conversion.

The use of the existing DSP has the advantage that no additional hardware is required, and that the clipping can be performed to a very high degree of precision. The disadvantage is that the DSP task has now become

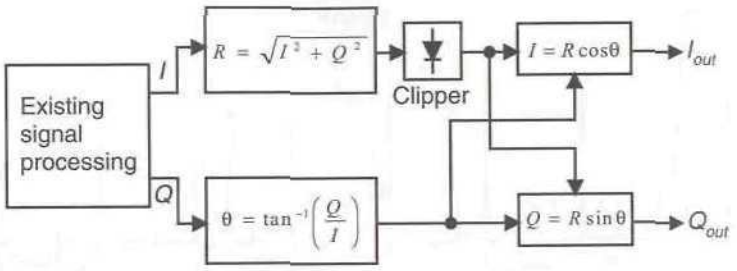


Figure 4.73 One form of DSP algorithm which could be used to provide amplitude clipping; *there are a number of other (possibly more efficient) alternatives.*

more complex; with more processing power being required and new algorithms needing to be added. The additional processing power is not, however, very large and hence should not prove a major drawback.

The form of the processing which could be employed in order to provide this type of amplitude clipping is shown in Figure 4.73. The original I and Q signals, generated by conventional means, are supplied to a Cartesian to polar co-ordinate conversion system. The resulting amplitude signal is then clipped, either by multiplying it by a fixed value ('gain') in order to precipitate numerical overload and hence 'saturation' (this would need to be followed by an appropriate 'attenuation' of the same value as the gain in order to return the signal to its original relative level), or by comparing the absolute numerical amplitude values with a fixed level (i.e., number; the 'clipping' level), and setting to this level any values which were greater (with the addition of the appropriate sign). Once the clipping function had been performed, the resulting polar signals could then be returned to Cartesian form before D/A conversion and feeding to the Cartesian loop.

The polar amplitude vector would thus be constrained to a level consistent with meeting the appropriate mask and with not requiring the PA to go beyond saturation (i.e., such that it *just* entered saturation, *but no* more). This clipping need, of course, only be employed for the quasi-linear modulation formats required of the flexible architecture radio. When truly linear formats are being used (e.g., 16-QAM), the PA must be backed-off by an appropriate amount to allow the peaks of the modulation to be accommodated within the peak power rating of the amplifier. The clipping code should then be bypassed as it is, of course, unnecessary.

4.8.10 Illustration of Practical Problems Using a Frequency-Offset Two-Tone Test

The creation of an I and Q two-tone test is detailed in Chapter 2 and summarised in Table 2.4. If such a test is configured so as to create an unequal spacing between the carrier and the upper and lower tones, a number of useful diagnostic performance indications for a Weaver-based Cartesian loop transmitter can be found in the resulting two-tone spectrum. These are summarised in Figure 4.74.

The wanted two-tones are A_0 and A_1 and are non-symmetrically placed around the carrier, B. The carrier frequency, F_{LO} is determined purely by the local oscillator frequency. The level of B relative to A_0 and A_1 provides an indication of the loop carrier suppression and hence of DC-null performance (where applicable).

The tones C_0 and C_1 are the imperfectly suppressed in-band image, due to gain and phase imbalances in the feedback path. Note that they are 'mirrored' with respect to the original tones.

D_0 and D_1 are spaced equally about the center of the channel (as denoted by the imperfectly-suppressed carrier, B) and have a spacing equal to the tone spacing. In other words they are separated from the carrier by the envelope frequency. These tones normally indicate that amplitude modulation of the carrier is taking place; this is usually caused by insufficient power supply decoupling (or poor grounding), thus allowing the supply to modulate the output spectrum directly.

E_0 and E_1 are separated from the wanted tones by the spacing of the wanted tones. These are therefore third-order intermodulation products not

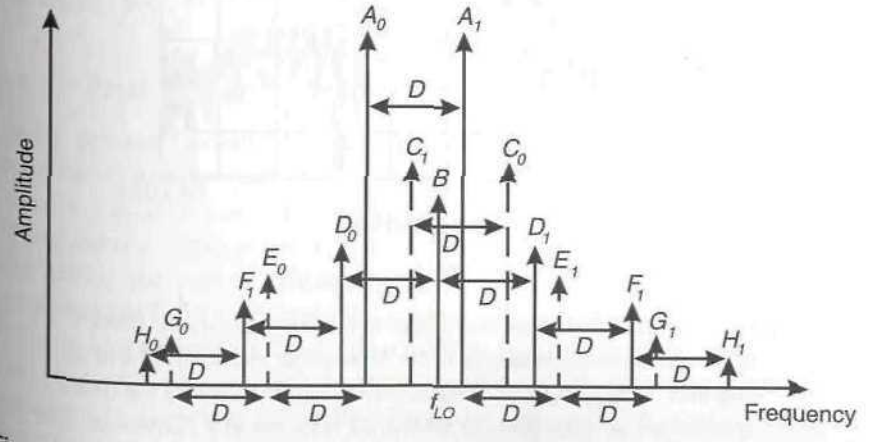
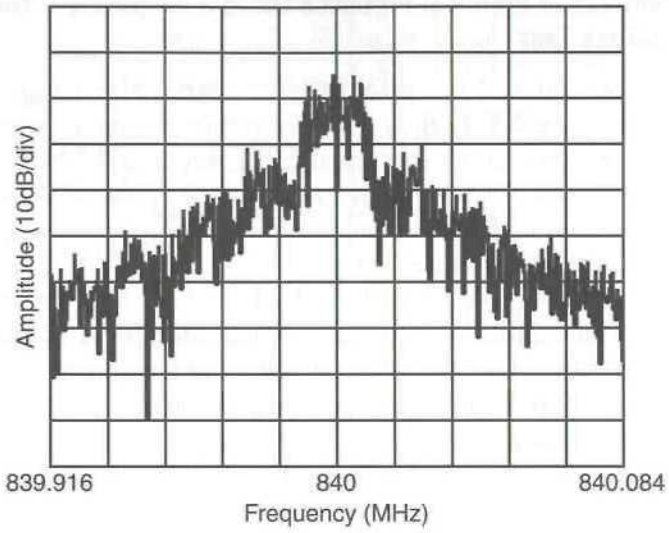
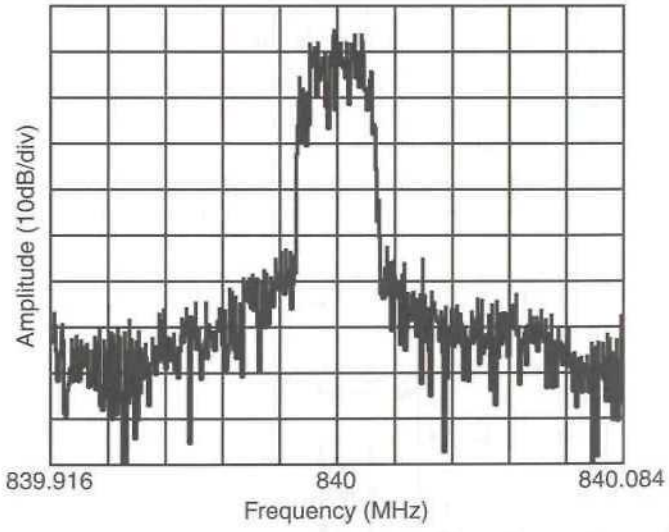


Figure 4.74 Offset two-tone test applied to a Cartesian transmittter.

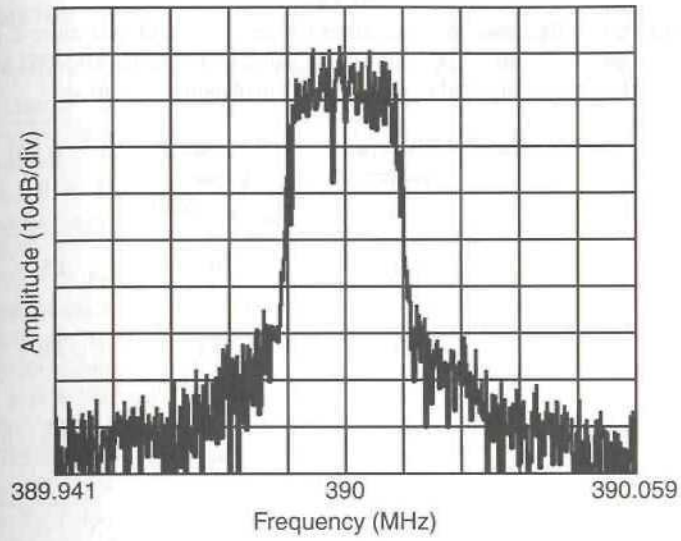


(a)



(b)

Figure 4.75 Practical test results from two Cartesian-loop transmitters: (a) DAMPS π/4-DQPSK, uncorrected response of the RF amplifier chain. (b) π/4-DQPSK response of the complete Cartesian-loop transmitter meeting the DAMPS specification at +33.5 dBm. (c) π/4-DQPSK response of a TETRA-compliant Cartesian-loop transmitter at +40.5 dBm. © Wireless Systems International Ltd.



(c)

completely suppressed by the action of the loop, or resulting from intermodulation distortion in the downconversion chain.

F_0 and F_1 and H_0 and H_1 are higher-order modulation or modulation distortion products caused by the same mechanism as D_0 and D_1.

G_0 and G_1 are imperfectly-suppressed fifth-order intermodulation products and may be caused by the same mechanisms as E_0 and E_1.

Thus it can be seen that a correctly-designed two-tone test can act as an extremely useful diagnostic aid in tracking down problems with a practical Cartesian loop.

4.8.11 Practical Results from a Cartesian-Loop Transmitter

Some practical results for π/4-DQPSK systems using a Cartesian-loop transmitter are provided in Figure 4.75. These illustrate the performance in both a modest linearity, but highly power efficient, application and in a high-linearity application. For the former case (American digital cellular, DAMPS), the power efficiency and adjacent channel power figures are provided in Tables 4.6 and 4.7, respectively. Note that these results were obtained by employing back-off to prevent saturation of the power amplifier and not the clipping mechanism mentioned above. The high-linearity application illustrated is TETRA, and the result shown meets the required specification by more than 3 dB in both the first and second adjacent

Table 4.6
Power efficiency of a Cartesian-loop linearised transmitter, based on a class-C power amplifier, when operating with NADC (North American Digital Cellular, DAMPS) and PDC (Japanese 'Personal Digital Cellular') modulation formats.

Cellular scheme	Supply voltage (V)	DC input power (W)	RF output power (into 50Ω)	η (%)
NADC	6.0	2.04	1.26 W (+31 dBm)	62
NADC	4.8	1.25	0.776 W (+29 dBm)	64
PDC	6.0	2.07	1.26 W (+31 dBm)	61

Table 4.7
Adjacent channel power for the unlinearised and linearised class-C power amplifier, measured according to the North American Digital Cellular (DAMPS) standard.

Adjacent channel	Open loop ACP (dBc)	Closed loop ACP (dBc)	Improvement (dB)
First	−18	−53	35
Second	−28	−62	34
Third	−39	−64	25

channels. Both designs also easily met the required signal vector error (SVE) specifications of the relevant system.

Note that the power efficiency figures quoted in Table 4.6 do not take account of the power consumption of the Cartesian loop linearisation circuitry. This assumption may be justified on the grounds that much of the linearisation circuitry duplicates functions which are already present in the transmitter (e.g., upconversion, local oscillator quadrature generation and power control); the *additional* circuitry required is therefore relatively modest. The power consumption of these additional items, in, for example, an ASIC implementation, is therefore negligible with respect to the power consumed by the PA itself.

References

1. Black, H. S., U.S. Patent no. 1,686,792, 9 October 1928.
2. Gardner, F. M., *Phaselock Techniques*, New York: Wiley, 1979, pp. 168–170.
3. Dye, N., and H. Granberg, *Radio Frequency Transistors—Principles and Practical Applications*, Chapter 12, Butterworth Heinemann, MA 02180, USA, 1993.
4. Mitchell, A. F., "A 135 MHz feedback amplifier," *IEE Colloquium on Broadband High Frequency Amplifiers—Practice and Theory*, London, England, 22 November 1979, pp. 2/1–6.
5. Perez, F., E. Ballesteros, and J. Perez, "Linearisation of microwave power amplifiers using active feedback networks," *IEE Electronics Letters*, Vol. 21, January 1985, pp. 9–10.
6. Ballesteros, E., F Perez, and J. Perez, "Analysis and design of microwave linearized amplifiers using active feedback," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 36, March 1988, pp. 499–504.
7. Pedro, J., and J. Perez, "An MMIC linearized amplifier using active feedback," in *IEEE International Microwave Symposium Digest (MTT-S)*, Atlanta, Georgia, Vol. 1, June 1993, pp. 95–98.
8. Hu, Y., J. Mollier, and J. Obregon, "A new method of third-order intermodulation reduction in nonlinear microwave systems," *IEEE Trans. on Microwave Theory and Techniques*, MTT-34, February 1986, pp. 245–250.
9. Gajda, G., and R. Douville, "A linearization system using RF feedback," in *IEEE International Electrical and Electronics Conference*, Toronto, Canada, 1983, pp. 30–33.
10. Ezzeddine, A. K., H. A. Hung, and H. C. Huang, "An MMIC C-band FET feedback power amplifier," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 38, No. 4, 1990, pp. 350–357.
11. Faulkener, M., D. Contos, and M. Johansson, "Linearization of power amplifiers using RF feedback," *IEE Electronics Letters*, Vol. 31, 1995, pp. 2023–2024.
12. Arthanayake, T., and H. B. Wood, "Linear amplification using envelope feedback," *IEE Electronics Letters*, Vol. 7, No. 7, 8 April 1971, pp. 145–146.
13. Wood, H. B., British Patent No. 1 005 073.
14. Bruene, W. B., "Distortion reducing means for single-sideband transmitters," *Proc. of the Institute of Radio Engineers*, December 1956, pp. 1760–1765.
15. Petrovic, V., and W. Gosling, "A high efficiency, low-cost VHF/AM transmitter using V-MOS technology," *Communications 78, IEE Conference on Communications Equipment and Systems*, Birmingham, England, 4–7 April 1978, pp. 281–285.
16. Petrovic, V., and W. Gosling, "Polar-loop transmitter," *IEE Electronics Letters*, Vol. 15, No. 10, 10 May 1979, pp. 286–288.
17. Petrovic, V., and C. N. Smith, "Reduction of intermodulation distortion by means of modulation feedback," *IEE Colloquium on Intermodulation—Causes, Effects and Mitigation*, London, England, 9 April 1984, pp. 8/1–8/8.
18. Petrovic, V., "Reduction of spurious emission from radio transmitters by means of modulation feedback," *IEE Conference on Radio Spectrum Conservation Techniques*, September 1983, pp. 44–49.
19. Bateman, A., D. M. Haines, and R. J. Wilkinson, "Direct conversion linear transceiver design," *IEE 5th International Conference on Mobile Radio and Personal Communications*, Warwick, UK, December 1989, pp. 53–56.
20. Weaver, D. K., "A third method of generation and detection of SSB signals," *Proc. of the Institute of Radio Engineers*, No. 44, 1956, pp. 1703–1705.

21. Bateman, A., R. J. Wilkinson, and J. D. Marvill, "The application of digital signal processing to transmitter linearisation," *IEEE 8th European Conference on Electrotechnics*, Stockholm, Sweden, 13–17 June 1988, pp. 64–67.
22. Bateman, A., and D. M. Haines, "Direct conversion transceiver design for compact low-cost portable mobile radio terminals," *39th IEEE Vehicular Technology Conference*, San Fransisco, California, 1–3 May 1989, Vol. 1, pp. 57–62.
23. Johansson, M., and T. Mattsson, "Linearised high-efficiency power amplifier for PCN," *IEE Electronics Letters*, Vol. 27, No. 9, 25 April 1991, pp. 762–764.
24. Johansson, M., and T. Mattsson, "Transmitter linearization using cartesian feedback for linear TDMA modulation," *IEEE Vehicular Technology Conference*, St. Louis, Missouri, USA, 19–22 May 1991, pp. 439–444.
25. Briffa, M. A., and M. Faulkner, "Stability analysis of Cartesian feedback linearisation for amplifiers with weak nonlinearities," *IEE Proc. on Communications*, Vol. 143, No. 4, August 1996, pp. 212–218.
26. Boloorian, M., and J. P. McGeehan, "The frequency-hopped Cartesian feedback linear transmitter," *IEEE Trans. on Vehicular Technology*, Vol. 45, No. 4, November 1996, pp. 688–706.
27. Briffa, M. A., "Linearization of RF Power Amplifiers," Ph.D. thesis, Victoria University of Technology, Melbourne, Australia, December 1996.
28. Boloorian, M., and J. P. McGeehan, "Phase-lag compensated Cartesian feedback transmitter," *IEE Electronics Letters*, Vol. 32, No. 17, 15 August 1996, pp. 1547–1548.
29. Briffa, M. A., and M. Faulkner, "Gain and phase margins of Cartesian feedback RF amplifier linearisation," *Journal of Electrical and Electronics Engineering, Australia*, vol. 14, no. 4, December 1994, pp. 283–289.
30. Johansson, M., and T. Mattsson, "Transmitter linearization using Cartesian feedback for linear TDMA modulation," *IEEE Vehicular Technology Conference*, St. Louis, Missouri, USA, 19–22 May 1991, pp. 439–444.
31. Martin, P. M., and A. Bateman, "Practical results for a generic modem using linear mobile radio channels," *IEEE 41st Vehicular Technology Conference*, St. Louis, Missouri, USA, 19–22 May 1991, pp. 386–392.
32. Bateman, A., and J. P. McGeehan, "Phase-locked transparent tone-in-band (TTIB): a new spectrum configuration particularly suited to the transmission of data over SSB mobile radio networks," *IEEE Transactions on Communications*, Vol. COM–32, 1984, pp. 81–87.
33. Kenington, P. B., "Emerging Technologies for Software Radios," *IEE Electronics and Communications Engineering Journal*, Vol. 11, no. 2, April 1999, pp. 69–83.
34. ETS 300 392-2: "Radio Equipment and Systems (RES); Trans-European Trunked Radio (TETRA); Voice plus Data (V + D)," Part 2: Air Interface (AI), European Telecommunications Standards Institute (ETSI), 1996.
35. Wray, A. J., and D. M. Boscovic, "Overcoming the implementation issues of linear transmitter technology—Motorola's experience," *IEE Colloquium on Linear RF Amplifiers and Transmitters, Digest No. 1994/089*, 11 April 1994, pp. 1/1–1/6.
36. Kenington, P. B., R. J. Wilkinson, and K. J. Parsons, "Noise Performance of a Cartesian Loop Transmitter," *IEEE Trans. on Vehicular Technology*, Vol. 46, No. 2, May 1997, pp. 467–476.
37. Steffes, M., "Noise analysis of Comlinear's op-amps," Application Note OA-12, January 1993 (Comlinear are now part of National Semiconductor).

38. Faulkner, M., T. Mattsson, and W. Yates, "Automatic adjustment of quadrature modulators," *IEE Electronics Letters*, Vol. 27, No. 3, 31 January 1991, pp. 214–216.
39. Trans-European Trunked Radio (TETRA); Conformance testing specification; Part 1: Radio, ETS 300 394-1, European Telecommunications Standards Institute (ETSI).
40. Hull, C. D., J. L. Tham, and R. R. Chu, "A direct-conversion receiver for 900 MHz (ISM band) spread-spectrum digital cordless telephone," *IEEE Journal of Solid-State Circuits*, Vol. 31, No. 12, December 1996, pp. 1955–1963.
41. Hilborn, D. S., S. P. Stapleton, and J. K. Cavers, "An adaptive direct-conversion transmitter," *IEEE Trans. on Vehicular Technology*, Vol. 43, No. 2, May 1994, pp. 223–233.
42. Cavers, J. K., "The effect of quadrature modulator and demodulator errors on adaptive digital predistorters for amplifier linearisation," *IEEE Trans. on Vehicular Technology*, Vol. 46, No. 2, May 1997, pp. 456–466.
43. Faulkener, M., and T. Mattsson, "Spectral sensitivity of power amplifiers to quadrature modulator misalignment," *IEEE Trans. on Vehicular Technology*, Vol. 41, November 1992, pp. 516–525.
44. Adachi, F., and M. Sawahashi, "Error rate analysis of MDPSK/CPSK with diversity reception under very slow rayleigh fading and cochannel interference," *IEEE Trans. on Vehicular Technology*, Vol. 43, No. 2, May 1994, pp. 252–263.

5

Feedforward Systems

5.1 Introduction

It will be evident from the discussions of Chapter 4 that although feedback techniques have found wide application in low frequency systems, their use at radio frequencies is much more restricted.

The inventor of the feedback technique, H. S. Black, had also patented the feedforward system some nine years before [1,2] although subsequent to the invention of feedback, interest in his earlier invention has been limited. The inherent simplicity and elegance of feedback ensured its success, whilst feedforward was virtually ignored, with only minor exceptions, until the late 1960s.

The use of both feedforward and feedback in the same system received some attention by McMillan [3] and van Zelst [4]. There then followed a number of papers applying the techniques outlined by McMillan and van Zelst to particular scenarios, amongst which are work by Deighton *et al.* [5] and Golembeski *et al.* [6].

The modern interest in the radio frequency use of feedforward, however, began in earnest with the work of Seidel *et al.* at Bell Laboratories [7]. They were working on the construction of very high specification VHF amplifiers requiring high output powers, extremely high degrees of linearity, good time stability and broad bandwidths (40% of centre frequency). The only reasonable solution to the extremely tight specification with which they were faced was the use of feedforward.

Work on feedforward systems continued at Bell Laboratories and a flat

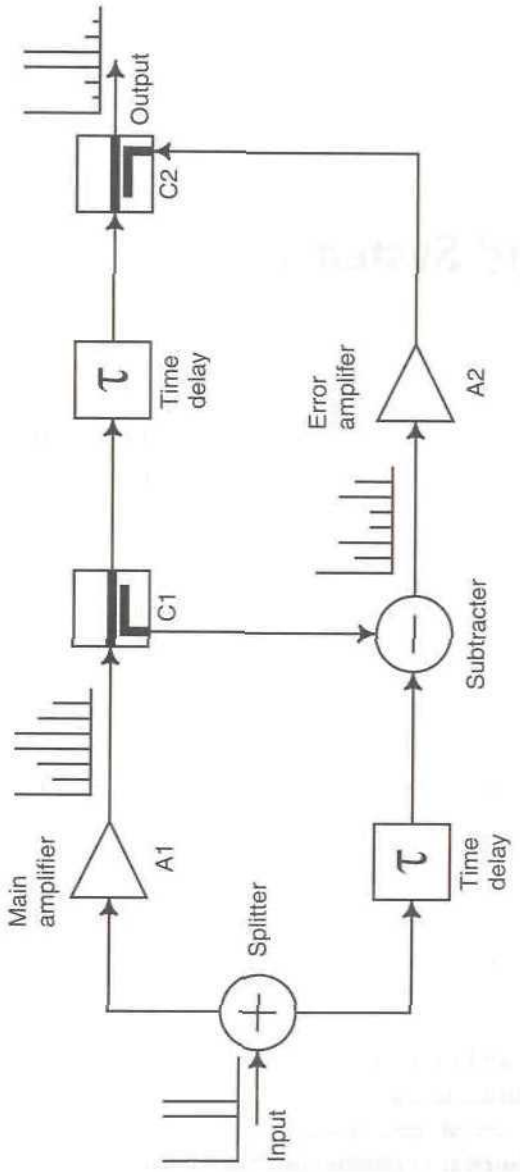


Figure 5.1 Configuration of a basic feedforward amplifier.

gain coaxial repeater amplifier operating in the frequency range 0.5 MHz to 20 MHz was constructed [8]. This system demonstrated the extremely wide bandwidths over which feedforward systems can achieve their correction and together with its predecessor brought a resurgence of interest in feedforward techniques in the early 1970s.

More recently, feedforward techniques have been applied to military HF communications [9], integrated cable TV amplifiers [10], satellite amplifiers [11] and cellular radio systems (both for the base station and for the mobile) [12,13].

5.2 Basic Operation

In its simplest form, a feedforward amplifier consists of the elements shown in Figure 5.1. Its operation may be clearly seen by referring to the two-tone test spectra shown at various points throughout the diagram.

The input signal is split to form two identical paths, although the ratio used in the splitting process need not be equal. The signal in the top path is amplified by the main power amplifier and the nonlinearities in this amplifier result in intermodulation and harmonic distortions being added to the original signal. Noise is also added by the main amplifier, although this is generally neglected in most applications.

The directional coupler, $C1$, takes a sample of the main amplifier output signal and feeds it to the subtracter (180° hybrid) where a time-delayed portion of the original signal, present in the lower path, is subtracted. The result of this subtraction process is an *error signal* containing substantially the distortion information from the main amplifier; ideally none of the original signal energy would remain.

The error signal is then amplified linearly to the required level to cancel the distortion in the main path and fed to the output coupler. The main-path signal through coupler, $C1$, is time delayed by an amount approximately equal to the delay through the error amplifier, $A2$, and fed to the output coupler in antiphase to the amplified error signal. The error signal will then cancel the distortion information of the main path signal leaving substantially an amplified version of the original input signal. Note that the output coupler will have a through-path loss and the implications of this will be considered in detail in Section 5.7.

For an ideal system, and assuming that the subtracter and coupler $C2$ provide the required signal inversion for subtraction to take place internally, the following equations may be derived.

If it is assumed that the input splitter is an ideal 3 dB hybrid, then the

output of the main amplifier, $V_{A1}(t)$, for a system input signal, $V_{in}(t)$, is:

$$V_{A1}(t) = \frac{A_{A1}}{2} V_{in}(t)e^{-j\omega\tau_{A1}} + V_d(t) \tag{5.1}$$

where τ_{A1} is the main amplifier time delay at an angular frequency ω, A_{A1} is the main amplifier gain, and $V_d(t)$ is the distortion added by the main amplifier.

The proportion of this signal which reaches the subtracter is determined by the coupling factor of coupler, $C1$. If this factor is $1/C_{C1}$, then the signal reaching one input to the subtracter is:

$$V_{sub1}(t) = \frac{A_{A1}}{2C_{C1}} V_{in}(t)e^{-j\omega\tau_{A1}} + \frac{V_d(t)}{C_{C1}} \tag{5.2}$$

The signal reaching the other input, assuming the time delay element to be lossless is:

$$V_{sub2}(t) = \frac{V_{in}(t)}{2} e^{-j\omega\tau_{T1}} \tag{5.3}$$

where τ_{T1} is the delay in the time delay element.

Thus the output of the subtracter (assumed lossless) is:

$$\begin{aligned} V_{err}(t) &= V_{sub1}(t) - V_{sub2}(t) \\ &= \frac{A_{A1}}{2C_{C1}} V_{in}(t)e^{-j\omega\tau_{A1}} + \frac{V_d(t)}{C_{C1}} - \frac{V_{in}(t)}{2} e^{-j\omega\tau_{T1}} \end{aligned} \tag{5.4}$$

It can be seen from (5.4) that for the original input signal to be completely removed from this error signal, the following conditions must hold:

$$\tau_{T1} = \tau_{A1} \tag{5.5}$$

and

$$C_{C1} = A_{A1} \tag{5.6}$$

The resulting error signal is then:

$$V_{err}(t) = \frac{V_d(t)}{C_{C1}} \tag{5.7}$$

A similar process takes place in the second part of the loop in which the error signal components are removed from the main amplifier output signal to leave substantially an amplified version of the original input signal.

The output from the main amplifier, having passed through the top-path time delay element is:

$$V_{T2}(t) = \frac{A_{A1}}{2} V_{in}(t)e^{-j\omega(\tau_{A1}+\tau_{T2})} + V_d(t)e^{-j\omega\tau_{T2}} \tag{5.8}$$

where τ_{T2} is the time delay in the top-path delay element and this element is assumed lossless. This signal forms the main through-path signal for the output coupler, $C2$. The signal to be injected into the coupled port of this coupler is derived from the error signal, $V_{err}(t)$, having been amplified by the error amplifier.

$$V_{A2}(t) = \frac{A_{A2}}{C_{C1}} V_d(t)e^{-j\omega\tau_{A2}} \tag{5.9}$$

If the coupler is assumed to possess the necessary phase inversion to facilitate subtraction of its two signals, then, given that the coupling factor is $1/C_{C2}$, the final output signal is:

$$V_{out}(t) = V_{T2}(t) - \frac{V_{A2}(t)}{C_{C2}} \tag{5.10}$$

or

$$V_{out}(t) = \frac{A_{A1}}{2} V_{in}(t)e^{-j\omega(\tau_{A1}+\tau_{T2})} + V_d(t)e^{-j\omega\tau_{T2}} - \frac{A_{A2}}{C_{C1}C_{C2}} V_d(t)e^{-j\omega\tau_{A2}} \tag{5.11}$$

In order for the distortion products, $V_d(t)$, to cancel perfectly, the following conditions must hold:

$$\tau_{T2} = \tau_{A2} \tag{5.12}$$

and

$$A_{A2} = C_{C1}C_{C2} \tag{5.13}$$

The final output signal is then:

$$V_{out}(t) = \frac{A_{A1}}{2} V_{in}(t)e^{-j\omega(\tau_{A1}+\tau_{T2})} \tag{5.14}$$

or, alternatively:

$$V_{out}(t) = \frac{A_{A1}}{2} V_{in}(t) e^{-j\omega(\tau_{A1}+\tau_{A2})} \tag{5.15}$$

Thus the output signal is an amplified and time-delayed replica of the input signal with the distortion from the main amplifier removed. Note that the above argument assumes that the error amplifier is distortion-free and hence contributes no additional distortion to the output spectrum. In reality, this will not be the case, however, the error amplifier will generally be operating at a much lower power level than the main amplifier, and hence can usually be designed to have a greater degree of linearity than that of the main amplifier. It is also predominantly operating on the distortion information from the main amplifier, which is at a much lower level than the main signal energy, and hence the distortion produced by the error amplifier will be relative to the main amplifier's distortion level and thus very significantly lower than the wanted signal energy. It is therefore reasonable, to a first approximation, to ignore the distortion added by the error amplifier in the above analysis.

5.3 Multiple Feedforward Loops

5.3.1 Limitations of the Single Loop

The derivation presented above assumed that the error amplifier was perfectly linear and that the gain and phase balance of the system components was perfect. Neither of these circumstances will exist in a practical amplifier.

If a real error amplifier is substituted for the perfect one assumed above, then (5.9) becomes:

$$V_{A2}(t) = \frac{A_{A2}}{C_{C1}} V_d(t) e^{-j\omega\tau_{A2}} + V_{d2}(t) \tag{5.16}$$

where $V_{d2}(t)$ is the distortion added by the error amplifier. If this is then followed through to the output, the final output signal becomes:

$$V_{out}(t) = \frac{A_{A1}}{2} V_{in}(t) e^{-j\omega(\tau_{A1}+\tau_{A2})} + V_{d2}(t) \tag{5.17}$$

In other words the distortion added by the error amplifier translates directly to the output signal. For example, if the main amplifier has intermodulation distortion (IMD) products 25 dBc relative to the wanted signals (neglecting harmonic distortion) and the error amplifier has a slightly better performance of 35 dBc for its IMD suppression, then the overall feedforward amplifier can, at best, achieve an IMD performance of 60 dBc.

If the gain- and phase-matching problems discussed below are also taken into account, it can be seen that the performance of a simple single-loop system is quite limited.

5.3.2 Use of Additional Loops

The problems of the single loop system may be alleviated to an arbitrarily large extent by the addition of further loops. This is shown diagrammatically in Figure 5.2. The single feedforward correction system may be considered as the main amplifier, for example, in a subsequent feedforward system. The loops can be nested as many times as is necessary to obtain the desired performance, as on each occasion the error amplifier is required to amplify a smaller level of error signal and hence can be designed for greater linearity and a lower power handling capacity (in theory—practice is often different in this area, as will be discussed below).

An alternative configuration is shown in Figure 5.3. In this case the error amplifier is itself linearised by a feedforward loop with the remainder of the system being unchanged. Since the distortion and noise of the error amplifier appear at the output directly, linearising this stage improves the ultimate performance which may be achieved from the system. The major disadvantage of this configuration is that it does not improve the overall cancellation of the main amplifier distortion as this amplifier is not included in the second loop. However, if the limitation in performance of a feedforward system is not being caused by the gain- and phase-matching of the system, but rather by the error amplifier linearity, then feedforward linearisation of the error amplifier is an appropriate solution. The components needed to construct a feedforward loop around the error amplifier would require a lower power rating than those necessary for a second loop involving the main amplifier (i.e., the whole first feedforward loop as described above). Since these high power components are expensive, linearising the error amplifier is sometimes an appropriate solution.

There are, however, some practical drawbacks of the multi-loop technique. Imperfect main signal cancellation will occur in the error loop of the second feedforward process of the type shown in Figure 5.2. This will result in the input signal to the second error amplifier being dominated by residual main signal energy and not main amplifier IMD as discussed above. Consequently, its power rating may well need to be similar to that of the first

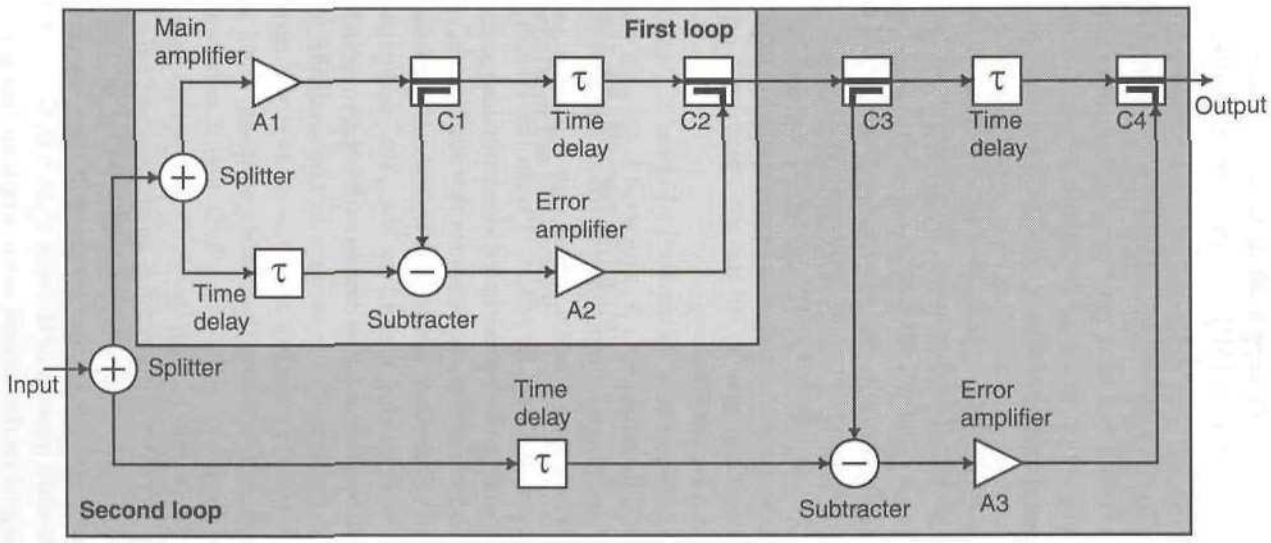


Figure 5.2 Dual-loop feedforward architecture.

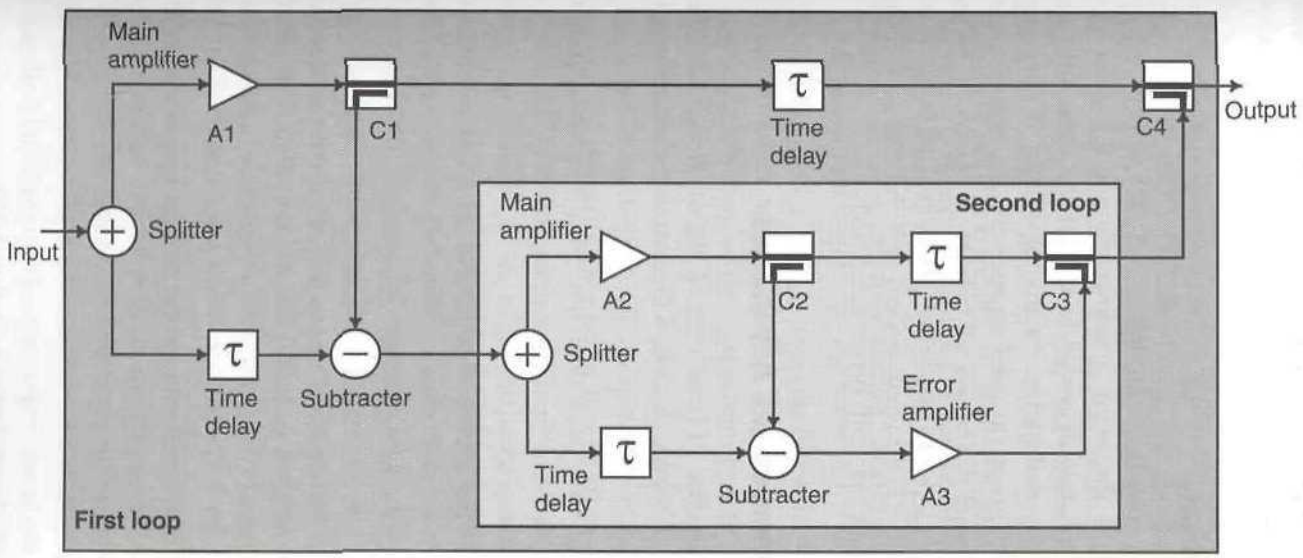


Figure 5.3 Alternative configuration for a dual-loop feedforward system.

error amplifier, since a high degree of back-off will be required in order to reduce the second error amplifier's IMD contribution, at the system output, to an appropriate level.

5.3.3 System Reliability

A further advantage of using multiple loops is that of fault tolerance. In a single-loop system, if the error amplifier should fail then the overall performance is immediately degraded to that of the main amplifier. However, in the case of a multi-loop system the failure of any single-error amplifier will only degrade the performance by the contribution of that amplifier, which is equal to the overall correction produced by the system divided by the number of loops employed (assuming that each contributes an equal IMD reduction). The overall system will thus fail gracefully and its remaining performance may still prove acceptable.

5.4 General Properties and Advantages

Some of the advantages of feedforward as an amplifier linearisation technique are detailed below; Section 5.17 discusses how these advantages are used in practical applications.

1. Feedforward correction does not (ideally) reduce amplifier gain. This is in contrast to feedback systems in which linearity is achieved at the expense of gain.
2. Gain-bandwidth is conserved within the band of interest. This is again in contrast to feedback systems which often require very wide feedback bandwidths in order to provide the required levels of correction.
3. Correction is independent of the magnitude of the amplifier delays within the system. A high-gain RF amplifier will often have a significant group delay and this is potentially disastrous for any form of feedback system due to the large potential for instability.
4. Correction is not attempted based on past events (unlike feedback). The correction process is based on what is currently happening rather than what has happened in the recent past.
5. The basic feedforward configuration is unconditionally stable. This is one of the most important advantages and follows from the points raised above (but see Section 5.15).
6. Cost is the main limiting factor to the number of stages (or 'loops') and hence the level of correction which may be achieved, although size and efficiency may also be important in some applications. In other words, an arbitrarily high level of correction is possible, as there is no theoretical limitation on the number of times which feedforward correction may be applied. In an ideal system, perfect correction could be achieved with just the basic system shown in Figure 5.1, however, in reality the error amplifier itself will distort the error signal and this will appear directly at the output. Gain- and phase-matching throughout the system also affect the performance and this is discussed in Section 5.5.
7. The error amplifier, ideally, need only process the main amplifier distortion information and hence can be of a much lower power than the main amplifier. Thus it is likely that a more linear and lower noise error amplifier can be constructed. This in turn will result in a lower overall system noise figure (see Section 5.13.1.1).
8. Fault tolerance. In a single loop feedforward system, the failure of either amplifier will result in a degradation of performance and possibly a lowering of the final output power; however, the system will not fail altogether. In the case of a feedback system there is only one forward-path amplifier and if this fails then the whole system has failed. If multiple feedforward loops are used then the overall system will degrade gracefully if one or more amplifiers should fail (see Section 5.3.3).

Feedforward also suffers from some major disadvantages which have generally led to its relative unpopularity when compared with feedback. These may be summarised as follows:

1. Changes of device characteristics with time and temperature are not compensated. The open-loop nature of the feedforward system does not permit it to assess its own performance and correct for time variations in its system components. Thus the performance of a basic (uncompensated) feedforward system can be expected to degrade with time.
2. The matching between the circuit elements in both amplitude and phase must be maintained to a very high degree over the correction bandwidth of interest. The precise levels of matching required for a given level of correction are investigated in Section 5.5.
3. Circuit complexity is generally greater than that of a feedback system, particularly with the requirement for a second (error) amplifier. This usually results in greater size and cost.

5.5 Gain- and Phase-Matching

The advantages mentioned above come at the expense of a requirement for a high degree of matching in both amplitude and phase for virtually all of the system components. This matching must also be maintained over the bandwidth of interest, which is generally large relative to that of feedback systems. Feedforward systems are generally applied at frequencies or over operating bandwidths where feedback systems are inappropriate (e.g., microwave frequencies [14]). Amplitude and phase matching within the feedforward loop must therefore be maintained over a wide frequency range.

The lack of an intrinsic feedback path in a basic feedforward system means that it cannot monitor its own performance and hence correct for gain or phase changes due to temperature or aging effects. Although near perfect cancellation may be set up manually upon manufacture (say), this will drift over time and degrade performance.

The techniques presented in this section enable the required degree of amplitude and phase matching to be determined in order to achieve any given, realisable, specification. In addition to their obvious application in the cancellation loop, they may also be used to predict the degree of matching required within the error loop. This will determine the degree of residual main signal energy present in the error signal and hence, in many cases, the power rating required of the error amplifier.

5.5.1 Derivation

The feedforward amplifier shown in Figure 5.1 will have two signals arriving at the output coupler, $C2$, as a result of an applied two-tone test. The first, $V_{A1}(t)$, resulting from the output of the main amplifier, $A1$, will consist of the original tones, together with intermodulation distortion (IMD) products. This signal passes via the coupler, $C1$, and the main-path time delay element to the output coupler, $C2$. This signal, $V_2(t)$, will differ in amplitude and phase from $V_{A1}(t)$ since the loss in the main-path time delay element is no longer assumed to be negligible. The coupler, $C1$, takes a sample of the main amplifier output signal, $V_{A1}(t)$, and this is then subtracted from the reference signal, resulting in the error signal, $V_{err}(t)$. The error signal is amplified by the error amplifier to form $V_{A2}(t)$ and fed to the coupled port of the output coupler, $C2$. The IMD products from the main amplifier output signal are suppressed at the feedforward amplifier output by anti-phase cancellation (after the coupler and delay elements), of $V_2(t)$, with the error signal, $V_{A2}(t)$, in the output coupler, $C2$.

For simplicity, consider a single intermodulation product generated by the main amplifier. This will be of the form:

$$V_{IM}(t) = V\cos(\omega t + \varphi) \tag{5.18}$$

After the coupler, $C1$, and time delay element, this becomes:

$$V_2(t) = A\cos(\omega t + \phi) \tag{5.19}$$

The error signal component of the IMD product, after amplification by the error amplifier, will be:

$$V_{A2}(t) = (A + \delta A)\cos(\omega t + \pi + \phi + \delta\phi) \tag{5.20}$$

where the additional π term arranges the signals to be approximately in anti-phase.

When these two signals ((5.19) and (5.20)) are added with equal weight at the output combiner (assumed to be a 3 dB coupler for simplicity), cancellation occurs. If the coupler was other than a 3 dB coupler (10 dB being typical in practice) then this would be taken into account by the error amplifier and would result in an additional amplitude weight in the error signal, $V_{A2}(t)$. The resultant output of the summation is:

$$V_{out}(t) = A\cos(\omega t + \phi) + (A + \delta A)\cos(\omega t + \pi + \phi + \delta\phi) \tag{5.21}$$

Expanding gives:

$$\begin{aligned} V_{out}(t) = {} & A\cos(\omega t + \phi) + (A + \delta A)\cos(\omega t + \pi)\cos(\phi + \delta\phi) \\ & - (A + \delta A)\sin(\omega t + \pi)\sin(\phi + \delta\phi) \end{aligned} \tag{5.22}$$

and hence:

$$\begin{aligned} V_{out}(t) = {} & A\cos\omega t\cos\phi - A\sin\omega t\sin\phi \\ & - A\cos\omega t\cos\phi\cos\delta\phi + A\cos\omega t\sin\phi\sin\delta\phi \\ & - \delta A\cos\omega t\cos\phi\cos\delta\phi + \delta A\cos\omega t\sin\phi\sin\delta\phi \\ & + A\sin\omega t\sin\phi\cos\delta\phi + A\sin\omega t\cos\phi\sin\delta\phi \\ & + \delta A\sin\omega t\sin\phi\cos\delta\phi + \delta A\sin\omega t\cos\phi\sin\delta\phi \end{aligned} \tag{5.23}$$

This resultant output signal is a single tone (since only a single intermodulation product is being considered) of the form:

$$V_{out}(t) = E\cos(\omega t + \theta) \tag{5.24}$$

Equating the above expressions and simplifying gives:

$$\begin{aligned} E\cos(\omega t+\theta) = \cos\omega t[&A\cos\phi - \cos\phi\cos\delta\phi + A\sin\phi\sin\delta\phi \\ &- \delta A\cos\phi\cos\delta\phi - \sin\phi\sin\delta\phi] \\ &+ \sin\omega t[A\sin\phi - A\sin\phi\cos\delta\phi + A\cos\phi\sin\delta\phi \\ &+ \delta A\sin\phi\cos\delta\phi + \delta A\cos\phi\sin\delta\phi] \end{aligned} \quad (5.25)$$

In this case, it is only the amplitude of the resultant signal, E, which is important, hence:

$$\begin{aligned} E^2 = [&A\cos\phi - \cos\phi\cos\delta\phi + A\sin\phi\sin\delta\phi \\ &- \delta A\cos\phi\cos\delta\phi - \sin\phi\sin\delta\phi]^2 \\ &+ [A\sin\phi - A\sin\phi\cos\delta\phi + A\cos\phi\sin\delta\phi \\ &+ \delta A\sin\phi\cos\delta\phi + \delta A\cos\phi\sin\delta\phi]^2 \end{aligned} \quad (5.26)$$

Since the initial phase, ϕ, is arbitrary and only the relative phase, $\delta\phi$ is of importance, it is permissible to set $\phi = 0$, giving the amplitude error, E as:

$$E = \sqrt{(A - A\cos\delta\phi - \delta A\cos\delta\phi)^2 + (A\sin\delta\phi + \delta A\sin\delta\phi)^2} \quad (5.27)$$

If $A = 1$, δA becomes the difference in amplitude of the main path and error path vectors. In the special case where the phases are perfectly matched, that is, $\delta\phi = 0$, this is then the magnitude of the resultant error vector. In other words:

$$E = \delta A \quad (5.28)$$

When $\delta\phi$ is not equal to zero, the magnitude of the resultant is:

$$E = \sqrt{(1 - \cos\delta\phi - \delta A\cos\delta\phi)^2 + (\sin\delta\phi + \delta A\sin\delta\phi)^2} \quad (5.29)$$

Thus the cancellation of the intermodulation product achieved at the output of the feedforward amplifier is:

$$S_{dB} = 20\log_{10}\sqrt{(1 - \cos\delta\phi - \delta A\cos\delta\phi)^2 + (\sin\delta\phi + \delta A\sin\delta\phi)^2} \quad (5.30)$$

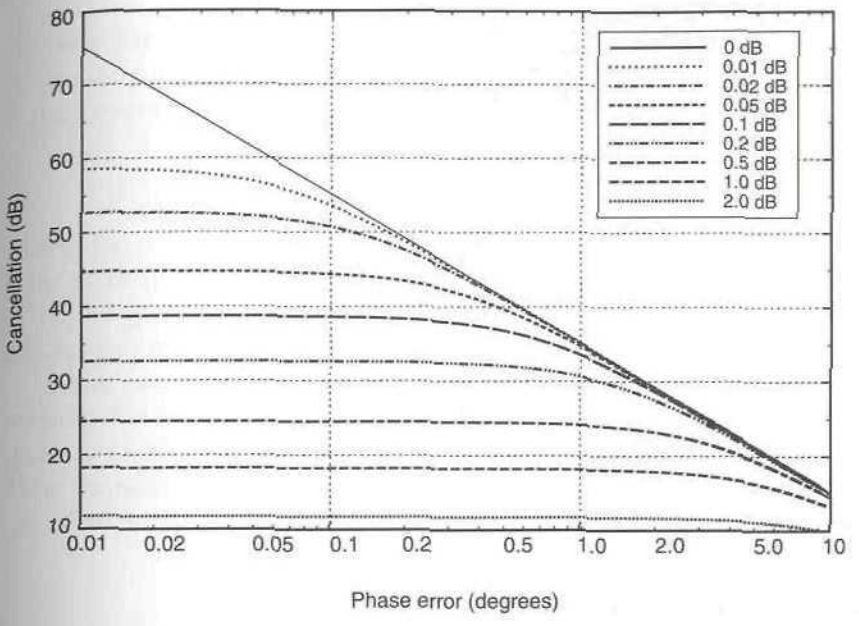


Figure 5.4 Distortion cancellation achievable by a feedforward amplifier with various values of amplitude and phase error.

5.5.2 Matching Characteristics

Figure 5.4 shows the amount of suppression which can be achieved for a given degree of phase and amplitude matching in a feedforward amplifier [15]. The characteristics shown are plotted using (5.30).

In order to achieve around 25 dB of cancellation, as would be required in some mobile radio applications, an amplitude error of better than 0.5 dB and a phase error of better than 5° would need to be achieved. In some applications more than 40 dB of suppression is required and the accuracy of the amplitude and phase matching becomes more stringent at less than 0.1 dB and 0.1°, respectively. This would be extremely difficult to achieve in a broadband amplifier system.

5.5.3 Vector Representation

Figure 5.5 shows a vector representation of the cancellation process. Note that the resultant vector, assuming that the amplitude and phase of the cancellation process are both imperfect, is at a phase different from both the original signal and the cancelling signal.

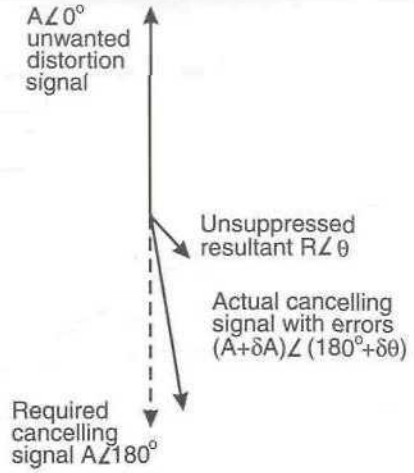


Figure 5.5 Vector representation of feedforward cancellation process.

5.6 Error Amplifier Design

It is instructive, here, to examine the trade-offs in the design of the error amplifier, in terms of both the gain- and phase-match and the IMD level. There is, for example, little point in designing an error amplifier with an extremely high level of linearity (and hence low IMD level) if its gain and phase characteristics deviate to such a large extent that the feedforward cancellation is poor. Both aspects of error amplifier design must be examined together and an appropriate compromise selected to achieve the required specification.

Any practical error amplifier will have an imperfect gain and phase characteristic with respect to frequency and hence perfect cancellation will only be achieved at a single frequency. At all other frequencies within its operational bandwidth the performance will be degraded, yielding a finite suppression of the main amplifier distortion specified by (5.30) above. Considerable design effort is required to improve the gain and phase characteristics and there is little point in attempting to achieve perfection.

A practical error amplifier will also distort the error signal it is required to amplify and will thus have a finite level of intermodulation ratio (level of IMD products relative to the carriers in a two-tone test). In the operation of a single-loop feedforward system, this distortion is summed directly into the output signal and receives no cancellation. It is thus also a fundamental limit on performance.

An optimum design would therefore balance these two limitations in order to just meet the desired specification (with an assumed safety margin).

If the gain and phase characteristics of the error amplifier were perfectly flat with frequency, then the level of the IMD products at the system output would simply be that due to the error amplifier. Thus, the desired overall IMD ratio for the feedforward system would be:

$$S_F = S_{A1} + S_{A2} \tag{5.31}$$

where: S_{A1} is the main amplifier intermodulation ratio (open-loop) and S_{A2} is the error amplifier intermodulation ratio (open-loop).

This gives a specification for the error amplifier performance (in terms of linearity) required to meet the overall system specification.

The gain and phase characteristics will, as outlined above, not be perfectly flat in practice and the level of suppression obtained will be that given by (5.30). This may impinge upon the IMD ratio given by (5.31) and cause the design to fail.

The final degree of suppression may be found from:

$$S_F = \begin{cases} S_{A1} + S_{dB} & \forall \ S_{A2} \geq S_{dB} \\ S_{A1} + S_{A2} & \forall \ S_{A2} < S_{dB} \end{cases} \tag{5.32}$$

This discussion on error amplifier design assumes that the intermodulation products generated by the error amplifier do not sum in phase with the residual intermodulation from the main amplifier, that is, that the two may be considered to be independent. This assumption will be valid in most practical systems, particularly where the specification is exceeded by more than a few dB in either parameter. The worst-case error, if the two largest products were of equal amplitude and added perfectly in phase, is 3 dB. This condition is extremely unlikely in practice.

5.7 Power Efficiency

The effect of the coupling factor of the output coupler in terms of its coupling of the error amplifier's intermodulation distortion into the feedforward amplifier's output [16] or on the power-handling of the error amplifier [14] is clearly important. Of equal, or perhaps greater concern, however, is its effect on the overall power efficiency of the feedforward system. The techniques presented below [17] allow the basic efficiency of a simple feedforward system to be calculated, for a variety of system configurations, as a function of the output coupling factor. The resulting efficiencies are only an approximation, as they depend heavily upon the

characteristics of the error signal and the power level of the error amplifier (i.e., how much back-off is required to prevent significant IMD contribution from the error amplifier). The use of an appropriate figure for the error amplifier efficiency can, however, help to improve the accuracy of the results.

For the feedforward amplifier shown in Figure 5.1, the following parameters may be defined:

For the main amplifier, $A1$:

Output power of $A1$ (per carrier):	P_{A1} (W)
Efficiency:	η_{A1} (dimensionless)
Third-order intermodulation level relative to carriers: (two-tone test)	S_{A1} (dB)

For the error amplifier, $A2$:

Output power of $A2$ (per carrier):	P_{A2} (W)
Efficiency:	η_{A2} (dimensionless)

For the output coupler, $C2$:

Coupling factor:	C_f (dB)
Through-path insertion loss:	L_{tp} (dB)

It will be assumed that the third-order intermodulation (IMD) products are dominant, and that the power required by the error amplifier, $A2$, in order to amplify the higher-order intermodulation products (or residual main-tone energy) may be neglected. This is a reasonable assumption in practice, due to the classical shape of the two-tone intermodulation spectrum of a typical amplifier and the fact that a good error loop design will provide appropriate main-tone suppression. Making this assumption will in any case not affect the optimum coupling factor; it will merely have an effect on the absolute efficiency figures. In particular, poor main-tone suppression will have a major effect in this respect and it is therefore important to ensure that good error loop cancellation is achieved.

When considering a two-tone test, the fraction of the per-carrier power of the main amplifier present in a single third-order product, F_{IM} is:

$$F_{IM} = 10^{(S_{A1}/10)} \tag{5.33}$$

The coupling factor of the output coupler ($C2$), C_{DC}, (dimensionless) is:

$$C_{DC} = 10^{-(C_f/10)} \tag{5.34}$$

The absolute power level of the third-order IMD products (per-tone) is therefore:

$$P_{3rd} = P_{A1}F_{IM} \tag{5.35}$$

The effective power which must be generated by $A2$, referenced to the output of coupler $C2$, in order to cancel these IMD products is:

$$P_{A2,C} = P_{A2}C_{DC} \tag{5.36}$$

The equivalent effective power of the IMD from the main power amplifier, $A1$, at the output (neglecting the insertion loss of the coupler $C1$) is:

$$P_{A1,C} = P_{A1}F_{IM}L_{DC} \tag{5.37}$$

where L_{DC} is the insertion loss of the coupler $C2$ (dimensionless) and is given by:

$$L_{DC} = 10^{-(L_{tp}/10)} \tag{5.38}$$

For full cancellation of the IMD products from the main amplifier, the effective main amplifier IMD and error amplifier output error signal levels ((5.36) and (5.37)) must be equal (and in anti-phase), hence:

$$P_{A1}F_{IM}L_{DC} = P_{A2}C_{DC} \tag{5.39}$$

or

$$P_{A2} = \frac{P_{A1}F_{IM}L_{DC}}{C_{DC}} \tag{5.40}$$

Taking, as an example, a 3 dB coupler: $C_{DC} = L_{DC} = 0.5$ giving:

$$P_{A2} = P_{A1}F_{IM} \tag{5.41}$$

as expected.

The overall efficiency of a feedforward amplifier, η_{ff}, is based upon the wanted RF output power and the DC power consumed by both the main and error amplifiers:

$$\eta_{ff} = \frac{P_{OUT,RF}}{P_{DC,A1} + P_{DC,A2}} \tag{5.42}$$

where $P_{OUT,RF}$ is the RF output power of the complete feedforward amplifier (i.e., following the output coupler), $P_{DC,A1}$ is the supply power to the main amplifier and $P_{DC,A2}$ is the supply power to the error amplifier. Now:

$$P_{DC,A1} = \frac{P_{A1}}{\eta_{A1}} \tag{5.43}$$

and

$$P_{DC,A2} = \frac{P_{A2}}{\eta_{A2}} \tag{5.44}$$

Also:

$$P_{OUT,RF} = P_{A1} L_{DC} \tag{5.45}$$

Thus:

$$\eta_{ff} = \frac{P_{A1} L_{DC}}{P_{A1}/\eta_{A1} + P_{A2}/\eta_{A2}} \tag{5.46}$$

The insertion loss (strictly, the proportion of incident power reaching the output) of a directional coupler is related to its coupling factor by:

$$L_{DC} = 1 - C_{DC} \tag{5.47}$$

Combining (5.46) and (5.47) and substituting for P_{A2} from (5.40) gives:

$$\eta_{ff} = \frac{P_{A1}\eta_{A1}\eta_{A2}(1 - C_{DC})}{P_{A1}\eta_{A2} + P_{A1}\eta_{A1}F_{IM}L_{DC}/C_{DC}} \tag{5.48}$$

Finally, substituting for the insertion loss of the output coupler, L_{DC}, from (5.47) and simplifying, gives:

$$\eta_{ff} = \frac{\eta_{A1}\eta_{A2}C_{DC}(1 - C_{DC})}{\eta_{A2}C_{DC} + \eta_{A1}F_{IM}(1 - C_{DC})} \tag{5.49}$$

5.7.1 Optimum Coupling Factor

Differentiation of (5.49) may be used to determine the optimum value of coupling factor for the output directional coupler, required to obtain the maximum DC to RF power conversion efficiency from a feedforward amplifier system.

Differentiating this (and setting it equal to zero) will yield the stationary points on the efficiency characteristic, that is,

$$\frac{d\eta_{ff}}{dC_{DC}} = 0 \tag{5.50}$$

Differentiation, using the quotient rule, gives:

$$\frac{\eta_{A1}\eta_{A2}P_{A1}(1 - 2C_{DC})[P_{A1}\eta_{A2}C_{DC} + \eta_{A1}P_{A1}F_{IM}(1 - C_{DC})]}{[P_{A1}\eta_{A2}C_{DC} + \eta_{A1}P_{A1}F_{IM}(1 - C_{DC})]^2} - \frac{\eta_{A1}\eta_{A2}P_{A1}C_{DC}(1 - C_{DC})[P_{A1}\eta_{A2} + \eta_{A1}P_{A1}F_{IM}]}{[P_{A1}\eta_{A2}C_{DC} + \eta_{A1}P_{A1}F_{IM}(1 - C_{DC})]^2} = 0 \tag{5.51}$$

which simplifies to:

$$C_{DC}^2(\eta_{A1}F_{IM} - \eta_{A2}) - 2\eta_{A1}F_{IM}C_{DC} + \eta_{A1}F_{IM} = 0 \tag{5.52}$$

Solving this using the standard quadratic formula, gives:

$$C_{DC} = \frac{2\eta_{A1}F_{IM} \pm \sqrt{4\eta_{A1}\eta_{A2}F_{IM}}}{2(\eta_{A1}F_{IM} - \eta_{A2})} \tag{5.53}$$

Equation (5.53) thus permits the calculation of the optimum coupling factor for a specific third-order IMD level in the main amplifier, from the main and error amplifier efficiencies.

5.7.2 Typical Efficiency Characteristics

Figure 5.6 shows a range of typical efficiency characteristics obtained from (5.49). The six cases considered gradually change from high-linearity class-A main and error amplifiers, through class-AB and -B, to class-C. Although the efficiencies and IMD levels quoted will not be correct at all frequencies and power levels (due to device technology differences), they allow some insight to be gained of the various compromises involved. In particular, note that for the lower-efficiency, more linear systems, the characteristics are less sharply defined, making the choice of coupling factor less critical. By contrast, the class-C based characteristics indicate that output coupling factor is a critical choice in obtaining optimum efficiency.

The peak efficiencies indicated in Figure 5.6 may be optimistic for a high-linearity system due to the residual main tone level present in the error

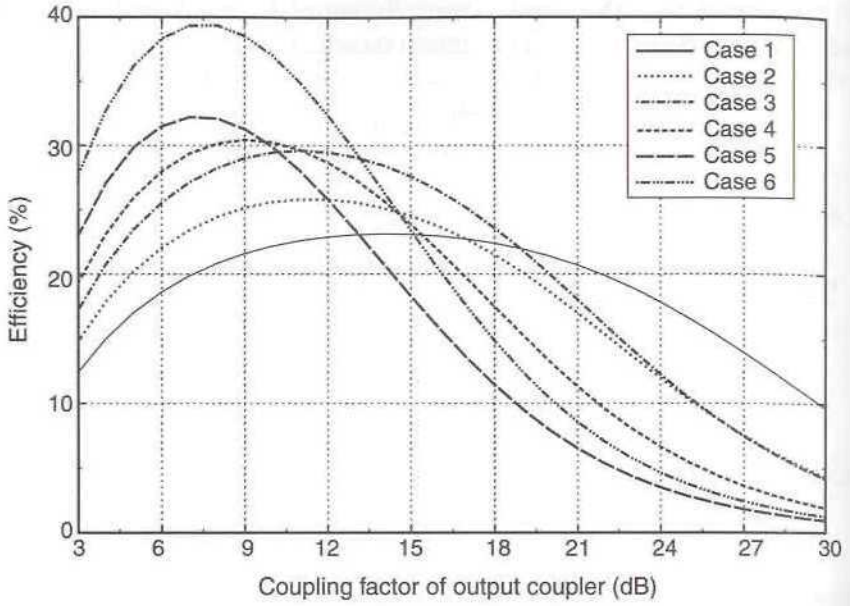


Figure 5.6 Feedforward amplifier efficiency characteristic with a range of system parameters (given in Table 5.1).

signal. The optimum power rating and power efficiency of the error amplifier must be compromised in order to ensure that its IMD contribution at the output is suitably small. Typical error amplifier efficiencies in this case could be as low as 1% or 2% (not the 5% assumed here). This effect is discussed in more detail in Section 5.8.2.

Table 5.1
Parameters for Figure 5.6

Case	Main amplifier IMD level (dBc)	Main amplifier efficiency (%)	Error amplifier efficiency (%)
1	−35	25	5
2	−30	30	5
3	−25	35	15
4	−20	40	20
5	−15	50	30
6	−15	60	40

5.8 Effect of Power Loss in the Main-Path Delay Element

The above arguments on power efficiency and optimum coupling value assumed that the components following the main amplifier were lossless, with the exception of the final output coupler. This is a reasonable first approximation, but will not be true in a practical system; the sampling coupler and the main-path delay element will both introduce loss and hence lower the available power efficiency.

There are a number of methods by which the main-path delay element may be realised and these will be covered in more detail in Section 5.13.4. The loss in these elements depends to a great degree on the fabrication method employed, the number of transitions required (e.g., from microstrip at the output of the PA to coax for the delay line and back again for the output coupler) and the delay of the error amplifier (which largely determines the length of main-path delay required). A typical high-power low-loss coaxial delay, complete with two transitions, will have a loss in the region of 2 dB to 2.5 dB at 2 GHz, including the through-path loss of the output coupler (assuming a typical multi-stage error amplifier and a high-power application). This may be reduced to 1.5 dB or less with the use of a filter delay line. A typical microstrip delay line will have a loss much greater than either of these figures, for an equivalent length and operational frequency, and will usually suffer from a poorer frequency response ripple.

The previous analysis can be modified [18] to take account of these effects, by combining them, and assuming that they occur as a loss in the main-path delay element (which is where the bulk of the loss will actually occur in practice). A new parameter must now be included:

Insertion loss of the main-path delay element: L_{TD2} (dB).

Again, it is assumed that the main amplifier has only third-order intermodulation distortion and that this distortion can be perfectly cancelled across the band of interest at the output coupler. It is also assumed that the error amplifier contributes no distortion of its own to the output of the system.

5.8.1 Overall Efficiency of a Feedforward Amplifier Incorporating Loss in the Main-Path Delay Element

The fractional loss of signal power caused by the loss in the delay element, L (dimensionless), is given by:

$$L = 10^{(L_{TD2}/10)} \tag{5.54}$$

The (wanted) main signal energy and the IMD products are both attenuated by this factor, giving:

$$P_{A1,L} = \frac{P_{A1}}{L} \tag{5.55}$$

and

$$P_{3rd,L} = \frac{P_{A1}F_{IM}}{L} \tag{5.56}$$

Equation (5.37) therefore becomes:

$$P_{A1,C} = \frac{P_{A1}F_{IM}L_{DC}}{L} \tag{5.57}$$

where L_{DC} is the through-path loss of the output coupler C2.

Substituting for L_{DC} from (5.47) gives:

$$P_{A1,C} = \frac{P_{A1}F_{IM}(1 - C_{DC})}{L} \tag{5.58}$$

For full cancellation of the IMD products from the main amplifier, the distortion power from the main and error amplifiers ((5.58) and (5.36)) must be equal (and in anti-phase), hence:

$$\frac{P_{A1}F_{IM}(1 - C_{DC})}{L} = P_{A2}C_{DC} \tag{5.59}$$

or

$$P_{A2} = \frac{P_{A1}F_{IM}(1 - C_{DC})}{LC_{DC}} \tag{5.60}$$

The overall efficiency of a feedforward amplifier is given by (5.42) as:

$$\eta_{ff} = \frac{P_{OUT,RF}}{P_{DC,A1} + P_{DC,A2}} \tag{5.61}$$

where $P_{OUT,RF}$ is the RF output power of the whole feedforward amplifier, $P_{DC,A1}$ is the supply power to the main amplifier and $P_{DC,A2}$ is the supply power to the error amplifier. If, in this case, account is taken of the power

drawn from the power supply in order to generate the intermodulation products, in addition to that needed to supply the power for the wanted signals, then:

$$P_{DC,A1} = \frac{P_{A1}(1 + F_{IM})}{\eta_{A1}} \tag{5.62}$$

and, as before:

$$P_{DC,A2} = \frac{P_{A2}}{\eta_{A2}} \tag{5.63}$$

Substituting for P_{A2} from (5.60) gives:

$$P_{DC,A2} = \frac{P_{A1}F_{IM}(1 - C_{DC})}{\eta_{A2}LC_{DC}} \tag{5.64}$$

Also:

$$P_{OUT,RF} = P_{A1,L}(1 - C_{DC}) \tag{5.65}$$

Substituting into (5.55) gives:

$$P_{OUT,RF} = \frac{P_{A1}(1 - C_{DC})}{L} \tag{5.66}$$

The overall efficiency of the feedforward amplifier, incorporating the insertion loss of the delay element, and taking account of the additional power drawn by the main amplifier in generating its IMD products, is therefore:

$$\eta_{ff} = \frac{\eta_{A1}\eta_{A2}C_{DC}(1 - C_{DC})}{\eta_{A1}F_{IM}(1 - C_{DC}) + \eta_{A2}C_{DC}L(1 + F_{IM})} \tag{5.67}$$

The optimum coupling factor may be derived as before by differentiating the efficiency (5.67) with respect to the coupling factor and setting to zero to find the maximum efficiency. Differentiation by means of the quotient rule and subsequent simplification [18] yield:

$$C_{DC}^2(\eta_{A1}F_{IM} - \eta_{A2}L(1 + F_{IM})) - 2C_{DC}\eta_{A1}F_{IM} + \eta_{A1}F_{IM} = 0 \tag{5.68}$$

This can be solved for C_{DC} to find the optimum coupling factor:

$$C_{DC,OPT} = \frac{\eta_{A1}F_{IM} \pm \sqrt{\eta_{A1}\eta_{A2}F_{IM}L(1+F_{IM})}}{\eta_{A1}F_{IM} - \eta_{A2}L(1+F_{IM})} \tag{5.69}$$

5.8.2 Typical Efficiency Characteristics

It is possible to derive a similar set of efficiency characteristics to those of Figure 5.6, but with the loss of the main-path delay element included. Two examples are given in Figures 5.7 and 5.8, illustrating both case 5 (class-C) and case 1 (class-A) of Table 5.1. It can be seen that even a small amount of loss following the main amplifier has quite a marked affect on the overall power efficiency. In the case of the case 5 example, illustrated in Figure 5.7, the maximum achievable efficiency is 32% (assuming no loss after the main amplifier, except for that in the output coupler). This reduces to 26% with 1 dB of delay-element loss and 22% with 2 dB of loss (i.e., a reduction of approximately one-sixth in efficiency for each 1 dB increase in insertion loss).

The power rating of the main amplifier will also need to be increased to overcome the delay element losses and hence maintain the overall power rating required of the linearised system.

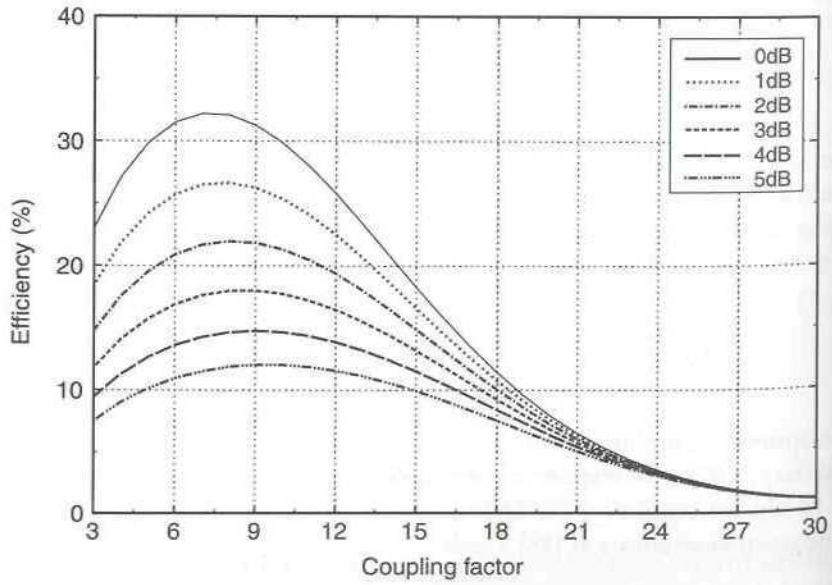


Figure 5.7 Feedforward amplifier efficiency characteristic, incorporating various values of main-path delay element loss, with $\eta_{A1} = 0.5$, $\eta_{A2} = 0.3$ and $S_{A1} = 15$ dB.

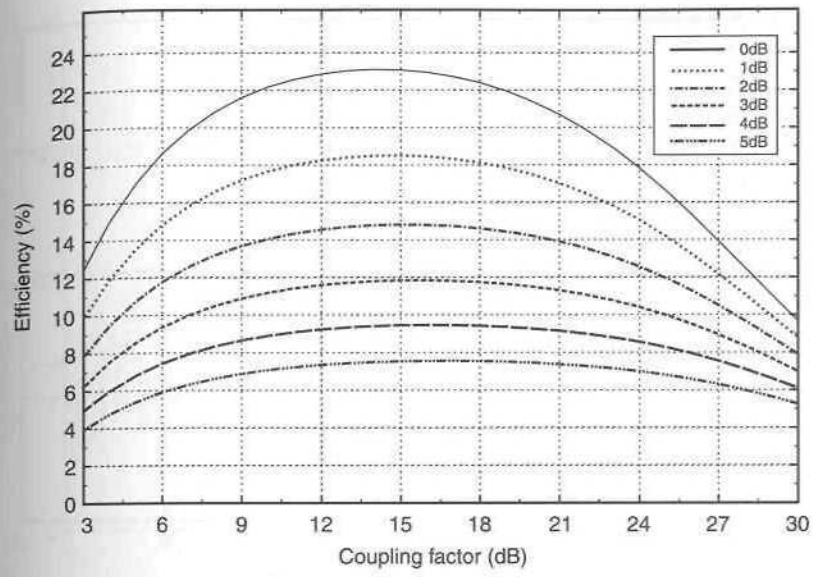


Figure 5.8 Feedforward amplifier efficiency characteristic, incorporating various values of main-path delay element loss, with $\eta_{A1} = 0.25$, $\eta_{A2} = 0.05$ and $S_{A1} = 35$ dB.

Figure 5.8 illustrates the efficiency obtainable from a feedforward system employing class-A amplifiers which have been optimised for linearity, rather than power efficiency (case 1 in Table 5.1). It is evident that the efficiency of the overall feedforward amplifier is little different from that of the main amplifier in the case where no loss is assumed in the delay element. However, as was the case for the class-C system discussed above, the overall efficiency of the feedforward system decreases quite obviously (although to a lesser extent) when small amounts of delay-element loss are incorporated.

If it is assumed that the optimal coupling factor may be chosen and employed in a feedforward amplifier, then the decrease in efficiency with increasing main-path delay element insertion loss may be derived. Examples of this are shown in Figures 5.9 and 5.10 for the class-C (case 5) and class-A (case 1) amplifier examples discussed above.

Both of these characteristics illustrate how rapidly the overall efficiency of a feedforward system falls as a result of the insertion loss of the main-path delay element. It is particularly marked in the case of the class-C amplifier-based system; this problem must be carefully considered since the main purpose of employing class-C amplifiers in such a system is in order to improve power efficiency.

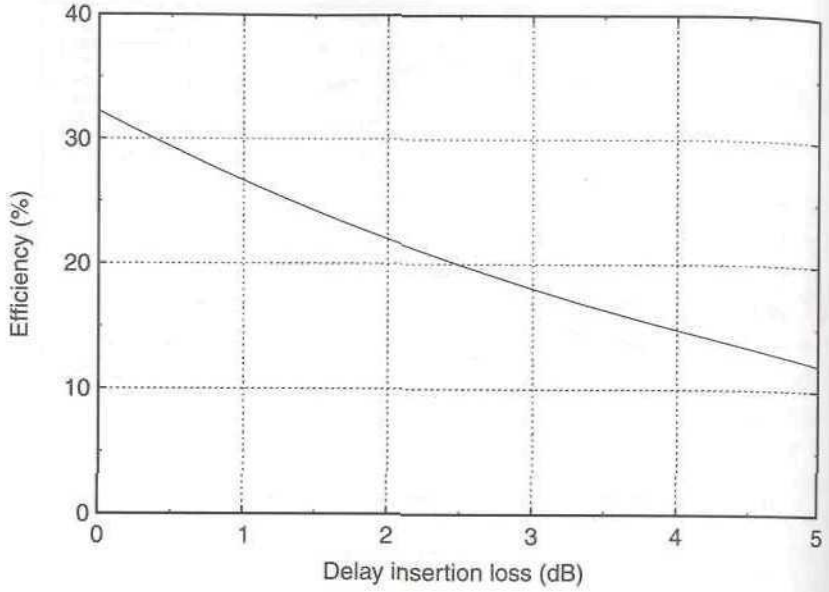


Figure 5.9 Maximum feedforward amplifier efficiencies obtainable for various values of main-path delay element loss ($\eta_{A1} = 0.5$, $\eta_{A2} = 0.3$ and $S_{A1} = 15$ dB).

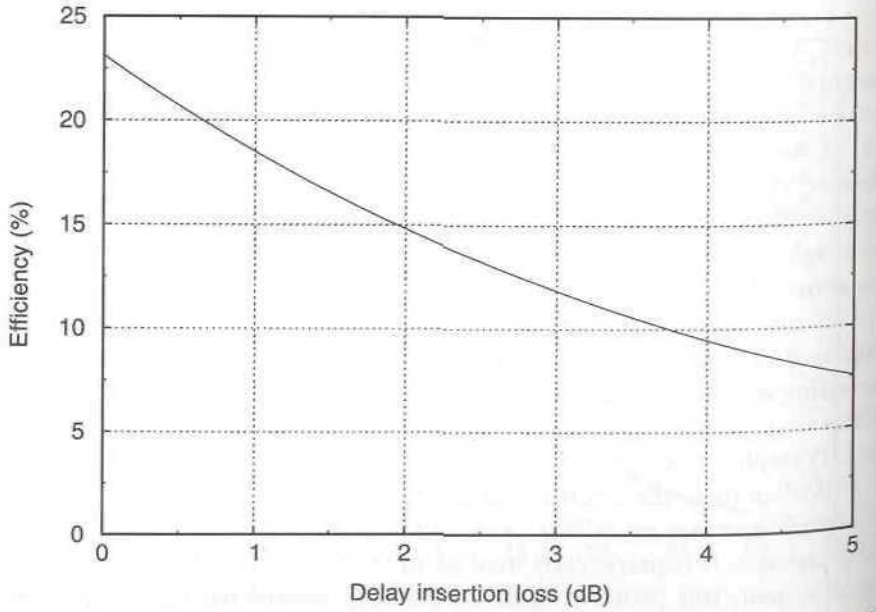


Figure 5.10 Maximum feedforward amplifier efficiencies obtainable with various values of main-path delay element loss ($\eta_{A1} = 0.25$, $\eta_{A2} = 0.05$ and $S_{A1} = 35$ dB).

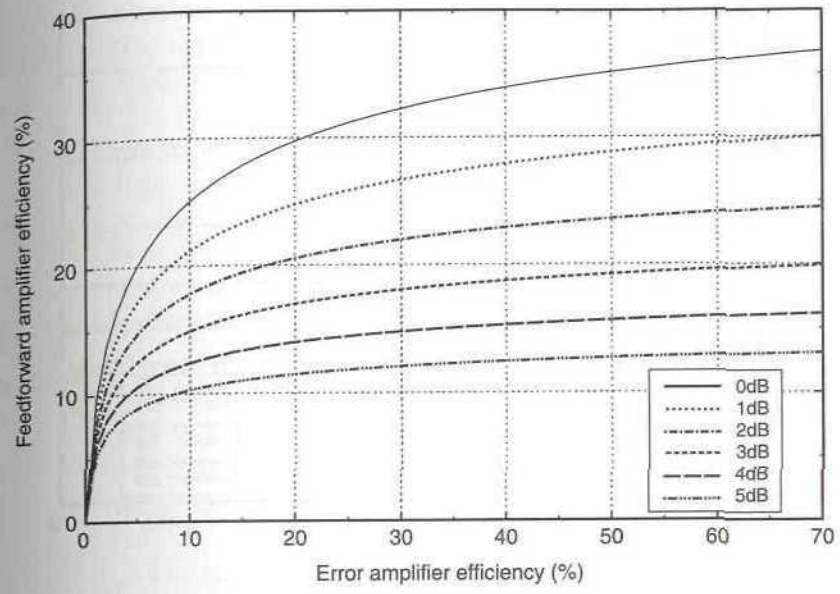


Figure 5.11 Maximum feedforward amplifier efficiencies obtainable for various values of error amplifier power efficiency ($\eta_{A1} = 0.5$ and $S_{A1} = 15$ dB). Note that the effect of losses in the main signal path may also be deduced.

One final result which it is worth noting concerns the effect of the error amplifier efficiency upon the overall efficiency of the feedforward system. Looking again at the high-efficiency feedforward system based on the use of a class-C main amplifier (case 5), it is evident from Figure 5.11 that the efficiency of the error amplifier has relatively little effect on the overall system efficiency (provided that its efficiency is not excessively low). Thus, for example, a reasonably efficient class-A amplifier could be employed as the error amplifier with a class-C main amplifier, and this would result in a relatively minor loss of overall efficiency. The resulting amplifier would have both an efficiency and a linearity far better than that of an equivalent class-A amplifier, even assuming a modest amount of delay-element insertion loss.

Referring to Figure 5.12 it is evident that the error amplifier efficiency in a class-A amplifier-based feedforward system has even less effect on the resulting overall system efficiency. Of far greater concern in this type of system is the through-loss in the main-path elements (e.g., delay line, or output coupler).

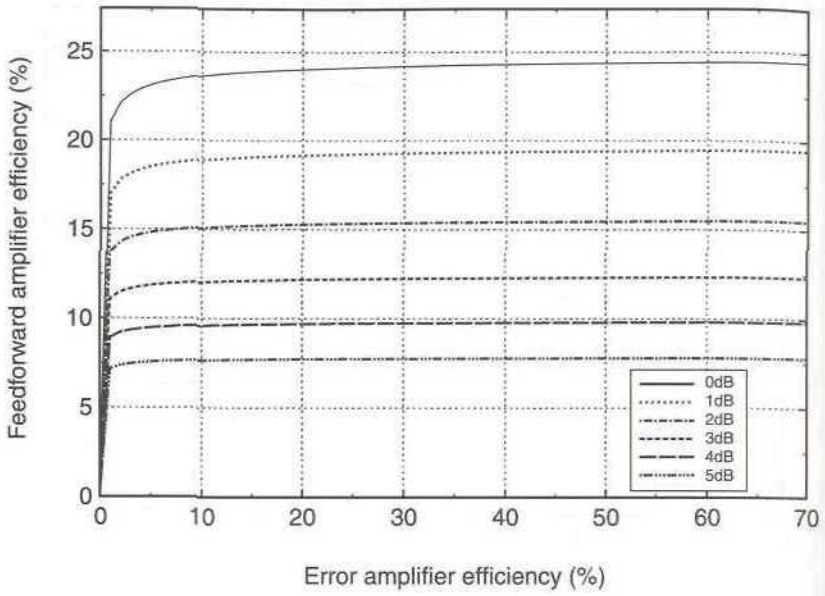


Figure 5.12 Maximum feedforward amplifier efficiencies obtainable for various values of error amplifier power efficiency ($\eta_{A1} = 0.25$ and $S_{A1} = 35$ dB). Note that the effect of losses in the main signal path may also be deduced.

5.9 Efficiency Improvement of a Feedforward Amplifier

It is evident from the discussion of Section 5.8 that the loss in the main-path delay element can be a significant factor in determining the overall power efficiency of a feedforward system. It follows therefore that the loss in this element should be kept to a minimum, in order to provide maximum efficiency and in order to minimise the power rating required for the main amplifier (to meet a given output power specification for the overall feedforward system).

There are two methods by which this aim may be achieved:

1. Utilise ultra low-loss coaxial cable (or, e.g., dielectric material for an etched delay line). The major disadvantage with this (otherwise successful) method of improving efficiency lies in its size and weight penalty. The use of high-power, low-loss cables results in an inevitable size penalty, as such cables tend to be very thick, employing a large quantity of copper and consequently have a large mass per unit length. This is not a particular problem in, for example, a terrestrial base station transmitter, but is potentially a problem in the mass-conscious world of satellite communications. The size issue is also of particular concern in any portable or mobile applications.

   A further problem is that of cost. Due to the relatively large quantities of expensive materials required in their construction, high-power, low-loss cables tend to be very expensive, particularly at higher frequencies. Although again not a problem in large *terrestrial (or even, in this case satellite) transmitters, this is a* major disadvantage in medium-power mass-produced mobile terminals.

2. Reduce (or eliminate altogether) the main-path delay element and consequently its loss. It may, at first sight, appear that this would also lead to the reduction or elimination of the linearising benefits of the feedforward system and this is to some extent true; however, there are benefits to this approach and these will be examined below.

   The reduction (in length or value) or elimination of the main-path delay element will consequently reduce the size, weight and cost of a feedforward system; it will also make the system easier to integrate, making the technique potentially applicable for mobile, or even portable, equipment. It will also, as has already been mentioned, improve the overall power efficiency, which again will make the technique more attractive for mobile or portable equipment.

The option of reduction or elimination of the main-path delay element is therefore worthy of serious consideration and in particular the effect of such a move on the level of linearity achievable by the system will be examined.

5.9.1 Theoretical Analysis

Both loops *of the feedforward system contain delay elements in order to* ensure that their relevant signals arrive at the subtraction point with the correct time and phase relationship. The analysis presented below is therefore applicable to the elimination of either time delay within the loop (main-path or reference-path), although the derivation will concentrate on the main-path delay as this has the greater effect on efficiency.

For the sake of simplicity, it is assumed that neither amplifier exhibits a frequency-dependent characteristic (i.e., *linear distortion*) and that the amplitudes of the various frequency components in the main-path and error-path

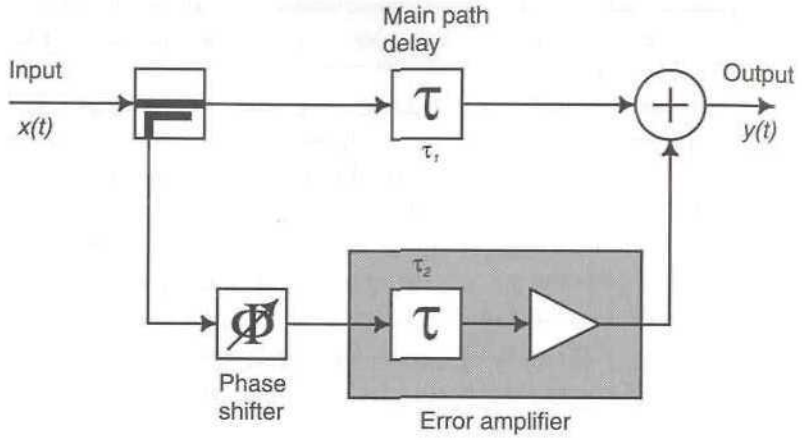


Figure 5.13 Simplified diagram of the correction loop in a feedforward system.

signals are the same. Thus, if the correct length of time delay element were to be incorporated, then the loop would exhibit perfect cancellation.

The error path consists of a phase-shift element, with phase-shift, θ, followed by an error amplifier which is modelled by a time delay, τ_2, and an ideal gain block (with no time delay). The main path consists of a delay element, with time delay τ_2, and an ideal summing junction (in place of the more usual directional coupler). This configuration is shown in Figure 5.13.

Assume that the input, $x(t)$, is a single CW tone of amplitude, A, that is,

$$x(t) = A\cos(\omega_0 t + \phi) \tag{5.70}$$

The final output signal, $y(t)$, is therefore given by:

$$y(t) = A\cos\{\omega_0(t + \tau_1) + \phi\} + A\cos\{\omega_0(t + \tau_2) + \phi + \theta\} \tag{5.71}$$

In examining the effect of the reduction or elimination of the main-path delay element value (and hence loss), it is convenient to assume that the error-path delay is therefore longer than the main-path delay by a time of $\Delta\tau$. The error-path time delay is therefore given by:

$$\tau_2 = \tau_1 + \Delta\tau \tag{5.72}$$

Combining the latter two equations gives:

$$y(t) = A\cos\{\omega_0(t + \tau_1) + \phi\} + A\cos\{\omega_0(t + \tau_1 + \Delta\tau) + \phi + \theta\} \tag{5.73}$$

Expanding this gives:

$$\begin{aligned} y(t) = {} & A\cos(\omega_0 t + \phi)\cos\omega_0\tau_1 - A\sin(\omega_0 t + \phi)\sin\omega_0\tau_1 \\ & + A\cos(\omega_0 t + \phi)\cos\omega_0\tau_1\cos(\omega_0\Delta\tau + \theta) \\ & - A\sin(\omega_0 t + \phi)\sin\omega_0\tau_1\sin(\omega_0\Delta\tau + \theta) \\ & - A\sin(\omega_0 t + \phi)\cos\omega_0\tau_1\sin(\omega_0\Delta\tau + \theta) \\ & - A\sin(\omega_0 t + \phi)\sin\omega_0\tau_1\cos(\omega_0\Delta\tau + \theta) \end{aligned} \tag{5.74}$$

Since the various system components are assumed linear for the purposes of this derivation, the output signal will consist purely of a single CW tone, hence:

$$y(t) = R\cos(\omega_0 t + \gamma) \tag{5.75}$$

Combining (5.74) and (5.75) gives:

$$\begin{aligned} R\cos(\omega_0 t + \gamma) = {} & A\cos(\omega_0 t + \phi)\{\cos\omega_0\tau_1(1 + \cos(\omega_0\Delta\tau + \theta)) \\ & - \sin\omega_0\tau_1\sin(\omega_0\Delta\tau + \theta)\} \\ & - A\sin(\omega_0 t + \phi)\{\sin\omega_0\tau_1(1 + \cos(\omega_0\Delta\tau + \theta)) \\ & + \cos\omega_0\tau_1\sin(\omega_0\Delta\tau + \theta)\} \end{aligned} \tag{5.76}$$

The magnitude of the resultant output signal may be found from:

$$\begin{aligned} R^2 = {} & [A\cos\omega_0\tau_1(1 + \cos(\omega_0\Delta\tau + \theta)) - A\sin\omega_0\tau_1\sin(\omega_0\Delta\tau + \theta)]^2 \\ & + [A\sin\omega_0\tau_1(1 + \cos(\omega_0\Delta\tau + \theta)) + A\cos\omega_0\tau_1\sin(\omega_0\Delta\tau + \theta)]^2 \end{aligned} \tag{5.77}$$

Simplifying this expression gives:

$$R^2 = [A + A\cos(\omega_0\Delta\tau + \theta)]^2 + [A\sin(\omega_0\Delta\tau + \theta)]^2 \tag{5.78}$$

After normalisation (i.e., setting $A = 1$), the amplitude of the resultant output signal becomes:

$$R = \sqrt{[1 + 1\cos(\omega_0\Delta\tau + \theta)]^2 + \sin^2(\omega_0\Delta\tau + \theta)} \tag{5.79}$$

This simplifies to:

$$R = 2\cos\{(\omega_0\Delta\tau + \theta)/2\} \tag{5.80}$$

The IMD cancellation, expressed in dB, is therefore:

$$R_{dB} \approx 20 \log_{10} \left| \frac{1}{[2 \cos\{(\omega_0 \Delta\tau + \theta)/2\}]} \right| \tag{5.81}$$

where ω_0 is the (angular) frequency offset from the band center frequency (in rad./s) and the delay $\Delta\tau$, is given by:

$$\Delta\tau = \frac{N}{f_C} \tag{5.82}$$

with N being the number of cycles of delay mismatch and f_C being the channel center frequency (in Hz). Note also that θ is in radians.

Finally, it is also possible to determine the phase difference, $\Delta\gamma$, between the input and output tones by referencing the output phase to the input phase. This is done by setting $\phi = 0$ and results in:

$$\Delta\gamma = \tan^{-1}[\sin(\omega_0 \Delta\tau + \theta)/\{1 + \cos(\omega_0 \Delta\tau + \theta)\}] \tag{5.83}$$

5.9.2 Typical Characteristics

Equations (5.81) and (5.83) may be shown graphically for a number of cycles of delay mismatch, and Figure 5.14 illustrates this at an example center frequency of 1 GHz. The phase-shift, θ, is chosen such that perfect cancellation occurs at the center of the band, that is, $\theta = 180°$. It can be seen that the degree of cancellation decreases with increasing frequency separation from the band center and also with increasing numbers of cycles of delay mismatch. With no delay mismatch, the simplified model assumed in the derivation would produce ideal (that is $-\infty$ dB) cancellation across the band.

It can be seen that greater than 30 dB of cancellation may be achieved across the band shown (10 MHz) with one cycle of delay mismatch and this would correspond to an intermodulation product level of less than −45 dB assuming that a class-C main amplifier (with a class-A error amplifier) was employed in the loop. Even with four cycles of delay mismatch, almost 20 dB of cancellation may be achieved across the band; this would still result in a class-C amplifier being improved to be on a par with, or better than a good class-A amplifier. In this case, the removal of most, if not all, of the delay would result in an efficiency rating comparable with (but nowhere near equalling) that of the original class-C amplifier.

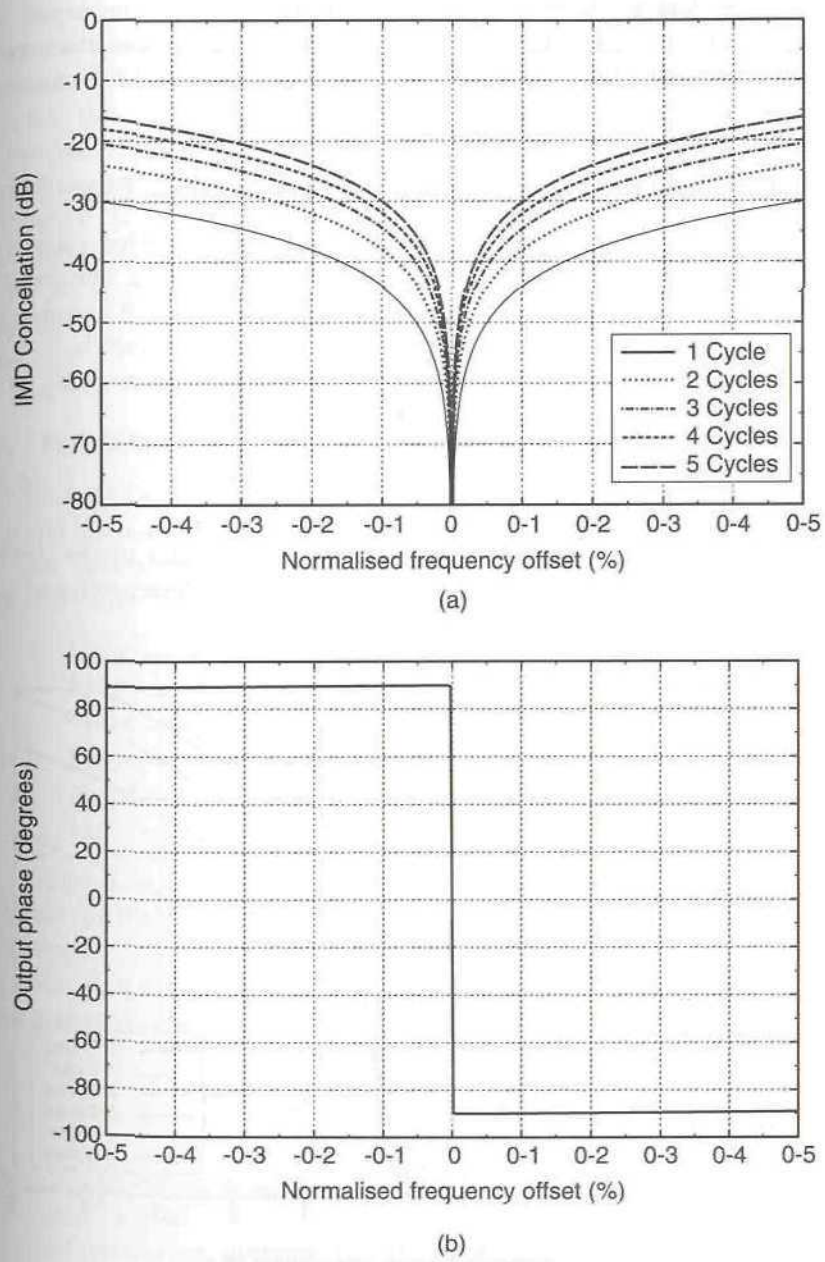


Figure 5.14 Effect of delay mismatch on achievable cancellation (a) and phase shift (for one cycle of delay mismatch), (b) in a feedforward amplifier.

Figure 5.14(b) illustrates the phase change between the input and output tones with one cycle of delay mismatch. It can be seen that across most of the band, the phase-shift tends to $\pm 90^\circ$, with the 180° transition occurring at the center of the band.

5.9.3 Combination of Delay Mismatch and Gain and Phase Error

It is clearly impossible for a practical feedforward system to achieve a perfect gain and phase balance and hence some error must be assumed. The result of this effect alone was discussed in Section 5.5; this section will examine the effect of such errors on a system with a deliberate delay mismatch.

The effect of delay mismatch may be incorporated into (5.30) as an additional phase error term:

$$\Delta\Phi = \delta\omega_0 \frac{N}{f_C} \tag{5.84}$$

where: $\delta\omega_0$ is the angular frequency offset from the center of the desired correction bandwidth, N is the number of cycles of delay mismatch and f_C is the band center frequency (in Hz).

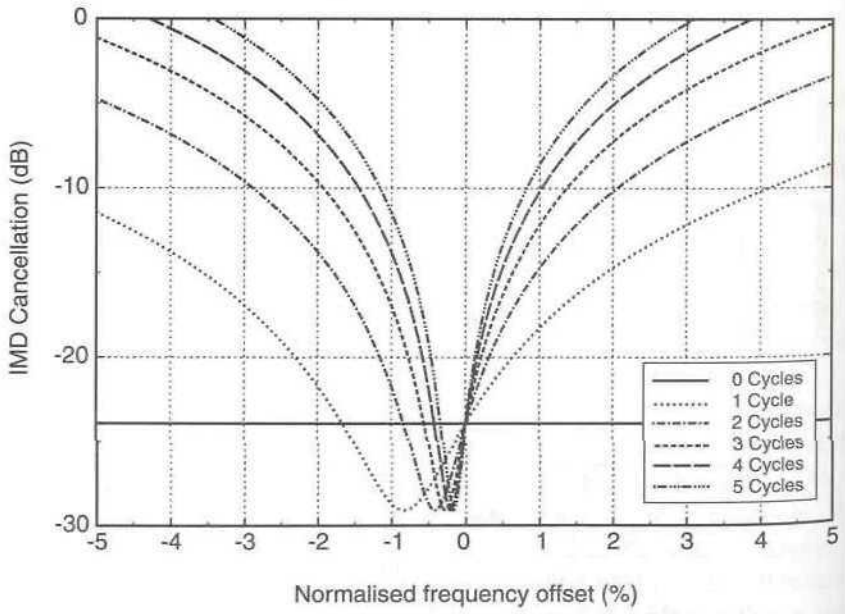


Figure 5.15 Effect of a gain error of 0.3 dB and a phase error of 3° on the cancellation achievable from a feedforward system with various amounts of delay mismatch.

As an example, Figure 5.15 illustrates the case of a typical set of values for gain and phase error of 0.3 dB and 3°, respectively. Notice that the phase error introduces a 'shift' in the frequency at which optimum cancellation is achieved; this is quite significant (note that the frequency scale is an order of magnitude greater in this figure than in Figure 5.14). Note also that a greater level of cancellation is achievable at the 'optimum' frequency than would be predicted by considering the gain and phase errors alone (the '0 Cycles' line). This apparent anomaly arises due to the fact that the deliberate delay error allows an optimal phase setting to be achieved at one frequency, hence resulting in the level of cancellation being determined by the amplitude error alone.

5.9.4 Path Difference and Subtraction Issues in a Feedforward System

It is evident from the above analysis and discussion that there are two potential methods of generating the error signal and subsequently matching the time delays in order to provide cancellation of the distortion in the output coupler:

1. Generate the necessary 180° phase difference by utilising a 180° hybrid to derive the error signal and match (exactly) the time delays in the main and error paths (including the error amplifier).
2. Utilise a conventional 0° hybrid (or coupler) to generate the error signal and mismatch the time delays in the main and error paths by half a wavelength (of the center frequency).

In practice, neither of these methods will be exact; however, it will be assumed for the present that a close match has been achieved and can be maintained.

For ideal cancellation to be achieved at the center frequency, the time delay mismatch must be given by:

$$\Delta\tau = N + \frac{180^\circ - \theta}{360 f_0} \tag{5.85}$$

where N is an integer number of cycles of the center frequency, f_0, and θ is in the range 0° to 180°.

The results of utilising the two potential schemes for providing cancellation (outlined above) are illustrated in Figures 5.16 and 5.17. It is evident from these figures that the use of the half-wavelength path difference is inferior to the 180° hybrid technique as it results in a poorer level of cancellation across the band. This is due to the phase difference between the

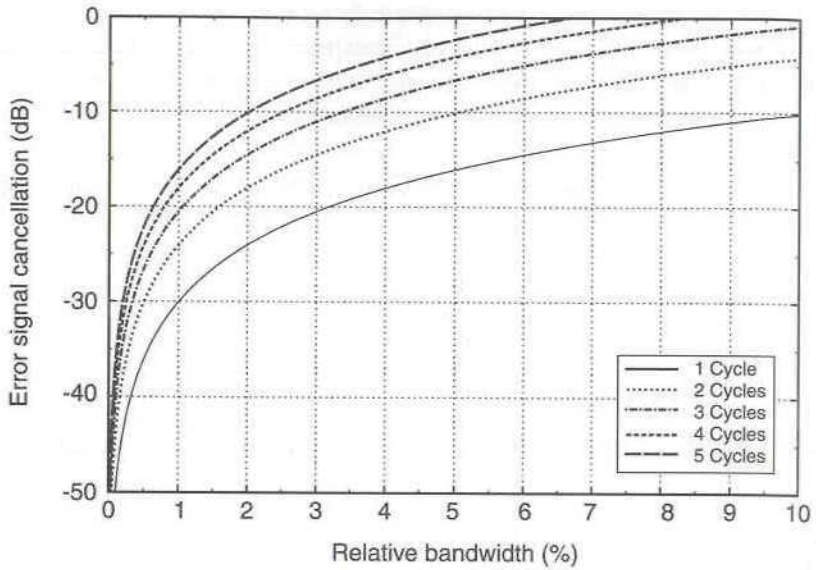


Figure 5.16 Effect of relative bandwidth on the achievable cancellation in a feedforward system with varying degrees of delay mismatch. The necessary antiphase relationship is provided in this case by employing a 180° hybrid as the subtracter for deriving the error signal.

two paths only being equal to 180° at the center frequency; for the case of a 180° hybrid, this is achieved across a relatively broad bandwidth (assumed to be larger than the bandwidth of interest in Figure 5.16).

5.9.5 Example Systems

There are many potential applications of the delay-element reduction or removal techniques outlined above. In general, applications will concentrate on systems where efficiency is a crucial factor and where the broadband cancellation required is either relatively modest, or else the bandwidth over which cancellation is required is modest. An example of the former system would be a satellite transponder and an example of the latter would be a mobile transmitter for a narrow-band modulation scheme (e.g., 5 kHz SSB linear modulation).

5.9.5.1 Satellite Transponder

As an illustration, consider a 200 MHz bandwidth satellite transponder operating at a center frequency of 10 GHz. This corresponds to a relative bandwidth of 2% and Figure 5.16 shows that under these circumstances it is

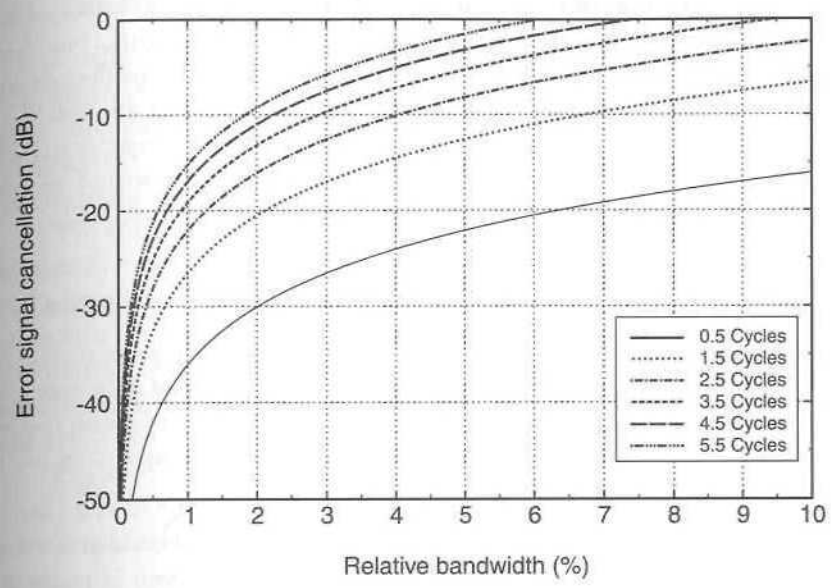


Figure 5.17 Effect of relative bandwidth on the achievable cancellation in a feedforward system with varying degrees of delay mismatch. The necessary antiphase relationship is provided in this case by the use of an odd number of half-cycles (of the center frequency) as a delay mismatch.

possible to achieve a distortion cancellation of greater than 20 dB. Thus, a feedforward system employing a class-C main amplifier and a class-C error amplifier could achieve a distortion level comparable with or better than that of a good class-A amplifier, with an efficiency rating approaching (but nowhere near equalling) that of the class-C amplifier.

5.9.5.2 Mobile Radio PA (Single-Channel)

As in the case of the Cartesian-loop amplifier discussed in Chapter 4, the use of the term 'single-channel' refers to the amplifier only being required to amplify one channel at a time. The channel to be amplified may, of course, appear anywhere within the rated bandwidth of the amplifier, assuming that a control technique is employed that is able to provide suitable adjustment of the loop gain and phase parameters to effect re-optimisation at the new frequency.

As an example, consider a 5 kHz SSB channel at 900 MHz: this corresponds to a relative bandwidth of 0.00056%, and Figure 5.16 shows that greater than 80 dB of cancellation may theoretically be achieved even with four cycles of delay mismatch (i.e., probably corresponding to a complete elimination of the delay element). This analysis does not, of

course, take into account the errors in the control system, leakages, or frequency-dependant characteristics of the system components, but does illustrate that the removal of the main-path delay has almost no effect on the practical cancellation of distortion which may be achieved in this case.

Feedforward may therefore prove to be an attractive alternative to both Cartesian loop and adaptive predistortion for use in narrowband mobile transmitters, particularly if aided by predistortion (see Section 5.9.7).

5.9.6 Reduction or Elimination of the Reference-Path Delay Element

There may appear, at first sight, to be little to be gained by reducing or eliminating the reference-path delay element, as it is concerned with very low power signals (in general) and consequently has a negligible effect on system power efficiency. There are, however, two reasons why such a move is worthy of consideration:

1. Improvement in system noise figure. It will be shown in Section 5.13.1 that the noise figure of a feedforward system is primarily dependent upon the loss in the reference path up to the input of the error amplifier. The delay element in the reference path must compensate for the time delay introduced by the main amplifier and in many systems this delay will be considerable due to the large gain required of that amplifier. It follows therefore that the reference-path time delay must be significant in length and for this reason, together with the almost inevitable use of thin (and consequently lossy) coaxial cable, the loss introduced by this element will be relatively large. It is therefore evident that such a large loss will significantly degrade the noise figure of the system.

   In many systems, where power amplification is the primary purpose, this degradation in noise figure is relatively unimportant. However, in lower power systems, the effect may be undesirable and consequently the artificial reduction or complete removal of the reference-path delay element would prove beneficial.

2. Reduction in overall size of the system. In a similar manner to that for the main-path delay element, the removal of the reference-path delay will result in a size reduction. This again will be of benefit in mobile or portable equipment and will also aid in integration of the system. Although the benefits in this area will be much smaller than the corresponding benefit of eliminating the main-path delay element, they are still worth considering, due to the comparatively large size of even the lowest power delay lines.

5.9.7 Efficiency Improvement Using Predistortion

It is possible to incorporate an RF predistorter (see Chapter 6) prior to the main amplifier in a feedforward system. This has the effect of improving the main amplifier IMD level without significantly degrading its power efficiency (in most systems). It is therefore a powerful tool in improving the efficiency of a feedforward system.

The efficiency levels which can be obtained by this technique may be predicted from (5.33) and (5.49) simply by an appropriate choice of assumed main amplifier IMD suppression, S_{A1}. Thus, for example, for a 10 dB predistorter, the value of S_{A1} should be increased by 10 dB, with all other parameters remaining unchanged. It is also possible to improve the error amplifier IMD performance using predistortion and this will generally have a practical improving effect on efficiency in many systems.

Figure 5.18 shows the efficiency improvement realised by using predistortion for a feedforward amplifier employing class-C main and error amplifiers (case 5). Notice that with 20 dB or more of predistorter benefit, the overall power efficiency of the system approaches that of the main amplifier (within 12%). Even a 10 dB predistorter, which is relatively straightforward to realise practically, yields an efficiency gain of 8% (i.e.,

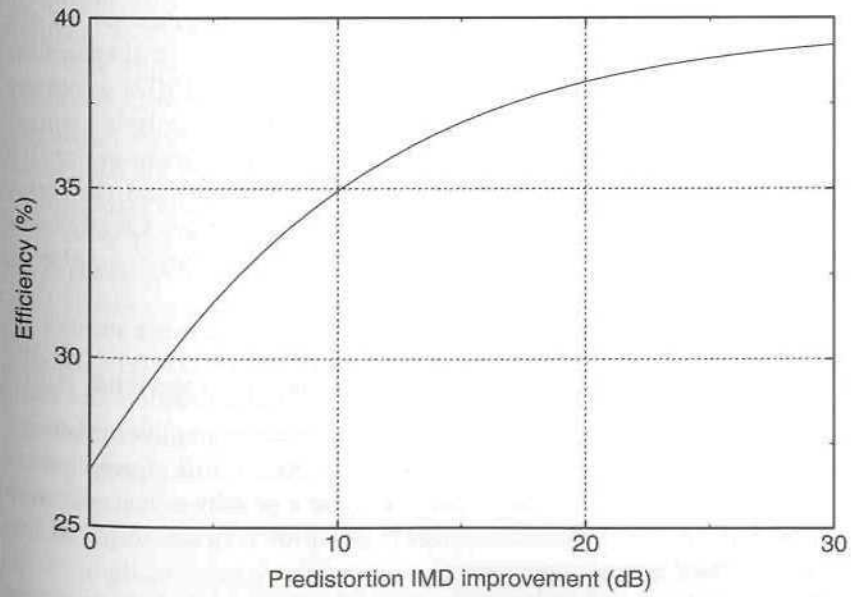


Figure 5.18 Feedforward efficiency characteristic with a predistorted main amplifier ($\eta_{A1} = 0.5$, $\eta_{A2} = 0.3$ and $S_{A1} = 15$ dB before predistortion).

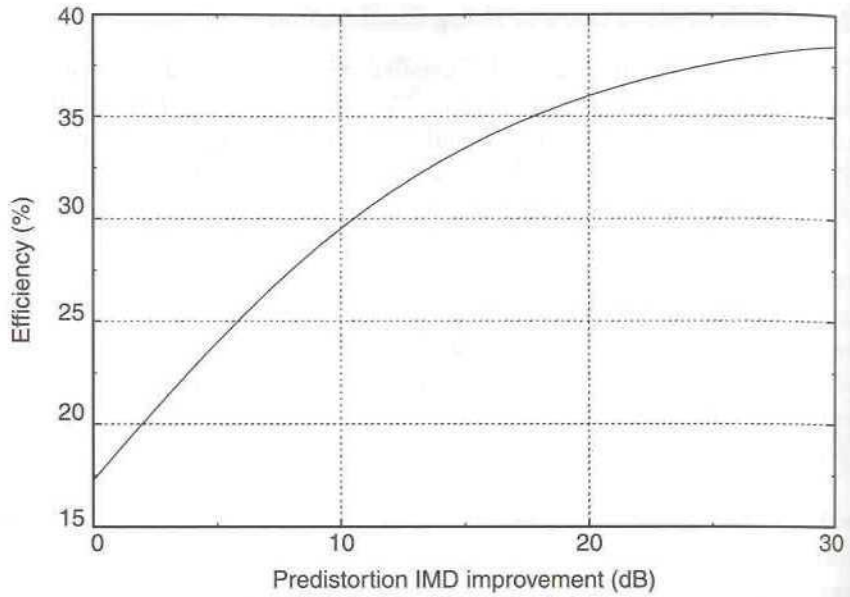


Figure 5.19 Feedforward efficiency characteristic with a predistorted main amplifier ($\eta_{A1} = 0.5$, $\eta_{A2} = 0.05$ and $S_{A1} = 15$ dB before predistortion).

around a 30% improvement). The theoretical IMD specification of the overall system, in this case, would be: 15 + 10 + 15 = 40 dBc (main amplifier IMD + predistorter improvement + error amplifier IMD) assuming >15 dB of cancellation from feedforward. This simple analysis assumes that the third-order distortion dominates, even after predistortion (or that the predistorter acts on all significant orders of distortion) and that error amplifier IMD dominates, rather than imperfect cancellation. Clearly some or all of these assumptions may be invalid in the case of class-C amplifiers, however, the basic principles are illustrated.

Figure 5.19 modifies the above by assuming that the error amplifier is class-A, rather than class-C. This could yield an overall IMD level of 60 dBc and hence is closer to the levels required in mobile radio handsets and base-stations. It can be seen from Figure 5.19 that the error amplifier efficiency quickly becomes unimportant, with greater than (say) 10 dB of predistorter improvement. This combination is therefore clearly worthy of consideration for use with the newer, broadband linear modulation formats proposed for, for example, third generation systems.

Note that the above analysis assumes perfect (or significant) main-tone cancellation in the error loop and this will not always be the case in practice. In broadband systems in particular, the main tone cancellation will be more

modest, such that the figures presented above are optimistic. They should therefore be interpreted with care, in that application.

5.10 Linear Distortion Correction in a Feedforward System

5.10.1 Introduction

The nonlinear frequency and phase response characteristics, present in any practical RF amplifier, are referred to as *linear distortion*. Amplitude 'ripple' across the bandwidth of the amplifier and 'roll-off' at the band edges are the most commonly observed manifestations of linear distortion. The phase response can also suffer from 'ripple' within the amplifier bandwidth, together with a 'flattening' of the response at the band edges. Whilst these effects are undesirable in a practical amplifier, they are usually considered more acceptable than nonlinear (i.e., harmonic and intermodulation) distortions.

A number of applications exist, however, in which a very linear frequency and phase response are essential. Examples include adaptive antenna systems, test and measurement instrumentation, calibration systems, and some electronic warfare systems. Both linear and nonlinear distortion properties are important in these systems and this adds to the difficulty of their design.

The application of feedforward to an amplifier with both linear and nonlinear distortions has a beneficial effect on each. As has already been outlined, considerable improvements in both intermodulation and harmonic distortions are possible, and it will be shown below that linear distortion is also dramatically improved by a correctly designed feedforward system.

It is possible to analyse the effect of various feedforward system parameters on the degree of linear distortion correction produced [19]. This process is detailed below.

5.10.2 Analysis of Linear Distortion Removal Characteristics

The feedforward system illustrated in Figure 5.20 is assumed to contain ideal system components, with the exception of the main and error amplifiers. These are assumed to have large, constant gain, upon which is superimposed a small sinusoidal gain variation ('ripple') across the band of interest. For the error amplifier, this ripple is assumed to have a different peak-to-peak amplitude and a phase-offset, ϕ, relative to that of the main amplifier.

The frequency response of the complete feedforward system can be assessed by examining the frequency responses of the main and error paths,

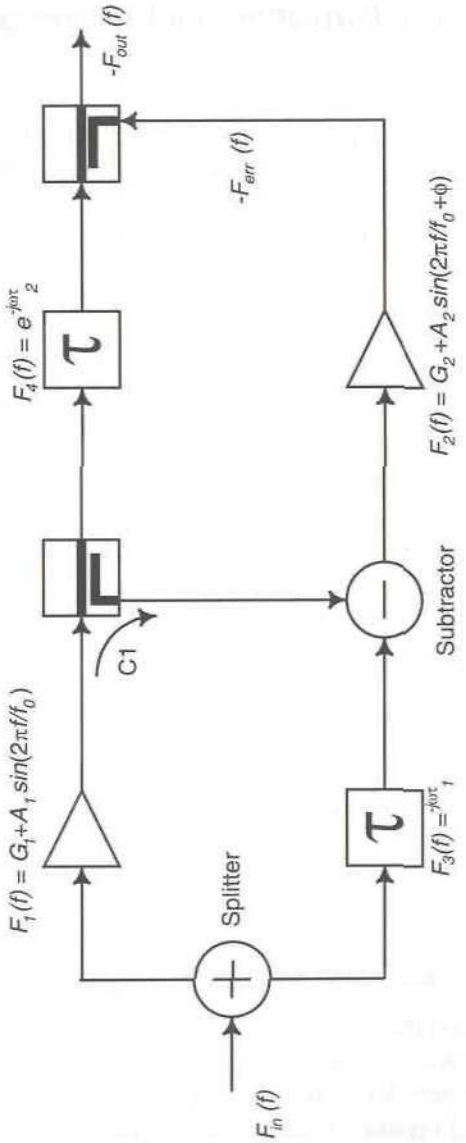


Figure 5.20 Simplified block diagram of a feedforward system, detailing the analysis of linear distortion properties.

and subtracting these in the output coupler. This technique can be explored as follows. Consider a signal (**S**), at a given frequency within the pass-band of the system: it will arrive at the output coupler (after amplification by the main amplifier), together with the error signal (which will be subtracted in the output coupler); the error signal representing the difference between the original signal (**S**) and an ideal amplified version of this original signal. Consider now that the signal (**S**) is swept across the bandwidth of the amplifier, the resulting response is formed by an infinite number of these subtractions in the output coupler. The result is then effectively a subtraction of the frequency response characteristics of the main and error paths and this therefore provides the overall frequency response of the feedforward system.

A basic feedforward system with an input signal, $F_{in}(f)$, has a transfer function up to the output of the error amplifier, $F_{err}(f)$, (and assuming a 3 dB loss in the input splitter and the subtracter) given by:

$$F_{err}(f) = \frac{F_{in}(f)}{2}\left[(G_1 + A_1 \sin 2\pi f/f_0)C_1 - e^{-j\omega\tau_1}\right] [G_2 + A_2 \sin(2\pi f/f_0 + \phi)] \tag{5.86}$$

where $\omega = 2\pi f$ and C_1 is the fractional coupling factor of the main-path sampling coupler (e.g. for a 20 dB coupler, C_1 would equal 0.01).

The complete transfer function, assuming a 3 dB hybrid output coupler and negligible loss in the sampling coupler, is given by:

$$F_{out}(f) = \frac{1}{\sqrt{2}}\left(F_{in}(f)[G_1 + A_1 \sin(2\pi f/f_0)]\frac{e^{-j\omega\tau_2}}{\sqrt{2}} - F_{err}(f)\right) \tag{5.87}$$

The delay elements are assumed to be ideal and to perfectly match the corresponding amplifier delays. It is therefore convenient to assume that they are zero and consequently that the amplifiers are ideal and contribute no delay to the loop. Therefore:

$$\tau_1 = \tau_2 = 0 \tag{5.88}$$

and hence (5.87) becomes:

$$H(f) = \frac{F_{out}(f)}{F_{in}(f)} = \frac{G_1 + A_1 \sin(2\pi f/f_0)}{2} - \frac{[\{G_1 + A_1 \sin(2\pi f/f_0)\}C_1 - 1][G_2 + A_2 \sin(2\pi f/f_0 + \phi)]}{2\sqrt{2}} \tag{5.89}$$

Expanding the above gives:

$$H(f) = \frac{G_1}{2} + \frac{A_1}{2}\sin(2\pi f/f_0) - \frac{C_1G_1G_2}{2\sqrt{2}} - \frac{C_1G_2A_1}{2\sqrt{2}}\sin(2\pi f/f_0) + \frac{G_2}{2\sqrt{2}} - \frac{C_1G_1A_2}{2\sqrt{2}}\sin(2\pi f/f_0 + \phi) - \frac{C_1A_1A_2}{2\sqrt{2}}\sin(2\pi f/f_0)\sin(2\pi f/f_0 + \phi) + \frac{A_2}{2\sqrt{2}}\sin(2\pi f/f_0 + \phi) \quad (5.90)$$

If it is assumed, for the present, that the error amplifier has a flat frequency response ($A_2 = 0$), then (5.90) reduces to:

$$H(f) = \frac{G_1}{2} + \frac{A_1}{2}\sin(2\pi f/f_0) - \frac{C_1G_1G_2}{2\sqrt{2}} - \frac{C_1G_2A_1}{2\sqrt{2}}\sin(2\pi f/f_0) + \frac{G_2}{2\sqrt{2}} \quad (5.91)$$

Assuming that the feedforward loop is balanced to achieve perfect cancellation (i.e., $C_1 = 1/G_1$), then:

$$A_1 = \frac{C_1G_2A_1}{\sqrt{2}} \quad (5.92)$$

Giving, $C_1 = \sqrt{2}/G_2$ and (from (5.91)):

$$H(f) = \frac{G_1}{2} + \frac{A_1}{2}\sin(2\pi f/f_0) - \frac{G_1}{2} - \frac{A_1}{2}\sin(2\pi f/f_0) + \frac{G_2}{2\sqrt{2}} = \frac{G_2}{2\sqrt{2}} = \frac{G_1}{2} \quad (5.93)$$

which is constant with frequency, hence resulting in the overall transfer function of the feedforward system being frequency flat.

5.10.3 Incorporation of Error Amplifier Ripple

If the error amplifier ripple is no longer neglected, in other words $A_2 \neq 0$ in (5.90), it is possible to analyse the effect of this ripple upon the overall system response.

For perfect matching within the feedforward loop, $C_1 = \sqrt{2}/G_2$ as before and (5.90) reduces to:

$$H(f) = \frac{G_2}{2\sqrt{2}} - \frac{G_1A_2}{2G_2}\sin(2\pi f/f_0 + \phi) - \frac{A_1A_2}{2G_2}\sin(2\pi f/f_0)\sin(2\pi f/f_0 + \phi) + \frac{A_2}{2\sqrt{2}}\sin(2\pi f/f_0 + \phi) \quad (5.94)$$

Combining terms gives:

$$H(f) = \frac{G_2}{2\sqrt{2}} - \frac{\sqrt{2}G_1A_2 - G_2A_2}{2\sqrt{2}G_2}\sin(2\pi f/f_0 + \phi) - \frac{A_1A_2}{2G_2}\sin(2\pi f/f_0)\sin(2\pi f/f_0 + \phi) \quad (5.95)$$

In virtually all practical systems, the amplitude of the frequency response ripple for the main and error amplifiers is small (usually a few dB or less) compared to the gain of the error amplifier (usually $>>$ 40 dB). The final term in (5.95) therefore tends to zero, and simplifies to give:

$$H(f) = \frac{G_2}{2\sqrt{2}} - \frac{\sqrt{2}G_1A_2 - G_2A_2}{2\sqrt{2}G_2}\sin(2\pi f/f_0 + \phi) \quad (5.96)$$

In almost any practical feedforward system, G_2 will be of the same order as, or greater than, G_1. In addition, the peak value of the error amplifier gain ripple, A_2, will generally be small, as this amplifier is generally low-power and designed for high linearity performance. Combining these two factors and examining (5.96), it is evident that the output ripple from the complete feedforward amplifier will be small and of the same form as that of the error amplifier. The ripple present in the main amplifier transfer characteristic having been completely removed (in a perfectly balanced system).

The peak value of the ripple is given by:

$$H = A_2\left(\frac{1}{2\sqrt{2}} - \frac{G_1}{2G_2}\right) \quad (5.97)$$

Note that if it were possible to set $G_1/G_2 = 1/\sqrt{2}$, then even this residual ripple would be eliminated. This is generally not possible in a practical system; however, the two amplifier gains may be designed to be quite close to this value, thus keeping the linear distortion to a minimum. This is possible even when utilising relatively poor (in terms of linear distortion) main and error amplifiers.

In some low-gain applications, it may be possible for the main and error amplifier ripple amplitudes to remain significant with respect to the gain of the error amplifier. In such circumstances, it is not possible to neglect the final term of (5.95) and hence, even under the conditions where:

$$\frac{G_1}{G_2} = \frac{1}{\sqrt{2}} \tag{5.98}$$

a residual ripple will result. Under these circumstances, assuming that (5.98) is satisfied, the resulting transfer function becomes:

$$H(f) = \frac{G_1}{2} - \frac{A_1 A_2}{2G_2} \sin(2\pi f/f_0) \sin(2\pi f/f_0 + \phi) \tag{5.99}$$

Therefore, as the relative ripple, A_2/G_2, present on the error amplifier transfer characteristic, tends to zero, so does the resultant ripple in (5.99), as expected.

Practical verification of the linear distortion correction properties of a feedforward system is provided in the literature [19], with a reduction in amplitude ripple from an initial value of 2.5 dB peak to peak to ±0.25 dB with feedforward correction, being reported. Greater improvements are possible with more careful selection and matching of components in the reference and error amplifier paths.

5.11 Temperature Drift and Component Aging

A considerable part of the design effort focused on feedforward amplifiers over the last 20 years has concerned the development of techniques to ensure the maintenance of optimum performance over time. Since the gain and phase matching characteristics of feedforward amplifiers are critical to their performance it is necessary to ensure that these can be maintained throughout the designed life of the amplifier.

It is here that one of the major design decisions of any feedforward system must be made. There is a compromise between the additional complexity of either environmental stabilisation, additional circuitry to monitor loop performance and correct for errors (i.e., a control scheme) or additional loops. It is instructive to examine each of these three alternatives in turn.

5.11.1 Environmental Stabilisation

Specialist environmental stabilisation is usually considered impractical due to the cost and size of the necessary equipment; both are usually prohibitive. For certain environments this stability may occur naturally, such as in submarine cable applications where the temperature of the sea bed at a great depth remains substantially constant throughout the year. However, even in this case, the effects of component aging are not considered and hence the system performance is likely to degrade with time.

One approach to environmental stabilisation is that of deliberately heating the loop and its components by mounting them on a temperature-controlled heating plate; the whole may then be thermally insulated from the outside world, thus providing (by two mechanisms) a degree of thermal stabilisation. This mechanism has been used in an outdoor repeater application, having as its main advantage relative simplicity and hence reliability.

5.11.2 Performance Monitoring

The idea of using additional circuitry to allow the feedforward system to monitor, and correct for, its own performance has received considerable interest. There are a number of patented configurations to fulfil this function as it is arguably the most elegant solution to the problem. A number of these configurations will be discussed in more detail towards the end of this section.

The idea of allowing the system to monitor its own performance and then perform the necessary correction implies some form of feedback system around the feedforward loop. This feedback system is required to control the gain- and phase-matching of the two halves of the feedforward loop, termed the *error loop* and the *correction loop* for convenience (see Figure 5.21). Thus it is evident that two separate feedback systems are required: one to correct for the gain- and phase-mismatches in the error loop in order to minimise the input signal component of the error signal, and the second to correct for the gain- and phase-mismatches in the correction loop in order to minimise the amount of distortion present in the final output signal. These two loops are shown conceptually in Figure 5.22.

The gain and phase adjustment components ('compensation circuits') shown in Figure 5.22 could appear in a number of different locations around the two halves of the feedforward loop. For example, these components in the error loop could be placed in the top path, for example, either before or after the main amplifier (along with a number of other positions as shown in

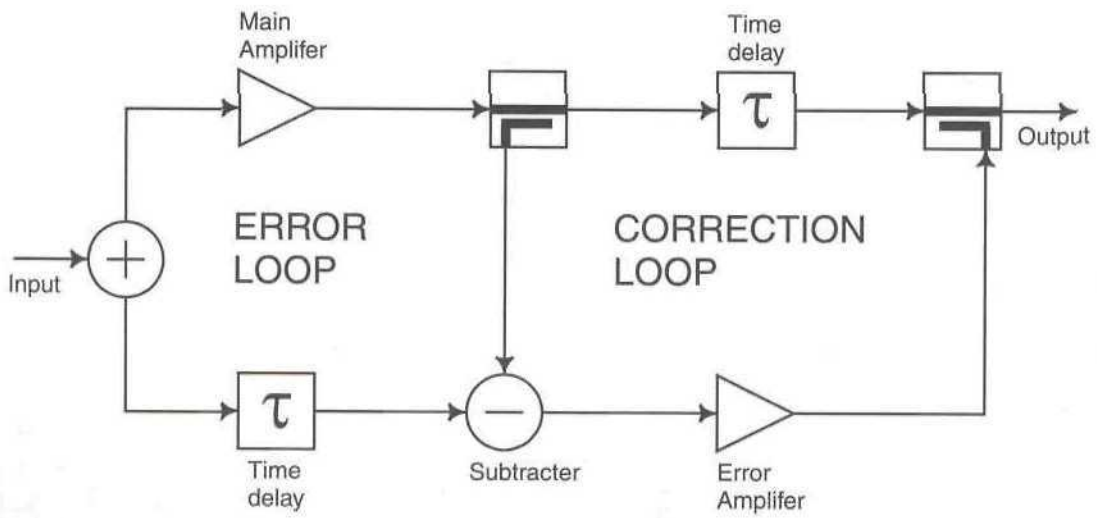


Figure 5.21 Error loop and correction loop in a feedforward system.

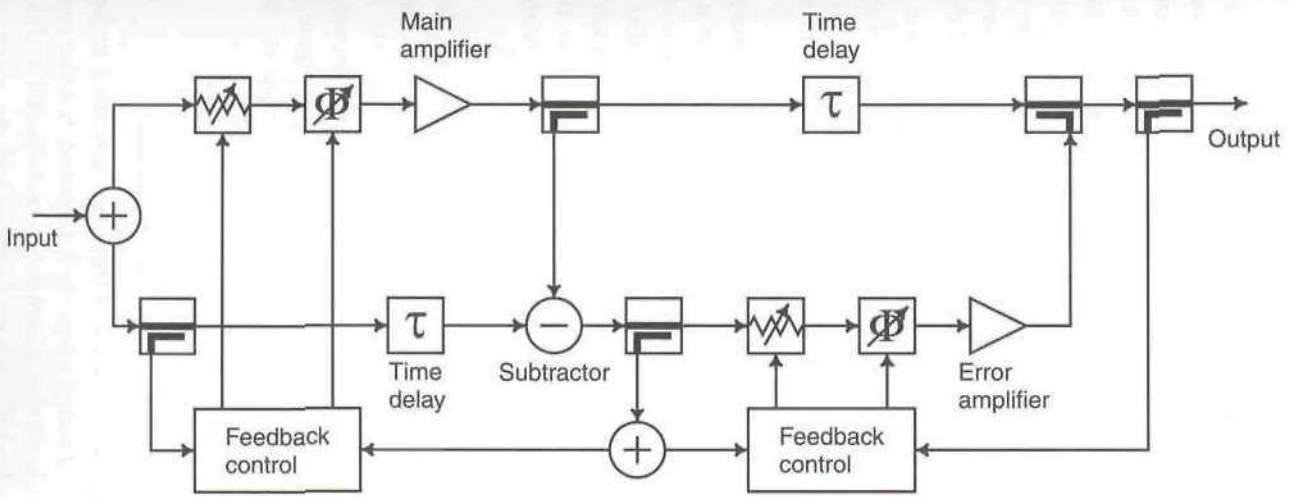


Figure 5.22 Feedback control applied to a feedforward amplifier.

Figure 5.23). Ultimately, the decision regarding the placement of these components rests with the system designer; however, a number of practical points must be considered.

1. *Power handling*. Since feedforward linearisation is most often applied to amplifiers having a significant power output, the power handling capabilities of the gain and phase compensation components must be considered. A common method of construction of these components involves the use of 3 dB quadrature hybrid couplers together with either PIN or varactor diodes respectively, however, both of these devices will only operate at low signal levels. It is thus necessary to place them in small-signal parts of the system, that is, at the inputs to rather than the outputs from either of the amplifiers.
2. *Linearity*. The linearity of the adjustment components is critical to the overall system performance since any distortion generated by either will translate to the output. In the case of the correction loop, distortion from these components will be amplified by the error amplifier (assuming that they are placed before it) and appear directly in the output signal. In the case of the error loop, if they are placed in the time-delay path then their distortion will appear in the error signal and be amplified by the error amplifier as above. If, however, they are placed before the main amplifier, then their distortion will appear as part of that from the main amplifier and it will be corrected as such at the final output. For this reason a position prior to the main amplifier is usually considered to be optimum for these components.

It is also possible to locate the adjustment components for the error loop between the sampling coupler and the subtracter (with suitable fixed attenuation as necessary). The signal levels at this point will in general be small, however, any distortion produced here will again appear in the error signal and will be amplified and translated to the output signal. A summary of the possible locations of the adjustment components is shown in Figure 5.23.

5.11.3 Additional Loops

The use of multiple feedforward loops has been discussed in Section 5.3. It was assumed that the additional loops were required in order to improve the absolute performance above that which could be achieved with a single loop, due to limitations in the linearity performance of the error amplifier, for example.

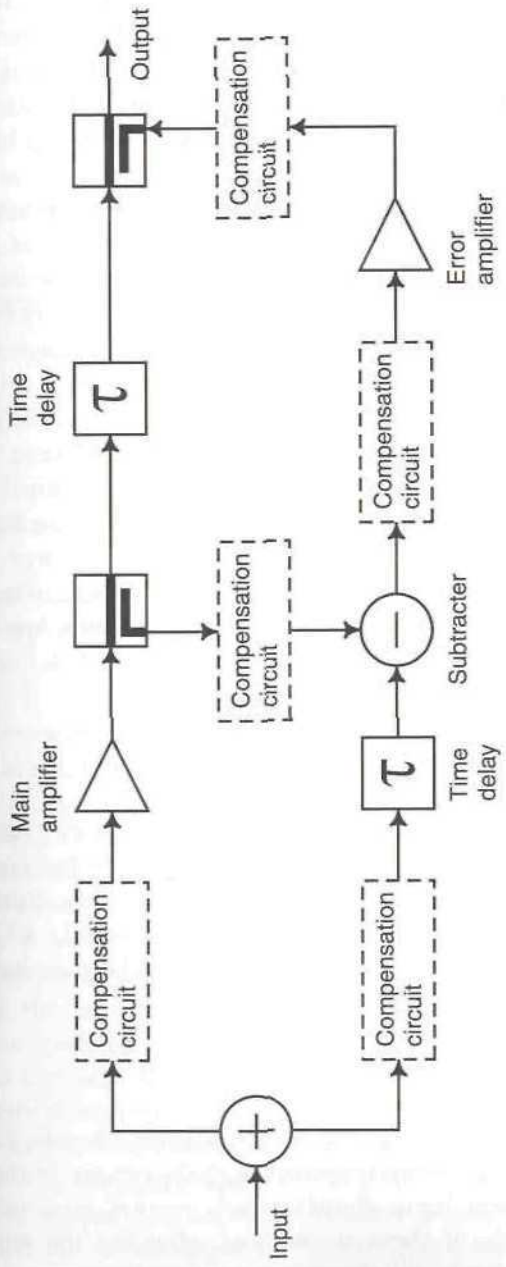


Figure 5.23 Summary of the available locations for the gain and phase adjustment components in the error and compensation loops.

However, it is possible to make certain assumptions (usually very pessimistic) about the degree of aging and temperature drift mismatch which will occur in the system over its designed lifetime. These assumptions will yield a level of performance below which the single loop system is unlikely to go during its lifetime; this may be as low as 6 dB of distortion rejection. It is then simply a matter of providing a sufficient number of additional loops in order to satisfy the specification, given that each is likely to have a similar performance.

As an example, consider a system in which the main amplifier has a linearity such that the intermodulation distortion products are 25 dB down on the wanted signals and where the desired system specification is for these products to be at least 50 dB down. If each loop in a multi-loop system can be considered to improve the distortion by 6 dB, then a minimum of five loops in total would be sufficient to easily meet the specification.

This technique is often applied to systems where extremely high levels of reliability are required over very long periods of time, such as in submarine cable repeaters. The additional expense is justified in terms of the greater reliability and guaranteed performance. Feedback systems, for example, could fail catastrophically if any one of a large number of components was to fail; with feedforward, only a single loop would fail, hence resulting in a degradation in IMD performance but not in, for example, instability and hence complete system failure.

5.12 System Examples

A large number of patents have been filed over the last 20 years relating to correction schemes designed to maintain the amplitude and phase balances discussed above over time and temperature. Some of the principles upon which these are based will be discussed in this section. In addition, some 'generic' analysis of control system performance has been undertaken and is detailed in the literature [20].

5.12.1 Adaptive Nulling

Many of the automatic correction systems rely upon adaptive nulling of the unwanted signals at the various stages through the system. In the error loop, it is the initial system input signals which require removal and hence monitoring the levels of these signals and adjusting the amplitude and phase controls accordingly will ensure optimum performance of this section.

The function of the correction loop is to cancel the distortion products

from the final output of the system and hence it is the distortion component of the final output signal which must be monitored and minimised in this case.

In either case, the principles involved are quite similar and can be examined together. For simplicity, it is sufficient to consider a two-tone test signal applied to the input of the system and to look at the cancellation of these tones in the error signal; this situation is illustrated in Figure 5.24. It will also be assumed that out-of-band products may be ignored as these are generally removed by filtering and thus only intermodulation products need be considered.

5.12.1.1 Energy Minimisation

The most straightforward method of assessing the level of the tones within the error signal is simply to detect the overall energy of the complete signal. In general, the level of the distortion products is small relative to the wanted signals, and thus the wanted signal energy will dominate. Thus a system can be envisaged in which the overall energy of the error signal is minimised by automatic adjustment of the gain and phase components using voltage control.

The use of energy minimisation in both parts of a feedforward correction loop is illustrated in Figure 5.25. The detector could be any form of broadband energy detector; a simple example being an envelope detector.

This somewhat crude approach may well be all that is required for the error signal in some systems, as large amounts of rejection of the remaining input signal energy below the level of the intermodulation products is not usually required. Once this level is low enough such that the error amplifier power is dominated by the intermodulation products then this is generally acceptable. If the error amplifier specification can be relaxed a little, such that it can be more powerful than is strictly necessary, then the input signal rejection specification can be relaxed still further. The ultimate consideration now becomes the cancellation of the wanted signal energy in the final output, which may be predicted using the methods outlined in Section 5.5. It may be acceptable to sacrifice a few tenths of a dB of output power to the cause of reduced system complexity.

In the case of the correction loop, two detectors are required. An output detector is required to ascertain the level of the signal-plus-distortion present at the output and an input detector is necessary to provide an accurate indication of the input power level, and more importantly, any changes in that level (e.g., due to power control variations). Whilst it would be possible to utilise an output detector alone, the extremely long

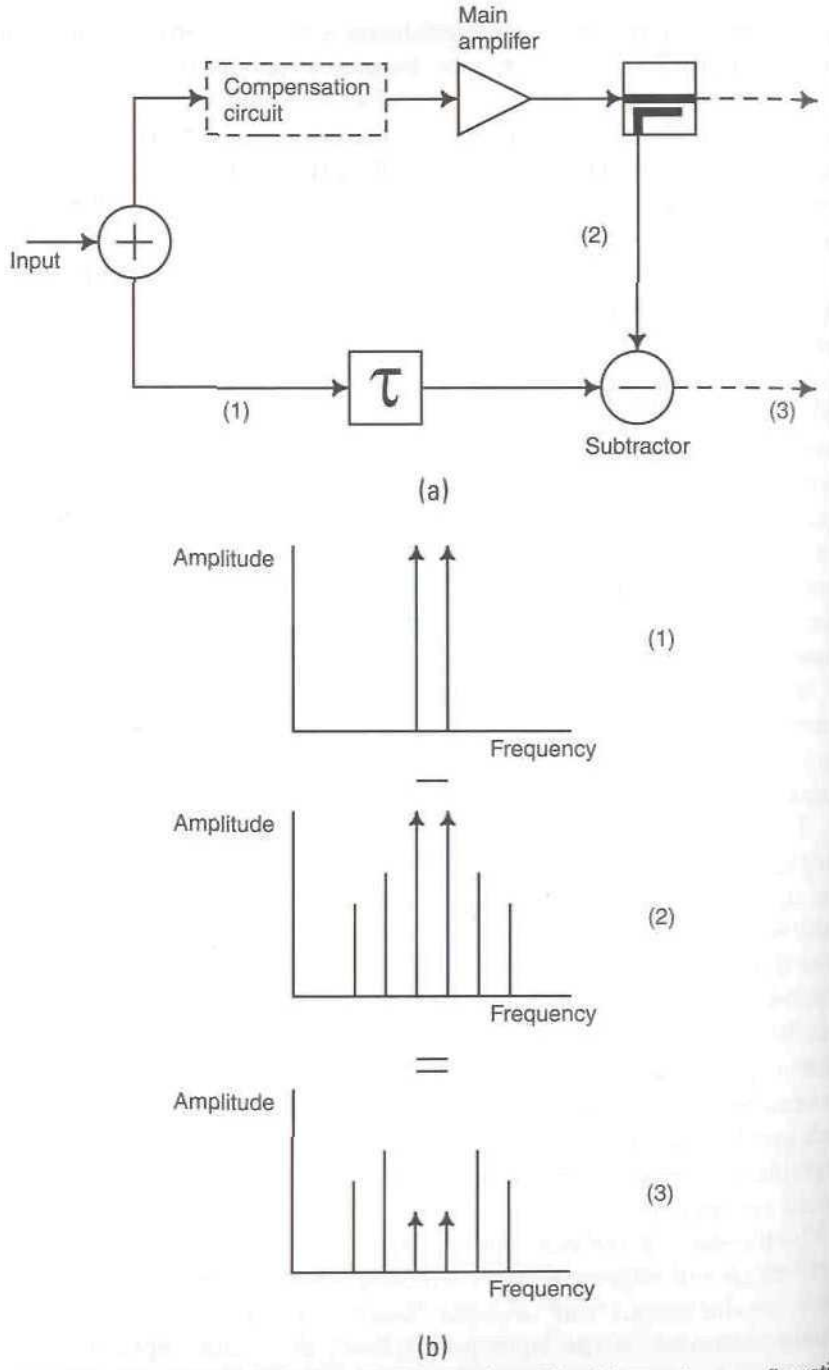


Figure 5.24 Two-tone test applied to a feedforward amplifier: (a) error-loop configuration and (b) cancellation of the input tones.

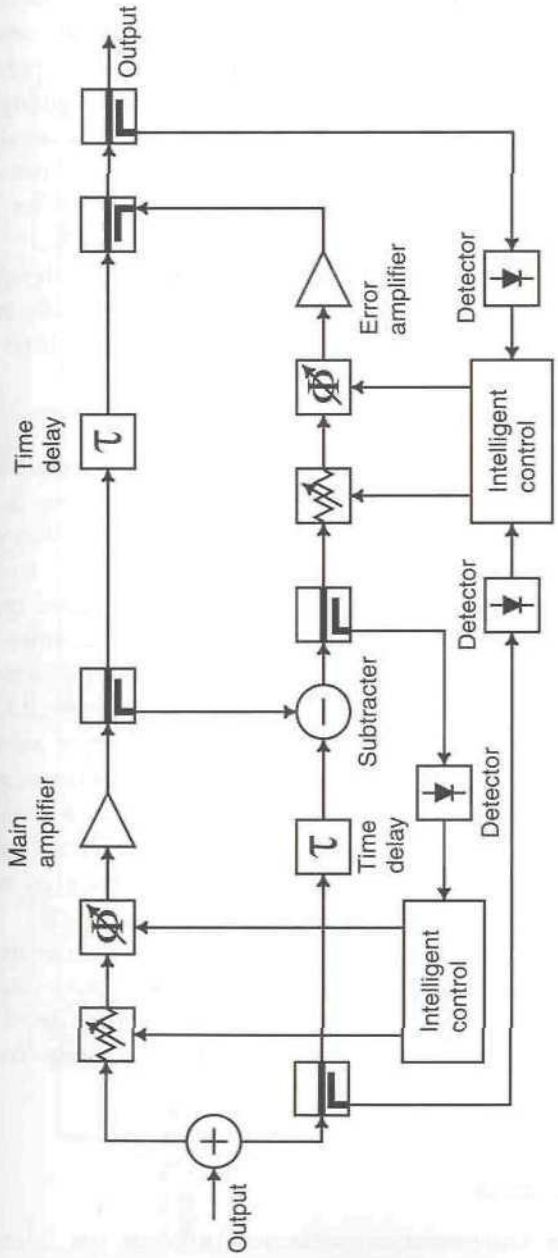


Figure 5.25 Compensation of a feedforward amplifier using energy minimisation techniques.

time-constant required in that detector (due to the extremely small changes in output level resulting from distortion level changes) mean that it could easily be fooled by even small changes in input (and hence output) level.

Utilising this technique for the correction loop is usually unsatisfactory for the reason outlined above. The wanted signal energy present at the output will be very large (hopefully) with respect to the remaining distortion and any changes in this distortion level will thus have an almost negligible effect on the output signal energy. As a result the problem of detecting these small changes in energy becomes almost impossible and an alternative solution is required.

One possibility is that of generating a second 'error signal' from the output signal. The main signal energy should be substantially cancelled in this case and hence energy detection could yield a more realistic result.

5.12.1.2 Coherent Detection

The obvious solution to the problems inherent in broadband energy detection systems is to employ some form of coherent detection or correlation process instead. Thus in the case of the error loop, the error signal could be correlated with the original input signals to generate a suitable feedback error signal to control the gain and phase components. This is the basic approach which many of the patented schemes employ.

The basic configuration of this approach is shown in Figure 5.26 and is similar to that published in a patent by the Marconi Company Ltd [21]. The correction loop correlation process utilises the final output signal and the error signal and this process assumes that the rejection of the original signal components from the error signal is very good. It is necessary to ensure good rejection of the original signal components in order to subsequently ensure that the result of the correlation process is influenced only by the distortion components and not by the wanted signals.

A number of other patented systems work on a similar principle but utilise different positions for their gain and phase adjusting circuits, along with techniques for reducing the tight constraints on the level of wanted signal components present in the error signal. These include: Bauman [22], Olver [23], King [24] and Kenington *et al.* [25].

5.12.2 Carrier Injection

The second main compensation technique involves the injection of an additional carrier (or carriers) into the error loop immediately prior to the main amplifier (although anywhere in that path will produce similar results).

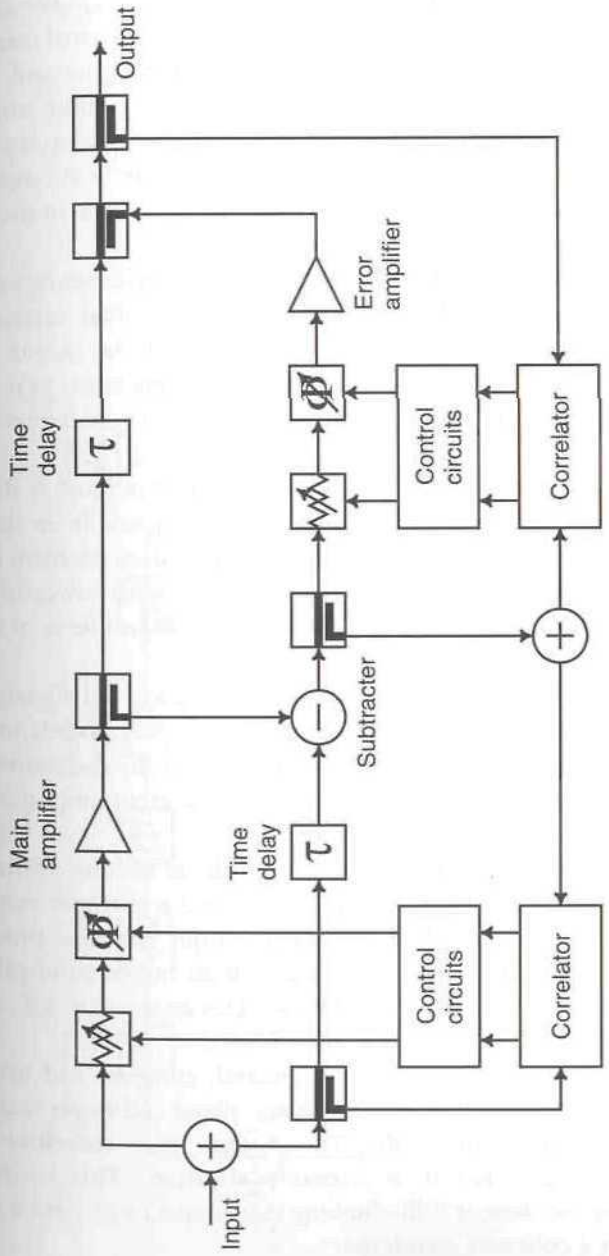


Figure 5.26 Compensation of a feedforward amplifier using correlation techniques.

The level of this signal is usually significantly less than the level of the main input signal(s); 10 dB or 20 dB being typical.

The injected carrier will be amplified by the main amplifier and will pass to the output coupler via the time-delay element in the usual manner. A sample of the carrier will also be present in the error signal and will be amplified by the error amplifier before being fed to the output coupler in anti-phase to its main-path counterpart. If the compensation loop is adjusted correctly, then this injected signal will be cancelled at the final output, and hence the compensation circuitry must aim to minimise its level in the output signal.

This may be achieved by the use of a narrow-band receiver tuned to the injected carrier frequency and hence receiving the residual carrier level. Some form of intelligent controller can then monitor the output of the narrow-band receiver and adjust the gain and phase components to minimise the residual carrier level. The configuration of this control strategy is shown in Figure 5.27.

There are a number of disadvantages with this technique. It does not help in any way with the elimination of the wanted signals from the error signal. This must still be done manually with the attendant problem of drift over time. The error amplifier must therefore be made more powerful than is strictly necessary for the required performance (or a different form of control must be used for this loop).

The injected carrier will intermodulate with the wanted signals in the main amplifier and create additional (unnecessary) intermodulation products. These should be eliminated from the final output by the action of the feedforward loop, but must still be amplified by the error amplifier, again adding to its power rating (usually negligible).

Incomplete removal of the pilot is also a problem in some systems and an out-of-band pilot plus additional output filtering is sometimes employed. This is clearly an expensive and inefficient technique and also potentially erroneous as it relies on the correlation between an out-of-band pilot and amplifier performance within the wanted band. This correlation will, at best, be limited.

The narrow-band receiver is not, in general, coherent and hence the controller must alternately adjust amplitude and phase and assess their effect on the output signal continuously. The system must therefore 'hunt' continuously to ensure that it is correctly adjusted. This process, in common with any incoherent 'hill-climbing' technique, will yield a poorer performance than a coherent counterpart.

An example of this type of pilot aided scheme has been patented by Myer [26]. The intelligent controller in this case is provided by a decreasing

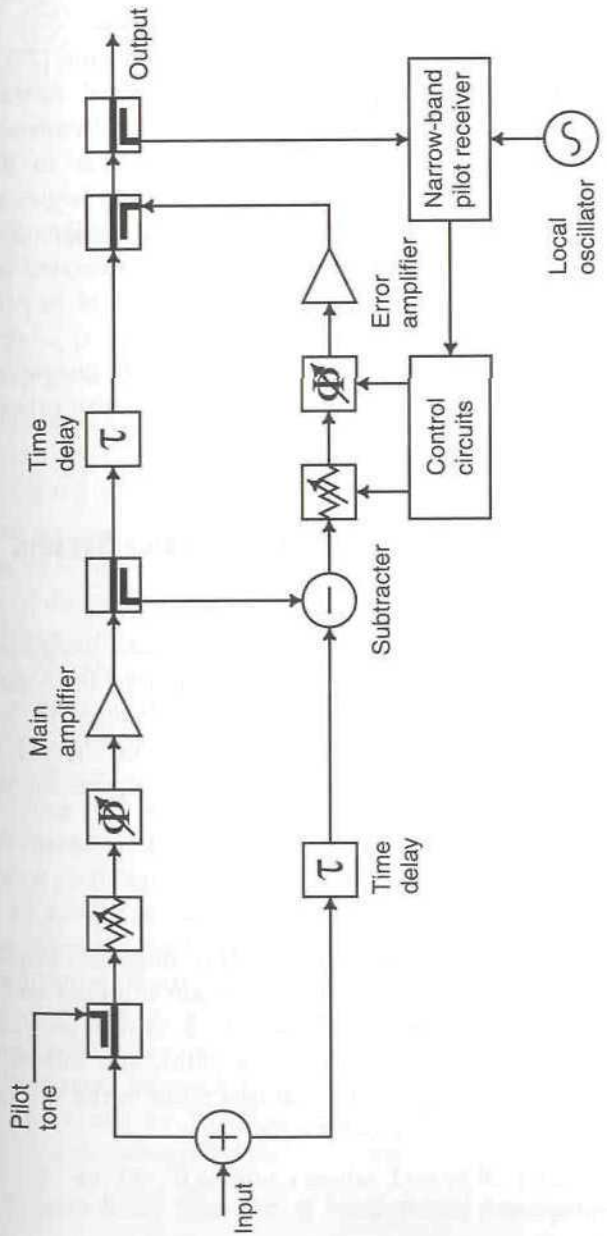


Figure 5.27 Compensation of a feedforward amplifier using pilot-injection techniques.

step-size algorithm, although a microprocessor or digital signal processor could be used to provide a similar function.

It is possible to employ a coherent detection system, based on the injected carrier, and this provides improved performance.

Myer has also patented a modification to this scheme [27] in which adjustment of the error loop components is also attempted. In this case the guarantee of at least one input carrier within a prescribed frequency range is required in order to permit the receiver and controller to detect and minimise its presence in the error signal. This obviously begins to restrict the generality of the feedforward system and hence is sacrificing one of its major advantages, although if this restriction can be tolerated in a given application (mobile radio base stations in the case of the Myer patent) then the technique can be made to work adequately.

There are many variations on the pilot control technique, including frequency hopping or spreading of the pilot, together with other forms of pilot modulation.

5.13 Summary of Requirements for the Major System Components

The earlier parts of this chapter have dealt with the basic design parameters for a feedforward system, together with the design and operation of automatic compensation networks. This section attempts to look more practically at the desirable characteristics for each of the major system components, how they might be achieved and what figures are obtainable in practice.

5.13.1 Input Splitter

The purpose of the input splitter is to divide the feedforward amplifier input signal in an appropriate ratio to feed both the main amplifier and reference paths. This splitter need not necessarily be a 3 dB hybrid, providing equal signal levels to both the main and reference paths, and indeed there are advantages in an unequal coupling ratio at this point in the system.

There are three options for this coupler:

1. Equal ratio 3 dB hybrid, which could be 0°, 90° or 180° (or a 6 dB resistive splitter; see Figure 5.28).
2. Directional coupler weighted in favour of the main amplifier (i.e., weighted to minimise the gain required of the main amplifier). Thus, the coupled port would form the reference path, as shown in Figure 5.29(a).
3. Directional coupler weighted in favour of the reference path; in this case, the coupled port would form the input to the main amplifier (Figure 5.29(b)).

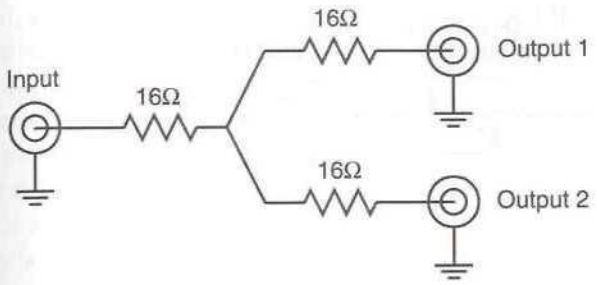


Figure 5.28 Simple resistive input splitter for a 50Ω system.

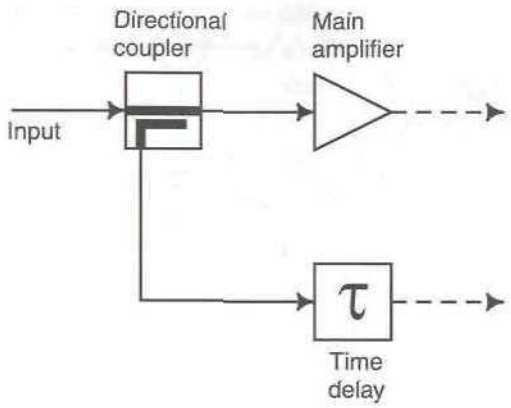
These various options all have their place, and it is a system design decision as to which is chosen. The two primary advantages to option 1 are simplicity of design and low cost (particularly if the 6 dB resistive splitter is chosen). Hybrid combiners for various broad frequency ranges are cheap and easily available and can thus be employed where cost is a strong factor, but with some concern for loss (and hence noise performance). If low loss and amplifier noise figure are not of concern, then a 6 dB resistive splitter is an obvious choice.

The latter two options can be distinguished by the trade-off between noise performance and main power amplifier complexity. Option 2 provides for the lowest gain requirement in the main amplifier, and hence provides the potential for a reduced number of stages and a more stable design (the main amplifier is often required to have a very high gain, for example, 1mW input for 1kW output = 60 dB; stability can therefore be a major concern).

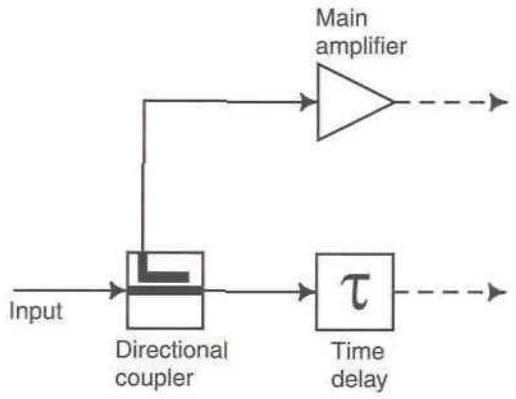
Option 3 allows the optimum noise figure to be realised in a feedforward system (see below) and hence is often chosen for that reason. It does, however, significantly increase the gain required of the main amplifier, frequently by 10 dB or more over either of the other options, and hence may not always be the most appropriate choice.

5.13.1.1 System Noise Figure for a Feedforward Amplifier

The main power amplifier in a feedforward system will often have a relatively high noise figure, due to its high power output requirement and



(a)



(b)

Figure 5.29 Alternative configurations for the input coupler; (a) optimised for minimum *main amplifier gain and (b) optimised for minimum noise figure of the complete* feedforward amplifier.

large gain. This is particularly true if an 'off-the-shelf' amplifier has been chosen for the main amplifier, as this is unlikely to have been conceived with low-noise as a primary design aim. The noise added by the main amplifier can be thought of as an additional signal, which is not also present on the reference path, and hence will appear as part of the error signal. It will therefore be corrected as part of the natural operation of the feedforward

process and, *assuming a perfect gain or phase balance for the overall system*, will be eliminated in the output of the complete feedforward amplifier. The feedforward process therefore not only eliminates distortion added by the main power amplifier, but also noise and indeed any other spurious signals present at the output of the main amplifier which are not also present in the reference path.

This is a very powerful and useful benefit of the feedforward system, as it allows relatively low-noise amplifiers to be constructed with extremely high third-order intercept points, hence resulting in a very high dynamic range system.

The noise figure of a feedforward amplifier is therefore determined by the elements of the system which are not included in the correction process, in a similar manner to the earlier arguments regarding distortion addition by the system. In other words, noise added in the reference path, or by the error amplifier and associated components, is not corrected for by the feedforward process and will be added to the output signal at the level it appears at the output of the error amplifier, less the coupling factor of the output coupler.

This may be summarised with reference to Figure 5.30 as follows: the total loss up to the error amplifier input is:

$$L_T = L_{C1} + L_{TD} + L_S + L_{CC} \quad \text{(dB)} \tag{5.100}$$

Since the components introducing this total loss may be assumed to be matched to the characteristic impedance of the system (50Ω), it can be shown that the resulting noise power is kTB (watts), where k is Boltzmann's constant (1.38×10^{-23} J/K), T is the system temperature in Kelvin and B is the bandwidth of interest (Hz) [28]. The system input noise is therefore:

$$N_{in} = kTB \quad \text{(W)} \tag{5.101}$$

The noise power at the output of the complete feedforward system may therefore be derived:

$$\begin{aligned} N_{out} &= G_{err}F_{err}N_{in} \\ &= kTB.10^{F_{A2}/10}.10^{(G_{A2}-C_{C2})/10} \quad \text{(W)} \\ &= kTB.10^{(F_{A2}+G_{A2}-C_{C2})/10} \end{aligned} \tag{5.102}$$

where: F_{err} is the error amplifier noise *factor*[1], and G_{err} is the gain of the error amplifier as seen at the output of the feedforward amplifier (i.e., incorporating the output coupler 'loss').

1. N.B. Noise-Figure = 10 $\log_{10}$(Noise Factor).

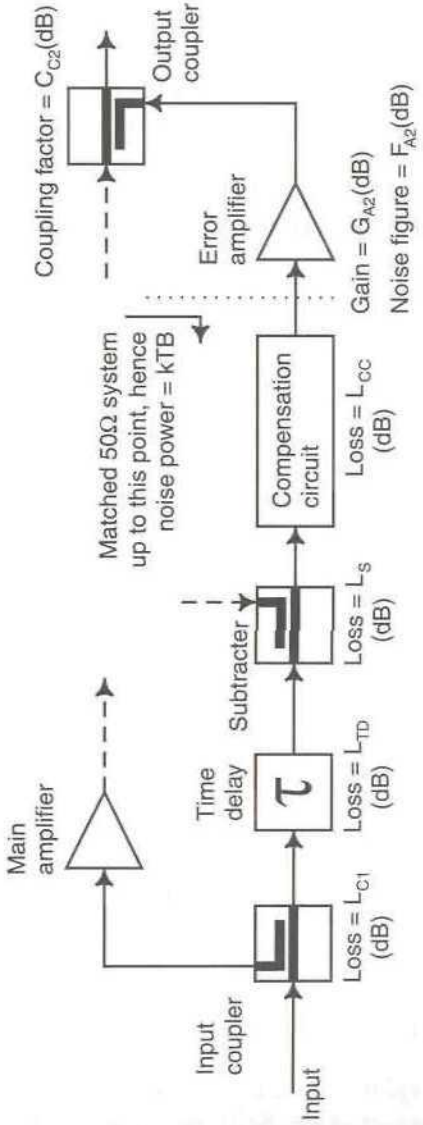


Figure 5.30 Noise figure of a feedforward system.

The system noise *factor* is therefore:

$$F = \frac{N_{out}}{G_T N_{in}} \tag{5.103}$$

where G_T is the total gain of the reference and error paths in the feedforward system and is given by:

$$G_T = 10^{(-L_T + G_{A2} - C_{C2})/10} \tag{5.104}$$

Hence, the system noise factor is:

$$F = \frac{kTB\left(10^{(F_{A2}+G_{A2}-C_{C2})/10}\right)}{10^{(-L_T+G_{A2}-C_{C2})/10}.kTB} \tag{5.105}$$

Simplifying gives:

$$F = 10^{(L_T+F_{A2})/10} \tag{5.106}$$

Hence the system noise figure is given by:

$$F_{dB} = 10\log_{10}(F) = L_T + F_{A2} \quad \text{(dB)} \tag{5.107}$$

The noise figure of the feedforward amplifier, in the case where perfect nulling of the main amplifier distortion and spurious signals is assumed, is therefore determined purely by the losses in the reference path and the noise figure of the error amplifier.

Note that the input splitter and the subtracter in Figure 5.30 are both shown as directional couplers, configured to provide minimum loss to the reference signal. This is the optimum configuration for minimum noise figure, although it does require a higher main amplifier gain than would a system based on 3 dB hybrid splitters.

5.13.2 Main Amplifier

The primary purpose of the main amplifier is to raise the signal level at its input to a level beyond that which is ultimately required at the system output. This additional power output requirement arises because of the inevitable losses in the main-path time delay element and also in the output

coupler and any additional sampling couplers for automatic gain or phase compensation purposes (see Section 5.11). For low-power portable equipment, the main amplifier may only require an output power of a few watts, but in many applications (e.g., cellular radio base stations) the peak output power may be in the kilowatt region. Regardless of power level, however, there are a number of essential characteristics which the main amplifier must possess in order to be suitable for inclusion in a feedforward system.

5.13.2.1 Flat Gain Response

It is important that the gain characteristic of the main amplifier in a feedforward system should deviate as little as possible from an ideal flat response with respect to frequency. This is essential in order to maximise the cancellation of the original (input) signal energy in the error signal, prior to its amplification by the error amplifier. The degree of cancellation which may be achieved at this point is governed by the peak-to-peak difference between the main amplifier gain response and the error-loop time-delay element response at each point across the band of interest. The error-loop time-delay element should have a relatively flat frequency response, and hence it is the main amplifier's frequency response which will be the primary determining factor of the 'quality' of the error signal.

The degree of cancellation to be expected at a given frequency may be predicted from the main amplifier's frequency response using Figure 5.4. Thus, for example, a main amplifier with an ideal phase response and an amplitude ripple of ±0.5 dB, would achieve a cancellation of the main signal energy in the error signal of just under 25 dB. A value of between 30 dB and 40 dB would be more desirable in a system employing a class-A main amplifier, with third-order IMD products around 30 dB down on the tones in a two-tone test. This would ensure that the main signal energy appeared at the same level as, or below, the IMD products in the error signal. This would suggest a gain response for the main amplifier which was flat to within ±0.2 dB or better. This may sound like a harsh requirement, but is not out of the question for the relatively small percentage bandwidths required in many cellular systems (e.g., 25 MHz at 900 MHz is only 2.8%).

If this level of gain response flatness cannot be achieved, then the rating of the error amplifier must be increased above that required to amplify purely the distortion present at the output of the main amplifier. This is undesirable because it will almost inevitably result in a poorer quality of error amplifier and hence a poorer overall IMD performance for the system. A second disadvantage is that the amplified residual main signal energy will cause partial cancellation of the wanted output signals to occur in the output coupler, hence lowering the power output of the system and hence its power

efficiency. For this to be a significant effect, however, the main amplifier gain response would have to be very poor (say ±1 to ±2 dB ripple) and hence could be improved as outlined below.

For wider bandwidth systems, where the gain response is unlikely to meet this requirement, a number of other techniques may be applied to alleviate the problems caused in the error signal. For example, series, or series and parallel *LCR* circuits (or microstrip equivalents) may be employed as gain flatteners prior to the main amplifier. Such techniques are particularly valuable for multi-octave systems, such as those employed at HF, and such frequencies lend themselves relatively easily to gain flattening techniques.

Alternatively, a higher power error amplifier may be employed, such that it is able to handle the increased signal level due to the poor suppression of the main signal energy. This amplifier need not be of particularly high quality, in terms of its IMD performance, as a second feedforward process may be subsequently employed to tidy up the remaining distortion. The advantage of such a 'two-loop' arrangement is that the first feedforward process will remove virtually all of the *linear distortion* (i.e., gain and phase nonlinearities) present in the main amplifier and leave an almost perfect 'main amplifier' for the second feedforward process to operate on, to eliminate the remaining intermodulation distortion.

The actual level of gain required from the main amplifier will be a little higher than that required of the overall feedforward system. This is necessary to take account of the losses in the input splitter (or coupler), any gain and phase compensation prior to the main amplifier, the output sampling coupler, the main-path time delay and the error signal injection coupler (plus any monitoring couplers thereafter). Thus, for example, the input splitter will have a loss of 3 dB, the gain and phase compensation will add a further 6 dB (say), the output sampling coupler will have a through-path loss (including connectors) of around 0.5 dB, the main-path time delay may add as much as 3 dB (although this may be reduced as will be outlined below) and the error injection coupler will add another 1 dB. The total additional gain required of the main amplifier will therefore be around 13.5 dB.

5.13.2.2 Linear Phase Response

For similar reasons to those outlined above, a linear phase response (or flat time-delay) with respect to frequency is required from the main amplifier. The relationship between time delay and phase-shift is given by:

$$\text{Time delay} = \frac{\text{phase shift}}{360 \times \text{frequency}} \tag{5.108}$$

where time delay is expressed in seconds, phase-shift in degrees and frequency in Hz.

Any deviation from this ideal characteristic will result in a reduction in suppression of the main signal energy in the error signal path; the degree of suppression may again be predicted from Figure 5.4.

It is also important to keep the amount of delay introduced by the main amplifier to as low a value as possible, as this will determine the value of the matching time delay element in the reference path. The larger the value of time delay required in the reference path, the physically larger the delay element must be and, often more importantly, the greater the loss this element will introduce. It is clear from the above derivation of the noise figure for a feedforward system that it is important to minimise the loss in the reference path, and hence in its time-delay element. Therefore lowering the delay of the main amplifier will indirectly reduce the overall system noise figure.

For this reason, it may often be desirable to place some of the system gain outside of the feedforward loop (i.e., at the input to the feedforward system) when this will not unduly degrade the overall system IMD response. Many low-power (and low-noise) amplifiers possess extremely good IMD characteristics, particularly when backed-off by a few dB, hence it is unnecessary to include such low-power stages as a part of the feedforward system. Removing them will decrease the main amplifier time delay, and should decrease the overall system noise figure, assuming that the external amplifiers have a suitably low-noise figure.

5.13.2.3 Good Distortion Performance

The IMD performance of the main amplifier is usually the primary determining factor governing the power required of the error amplifier and hence the likely IMD performance of that sub-system (assuming that the main amplifier's gain and phase characteristics are reasonably well behaved). The range of IMD performance levels likely to be encountered in practice for a high-power (>50W) solid-state PA is between −20 dBc (for a poor-quality class B amplifier) and −35 dBc for a good quality class-A design. The power required of the error amplifier at the two extremes of likely performance may be estimated using (5.40) given previously (and reproduced below):

$$P_{A2} = \frac{P_{A1} F_{IM} L_{DC}}{C_{DC}} \tag{5.109}$$

Substituting for the coupler through-path loss (from (5.47)):

$$P_{A2} \approx \frac{P_{A1} F_{IM} (1 - C_{DC})}{C_{DC}} \tag{5.110}$$

For a 1 kW main amplifier power ($P_{A1} = 1000$), and a 10 dB output coupling factor ($C_{DC} = 0.1$), the power rating of the error amplifier for the poor-quality class-B main amplifier would be in excess of 90W. In the case of the high-quality class-A design this reduces to around 2.8W. Hence it is obvious that careful design of the main amplifier will directly result in both significantly better system linearity performance (partly due to the good 'starting point' of a good main amplifier and partly due to the low-power and hence high-quality error amplifier which may therefore be employed) and significantly better power efficiency (due to the lower power requirement of the error amplifier).

It should be noted that the above power levels derived for the error amplifier are only a rough guide and are distinctly optimistic. This arises for two reasons: first, the derivation assumes that only one significant intermodualtion product exists, whereas, in practice, there will normally be at least two of significance (this is certainly the case for a two-tone test, but is less stringent for larger numbers of tones). Secondly, in a practical system, it is likely that multiple carriers will be utilising the amplifier simultaneously and hence the intermodulation distortion characteristic will be far from that of the simple case assumed above. It will more likely consist of many tens of 'tones' at various power levels which will result in a PEP requirement for the error amplifier somewhat above that alluded to in the derivation. Practical experience suggests that the worst-case is frequently encountered with a 'three-tone test', although this will depend to some extent on the characteristics of the particular main amplifier under consideration.

Harmonic distortion is less of a concern in most systems, since although the feedforward process will endeavour to eliminate it from the output spectrum (given a sufficiently wide frequency response for the various system components), it is often more easily removed by either filtering the output of the main amplifier or the output of the overall feedforward system. The latter is preferable from the point of view of preventing signals from co-located transmitters from entering the feedforward system by coupling into the transmit antenna. The former is preferable as a means of restricting the level of the signal entering the error amplifier, since it will then contain only IMD information.

It may therefore be desirable in certain situations to employ filtering in both locations.

5.13.3 Sampling Coupler and Subtracter

5.13.3.1 Sampling Coupler

The purpose of the sampling coupler following the main amplifier is to extract a small sample of the distorted output signal emanating from that amplifier so that it may be used in deriving the error signal. Since the gain of the main amplifier is usually large, and hence the difference between the level of the signal present on the reference path and that present at the output of the main amplifier is also large, the coupling factor of this coupler need only be small.

It is indeed highly desirable that this coupling factor be maximised (within reason) as this will minimise the loss introduced in the 'through'-path of the coupler, since (from (5.47)):

$$L_{DC} = 1 - C_{DC} \tag{5.111}$$

and hence will maximise the power delivered to the load from the overall feedforward amplifier. This in turn will help maximise the power efficiency of the system.

Thus, for example, if the coupling factor is 30 dB ($C_{DC} = 0.001$), then the through-path loss, L_{DC}, is 0.0043 dB. This figure, of course, neglects losses due to the connectors on the coupler, together with resistive and other losses in the through-path conductor and dielectric. It does, however, indicate that it is relatively straightforward to ensure that this component does not significantly contribute to the main-path power loss.

It is desirable that all of the couplers within the feedforward system have adequate 'directivity' to ensure that reflected signals (particularly those from the antenna or load) cannot corrupt the error signal or cause significant distortion in the output of the error amplifier (in the case of the output coupler). In addition, this directivity can have an effect on system stability (see Section 5.15) and hence must be adequate for this purpose also. For a coupler where the forward-path signal is being sampled (and hence signals reflected from the output of the coupler or beyond are undesirable), the directivity is defined as:

$$\text{Directivity} = 10\log_{10}\left(\frac{\text{Power at coupled port (signal at coupler input)}}{\text{Power at coupled port (signal at coupler output)}}\right) \quad \text{(dB)} \tag{5.112}$$

where the two powers measured at the sample port are expressed in watts. All ports are assumed to be terminated in reflectionless loads. Typical

directivity values for HF, VHF, and UHF systems are in the range 20 dB to 30 dB, with higher values being quite practical, but at a significantly increased cost.

In addition to a high directivity, the sampling coupler must have a flat frequency response across the band of interest, together with a linear phase response. Both of these characteristics are relatively easy to achieve with modern coupler designs, although they do become more difficult at very high power levels, low coupling factors (e.g., 10 dB), and for multi-octave designs.

Note also that the effect of the (imperfect) terminating impedances on the directivity of the coupler must be taken into account, as these will degrade the measured directivity figures for the coupler.

5.13.3.2 Subtracter

The role of the subtracter in a feedforward system is in forming the error signal as the difference between the sample of the main amplifier output signal (derived from the sampling coupler above) and the reference signal. The directivity of this coupler is less important than that of the sampling coupler above, assuming that the input match to the error amplifier is reasonable, but it may still require a low through-path loss (and hence a large coupling factor) if a low system noise-figure is required. A coupling factor of 20 dB or more will ensure that the through-path loss is small, and hence help to maintain a high noise-figure for the complete feedforward amplifier, whilst still allowing approximately the correct signal levels to be present for adequate subtraction in a high-gain (>50 dB) system.

Again, the frequency and phase responses need to be good in order to ensure a high degree of suppression of the main signal energy in the error signal. In both cases the quality of this component must be better than that of the match between the main amplifier response and the reference path characteristic (including any frequency compensation present in the reference path). This will ensure that any effort expended in producing a good match between the two paths, particularly in the design of the main amplifier, is not wasted.

5.13.4 Time-Delay Elements

Two time-delay elements are required in an uncompensated single-correction feedforward system: one in the reference path, which must match the average delay of the main amplifier, and one in the main signal path, after the sampling coupler, to match the average delay of the error amplifier and associated gain and phase adjustment components.

These time-delay elements are typically formed by lengths of transmission line and the precise nature of this line will vary depending upon the application, in particular the desired power level and the required operating frequency. Thus, low-power systems may utilise etched microstrip techniques appearing on the same board as both amplifiers and indeed the same board as the relevant (etched) couplers. At higher powers, the main-path delay line is generally fabricated from coaxial cable (of a suitable peak and mean power rating) or is a length of waveguide (at microwave frequencies). The loss of this delay line is also a critical factor in determining the efficiency of the overall feedforward system and is of particular concern at UHF and above (where transmission line losses are typically higher).

The power rating of the transmission line must be considered from two perspectives: peak power rating and mean power rating. The peak power rating is generally determined by the insulating properties of the dielectric material between the inner and outer conductors in the cable. The integrity of this insulation is also dependant upon the environment in which the cable is to be employed; particularly hot or humid conditions may de-rate the insulating properties of the dielectric and must be considered when specifying the cable.

The mean power rating of a cable is determined by its losses and their consequent heating effect upon the cable itself. Thus a cable with a good mean power rating will generally possess a low loss per unit length, and hence a low resistive loss, and will often be of a considerable diameter. A long delay-line made from such cable will therefore consume a large amount of space in the final amplifier. The dielectric material must also be low-loss and hence will usually be of a higher quality than that of standard coax. The resulting cable will therefore be expensive, although this is likely to be relatively insignificant when compared to, say, the cost of the main amplifier.

Where space is at a premium, the main-path delay element may be fabricated from a cable or etched transmission line based on a high dielectric constant substrate (or cable dielectric). It may be reduced in length (and hence value) or even eliminated altogether in some applications, as outlined earlier in this chapter. The use of these delay reduction techniques will inevitably compromise the broadband performance of the feedforward system, however, there are many applications in which this is not an overriding concern.

Recently, filter-based delay lines have become popular due to their low insertion loss characteristics and relatively small size. Achieving appropriate gain and phase flatness is usually the main challenge with this form of solution.

5.13.5 Error Amplifier

The error amplifier effectively provides the reference level of performance for the complete feedforward system; the basic purpose of the feedforward technique being to transfer the 'performance burden' in terms of linearity, flatness of frequency response and noise-figure from the main amplifier to the error amplifier. Since the error amplifier should, in virtually all cases, be of a lower power than the main amplifier, this is a reasonable expectation.

The distortion level resulting from the error amplifier is critical in determining the overall system performance, and in many dynamically-controlled systems, is the limiting factor in determining the overall distortion performance of the feedforward amplifier. Any distortion produced by the error amplifier is injected, uncompensated, into the output of the feedforward system, by the output coupler. The ultimate distortion floor, assuming perfect matching and hence cancellation of the main amplifier distortion, is therefore the level of the error amplifier distortion products less the output coupling factor. Similarly, any noise added to the error signal by the error amplifier will directly add to the output noise of the feedforward system (again, less the coupling factor of the output coupler). It is therefore important to minimise the noise figure of this amplifier.

The characteristics of the intermodualtion distortion added by the error amplifier are such that the bandwidth of the intermodualtion distortion from the feedforward system is extended beyond that which is initially produced by the main amplifier. This situation is illustrated in Figure 5.31. The error signal contains mainly the distortion information from the main

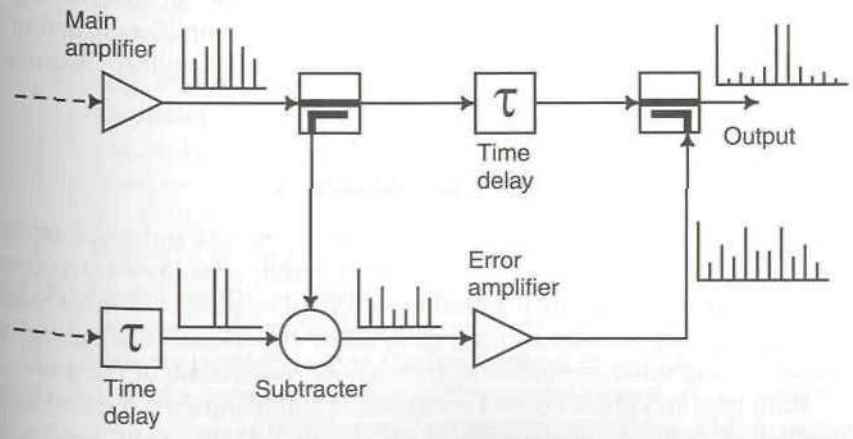


Figure 5.31 Distortion resulting from the error amplifier.

amplifier, and this in turn will be dominated (in a typical class-A or -AB amplifier) by the third-order products. Since the spacing of the third-order products will be three times that of the original tones which produced them, the bandwidth of the subsequent error amplifier intermodulation distortion will, to a first approximation, be trebled.

The power requirement of the error amplifier is determined by the main amplifier distortion level, the level of cancellation of the main signal energy in the error signal and the output coupling factor. Assuming that the main amplifier is class-A, and that the main signal cancellation is good, then this amplifier will be typically 10 dB to 15 dB less powerful than the main amplifier for a 10 dB output coupling factor.

In a high-power system, the gain of the error amplifier will typically be very high. This arises from the high gain of the main amplifier (and consequent low level of the reference signals relative to the main amplifier output signals) and the output coupling factor. For a main amplifier with a gain of 40 dB (10mW in, 100W out) and an output coupling factor of 10 dB, the error amplifier gain will need to be between 55 dB and 60 dB (to compensate for the losses in the subtracter and the gain and phase adjustment components). This gain must be achieved with as short a delay as possible, since this will directly affect the cancellation loop time delay and its associated physical size and loss. This in turn will have a marked effect on the overall efficiency which can be achieved by the complete feedforward amplifier.

Finally, a flat frequency and linear phase vs. frequency response are required in order to ensure a high degree of cancellation across the band of interest. This specification should be commensurate with the distortion specification of the error amplifier, as there is little point in achieving a perfect frequency or phase response such that the main amplifier distortion is perfectly removed, leaving a significant level of error amplifier distortion added into the output.

5.13.6 Gain and Phase Compensation Circuits

Numerous circuits exist for the control of the amplitude and phase (either separately or together) of the various signals within a feedforward system. Such elements are key in any practical feedforward amplifier either for initial setting-up (e.g., upon manufacture) or to allow continuous 'fine-tuning' to constantly re-optimise performance throughout the lifetime of the unit.

Both gain and phase control components (and combined circuits) have a number of desirable characteristics for use in a feedforward system, in addition to their individual requirements.

1. *Low distortion. This is particularly necessary if either or both* components is employed in the reference path, but is still necessary to some extent wherever they are placed. The use of PIN or varactor diodes to provide voltage control introduces one or more nonlinear elements into the system and care must be taken to ensure that their use does not degrade the overall linearity performance of the complete amplifier.

   The use of diodes with a long carrier lifetime with respect to the signal frequencies being processed, together with appropriate drive levels, should ensure that a very low level of distortion can be achieved. Levels of distortion commensurate with the use of these components in the reference path should result from employing such techniques. Further details regarding the use of GaAs MESFETS and silicon p-i-n diodes in attenuator applications, together with expressions for second- and third-order intercept points may be found in the literature [29,30].
2. Flat gain response with frequency. Again, these components should not degrade the cancellation performance of which the system would otherwise be capable, due to amplitude ripple. It should be relatively straightforward to ensure that one or other of the amplifiers in the system is the limiting factor in this regard.
3. Linear phase response (constant delay) with frequency. Similarly, the component's phase response should not degrade the achievable cancellation of the amplifier. A linear phase response is often a greater challenge, particularly over a wide bandwidth, due to the resonance effects and reflections caused by the operation of typical embodiments of these circuits (in particular the variable phase element).
4. Low insertion loss. Both of these components can suffer from a significant insertion loss, particularly when used at higher frequencies, and the cascade of the two elements (to provide both gain and phase control) *can significantly increase the gain requirement of the* main amplifier, or degrade the system noise performance, if employed in the reference path or prior to the error amplifier.
5. Linear control relationship. This is a less important requirement than those mentioned above, as it should not affect the ultimate linearity performance of the feedforward system. It will, however, affect the dynamic behavior of the control system and consequently such factors as pull-in performance and overshoot. This is particularly important where the amplifier will experience significant and

rapid changes in mean power loading, for example, in a cellular radio system where channels are assigned, handed-off and re-assigned on a regular basis. Such loading changes will require the control system to rapidly re-adjust and re-optimise to the new situation; a highly nonlinear control relationship will complicate this process.

6. Minimum interaction between gain and phase control. For control system stability and optimum pull-in performance, it is important that the interaction between the gain and phase controllers is kept to a minimum. In addition, it is important that the controllers themselves exhibit little or no parasitic behaviour (e.g., parasitic phase-shift with variations in amplitude control for the variable attenuator).

5.13.6.1 Variable Gain Components

Figure 5.32 illustrates one form of voltage variable gain element; in this case an attenuator. The PIN diodes are utilised to match and mismatch two of the

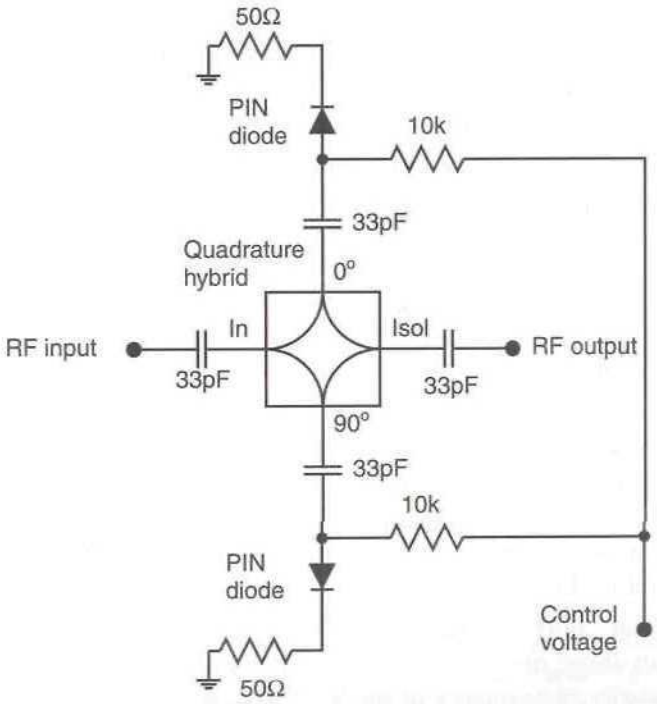


Figure 5.32 Simple voltage-variable attenuator (component values shown are for UHF operation). Note that it is possible to omit the 50Ω resistors, with the PIN diodes providing the necessary impedance.

ports of the quadrature hybrid, with the 50Ω resistors providing a low VSWR termination when the PIN diode is at a low impedance. Both diodes are controlled utilising the same control voltage to ensure that the system is as closely balanced as possible in operation. Differences in the characteristics of the two diodes will mean that this balance is never perfect; however, acceptable performance can usually be achieved in practice at VHF, UHF, and low SHF frequencies utilising this technique. The capacitors provide DC isolation for the quadrature hybrid to prevent damage (in the case of low-power transformer-based designs) or shorting of the control voltage (in the case of transmission-line based hybrids). Note that the 50Ω resistors may be omitted in some cases.

An alternative technique is to utilise a variable gain amplifier as part of the main or error amplifier chains, thus saving one element of the system. Such amplifiers are now available in MMIC form, although they generally operate at a very low power level, with maximum output powers of 0 dBm or less being typical. They are therefore only suitable for the first or perhaps second stage of any amplifier chain.

5.13.6.2 Variable Phase Components

A similar principle can be utilised to provide a variable phase-shift element and is illustrated in Figure 5.33. A quadrature hybrid is again employed, although in this case utilising a varactor (or varicap) diode to alter the matching on each of the ports. The diodes employed should be of the same type and appropriate for the frequency of operation of the phase-shifter.

Typical variations in phase of 90° to 120° across the control voltage range (approximately 30V for many diodes) may be obtained at UHF with this circuit, although most of this range is obtainable with a voltage range of 15V, or less.

Such techniques are not truly broadband, and other configurations may be employed if octave or multi-octave operation is required.

5.13.6.3 Vector Modulator

The use of separate gain- and phase-controls can introduce problems when they are employed directly following one another in cascade. The matching or mismatching provided by one element as part of its control process may have a significant effect on the other and may cause its characteristics to change. A similar effect can also be observed within an individual element as, for example, a variable attenuator will almost certainly introduce a small variable phase-shift across its attenuation range. This will typically be only a few degrees, but is likely to be sufficient to significantly degrade the overall cancellation of the feedforward system. Likewise, a variable phase-shift

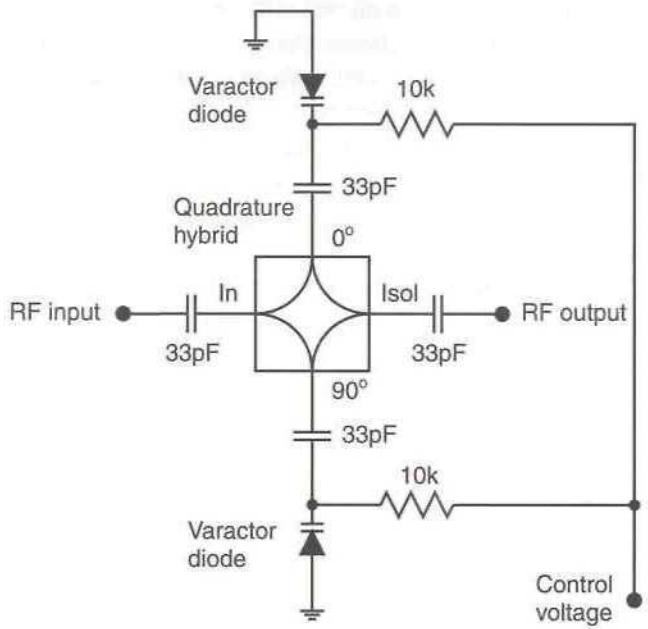


Figure 5.33 Simple voltage-variable phase-shift network (component values shown are for UHF operation).

element will introduce a small (typically < 1 dB) gain variation across its phase-shift range. This will again be sufficient to seriously degrade the performance of the feedforward system in a high-performance design.

An alternative technique for both gain and phase control is the vector modulator, illustrated in Figure 5.34. The basic principle of operation is to vary the signal in Cartesian component form, rather than in polar form, as represented by separate gain and phase elements in cascade.

An equal variation of both voltage-controlled attenuators (shown implemented using multipliers in Figure 5.34) will produce a pure signal loss, with no phase-shift (assuming ideal components), whereas unequal variation will produce a variable phase-shift and a signal loss. Such a system may be easier to control automatically in certain circumstances and overcomes many of the interaction and isolation problems of discrete variable attenuators and phase-shifters (due to the inherent isolation of the quadrature hybrids). It also has the significant advantage that a full 360° of phase control is provided; a cascade of at least three of the phase-shifters of the type illustrated in Figure 5.33 would be required to achieve this. The system does, however, have the drawback that it is less intuitive and hence often more difficult to adjust manually. Also, the intercept point of the multipliers is

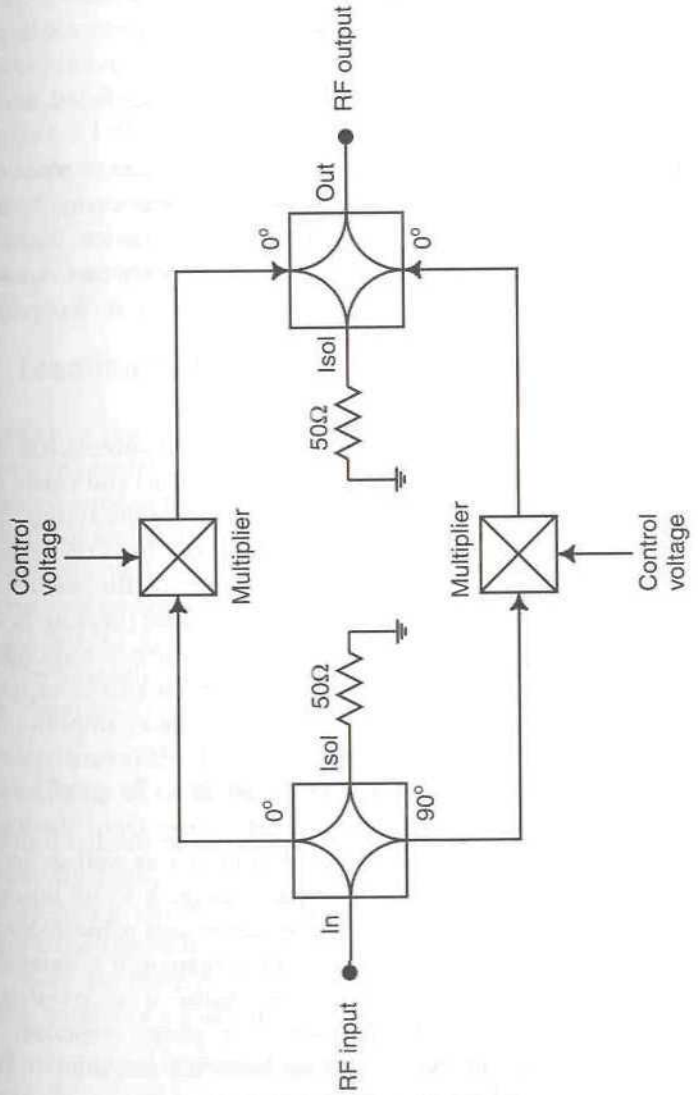


Figure 5.34 Vector modulator-based voltage-variable gain and phase-shift network.

likely to be less than that available from variable phase and attenuation circuits of the type just described.

The loss in a vector modulator is likely to be greater than that possible with separate variable gain and phase components. In addition, the probable lower intercept point will necessitate a lower drive level and either or both of these issues may lead to noise problems. When a vector modulator is employed therefore in either the reference path or prior to the error amplifier, the noise performance of the loop may be degraded and this could be significant in some circumstances.

It is possible to use PIN diode based variable attenuators in place of the multipliers (in Figure 5.34), however, the highly non-linear control voltage vs attenuation characteristic inherent in most of these devices can make both manual and automatic control yet more difficult (in the latter case due to the variation of the effective control loop gain as the attenuation changes).

5.13.7 Output Coupler

The purpose of the output coupler is to perform the final subtraction of the error signal from the distorted main amplifier output signal and hence it has a strong influence in determining the final performance of the system. Since it appears in the main signal path, its through-path loss is of concern, and this has a fundamental affect on the overall efficiency of the feedforward system (see Section 5.7). Whilst this loss must be minimised (bearing in mind the power requirements of the error amplifier), the frequency response and phase characteristic must not be compromised and it must also be capable of handling the required mean and peak powers of the main amplifier. This coupler is therefore a very important element in the feedforward system.

The directivity of the output coupler must also be good, as this parameter determines how much of the reflected energy from the load or antenna is coupled into the output of the error amplifier as well as, in part, how much of the error amplifier output signal gets back to its input (see Section 5.15). Since the spectral purity and frequency and phase flatness of this amplifier are critical to the performance of the system, it is essential to minimise any external disturbances which may cause it to overload or introduce additional ripple into its frequency or phase response. This becomes particularly important where the feedforward amplifier is prone to interference from co-sited services, since any energy from these signals, received on the system antenna and fed back down to the feedforward amplifier output, may cause intermodulation in the error amplifier output. It is the directivity of the output coupler which determines how much of this energy the error amplifier output sees and hence what level of intermodulation distortion is produced, in the absence of an external isolator (which adds cost and loss). Again the effect of imperfect terminating impedances on the coupler directivity (discussed earlier) will need to be considered.

Most systems operating between VHF and a few gigahertz utilise either stripline or microstrip couplers, with the former generally being preferred due to their excellent flatness, high directivity and low additional through loss (above that due to their coupling factor alone). Directivity figures in excess of 20 dB are typical, combined with gain and phase flatness figures of better than 0.1 dB and 1°, respectively, over a 100 MHz bandwidth at 2 GHz. More recently, the inclusion of the main-path sampling coupler and the output coupler along with a filter-based delay line has become common and a number of manufacturers offer this facility. This helps to minimise the number of transitions and hence keeps overall loss to an absolute minimum.

5.14 Location and Matching Considerations

A number of external factors can affect the operation and performance of a feedforward system. Signals can be injected into the transmit antenna by another co-sited antenna, and its signals may intermodulate with the output signals from the feedforward system, lowering its performance.

Alternatively, a poorly-matched transmit antenna, or damaged feeder cable may cause significant reflected power to impinge upon the output of the feedforward amplifier, again degrading performance.

The following sections discuss the likely consequences of such issues, in the absence of an isolator at the output of the feedforward system.

5.14.1 Effects of Co-Siting With Other Systems

The major concern when co-siting a transmitter incorporating a feedforward amplifier with other radio systems, is that of signals radiated by the co-sited antennas being injected into the output of the feedforward amplifier by means of the transmit antenna.

There are three potential effects of unwanted signals reverse injected into the output of a feedforward amplifier:

1. Increased distortion in the main amplifier output.
2. Increased distortion in the error amplifier output.
3. Reduction in performance of the adaption scheme (if fitted).

Each of these is illustrated in Figure 5.35 and all are dealt with separately below.

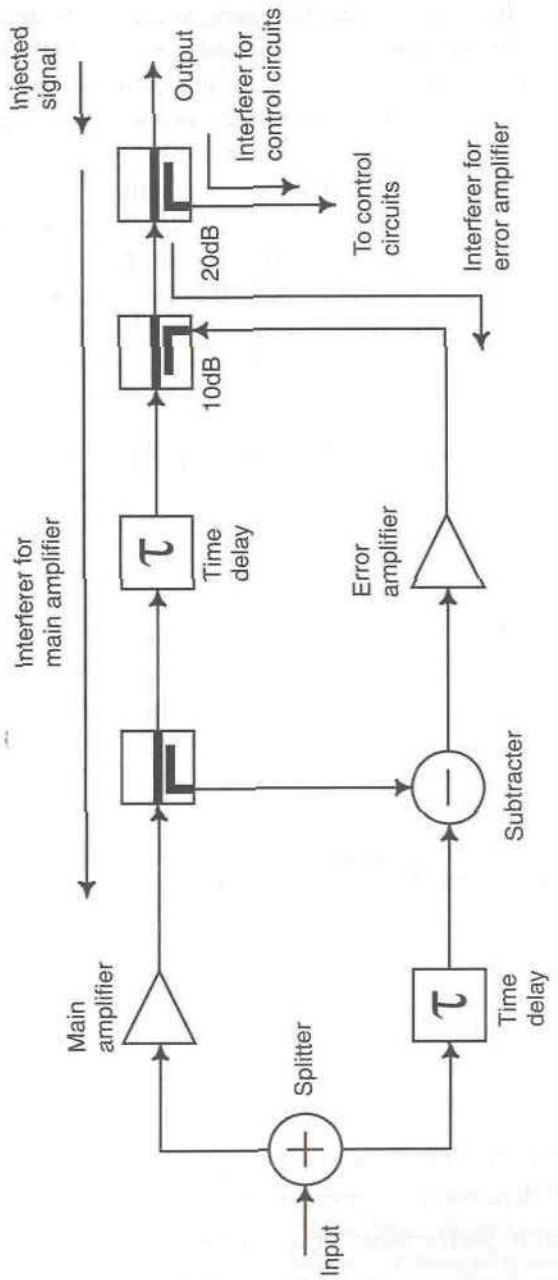


Figure 5.35 Potential paths for injected signals to cause interference.

5.14.1.1 Distortion in the Main Amplifier Output

This is potentially the most damaging problem, as the reverse-injected signal can pass from the feedforward system output directly through the various main-path couplers and the high-power time delay to the output of the main amplifier. It will suffer very little attenuation en-route and hence it can be assumed (pessimistically) that the full coupled signal level will impinge on the output devices of the main PA. This signal level is typically assumed to be around −30 dB in many type-approval specifications.

In the case of a multi-carrier linear amplifier, care must be taken in the interpretation of this '−30 dB' injection level. For example, if the overall amplifier is rated at 1 kW, this could imply an injected signal level of 1W. However, this would be a little unfair on the linear amplifier, as the individual carriers would be typically 25W or less and hence the injected carrier ought to be 30 dB down on each of these signals, that is, 25mW (+14 dBm).

This signal will generate a number of IMD products in the output devices of the main PA. As long as these IMD products are not greater than the intrinsic distortion products of the main amplifier itself (which they should not be, as the injected signal will be 30 dB down on the wanted signals, or about the same level as the IMD products from a good PA), this should not cause any problems. These new IMD products result from the main PA and hence will be treated in exactly the same manner as the existing products, that is, they will be eliminated from the overall amplifier output by the normal action of the feedforward system, assuming that they fall within the correction bandwidth of the feedforward loop. If they do not fall within the correction bandwidth, then an output filter will be required to ensure their elimination. These are normally provided anyway to remove broadband IMD and harmonic products.

5.14.1.2 Distortion in the Error Amplifier Output

The situation here is slightly different from that of the main amplifier above. The injected signal must now be reverse-coupled in the output coupler, in order to reach the output of the error amplifier. It will thus experience a minimum of 10 dB of attenuation (for a 10 dB output coupler) and probably more, due to the directivity of that coupler. Hence it will arrive at the output of the error amplifier some 40 dB or more below the individual carrier powers of the feedforward amplifier output signal (for a main amplifier with 30 dBc IMD performance). Here it will again cause intermodulation which will then be injected into the output of the feedforward system directly via the output (injection) coupler. This injected IMD will therefore experience a

further 10 dB of attenuation by the normal action of the output coupler. Thus if the required IMD specification is 70 dBc, the intermodulation caused in the output of the error amplifier by this interfering tone must be greater than 20 dB below the level at which the injected signal impinges on the error amplifier output. It is likely, in practice, to be significantly better than this figure and hence should not prove to be a problem.

5.14.1.3 Reduction in Performance of the Adaption Scheme

This effect, if it exists at all, will be heavily dependent on the specific adaption scheme chosen and hence can only be argued on that basis.

Most schemes involve the use of high-directivity sampling couplers at the output of the overall feedforward system, in order to monitor the performance of the system. Such couplers are also generally of a high coupling factor, and hence only a very small amount of any injected signal will appear at the sampled port of the coupler. A well designed scheme should be able to cope with such a small level of interfering signal, although the absolute performance of the system may be slightly affected.

5.14.2 Effect of Poor Antenna Matching

This is potentially a difficult problem with which a feedforward system must cope, as it has some far reaching implications on the power rating of the error amplifier. This, however, should be the only potential problem as, for example, the control system etc. should not be affected.

A poor antenna match can create four potential problems:

1. Error amplifier overload due to excessive (and erroneous) input signals.
2. Error amplifier overload due to its attempt to correct for increased linear distortion (frequency response ripple) in the overall system.
3. Error amplifier reverse power protection.
4. Correction-loop instability.

The first two problems are closely related, but will be examined separately below.

5.14.2.1 Error Amplifier Overload Due to the Presence of Reflected Signals

If the output of the main amplifier is not adequately matched to the system characteristic impedance (usually 50Ω), then the reflected signal from the mismatched load is re-reflected and subsequently sampled by the sampling coupler at the output of the main amplifier. This sampled signal may be of a

significant level and will still leave a large signal having had the reference signal energy cancelled in the subtracter. The input to the error amplifier may therefore be significantly larger than would be anticipated from a normal analysis (and consequent design) of the system.

The error amplifier is thus likely to overload in attempting to amplify this excessive signal level, with the resulting distortion being injected directly into the output of the feedforward system.

5.14.2.2 Error Amplifier Overload Due to Excessive Frequency-Response Ripple

The poor antenna match will create a significant ripple in the transfer characteristic of the main power amplifier, if that amplifier itself has a poor output match to 50Ω, and assuming that the antenna feed system is long with respect to the operating frequency and that the operating bandwidth is large. The result of this is that the error amplifier will attempt to supply sufficient power to partially cancel the peaks and 'fill-in' the troughs of this rippled characteristic, in order to provide a flat overall response. The power required of the error amplifier in order to do this could necessitate it having a power rating not dissimilar to that of the main amplifier (assuming, say, a 10 dB output coupling factor) and this is less than satisfactory. The only real solution to this problem is to improve the output match of the main power amplifier in order to try and prevent the second reflection process and its consequent effect upon the amplifier frequency response. One method of achieving this is by the use of an isolator at the output of the main amplifier and before the sampling coupler. The signal reflected by the antenna would then see a good 50Ω match and hence would be absorbed. An isolator could be used at this point without undue concern about the distortion it would introduce, as this would be corrected by the feedforward process (assuming that it was not excessive). Such an isolator would obviously be quite large (in a high-power system) and moderately expensive, but this must be considered against the overall cost and benefits of the system, rather than viewed independently. Alternatively, the amplifier itself could be modified to improve its output match.

5.14.2.3 Error Amplifier Reverse Power Protection

A portion of the reflected signal from the mismatched load will be sampled by the output coupler and fed to the output of the error amplifier. This level of this sampled signal will obviously be reduced by the coupling factor of the output (or injection) coupler (say 10 dB), but may still be significant in level.

The error amplifier must be capable of dissipating this power without excessive heating or disruption of its normal operation (i.e., without

increasing its distortion level).

5.14.2.4 Correction-Loop Instability

If the mismatched load contains a reactive component, then an oscillatory path may be created around the correction loop [31], with the mismatch providing the necessary feedback (by reflection) and the reactive element the necessary phase-shift.

Again the solution is to improve the output match of the main amplifier, either by a re-design of the output stage, or by the use of an isolator (if practical).

5.14.2.5 *Use of an Output Isolator*

From the above arguments, it is clear that the simplest solution in most cases is the use of an isolator at the output of the feedforward system. This isolator must possess a very high intercept point, as it must not degrade the high-linearity achieved from the amplifier. It will also, clearly, require a low through-loss, so as not to reduce the output power and efficiency of the feedforward system.

5.15 Loop Instability

One of the key advantages of a feedforward system is in its ability to achieve unconditional stability, hence allowing it to provide high degrees of linearity improvement over a broad bandwidth. It is, however, possible for the feedforward technique to be configured in such a manner that it will result in an unstable system [32]. The instability may be present in either the error loop or the correction loop or both (see Figure 5.21), although in a practical system, it is most likely to occur in the error loop (due to the gains and coupler values generally applied in that part of the system).

5.15.1 Instability in the Error Loop

In order for the error loop to be correctly balanced, the gain from the system input, to directional coupler 3 in the lower ('reference') path, by both the main-path and reference-path routes (see Figure 5.36), must be equal (taking account of the coupling factor of the coupler). Similarly, the phase difference between these two paths must be 180° in order for subtraction to occur at the coupler. Where these two conditions are fulfilled, perfect cancellation of the input signal energy in the error signal will result.

The situation described above is, of course, the desired operation of the

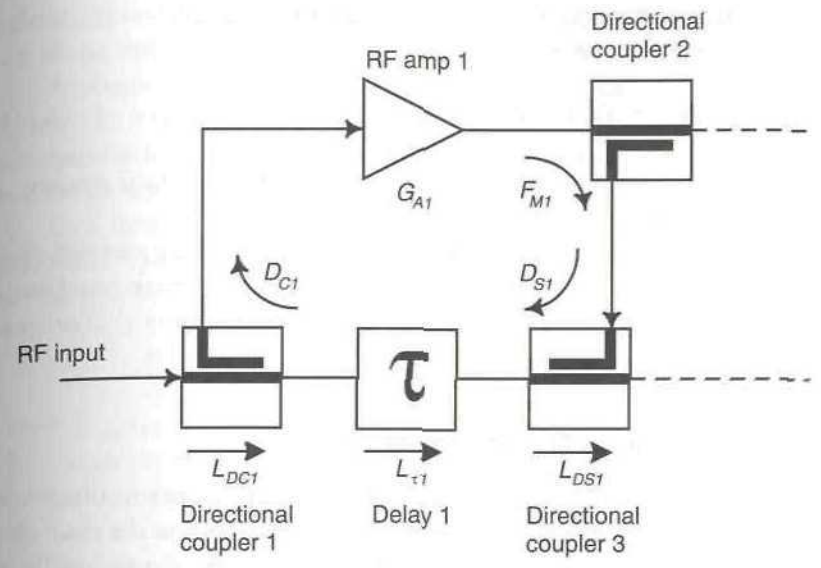


Figure 5.36 Error loop of a feedforward system, showing a possible oscillation mechanism.

system. It is also possible, however, for signals to propagate around the system in the direction indicated by the arrows in Figure 5.36. In this case, coupler 2 is operating normally, whilst the other two couplers are providing a reverse leakage path due to their imperfect directivity (or other undesired signal leakage in the system).

Under circumstances in which some of the signal is able to take this reverse path, then it is evident that a path exists from the output of the main amplifier back to its input. If the gain around this loop is greater than or equal to unity, then instability of the amplifier may result. Clearly, it is necessary to keep this gain well below unity in a well designed system, so that a margin of safety is provided.

In a feedforward system, the main amplifier gain (in dB) is determined by the loop coupling factors in the error loop (assuming that the loss in the time-delay element is negligible):

$$G_{A1} = L_{C1} + L_{S1} + F_{C1} + F_{M1} + F_{S1} \tag{5.113}$$

where: L_{C1} is the through-path loss of the input coupler, L_{S1} is the through-path loss of the subtraction coupler, F_{C1} is the coupling factor of the input coupler, F_{M1} is the coupling factor of the main-path coupler and F_{S1} is the coupling factor of the subtraction coupler (all in dB).

The gain coefficient, C_{O1} (in dB) around the 'reverse-loop', formed from the output to the input of the main amplifier, is given by:

$$C_{O1} = G_{A1} - F_{C1} - D_{C1} - F_{S1} - D_{S1} - F_{M1} \qquad (5.114)$$

where: D_{C1} is the directivity of the input coupler and D_{S1} is the directivity of the subtraction coupler (both in dB).

The stability criterion for this loop is therefore that the gain coefficient, C_{O1}, should be less than 0 dB in order for the system to be guaranteed stable and should preferably be say −5 dB (or less in some applications) in order to allow a margin of safety.

5.15.2 Instability in the Correction Loop

Analysis of the correction loop proceeds in a similar manner to that of the error loop above. In this case, it is the gain from the input to the main-path coupler 2, to the output directional coupler 4 in the upper ('main') path, by both the upper and lower routes in Figure 5.37, which is of concern. In order for the correction loop to be perfectly balanced, these two gains must be equal (taking account of the coupling factor of the coupler). Again, the phase

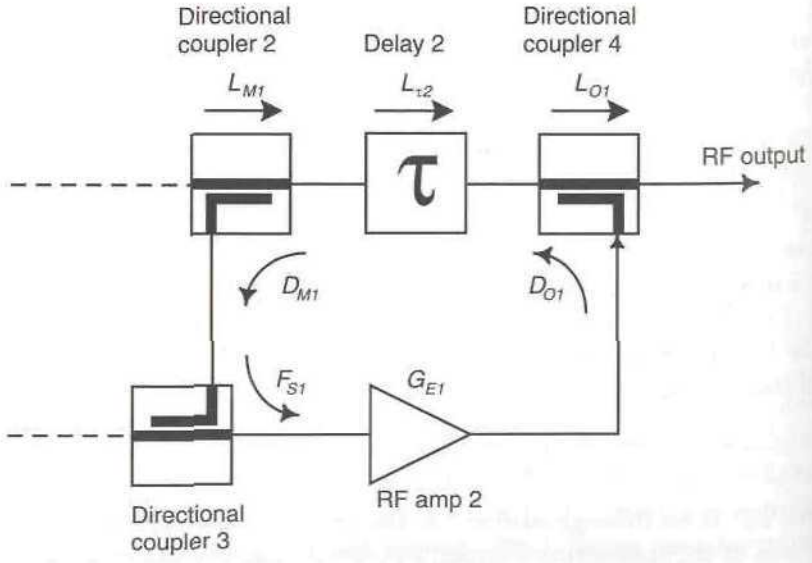


Figure 5.37 Correction loop of a feedforward system, showing possible oscillation mechanism.

difference between these two paths must be 180° to ensure perfect subtraction of the distortion produced by the main amplifier.

A reverse path also exists in this loop, as indicated by the arrows in Figure 5.37; it is therefore possible for the error amplifier to have a path from its output back to its input. If the gain around this loop is unity (or greater), then instability of the error amplifier is possible.

In a feedforward system, the gain of the error amplifier (in dB) is determined by the loop coupling factors in the cancellation loop (assuming that the loss in the time-delay element is negligible):

$$G_{E1} = L_{M1} + L_{O1} + F_{M1} + F_{S1} + F_{O1} \qquad (5.115)$$

where: L_{M1} is the through-path loss of the main-path coupler, L_{O1} is the through-path loss of the output coupler, F_{M1} is the coupling factor of the main-path coupler, F_{S1} is the coupling factor of the error subtraction coupler and F_{O1} is the coupling factor of the output coupler (all in dB).

The gain coefficient, C_{E1} (in dB) around the 'reverse-loop', formed from the output to the input of the error amplifier, is given by:

$$C_{E1} = G_{E1} - F_{M1} - D_{M1} - F_{O1} - D_{O1} - F_{S1} \qquad (5.116)$$

where: D_{M1} is the directivity of the main-path coupler and D_{O1} is the directivity of the output coupler (both in dB).

This gain coefficient should also be less than 0 dB by a suitable margin in order to ensure that the system is guaranteed to be stable.

5.15.3 Practical Implications

There are two categories of system which are prone to instability problems, unless attention is paid to the issues outlined above. The first is high-gain systems, in which the main amplifier gain is necessarily very high in order to overcome the various coupling factors in the system (perhaps 10 dB or more greater than the desired overall gain from the feedforward system). A significant burden is therefore placed on the directivity of the couplers, in order to avoid instability.

Consider, for example, a feedforward system with the following parameters: $L_{C1} = 6$ dB, $L_{S1} = 6$ dB, $F_{C1} = 6$ dB, $F_{M1} = 20$ dB, $F_{S1} = 6$ dB, $F_{O1} = 10$ dB, $L_{O1} = 0.46$ dB. The coupling values and losses chosen here imply resistive splitters for the input and subtraction couplers. Using these values results in a main amplifier gain of 32 dB (from (5.113)) and requires the combined directivity of the two reference-path couplers to be 12 dB in

order to avoid instability in the error loop. This is obviously impossible when using resistive splitters (without, for example, the aid of isolators). An isolator placed in the reference path would be an appropriate solution in this case.

In the case of the cancellation loop, the required error amplifier gain is 36.5 dB, and results in the coupler directivity (combined) for the two main-path couplers needing to be only 0.5 dB. The main path couplers are those between the output of the main amplifier and the output of the feedforward system. This requirement can easily be met in most systems.

The second category of concern is systems employing broadband RF amplifiers (e.g., broadband MMIC devices). In this case, the coupler directivity must be maintained over a very wide bandwidth, *even if the system is not required to provide cancellation over this bandwidth*. This can be difficult to achieve with a number of coupler designs. The alternative solution, applying RF filtering following the main amplifier, will introduce significant additional delay into the main path. This in turn will require a larger compensating delay in the reference path. Adding filtering therefore removes one of the principal benefits of using broadband MMICs in a feedforward system: namely their intrinsic low group delay, which results in a correspondingly short reference-path delay line.

5.15.4 Effect on Input and Output Match

The same basic mechanism which has been highlighted above as potentially causing instability can also cause a degradation in the input and/or output match of the amplifier. The directivity (or isolation in the case of a 3 dB splitter) of the input coupler (directional coupler 1 in Figure 5.36) is usually adequate, such that the input reflection coefficient of the main amplifier can be neglected. Similarly, the input reflection coefficient of the error amplifier may be neglected since, when the error signal generation loop is correctly balanced, no (or little) main signal energy reaches the input to the error amplifier and hence cannot be reflected.

The input return loss is therefore dominated by the signal passing *through the main amplifier, via directional coupler 2 (correctly)* and then via directional coupler 3 (due to its imperfect directivity) and the error loop delay line, to the input coupler. Assuming that the loops are balanced, the input return loss (in dB) becomes:

$$R_{LI} = 2L_{DC1} + 2L_{\tau 1} + L_{DS1} + D_{S1} \tag{5.117}$$

where L_{DC1} and L_{DS1} are the through-path losses in the input and

subtraction directional couplers respectively and $L_{\tau 1}$ is the loss in the reference-path delay line (all in dB).

Similarly, the output return loss may be calculated by examining the correction loop (Figure 5.37). In this case, it is the directivity of the main-path sampling coupler (D_{M1}) which is critical in obtaining a good output match in a correctly balanced system. The output return loss (in dB) is therefore:

$$R_{LO} = 2L_{O1} + 2L_{\tau 2} + L_{M1} + D_{M1} \tag{5.118}$$

where L_{O1} and L_{M1} are the through-path losses in the output and main path directional couplers respectively and $L_{\tau 2}$ is the loss in the main path delay line (all in dB).

Whilst it is possible to increase the various losses in (5.117) and (5.118) in order to improve the input or output match, this has undesirable consequences (e.g., the increase of system noise figure or output path losses and hence reducing overall output power). The only realistic option in most cases is therefore an increase in coupler directivity of the relevant coupler (which will also depend upon the input and output match which it sees in each case).

5.16 Application Areas

The feedforward amplifier linearisation technique has been shown to be appropriate for applications in the field of satellite communications due to its ability to operate with unconditional stability over the large bandwidths required for this application. The use of a broadband linearisation technique has the advantage that it removes the need for a high-power power combiner, with its associated loss, and a large number of individual, single-channel amplifiers. The losses associated with a power combiner large enough to sum the hundreds of carriers present in many satellite systems would be considerable.

The alternative strategy for satellite applications is the use of a backed-off conventional class-A or class-AB linear amplifier, since the intermodulation performance of such amplifiers is often satisfactory. The advantage of a correctly designed feedforward system in this case is an improvement in efficiency. The use of predistortion techniques has also been successful in satellite systems, with a corresponding improvement in efficiency, and this form of solution is outlined in the next chapter.

The use of a class-C stage as the main amplifier in a satellite system is

quite feasible as the intermodulation performance requirements for satellite amplifiers are, in general, less stringent than those of mobile radio amplifiers. This is due to the fact that near-far problems no longer predominate and hence uncorrected class-A amplifiers may often be used. The addition of feedforward linearisation to a class-C stage allows the overall system distortion specification to outperform a class-A stage whilst still providing a useful improvement in efficiency. Such efficiency improvements can, in turn, provide a considerable weight and hence cost saving for a satellite payload.

The increasing interest in the use of the feedforward linearisation process for cellular base station amplifiers has led to a requirement to develop a fundamental understanding of the system efficiency characteristic when using a linear main amplifier and feedforward techniques to improve its intermodulation performance [33]. Such techniques are required to overcome the near-far problems associated with cellular systems; this leads to a requirement for intermodulation distortion products at least 60 dB down on the wanted signals (much greater in some systems). Although efficiency has obviously been sacrificed in this case, due to the use of a linear main amplifier, it is still of concern in order to minimise the amount of cooling required for the complete system, for example, along with the size of the power supplies.

Feedforward has been used in a number of cable television repeater systems and has even appeared in hybrid form for this application [10]. Cable television requires reasonable linearity performance over a very large number of carriers and a very wide bandwidth (one octave or more). Feedforward is therefore ideal for this application.

Feedforward has also been suggested for use in ultra-high dynamic range, low-noise amplification in receiver systems [34]. So long as the noise figure issue is carefully considered, a significant overall improvement in dynamic range may be achieved (defined as the difference between the third-order intercept point and the noise figure for the amplifier). This is particularly beneficial in systems which are interference limited, rather than noise limited. This is increasingly the case in base station applications due to the increased use of mast space and increased number of channels in use generally.

The linear distortion correcting properties of feedforward may be employed to good effect in broadband instrumentation applications, particularly those where diode-based levelling loops are impractical due to the dynamics of the system (e.g., fast-hopping synthesisers). Front-end amplification for spectrum analysers may also benefit from both the IMD and linear distortion correction benefits of the feedforward technique.

5.17 Potential Advantages

A summary of the potential advantages of a broadband linear amplifier applied to a cellular or PMR base station is as follows:

5.17.1 Flexibility

Ideally a feedforward amplifier system should contain no frequency selective components, other than those defining the overall frequency band of operation. Any of the multiple inputs to the amplifier can therefore be used on any channel frequency within the allocated band and these frequencies can be re-allocated virtually instantaneously. This is a distinct advantage over cavity combiner techniques in which mechanical re-tuning is required. A number of cavity combiner manufacturers have recently proposed the use of stepper motors for automatic re-tuning of their cavities. This would allow a limited amount of flexibility, but is still far from satisfactory and its reliability under frequent use must also be questioned.

Cavity combining also places severe restrictions on the minimum channel spacing which may be used at a given site and also precludes the use of slow (or fast) synthesiser-based frequency-hopping, such as that used in GSM. 'Baseband hopping', that is, matrix switching of the baseband input signals, must be employed instead and this is often a more complex technique.

5.17.2 Size

Most high-power combination techniques involve physically large hardware (particularly at VHF), which generally requires expensive mechanical engineering in its construction. In a broadband linear amplifier system the combination process takes place at the input at a low power level and hence using physically small, low-cost components. Both of these scenarios are illustrated in Chapter 1.

The amplifier in a linear PA solution is potentially large due to the high power levels involved, however, only one such amplifier is required and hence there is generally a significant overall space saving.

5.17.3 Transparency

The use of a linear amplifier is often required to satisfy the multi-carrier aspects of the base station problem, however, it brings with it a number of additional benefits. The major benefit is that it is able to deal with any

modulation scheme, both present and future, within its operational bandwidth. This somewhat sweeping generalisation may be tempered slightly by the operation of the particular linearisation or control technique under consideration, but should certainly permit most forms of modulation to be utilised.

5.17.4 Future-Proof Design

This follows on from the flexibility and transparency argument outlined above. The current lifetime of a base station is dictated by a number of factors, only one of which is reliability. In many cases systems are replaced more because they use outdated modulation techniques with poor spectral efficiency than because the state of the art in power amplifier technology has advanced to such a degree that replacement of an older system is justified on those grounds alone.

The utilisation of linearised amplifier technology should enable the RF components of a base station to be replaced when reliability considerations dictate, rather than with the latest fashion in modulation techniques. This will enable users to update equipment smoothly and in a planned manner (be this on a technical or financial basis), rather than having to find a large investment in one go for an entire new system.

5.17.5 Dynamic Channel Allocation (DCA)

This is potentially of less interest in PMR systems, due to the small number of channels usually allocated at each base station site. It is, however, of great importance in cellular systems, as it has been shown in the literature [35] that the use of DCA can enable them to experience an approximately two-fold capacity increase without requiring any increase in their overall spectrum allocation.

The use of linear amplifier technology in a PMR scenario would, however, permit a site operator to offer a service to a number of smaller organisations, each requiring only a small number of channels. So long as these channels were in a similar portion of the frequency spectrum, they could all potentially be amplified by the same broadband linear amplifier.

5.17.6 Positioning Flexibility

The use of a single amplifier, with low-level combining, results in a very flexible configuration from the point of view of equipment positioning. The power amplification process can be separated by a considerable distance from the power combination process without loss of overall performance. The

inevitable cable losses now occur at a low power level, where they may be easily and cheaply recovered, and the overall power output of the base station remains unaffected.

This is a particular advantage when considering the siting of base stations within an urban area. The bulk of the equipment (for example, signal generation, up-conversion, supervisory functions) can be sited in (say) a basement where the cost of site rental is low, with only the amplifier requiring siting at the top of the building in expensive 'office' (cupboard) space. This offers a potential saving to the service provider and hence renders the equipment manufacturer's product more attractive.

5.18 Practical Results

Figure 5.38 shows the output spectrum from a very high-linearity feedforward power amplifier operating at full rated power (20W mean, in this case).

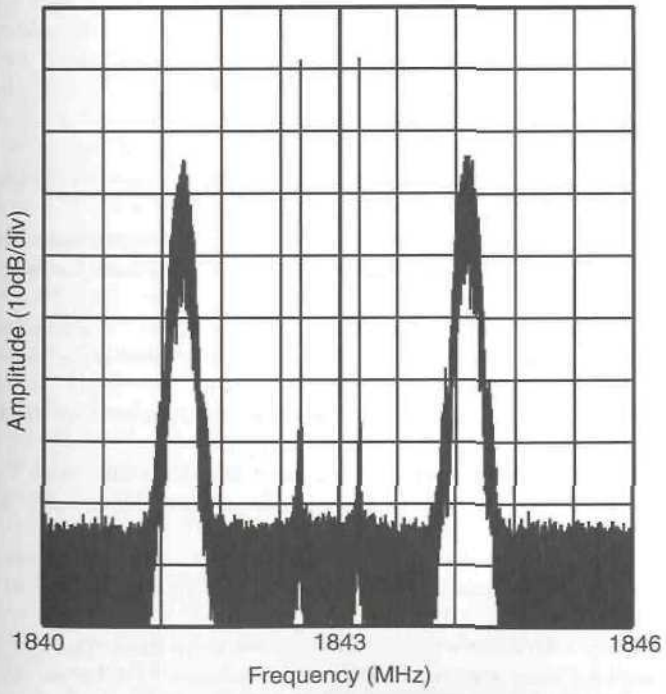


Figure 5.38 Output spectrum of a very high-linearity feedforward power amplifier with two GSM signals and two CW carriers. © Wireless Systems International Ltd.

The signals consist of two GSM carriers and two CW carriers, each at 5W mean power; note that the GSM signals appear lower due to the very narrow resolution bandwidths required to achieve the measurement dynamic range. The amplifier is achieving a >75 dBc intermodulation performance, in this case over a 6 MHz bandwidth (although the amplifier is specified to the same performance over 30 MHz).

This high level of performance is maintained using a control scheme, which monitors and corrects *all* loops in the system. The control scheme is DSP-based and operates in a manner which ensures that the ultimate system reference is digital within the DSP, hence providing a guarantee that performance cannot drift with, for example, temperature changes, or aging. It should be noted that although many feedforward control systems utilise DSP or microprocessor control, almost all work on DC input signals (e.g., from correlation mixers) and the offsets in these devices are prone to long-term (weeks or months) drift. Such changes in the offset voltages will have a consequent effect on achievable long-term linearity from the amplifier, and this has been cited as an objection to the practical use of feedforward by some system operators.

References

1. Black, H. S., "Translating system," U.S. Patent 1,686,792, Issued 9 October 1928.
2. Black, H. S., U.S. Patent 2,102,671, Issued December 1937.
3. McMillan, B., "Multiple feedback systems," U.S. Patent 2,748,201, Issued May 29 1956.
4. Zaalberg van Zelst, J. J., "Stabilised amplifiers," *Philips Technical Review*, No. 9, 1947, pp. 25–32.
5. Deighton, M. O., E. H. Cooke-Yarborough, and G. L. Miller, "A method of enhancing gain stability and linearity of transistor amplifier systems," *Proc. of the Northeast Regional Electrical Manufacturers Conference*, Boston, USA, 1964.
6. Golembeski, J. J., *et al.*, "A class of minimum sensitivity amplifiers," *IEEE Trans. on Circuit Theory*, Vol. CT–14, No. 1, March 1967, pp. 69–74.
7. Seidel, H., H. R. Beurrier, and A. N. Friedman, "Error Controlled High Power Linear Amplifiers at VHF," *The Bell System Technical Journal*, Vol. 47, May/June 1968, pp. 651–722.
8. Seidel, H., "A feedforward experiment applied to an L–4 Carrier System Amplifier," *IEEE Trans. on Communication Technology*, Vol. COM–19, No. 3, June 1971, pp. 320–325.
9. Olver, T. E., D. C. Andrews, B. S. Abrams, and E. E. Barr, "Linear wideband HF power amplifier using adaptive feedforward cancellation," *Proc. of the IEEE Military Communications Conference, MILCOM*, Vol. 1, 1982, pp. 21.6/1–8.
10. Motorola Semiconductors Ltd., *R.F. Device Data*, Vol. II, Fifth Edition, 1988, pp. 134–135.

11. Dixon, J. P., "A solid-state amplifier with feedforward correction for linear single-sideband applications," *Proc. of the IEEE International Conference on Communications*, Toronto, Canada, June 1986, pp. 728–732.
12. Johnson, A. K., and R. Myer, "Linear amplifier combiner," *Proc. of the 27th IEEE Vehicular Technology Conference*, Tampa, Florida, USA, June 1987, pp. 421–423.
13. Stewart, R. D., and F. F Tusubira, "Feedforward linearisation of 950 MHz amplifiers," *IEE Proc.*, Vol. 135, Pt. H, No. 5, October 1988, pp. 347–350.
14. Lubell, P. D., W. B. Denniston, and R. F Hertz, "Linearizing amplifiers for multi-signal use," *Microwaves*, April, 1974, pp. 46–50.
15. Kenington, P. B., and R. J. Wilkinson, "The specification of error amplifiers for use in feedforward transmitters," *IEE Proc. Part G*, Vol. 139, No. 4, August 1992, pp. 477–480.
16. Lubell, P. D., W. B. Denniston, and R. F Hertz, "Linearizing amplifiers for multi-signal use," *Microwaves*, April, 1974, pp. 46–50.
17. Kenington, P. B., "Efficiency of Feedforward Amplifiers," *IEE Proc. Part G*, Vol. 139, No. 5, October 1992, pp. 591–593.
18. Parsons, K. J., and P. B. Kenington, "Efficiency of a feedforward amplifier with delay loss," *IEEE Trans. on Vehicular Technology*, Vol. 43, No. 2, May 1994, pp. 407–412.
19. Kenington, P. B., and D. W. Bennett, "Linear distortion correction using a feedforward system," *IEEE Trans. on Vehicular Technology*, Vol. 45, No. 1, February 1996, pp. 74–81.
20. Cavers, J. K., "Adaptation behaviour of a feedforward amplifier linearizer," *IEEE Trans. on Vehicular Technology*, Vol. 44, No. 1, February 1995, pp. 31–40.
21. Gerard, R. E. J., and G. N. Hobbs, "Improvements in or relating to amplifiers," U.K. Patent No. GB 2107540B, published 26 June 1985.
22. Bauman, R. M., "Adaptive feed-forward system," U.S. Patent No. 4,389,618, published 21 June 1983.
23. Olver, T. E., "Adaptive feedforward cancellation technique that is effective in reducing amplifier harmonic distortion products as well as intermodulation distortion products," U.S. Patent No. 4,560,945, published 24 December 1985.
24. King, N. J. R., "Feedforward amplifiers," U.K. Patent Application No. GB 2167256A (withdrawn), application published 21 May 1986.
25. Kenington, P. B., M. A. Beach, A. Bateman and J. P. McGeehan, "Apparatus and Method for Reducing Distortion in Amplification," US Patent No. 5,157,345, 20 October 1992.
26. Myer, R. E., "Automatic reduction of intermodulation products in high power linear amplifiers," U.S. Patent No. 4,580,105, published 1 April 1986.
27. Myer, R. E., "Feed forward linear amplifier," U.S. Patent No. 4,885,551, published 5 December 1989.
28. Fish, P. J., *Electronic Noise and Low Noise Design*, Macmillan Press, 1993, Chapter 4.
29. Caverly, R. H., "Distortion in broad-band gallium arsenide MESFET control and switch circuits," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 39, No. 4, April 1991, pp. 713–717.
30. Caverly, R. H., and G. Hiller, "Distortion in p-i-n diode control circuits," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 35, No. 5, May 1987, pp. 492–501.
31. Sabin W. E., and E. O. Schoenike (editors), *Single-sideband Systems and Circuits*, McGraw-Hill, 1987, chapter 13.
32. Kenington P. B., P. A. Warr and R. J. Wilkinson, "Analysis of Instability in Feedforward Loop," *IEE Electronics Letters*, Vol. 33, No. 20, pp. 1669–1671, 25 September 1997

33. Johnson A. K., and R. Myer, "Linear amplifier combiner," *Proc. of the 27th IEEE Vehicular Technology Conference*, Tampa, Florida, USA, June 1987, pp. 421–423.
34. Warr, P. A., P. B. Kenington, and M. A. Beach, "Feedforward Linearisation in High Dynamic Range Receivers," *ACTS Mobile Summit*, 7–10 October 1997, Aalborg, Denmark, pp. 839–844.
35. *Proc. of the IEEE Global Communications Conference*, Orlando, Florida, USA, December 1992.

6

Predistortion Techniques

6.1 Introduction

Predistortion is conceptually the simplest form of linearisation for an RF power amplifier. It simply involves the creation of a distortion characteristic which is precisely complementary to the distortion characteristic of the RF PA and cascading the two in order to ensure that the resulting system has little or no input-output distortion. It is, of course, possible to cascade the complementary distortion element *after* the RF PA and this is referred to as *postdistortion*; however, there are a number of obvious drawbacks with this technique and few advantages (for high-power amplifiers). The result is that most complementary distortion systems are based around predistorting the input signal and fall into one of the following categories:

1. *RF predistortion*—the nonlinear predistorting element operates at the final carrier frequency.
2. *IF predistortion*—the predistorting element or network operates at a convenient intermediate frequency, thereby possibly allowing the same design to be utilised for a number of different carrier frequencies. Alternatively, the required predistortion components may not operate satisfactorily at the desired carrier frequency, hence necessitating the use of an IF.
3. *Baseband predistortion*—prior to the advent of digital signal processing (DSP) devices, this technique had few advantages over the IF technique. It is now, however, a useful tool. In this case, the

predistortion characteristic is typically stored as a table of gain and phase weighting values within a DSP (usually in quadrature form) in order to predistort the baseband information before upconversion. It is possible to use feedback to provide updating information for these coefficients; this technique is called *adaptive baseband predistortion*.

The techniques used in RF and IF predistortion are generally similar and hence will be discussed together.

6.2 RF and IF Predistortion

6.2.1 Introduction

A fundamental advantage of RF and IF predistortion is its ability to linearise the entire bandwidth of an amplifier or system simultaneously. It is therefore ideal for use in wideband multicarrier systems, such as satellite amplifiers or in cellular/PCN base station applications.

The degree of linearity improvement which can be achieved in practice depends upon a wide variety of considerations, and in particular on the form of the transfer characteristic of the amplifier. With traditional predistortion systems, the achievable linearity improvement is modest, by comparison with, say, a controlled feedforward system or a Cartesian feedback system, but is adequate for many applications. In general, the better behaved the transfer characteristic is, the greater the degree of improvement which can be achieved and, importantly, maintained over a variety of input conditions (most notably power level). This is not true, however, if the amplifier is already very linear, for example if it is operating backed-off, or is a low-power class-A amplifier. In this case, the dominant nonlinearity may not be due to compression, making it difficult for simple forms of predistortion to work well.

More recent advances, such as the APL™ technique [1], allow multiple orders of nonlinearity to be corrected and this paves the way for far higher degrees of linearity improvement to be achieved from predistortion techniques.

6.2.2 Theory of Operation

The basic form of a predistortion linearisation scheme is shown in Figure 6.1. The predistorting function, $\beta(V_i)$, operates on the input signal in such a manner that its output signal is distorted in a precisely complementary

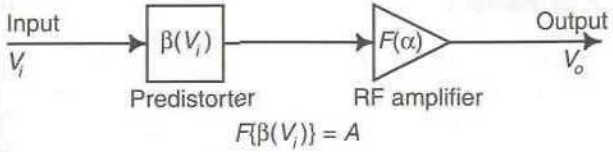


Figure 6.1 Schematic of an RF amplifier and predistorter.

manner to the distortion produced by the RF PA, $F(\alpha)$. The output signal is therefore an amplified, but undistorted replica of the input signal.

The problem, then, is to ascertain the required form of the predistortion characteristic and to fabricate a circuit with a transfer characteristic which closely resembles the required function (Figure 6.2). Note that Figure 6.2 is only an illustration of the form of operation and that a practical (e.g., cubic) predistorter will not result in a characteristic which is a direct mirror image of the amplifier characteristic, as suggested by this diagram. This is principally because it is not usual to predistort even-order elements of the amplifier transfer function, hence resulting in a marked difference between the predistorter and inverse amplifier characteristics.

This is not a trivial problem and a large number of different networks have been utilised over the years in an attempt to mimic various types of characteristic. The simplest, and most widely used, networks merely attempt to predistort the third-order characteristic and may, in the process, increase the level of higher-order distortion products.

Other networks attempt to 'curve-fit' the distortion characteristic and thereby improve the performance of a number of orders of distortion. For such networks to achieve a high level of performance, however, they often need to be designed, or at least adjusted, for each individual amplifier (even of the same design). Predistortion amplifiers can therefore benefit from the use of an automatic control technique in very much the same manner as a feedforward system.

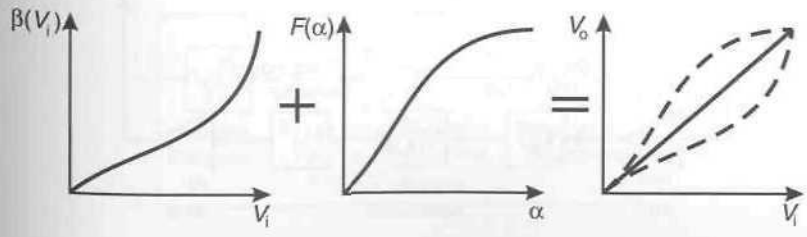


Figure 6.2 Operation of a predistortion system.

6.2.3 Cubic Predistorters

The aim of a cubic predistorter is to eliminate third-order distortion by means of the correctly-phased addition of a cubic component to the input signal. In the case of a bandpass system, it is only necessary (or beneficial) to reduce the third-order products to the same level as, or slightly below, the level of the next highest products (usually the fifth-order products). Improvements beyond this point are generally of little benefit, other than when predistortion is used in conjunction with feedforward, in which case, the power contained in the error signal (and hence the power rating required of the error amplifier) can be reduced.

Cubic predistortion is of particular benefit for the linearisation of travelling-wave-tube (TWT) amplifiers, as these usually possess a predominantly third-order characteristic. High degrees of linearity improvement can therefore be achieved from a suitably designed (and controlled or adjusted) cubic predistorter.

The use of this form of predistortion, termed *polynomial predistortion*, has been widespread for many years, particularly in the fields of satellite power amplification and high bit-rate point-to-point links [2,3]. It has traditionally been used where only modest degrees of linearity improvement are required, due to the poor matching usually achieved between the predistorter and amplifier characteristics. More recently, however, the use of predistortion as an aid to the feedforward correction process has been suggested [4] and systems have been fabricated with careful matching of the PA device and the predistorter [5] (usually by means of purpose-designed silicon for both the PA and predistorter devices). This paves the way for RF predistortion to achieve far greater levels of linearity improvement.

One form of cubic predistorter is shown in Figure 6.3. The RF (or IF) input signal is split, in this case by a directional coupler, to form a main path

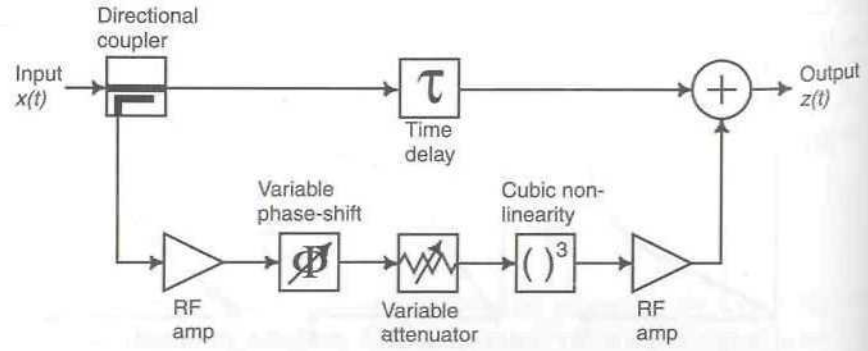


Figure 6.3 Cubic RF/IF predistorter.

and a secondary path. The main path contains a time-delay element, to compensate for the delay through the various elements in the secondary path, and ensure that the signals recombine with the correct time relationship. Since this delay element will typically be operating at a low power level, its insertion loss is usually not critical, although it will contribute to the overall noise figure of the system.

The secondary path contains a low-level buffer amplifier, followed by gain and phase control elements to ensure that the correct relationship is achieved at the summing junction (combiner). The predistortion element is formed by the cubic nonlinearity and the resulting signal is buffered and amplified by the post-distortion amplifier. Both of the amplifiers in the secondary path are small-signal devices and hence contribute negligible distortion.

The configuration shown in Figure 6.3 is that of a *scalar* predistorter and the phase relationship between the upper and lower paths is such that the lower path subtracts from the upper path, as outlined below.

6.2.4 Generation of an Expansive Characteristic

The general form of characteristic required to linearise almost all RF power amplifiers is expansive in form, that is, a characteristic whereby the gain increases with increasing input signal level. The simplest method of creating this form of characteristic, shown in Figure 6.4, is by subtraction of a

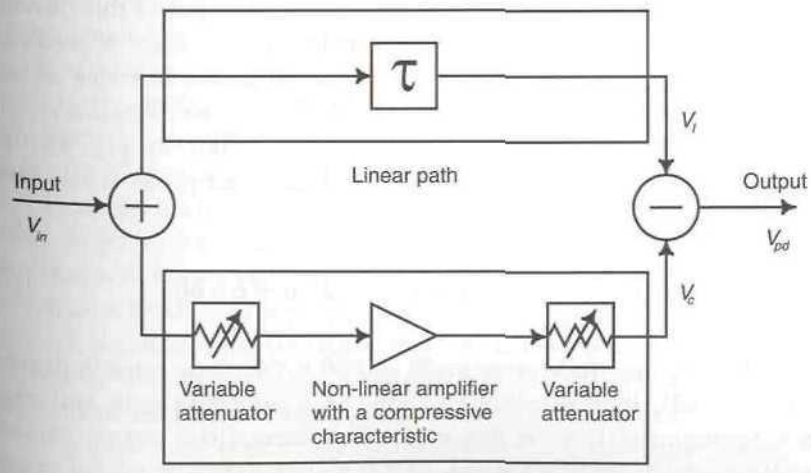


Figure 6.4 Generic method of creating an expansive characteristic.

compressive characteristic from a linear characteristic. This can be illustrated mathematically as follows.

The output of the linear path (typically just a time delay) is given by:

$$v_l(v_{in}) = a_1 v_{in} \tag{6.1}$$

and that of the compressive path (a low-power amplifier driven at an appropriate power level, for example) is given by:

$$v_c(v_{in}) = a_2 v_{in} - b v_{in}^3 \tag{6.2}$$

Subtracting the above equations gives:

$$v_{pd}(v_{in}) = (a_1 - a_2) v_{in} + b v_{in}^3 \tag{6.3}$$

This is now an expansive characteristic with a linear gain of $a_1 - a_2$, and may be used to predistort any compressive amplifier characteristic (cubic in this example) by appropriate choice of a_1, a_2, and b. These are usually set by the drive level of the lower path (compressive) amplifier and the gain or attenuation settings in both paths. Note that in many predistorters, a_1 will be approximately unity and a_2 will be less than unity.

6.2.5 Gain- and Phase-Matching Characteristics

It is instructive to examine the effects of gain- and phase-mismatch between the linear and polynomial paths of this type of predistorter as this provides characteristics which allow the ideal performance of this form of predistortion to be predicted. The results illustrate the similarity, in terms of ideal performance, between this type of system and that of a feedforward system.

An analysis of this form has been undertaken in [6], and the intermodulation suppression as a function of gain and phase error shown to be:

$$S_{IMD} = -10 \log\left(1 + 10^{\frac{\delta A}{10}} - 2 \cdot 10^{\frac{\delta A}{20}} \cos \delta\theta\right) \tag{6.4}$$

where δA is the amplitude error in dB and $\delta\theta$ is the phase error in degrees.

Using (6.4), it is possible to show how imperfect gain and phase balance settings for the predistorter will influence the degree of IMD cancellation which can be achieved. This is shown in Figure 6.5 for various gain (amplitude) and phase errors.

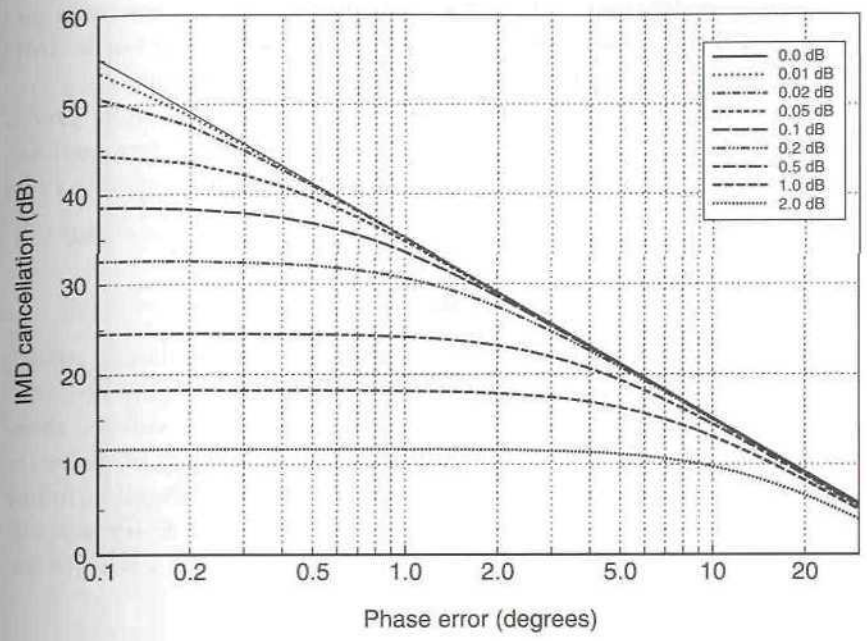


Figure 6.5 IMD cancellation which can be achieved from a basic cubic predistorter for various values of gain and phase error.

These results are identical to those derived in Chapter 5 for feedforward correction and examination of them leads to the conclusion that an ideal predistorter would require an extremely high degree of gain- and phase-matching accuracy in order to achieve a high level of IMD reduction. This is unlikely to be necessary for conventional diode or transistor-based predistorters, as their characteristics are not sufficiently accurate to provide high levels (> 30 dB) of IMD removal; hence a more relaxed gain- and phase-matching specification can be used.

These results have assumed a purely AM–AM characteristic for the RF power amplifier and this will not be the case in practice, as AM–PM distortion will have a significant effect. This will further degrade the likely performance level of a simple predistorter of the type considered here. It is, however, possible to use two such predistorters in a quadrature arrangement in order to correct for both AM–AM and AM–PM distortion and this arrangement should offer an improved overall performance.

It should, of course, be remembered that the suppression levels indicated in Figure 6.5 are additional to the raw IMD performance of the RF power amplifier. So, for example, if the uncompensated main amplifier

had a third-order IMD ratio of 30 dB and the predistortion system had a gain error of 0.5 dB and a phase error of less than 1°, the overall third-order IMD performance of the complete system would approach −55 dBc.

In practice, however, the performance of a purely third-order predistorter will be limited by the fifth-order IMD level of the RF power amplifier. There is therefore little point in suppressing the third-order products to a level much below the level of the fifth-order IMD.

6.2.6 Ideal Cubic Fit for a Class-C Amplifier

The result of using an ideal cubic predistorter on a highly nonlinear (class-C) amplifier is shown in Figures 6.6 and 6.7. The use of a highly nonlinear amplifier illustrates well the form and resultant of both a second- and third-order fit as well as the more usual case of a third-order only fit. Since the class-C amplifier fit characteristic was truncated to terms up to and including third-order, the resultant second- and third-order optimum fit is essentially perfect and the overall predistorted amplifier characteristic is a straight line

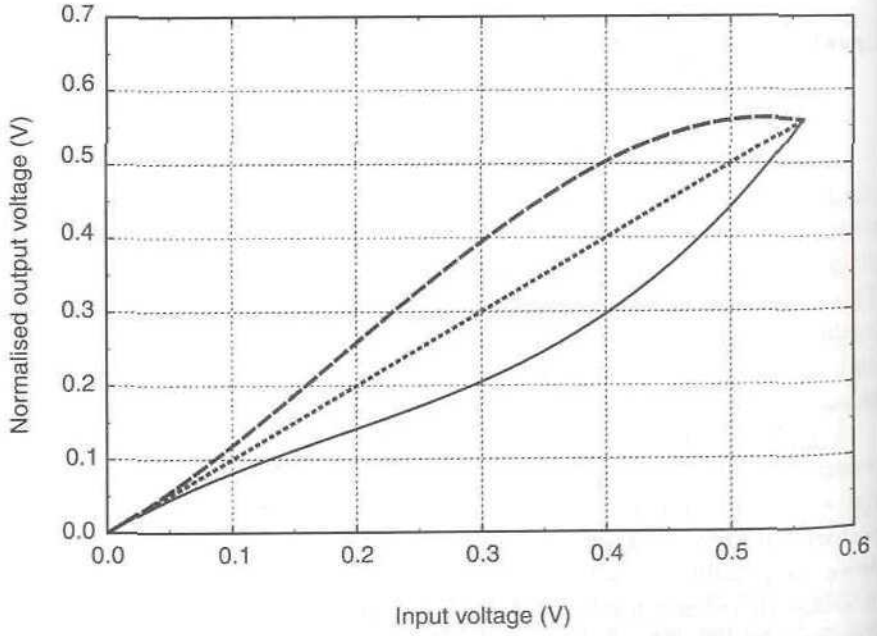


Figure 6.6 Cubic fit of a class-C amplifier characteristic (dashed line) and its ideal predistorter (second- and third-order) characteristic (solid line), with the resultant predistorted amplifier characteristic (dotted line).

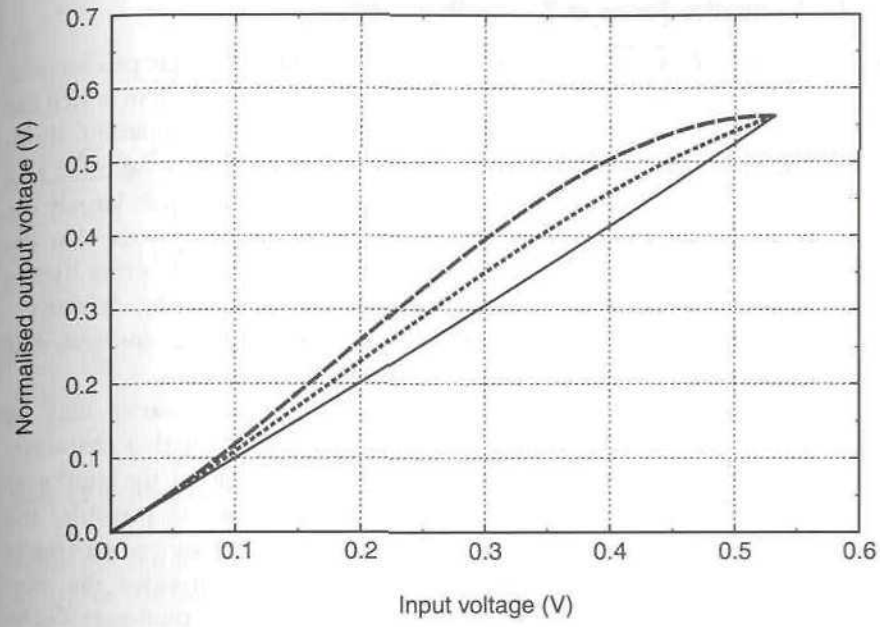


Figure 6.7 Cubic fit of a class-C amplifier characteristic (dashed line) and its ideal cubic-only predistorter characteristic (solid line), with the resultant predistorted amplifier characteristic (dotted line).

(Figure 6.6). Note that the class-C amplifier characteristic was derived from measurements on a real class-C power amplifier and hence the characteristics shown are representative of a real system (up to and including third-order). In practice, third-order only predistortion of a class-C amplifier would yield little benefit due to the high level of AM–PM conversion and the high-level of higher-order products.

The situation is somewhat different in the more typical case of a third-order only predistorter, as illustrated in Figure 6.7. In this case, no attempt is made to eliminate the second-order distortion present in the class-C characteristic and hence both the predistorter characteristic and the resultant predistorted amplifier response look markedly different from those of Figure 6.6. The predistorter characteristic appears to be almost linear and the resultant predistorted amplifier characteristic is clearly nonlinear (this is, of course, because it still contains a second-order component).

In both cases, using appropriate time-delay equalisation within the predistorter, a perfect two-tone in-band spectrum would result; in the latter case, however, harmonic zone components would also be present. Since these are filtered in most systems, the practical outcome in both cases is the same.

6.2.7 Alternative Forms of Simple Predistorter

The order and placement of many of the components in a cubic predistorter is not critical. An alternative version is shown in Figure 6.8 [7], in which the phase-shifter is incorporated in the delay-line path and the attenuator follows the cubic nonlinearity. The results presented above are also valid for this form of predistorter, since it is performing the same function in largely the same manner. The principal advantages of this configuration are that the gain and phase controllers are largely isolated from each other (some interaction can occur otherwise) and that the delay line can be shorter (or even eliminated) since the delay through the gain and phase controllers is likely to be similar.

There are many methods by which a cubic nonlinearity may be constructed, most of which involve a diode or transistor with a characteristic which closely matches the third-order characteristic of the nonlinear amplifier. Where access to a GaAs or Silicon foundry is possible, the characteristics of the predistorter device may be designed to accurately match the required characteristic and this method provides the best performance. Some typical predistorter devices include dual-gate GaAs FETs operating close to pinch-off [8] and Schottky diodes [9]. Further details on a range of circuit configurations for this element are provided below.

It is also possible to use nonlinear elements employing an expansive characteristic, as shown in Figure 6.9. Amplitude and phase adjustment is provided in the two parallel paths in a similar manner to that shown in

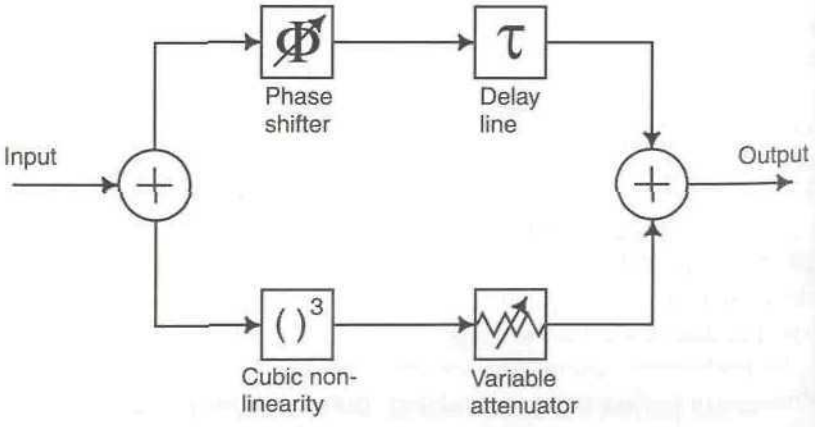


Figure 6.8 Alternative form of RF/IF predistorter.

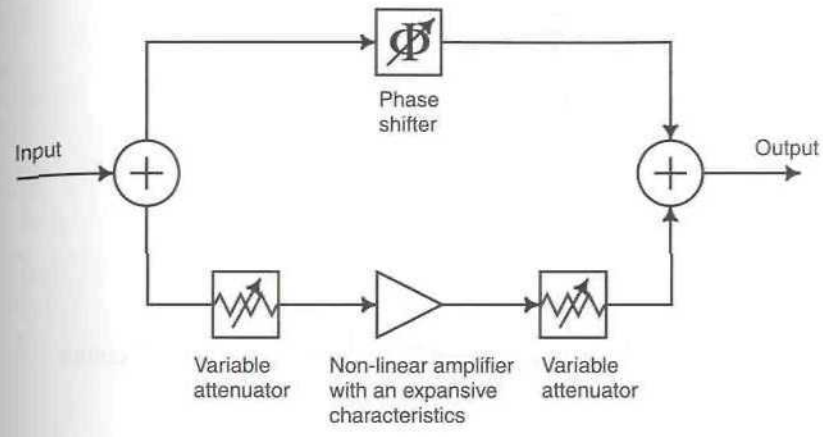


Figure 6.9 Predistorter employing an expansive amplifier.

Figure 6.8. An IF predistorter of this form, operating at 70 MHz, is reported in [10] for use with an 11 GHz TWT amplifier.

6.2.8 Single-Diode Predistorters

6.2.8.1 Series-Diode Predistorter

The simplest form of predistort nonlinear element is a series diode, with many examples appearing in the literature (e.g., [11]–[13]). The format of this type of predistorter is illustrated in Figure 6.10, with application being possible at both IF and RF.

The predistorter uses a Schottky diode with a separate parallel capacitor (C_P) in order to achieve a positive amplitude and a negative phase deviation at a low-bias condition (set by the bias resistor, R_{bias}). Adjustment of the bias resistor and the value of the parallel capacitor allow the predistorter characteristic to be matched as closely as possible to that of the power amplifier under consideration.

This type of predistorter will not result in spectacular linearity improvements nor large efficiency gains, but is nevertheless very simple and cost-effective. Adjacent channel power ratio (ACPR) improvements in the region of 4 dB for IS-95 CDMA have been reported at 1.9 GHz (in [11]).

6.2.8.2 Varactor Diode Predistorter

This type of predistorter is strictly a combination of two techniques: a varactor diode for AM–PM linearisation and second harmonic control (see Section 6.2.11) for AM–AM linearisation. This combination has the advan-

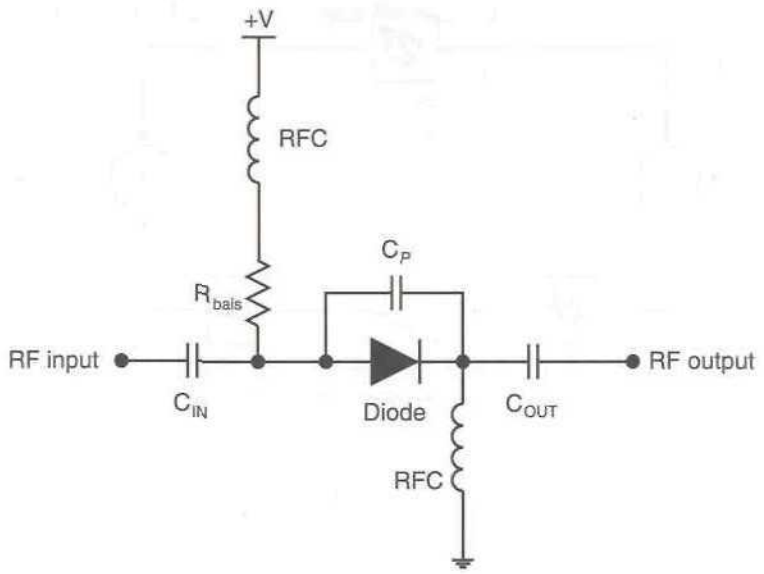


Figure 6.10 Series-diode predistorter.

tage, over the series diode technique outlined above, of a lower insertion loss (around 2 dB relative to perhaps 6 dB for the series-diode alternative).

The application of this technique to a GaAs FET single-stage amplifier is shown in Figure 6.11 [14]. The varactor diode functions as a compensation for the nonlinear capacitance at the input to the GaAs FET and hence serves to greatly reduce the AM–PM conversion in the resulting amplifier. The AM–AM characteristic is linearised using source second harmonic control [15] and hence the combination of the two techniques compensates for both the AM–AM and AM–PM conversion of the amplifier. The reported improvement in the first adjacent channel performance for a $\pi/4$-DQPSK signal (Japanese PHS specification) was around 15 dB.

6.2.9 FET-Based Predistorters

A number of variants on the basic use of a FET source-drain channel as a predistorter element are shown in Figure 6.12(a–c) [16,17], which illustrates the form of this type of nonlinearity; however, a wide variety of possible configurations exist.

Figure 6.12(a) shows a basic transmissive-mode nonlinearity, which could be employed as the nonlinear element in the predistorter shown in Figure 6.3. The bias voltage adjusts the degree (severity) of nonlinearity created and the variable capacitance acts to approximately adjust the phase of

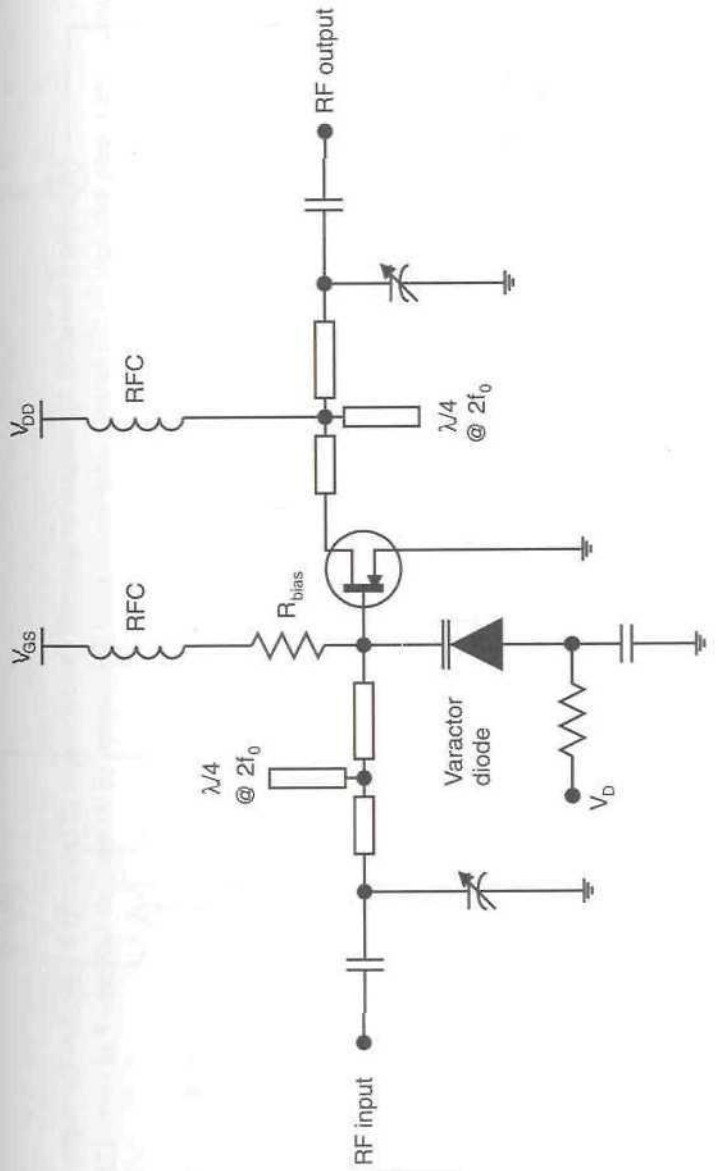
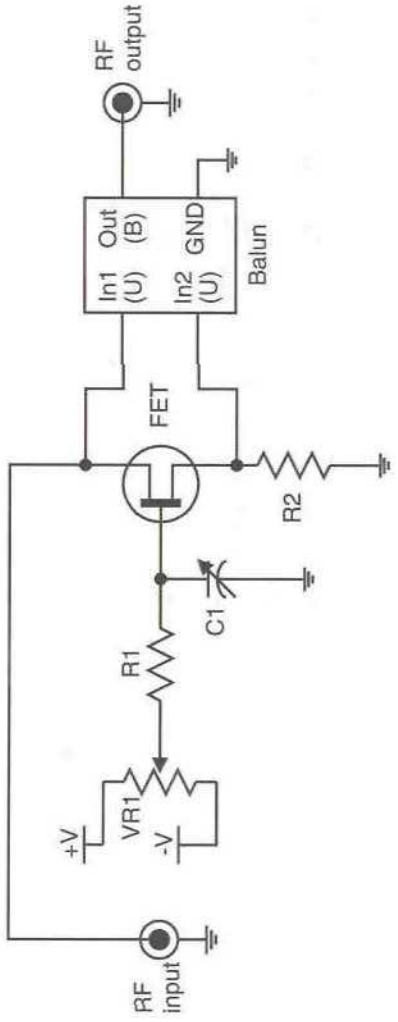
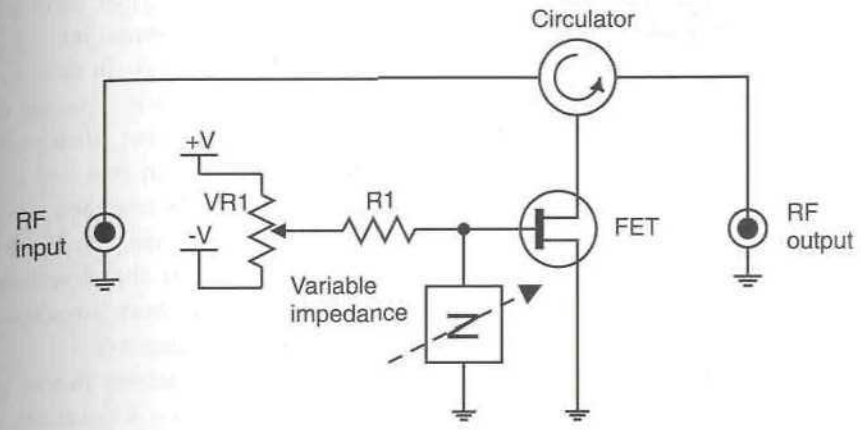


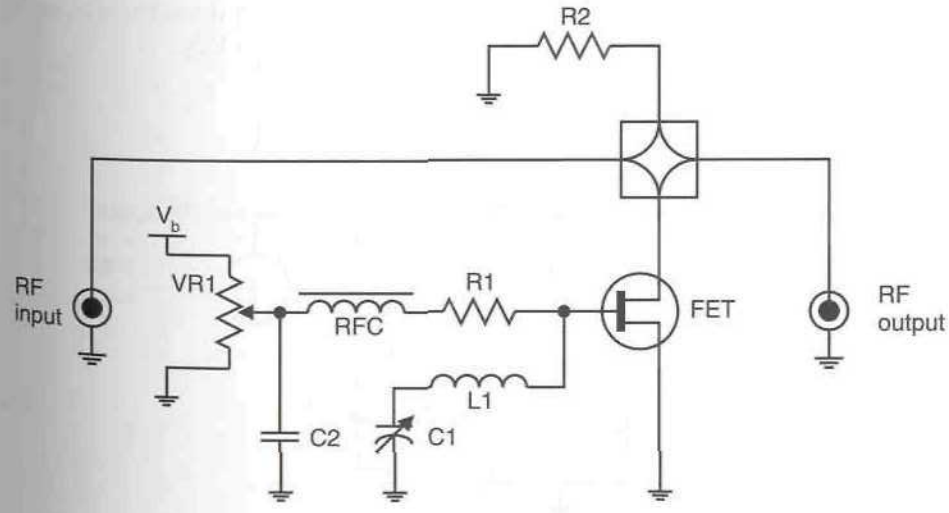
Figure 6.11 Varactor diode predistorter applied to a single-stage FET amplifier (from [14] © IEE 1999).



(a)



(b)



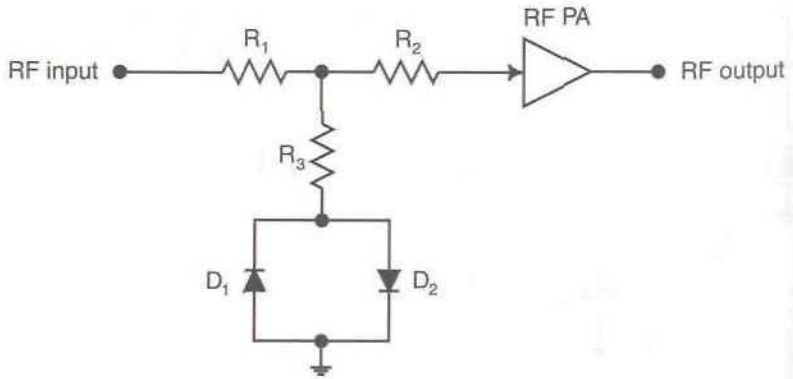
(c)

Figure 6.12 FET-based RF predistorter nonlinearity; (a) transmissive mode source-drain nonlinearity; (b) reflective mode source-drain nonlinearity employing a circulator; (c) reflective mode source-drain nonlinearity employing a hybrid.

the resulting IM products. The balun is used to extract the required signal across the source and drain terminals of the FET, although other methods could be employed to fulfil this function (e.g., a differential amplifier).

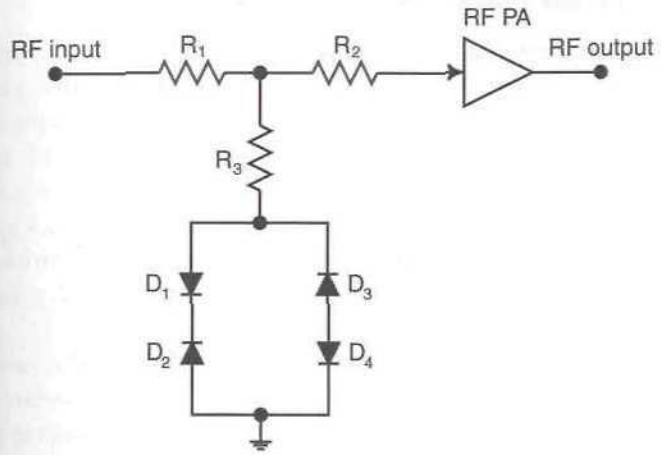
Figure 6.12(b) now employs the FET in a reflective mode, in this case on one terminal of a circulator. A signal entering the circulator is passed to the port connected to the FET nonlinearity and the resulting (distorted) reflected signal then passes to the output of the circulator. In this way the input and output match of the nonlinearity may be maintained at a reasonable level over a relatively broad bandwidth (governed, ideally, by the circulator). The variable bias voltage has a similar affect to that described above and the variable impedance can provide a degree of both amplitude and phase control of the signals emanating from the nonlinearity.

Finally, Figure 6.12(c) also employs the FET in a reflective mode, in this case in conjunction with a hybrid splitter or combiner. Operation is similar to that of the circulator-based approach described above, with the principal difference being that the resistance, R2 may be used to define an amount of undistorted input signal energy appearing in the output (by making R2 other than a matched load). The phase of this signal may also be varied by making this impedance other than purely resistive. If the hybrid used is a 180° type, then this system can be configured as a complete predistorter, rather than just a nonlinear element (in much the same way as the anti-parallel diode predistorter illustrated in Figure 6.13(c)).

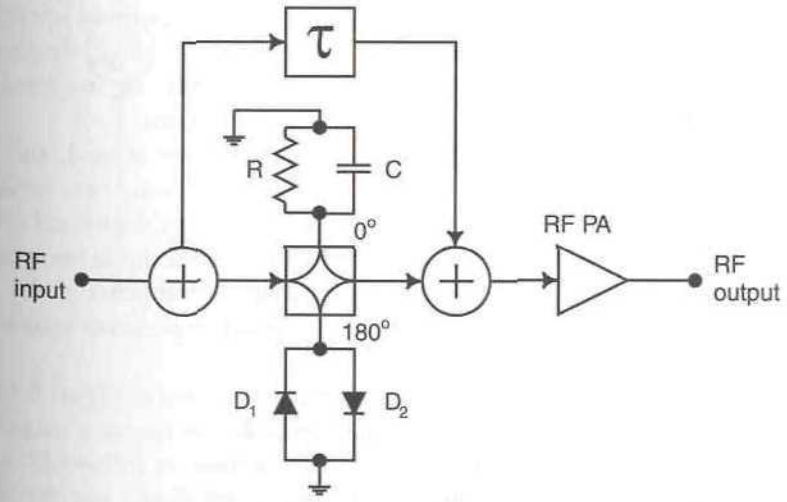


(a)

Figure 6.13 RF predistorter employing anti-parallel diodes (bias networks not shown): (a) conventional anti-parallel configuration, (b) bridge configuration, (c) hybrid-based configuration.



(b)



(c)

6.2.10 Anti-Parallel Diode-Based Predistorter

Arguably the simplest practical form of third-order diode-based predistorter is shown in Figure 6.13 in three different variants [18]–[23]. In Figure 6.13(a) and (b), the predistorter is formed around a T-attenuator (implemented by resistors R_1–R_3) which serves to sample the main RF signal path and also to re-inject the distortion component back in to that path. The nonlinearity itself is implemented by two anti-parallel diodes (Figure 6.13(a)) or by a diode bridge (Figure 6.13(b)); both configurations share the property that if the diodes are perfectly matched then (ideally) only third-order distortion is generated and re-injected into the main path.

If the diodes are not perfectly matched (as will be the case with all practical diodes) then second-order distortion will also be generated and this can interact with the nonlinearity of the amplifier and generate unwanted additional IMD products [24].

The use of an attenuator as the sampling and injection mechanism is a simple method of performing both functions whilst maintaining an acceptable 50Ω match for both the input and output signals. It will, however, have an affect on noise-figure performance of the overall amplifier and it may be preferable to employ directional couplers instead in some applications (with additional amplification in the nonlinearity path as necessary).

High-speed Schottky diodes are required for D_1–D_4, with an operating frequency range commensurate with the required RF operating frequency. Additional temperature (or other) compensation may be required to maintain performance over a range of operating conditions.

In Figure 6.13(c), a 180° hybrid splitter or combiner is used, and this arrangement has a number of advantages over the previous two variants. The linear impedance (resistance) in the zero degree path of the hybrid serves to cancel the residual linear component at the output of the diode branch; the capacitor appearing in parallel with this impedance compensates for the diode's reactance. The hybrid also presents a good impedance match for both the input and output signals.

A variation on the hybrid-based predistorter is shown in Figure 6.14, in this case employing a circulator to provide the isolation function originally provided by the hybrid. Its operation may be described as follows [25].

The current flowing into the pair of anti-parallel diodes is given by:

$$i(t) = I_0[\exp(a_d v(t)) - \exp(-a_d v(t))] \tag{6.5}$$

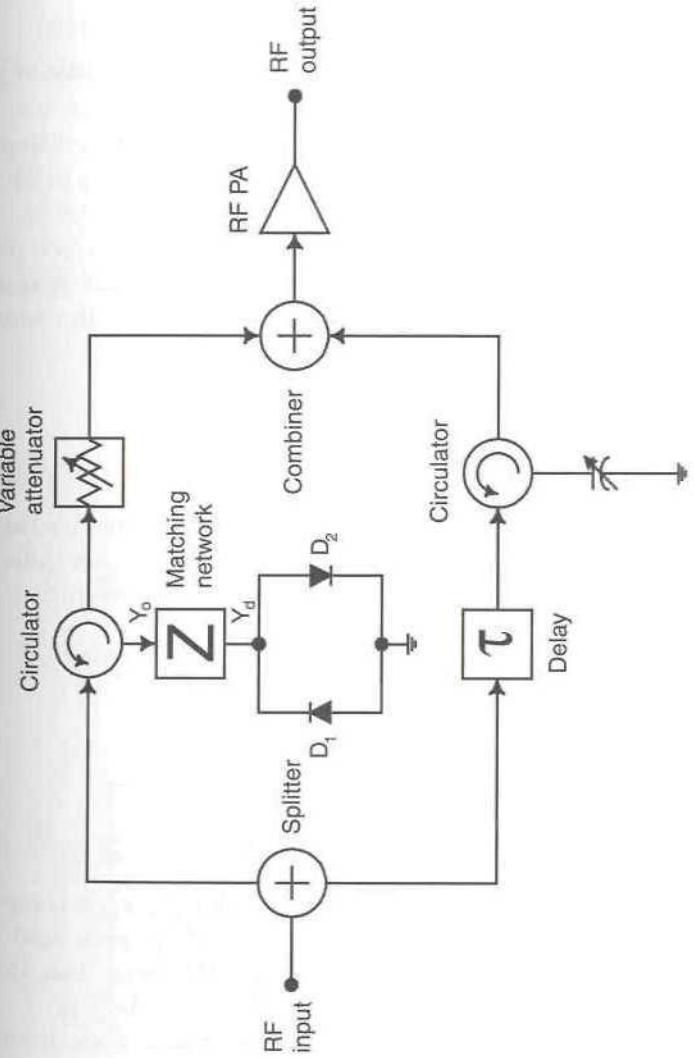


Figure 6.14 Circulator-based anti-parallel diode predistorter (from [25] © IEEE 1989).

This can be expanded by a Taylor series to become:

$$i(t) = 2I_0\left[a_d v(t) + \frac{a_d^3}{3}v^3(t) + \cdots\right] \tag{6.6}$$

Truncating this after the cubic term, the admittance of the pair of diodes now becomes:

$$Y_d(t) = \frac{i(t)}{v(t)} \cong 2I_0\left[a_d + \frac{a_d^3}{3}v^2(t)\right] \tag{6.7}$$

The matching network is designed to transform the linear term of the admittance ($2I_0a_d$) into the admittance of the circulator, Y_0. Equation (6.7) may then be written as:

$$Y_d(t) \cong Y_0 + ka_d^3v^2(t) \tag{6.8}$$

where k is a constant.

Thus, theoretically, the only signals which are reflected by the lower port of the circulator (and hence transferred to the output) are those based on the second term of (6.8). The reflection coefficient is therefore:

$$\Gamma \cong Ka_d^3v^2(t) \tag{6.9}$$

where K is a constant.

The output from the circulator is then:

$$V_{out}(t) = \Gamma v_{in}(t) \cong Ka_d^3v_{in}^3(t) \tag{6.10}$$

This indicates that (ideally) only the third-order distortion components will appear at the output and that the input signal component, $v_{in}(t)$ will be suppressed. In the practical system reported in [25], more than 15 dB of suppression of the input signal components was achieved.

Temperature compensation for the phase response of the lineariser is required, since the transmission phase-shift of a circulator decreases with increasing temperature. This would introduce a temperature-dependent phase error between the linear and distorted paths and hence must be corrected. The use of a second circulator, configured as a variable phase-shifter, can fulfil this role (see Figure 6.14). If the two circulators are identical and subject to the same temperature variations, then the two through-path phase responses will track and their effect will be cancelled at the lineariser output. The variable phase-shift function itself is then only necessary for setting up of the system or for compensation of changes due to loading, aging, or temperature changes of the power amplifier being linearised.

6.2.10.1 Improved Anti-Parallel Diode Configuration

An improved anti-parallel diode configuration has been proposed, in which the second-order distortion caused by mismatches in the two (or four) diode characteristics are substantially eliminated [26]. The configurations of both the modified anti-parallel circuit and the modified bridge circuit are shown in Figure 6.15. The reported values for R and R_d are 686Ω and 60Ω, respectively, for diode nonideality factors, η_1 and η_2 of 1 and 1.2, respectively. The values for R and R_d may be found from (6.11) and (6.12):

$$R_{d,\min} = \frac{g_{11}^{1/3} + g_{13}^{1/3}}{g_{11}^{2/3}\, g_{13}^{2/3}} - R_S \tag{6.11}$$

$$R = \frac{2\left[g_{11}^{2/3}(g_{13}R_S + 1) - g_{13}^{2/3}(g_{11}R_S + 1)\right]}{g_{13}^{2/3}\left[\frac{1}{R_d}(g_{11}R_S + 1) + g_{11}\right] - g_{11}^{2/3}\left[\frac{1}{R_d}(g_{13}R_S + 1) + g_{13}\right]} \tag{6.12}$$

Where R_S is the series resistance of one of the diodes, and R_d is the value

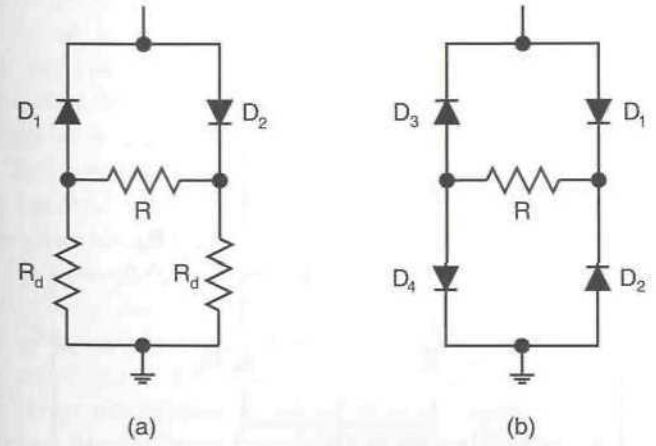


Figure 6.15 Improved anti-parallel diode configuration (a) and bridge configuration (b) for use in predistorters (bias circuitry not shown). From [26] © IEEE 1998.

chosen for the diode series resistors, which must be greater than, or equal to, $R_{d,\min}$ (from (6.11)). The g_{1n} values are found from:

$$g_{11} = I_B \alpha_1$$
$$g_{13} = \frac{1}{6} I_B \alpha_1^3 \tag{6.13}$$

Where I_B is the diode bias current and

$$\alpha_1 = \frac{q}{\eta_1 kT} \tag{6.14}$$

with T being the junction temperature, q the electronic charge, k, Boltzman's constant and η_1 the diode nonideality factor [27].

The primary disadvantage of the technique is a small reduction in third-order IM output from the predistorter, which is dependent upon the value chosen for R_d (larger values lead to a greater reduction in wanted third-order output). For the values stated above, a reduction of 1.5 dB will result.

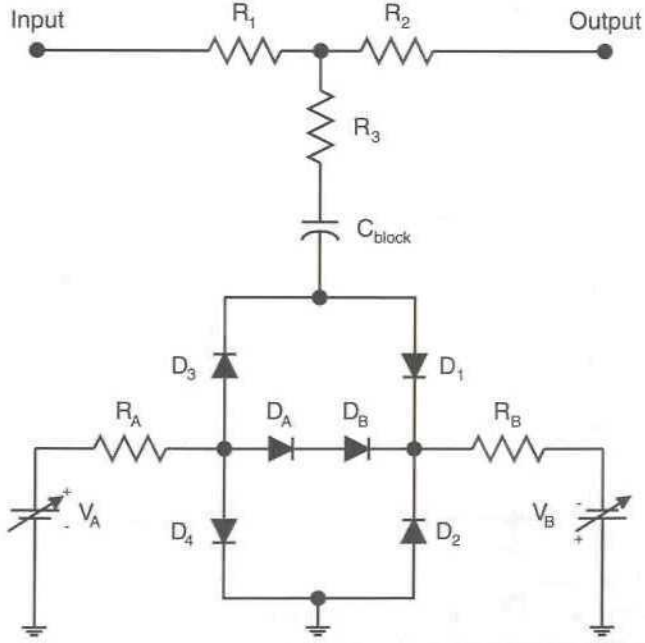


Figure 6.16 Modified bridge configuration employed in an attenuator-based third-order predistorter (from [26] © IEEE 1998).

Note that if R_d is chosen to be exactly equal to $R_{d,\min}$ (from (6.11)), then $R \rightarrow \infty$ and hence may be omitted.

A modification of Figure 6.15(b) is shown in Figure 6.16, in which R is replaced by two series diodes. The diode bias currents may be set by varying the voltage sources V_A and V_B.

6.2.11 Predistortion Using Harmonics

An alternative method of generating the required third-order nonlinearity is shown in Figure 6.17, in the form of a complete predistorted amplifier. The technique relies on the fact that most practical RF amplifiers, even low-power stages, generate a significant amount of second-harmonic distortion. This second-harmonic component may easily be extracted using a bandpass filter (and eliminated from the main-path using a low-pass filter), before undergoing appropriate gain and phase weighting and amplification prior to re-injection into the main path.

The second-order nonlinearity present in the main power amplifier will then act on the main and (weighted) second-order components, yielding a third-order component. If the gain and phase weightings are appropriately adjusted, this third-order component can be arranged to be in anti-phase to that naturally generated by the PA nonlinearity, hence providing overall third-order linearisation of the PA.

Some simple practical results from this technique are presented in [28] and indicate that an IMD improvement of some 20 dB can be achieved. It should be noted, however, that these results were obtained over a narrow bandwidth and using low-power MMIC amplifiers. Results with a high-power PA over a broad bandwidth can be expected to be somewhat less. A further problem is the third- and higher-order IMD generated by the first amplifier. This amplifier must be driven sufficiently hard to yield a reasonable level of second-order distortion, but not too hard to generate significant third-order distortion of its own. This preamplifier distortion will not be corrected by the system.

Finally, the preamplifier requires a frequency response (from the active device(s)) up to and including the second-harmonic of the desired input signal, in order for a reasonable signal to be generated. In particular, significant frequency response roll-off in this region would lead to poor gain and phase flatness (with frequency) of the cubic signal and hence poor broad-band cancellation.

An alternative version of the same concept is detailed in [15]. In this case, a controlled level of the second harmonic distortion produced by, and

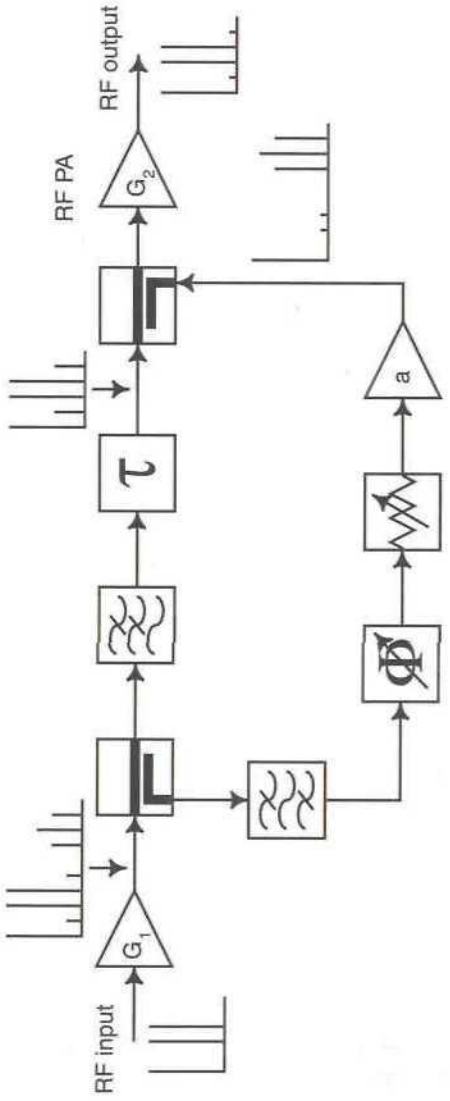


Figure 6.17 Second-harmonic based predistortion system.

reflected from, the input to the active device is rereflected at the correct phase to act as a predistorting signal when added to the wanted input signals. This technique is potentially simpler than that shown in Figure 6.17, but may be difficult to realise with some amplifier configurations due to the compromises required in the input matching network design.

6.2.12 IF Predistortion

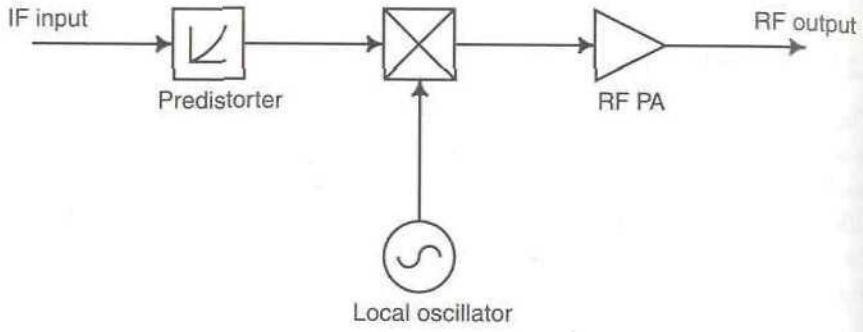
There are many system configurations which allow the use of predistortion at an intermediate frequency (IF). These have the advantage that a particular predistorter design may be employed in a family of similar amplifiers at different operating frequencies. The construction of a predistorter at IF may also be significantly simpler than an equivalent device at the final RF frequency, particularly where that frequency falls in the microwave region.

Three possible system configurations are shown in Figure 6.18. In Figure 6.18(a), the input is at the intermediate frequency and hence this configuration must be used as part of an overall transmitter design. In Figure 6.18(b) the input is at the final RF output frequency and hence this configuration forms a complete linearised amplifier. In Figure 6.18(c), an alternative RF amplifier configuration is used, with the advantage that no nonlinear devices appear in the main signal path. This configuration may therefore prove advantageous where a wide dynamic range is required.

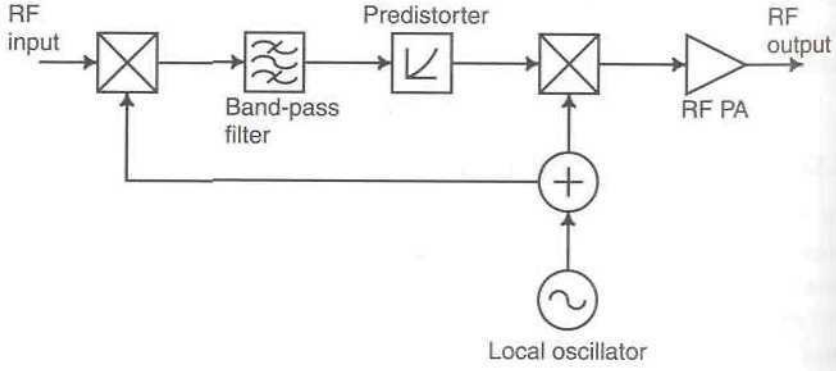
6.2.13 Curve-Fitting Predistorters

The types of predistorter described above rely on the use of a device with either a defined nonlinearity (e.g., cubic) or a carefully selected characteristic (usually the opposite to that of the RF amplifier). A more flexible form of predistorter can be formed by the use of a piecewise curve-fit to the amplifier transfer characteristic [29]. The predistorter characteristic is thus made up of a number of linear elements which approximate the inverse of the amplifier transfer characteristic; the approximation can be made arbitrarily close by the selection of a sufficient number of elements.

This is therefore a very general form of predistorter, as it can linearise any form of nonlinear characteristic and is not restricted in the number of orders of distortion which it can remove. The main disadvantage of the technique is in its relative complexity, particularly in terms of the initial adjustment procedure. The variables which must be adjusted (or correctly fixed) are the various slopes of the linear segments and the breakpoints between them. Even for the simple four-segment predistorter characteristic

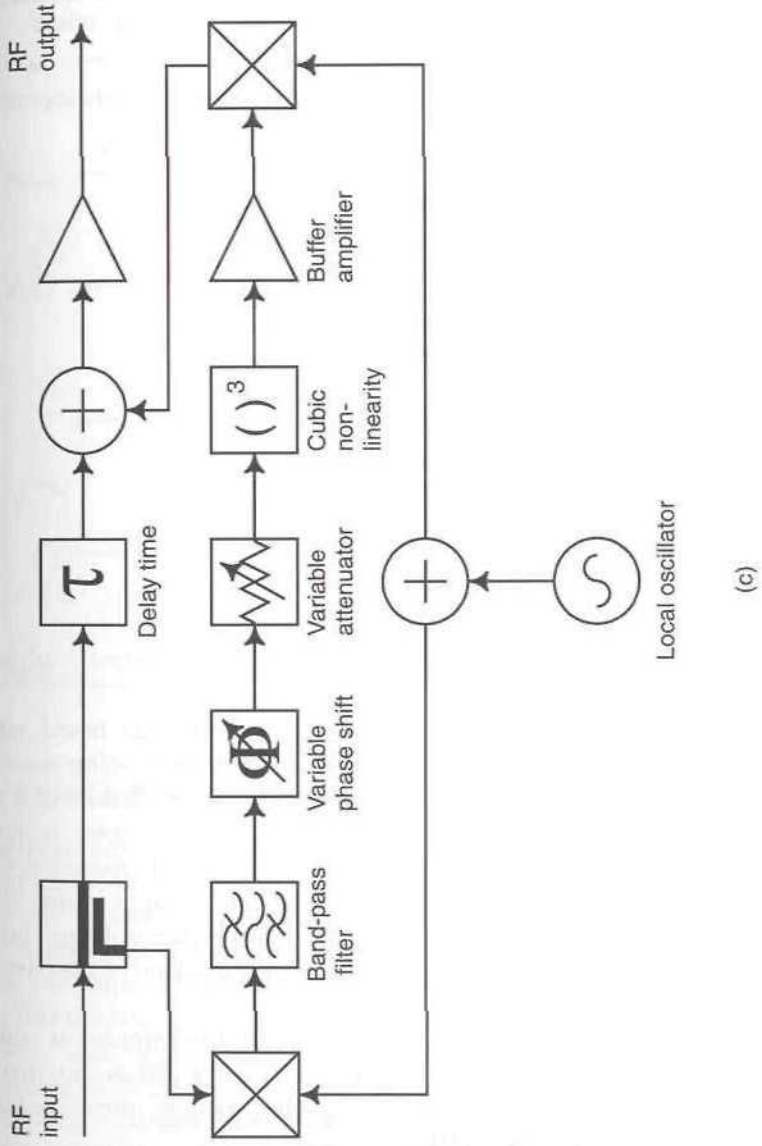


(a)



(b)

Figure 6.18 Alternative configurations for the use of an IF predistorter. Note that the amplifiers in each case are assumed to have a suitably narrow-band, bandpass characteristic.



(c)

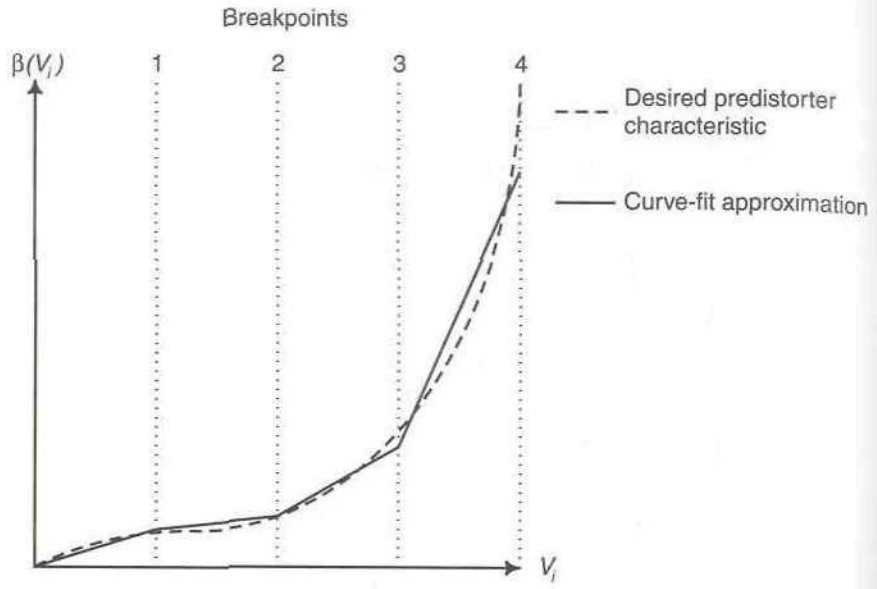


Figure 6.19 Example characteristic for a curve-fit predistorter.

shown in Figure 6.19, there are seven variables to correctly adjust (four segment gradients and three breakpoints).

Implementation of curve-fit predistorters may be based either on attenuator or amplifier configurations, with the former being most appropriate at lower RF or IF frequencies (due to the limited bandwidth of most differential amplifiers). An attenuator-based implementation is shown in Figure 6.20 for four segments [30]. The principle of operation involves using the diodes as switches with the various bias voltages setting the input level at which the corresponding diode switches. The bias voltages thus set the breakpoints and the resistor values determine the gradient of the corresponding segment.

The equivalent amplifier-based curve-fit predistorter is shown in Figure 6.21, with the principle of operation being similar to that of the attenuator version; the difference being that gain is introduced at each breakpoint rather than loss [31].

It is evident from these configurations that, as they stand, they are only applicable to unipolar signals, that is, the signal envelope. Operation with bipolar signals (i.e., the instantaneous RF signal) is possible by the introduction of additional diodes and biasing supplies.

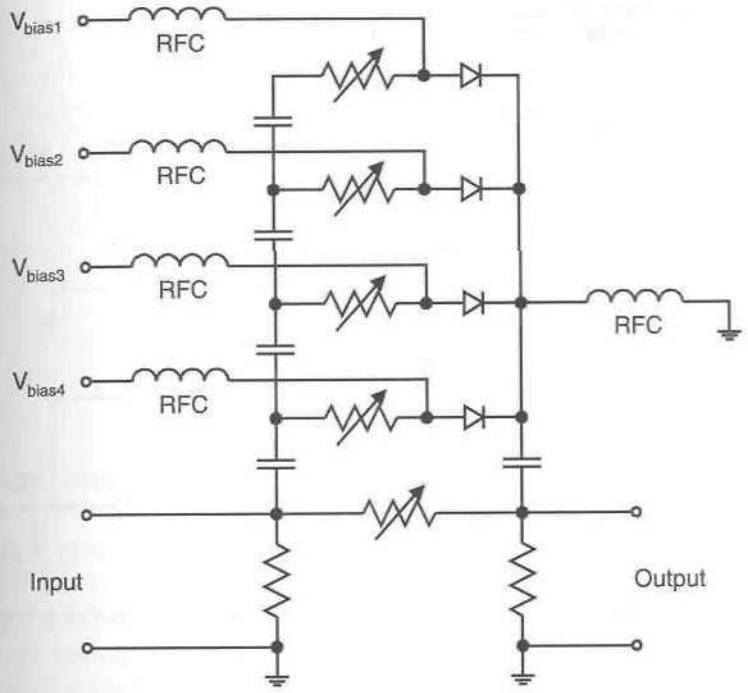


Figure 6.20 Attenuator-based curve-fit predistorter.

6.2.14 Digital IF Predistortion

This method of predistortion has seen relatively little activity to date due to the complexity and speed of the processing required to achieve a reasonable bandwidth. Predistortion is performed on a digitised version of the IF signal, in real time. The process may be made adaptive and is, in many respects, similar to the adaptive baseband predistortion methods described later in this chapter.

Horn and Egger [32] report an improvement in third-order IMD performance of 25 dB over a 45 MHz bandwidth, with both the bandwidth and IMD improvement being limited largely by the speed and resolution of the A/D and D/A converters. The efficiency of this method will also be limited by the large power consumption of such fast devices, unless the amplifier to be linearised is of a very high power level.

Improvements in device technology will, however, increase the attractiveness of this predistortion method with time.

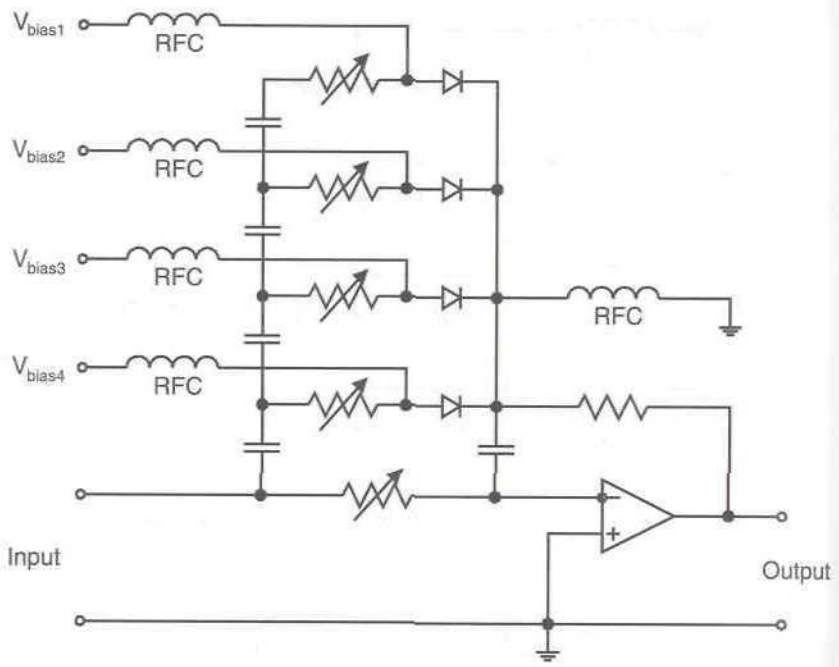


Figure 6.21 Amplifier-based curve-fit predistorter.

6.2.15 Complex Predistortion Techniques

Complex predistortion techniques attempt to compensate for both the AM/AM and AM/PM characteristics of an amplifier by operating on orthogonal versions of the input signal (either I/Q or amplitude and phase).

A block diagram of a Cartesian (quadrature) predistorter is shown in Figure 6.22 for operation either at IF or at RF. The input signal is split into quadrature paths using, for example, a quadrature hybrid coupler. Each path is then provided with a different amplitude nonlinearity to compensate for the characteristics in the Cartesian nonlinear model (see Chapter 2). The combined signals from both paths then form the input signal to the RF power amplifier.

The main disadvantage of this technique is in its difficulty of initial adjustment. This arises due to the AM/AM and AM/PM characteristics of the amplifier not being independent of each other; adaptive control can therefore be advantageous. The use of this technique has been reported in the literature, with a 7 dB reduction in the third-order IMD level over a 500 MHz bandwidth at 12 GHz having been achieved [33].

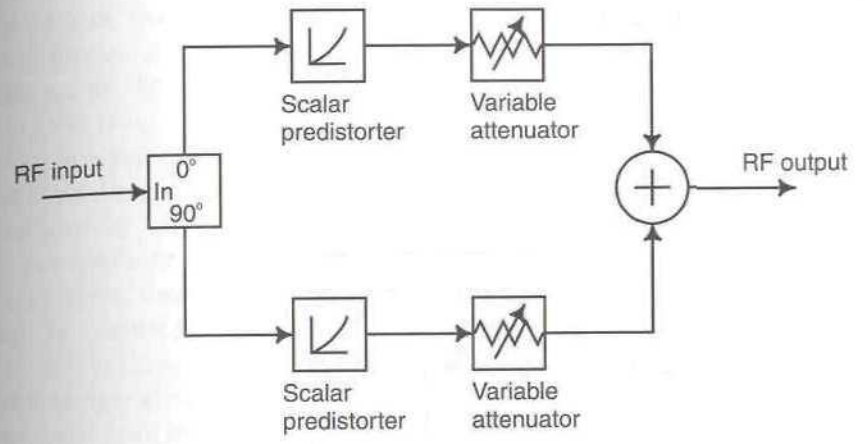


Figure 6.22 Cartesian predistorter.

It is possible to make this type of scheme adaptive and a version is described in the literature [34,35]. In this case, the two nonlinear predistorting functions are formed using operational amplifiers and linear four-quadrant multipliers and are adapted by a microcontroller, based on the amount of out-of-band energy detected at the amplifier output. Using this technique, an improvement in the third-order IMD performance of 15 dB was obtained.

The basic form of a polar (amplitude and phase) predistorter is shown in Figure 6.23. It operates on the amplitude of the signal in the same manner as the basic predistorters described earlier in this chapter, but provides linearisation of the AM/PM characteristic via the phase predistorter. This element provides a degree of phase modulation of the predistorted input signal proportional to the envelope amplitude of that signal.

The advantage of this method is that manual adjustment is relatively straightforward, due to the effectively separate controls for AM/AM and AM/PM distortion characteristics. Results from this technique have been reported in the literature by Egger *et al.*, with an improvement in the third-order IMD level of 25 dBc, over a 120 MHz bandwidth at 11.84 GHz, having been achieved [9].

6.2.16 Adaptive Control of Predistortion

The characteristics shown in Figure 6.5 indicate that the performance of a predistortion system depends upon the gain- and phase-matching accuracy of the two paths in the predistorter (particularly the gain matching). In order

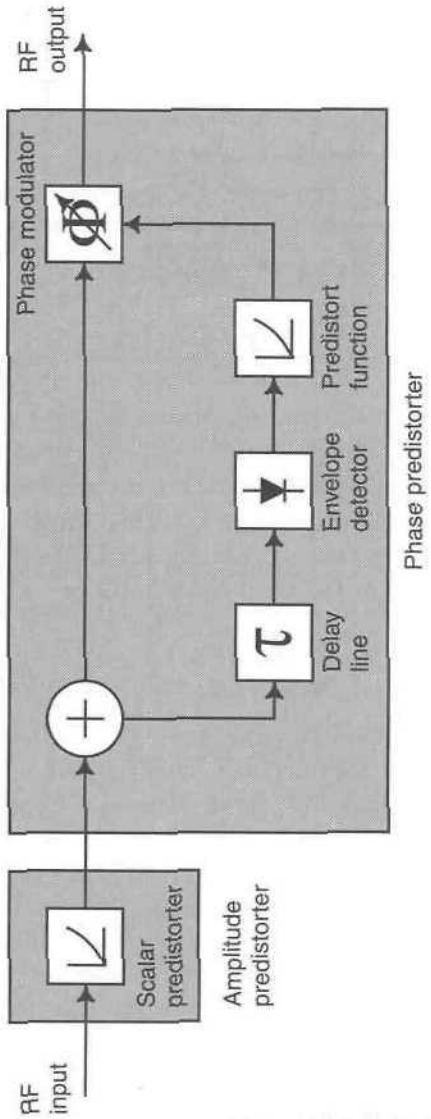


Figure 6.23 Polar predistorter.

to achieve and maintain optimum performance of the predistorter it is therefore necessary to ensure that this matching can be maintained over the lifetime of the amplifier and over its operational temperature range and dynamic range.

In many cases, the degree of improvement in IMD performance sought from predistortion is relatively modest, perhaps being in the order of 10B, and hence an initial setting up procedure is all that the product will require. In cases where more stringent specifications are placed upon the predistortion system, some form of adaptive control of the gain and phase matching will be required (in a similar manner to that described for feedforward systems in Chapter 5). In most cases, this control is only required to deal with temperature changes and component ageing and hence the adaption rate need only be slow.

Some adaption systems of this type have been described in the literature [6,36–38] and the basic approaches will be described below.

6.2.16.1 Predictive Temperature Compensation

In certain applications, notably fixed links and other non-mobile systems, it is possible to use a measured or predicted (or both) set of temperature characteristics to compensate for the change in performance of a predistorter with temperature and even aging. The format of such a system is shown in Figure 6.24 [38].

The system operates by measuring the temperature of the power amplifier (usually on the heatsink, close to the final stage device(s)), and utilising a predefined look-up table to relate the measured temperature to settings for the gain and phase controllers within the predistorter. This look-up table need not be 'intelligent', that is, it may simply be implemented by means of an EPROM or similar memory device, thus resulting in a relatively low complexity system. The addition of an elapsed-time clock would also allow aging effects to be taken into account, although it is questionable if this is necessary, as a better approach would be to use devices with a known, modest, aging characteristic. A degree of 'burn-in' is also usually necessary when using this approach, since significant changes in the amplifier's characteristic can occur in the first few hours of operation.

The coefficients for the look-up table are usually determined experimentally, with optimum performance being obtained (in the case of a large production run) by automatic measurement and optimisation over a number of samples of the final amplifier. It is, of course, possible to characterise each amplifier individually and to use appropriate (individual) look-up tables for each amplifier. This is usually prohibitive, in terms of test time, for most applications.

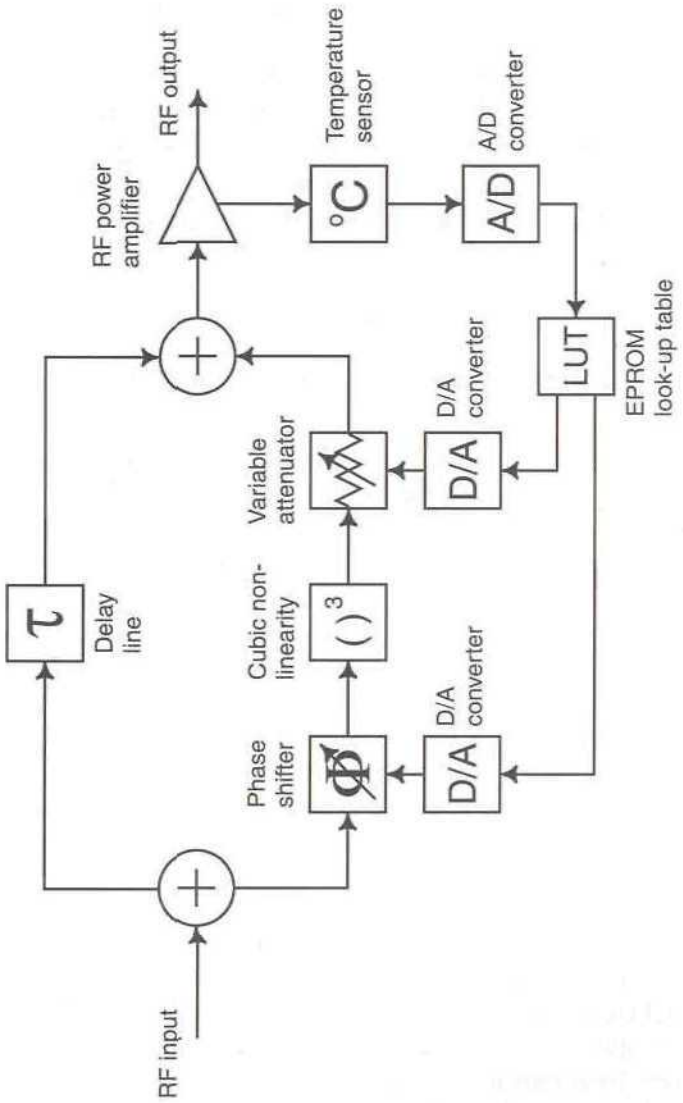


Figure 6.24 Predictive type of temperature compensation for a predistortion lineariser.

6.2.16.2 Feedback Control Based on Adjacent Channel Measurement

A closed-loop alternative to the above temperature-based compensation scheme is shown in Figure 6.25 (based on [38]). This system utilises a measurement of the adjacent channel energy, for both upper and lower adjacent channels, present in the output spectrum. This measurement is made using narrowband IF filters, which pass some or all of the adjacent channel energy (after downconversion to the IF at which the predistorter operates). This adjacent channel power is then summed and detected, with the resultant DC voltage providing an indication of the amount of signal energy which has spread into the adjacent channels, due to amplifier nonlinearity. An algorithm in the DSP (or micro-controller) then acts to minimise this energy by alternately incrementing the amplitude and phase controls and assessing if an improvement has been obtained.

This form of controller has a number of obvious disadvantages, particularly in 'generic' applications (i.e., applications where the signal bandwidth and characteristics are not known and can change). In this case, the IF filters would be required to adapt with the channel bandwidth, and this may be particularly problematic with multicarrier input signals. The subtraction of the input signal from the output spectrum, required in order to reduce the roll-off needed from the IF filters, may also itself require control, to ensure adequate performance over the lifetime (and temperature range) of the product. This process would introduce further (significant) complexity.

Finally, a purely RF-based implementation of this approach (i.e., an RF input and output system, with no IF stages) would be difficult to realise at other than low RF frequencies, due to the difficulty of fabricating suitably narrow-band adjacent channel filters.

6.2.16.3 Correlation Based Feedback Control

A further alternative control mechanism is shown in Figure 6.26, in which the amplitude control is provided by a direct correlation of the output spectrum IMD (after subtraction of the 'wanted' signals), with the third-order IMD provided by the cubic nonlinearity. The result of the correlation process feeds an integrator which controls the amplitude setting of the predistorter. This determines the relative level of the third-order components in the input signal to the nonlinear amplifier. The control scheme thereby acts to minimise the level of third-order IMD in the output spectrum.

This control scheme as shown does not permit control of the phase of the third-order component. In many systems this can be set up at

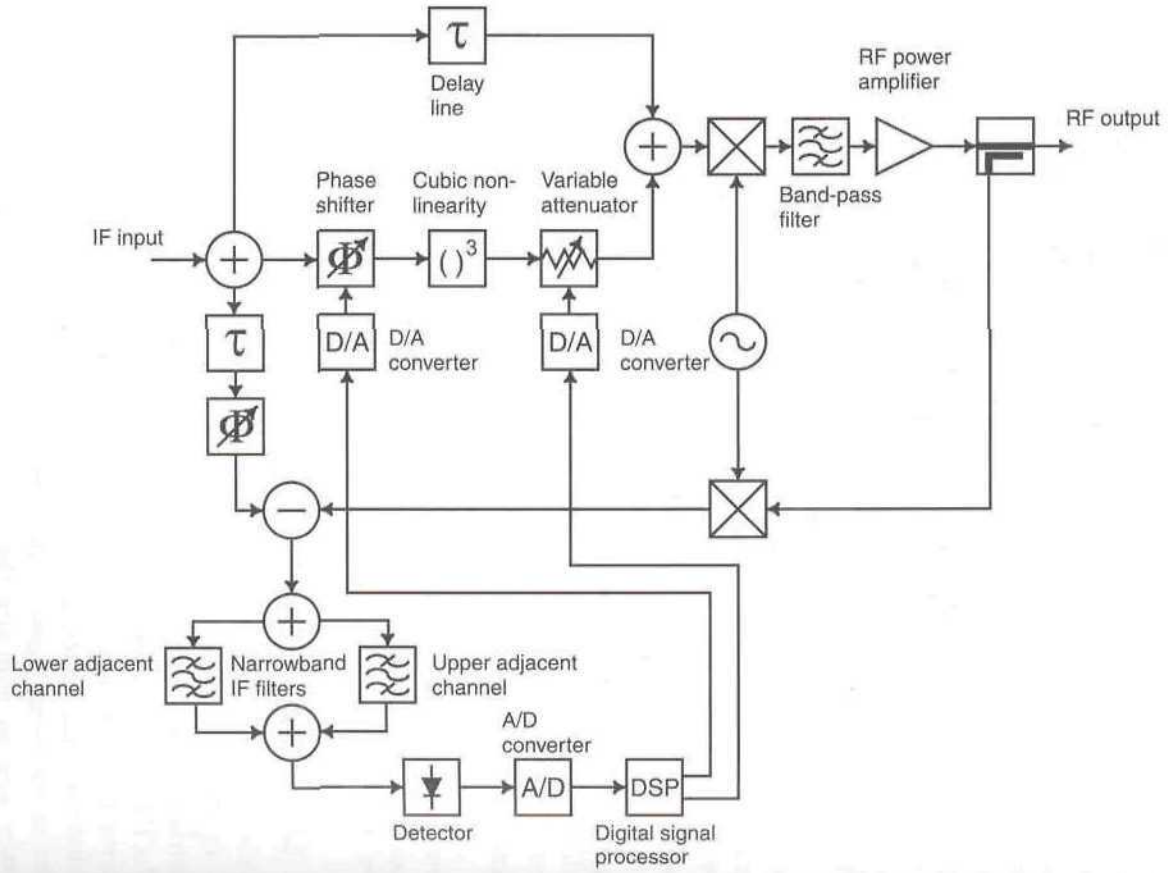


Figure 6.25 Predistorter control based on adjacent-channel measurement.

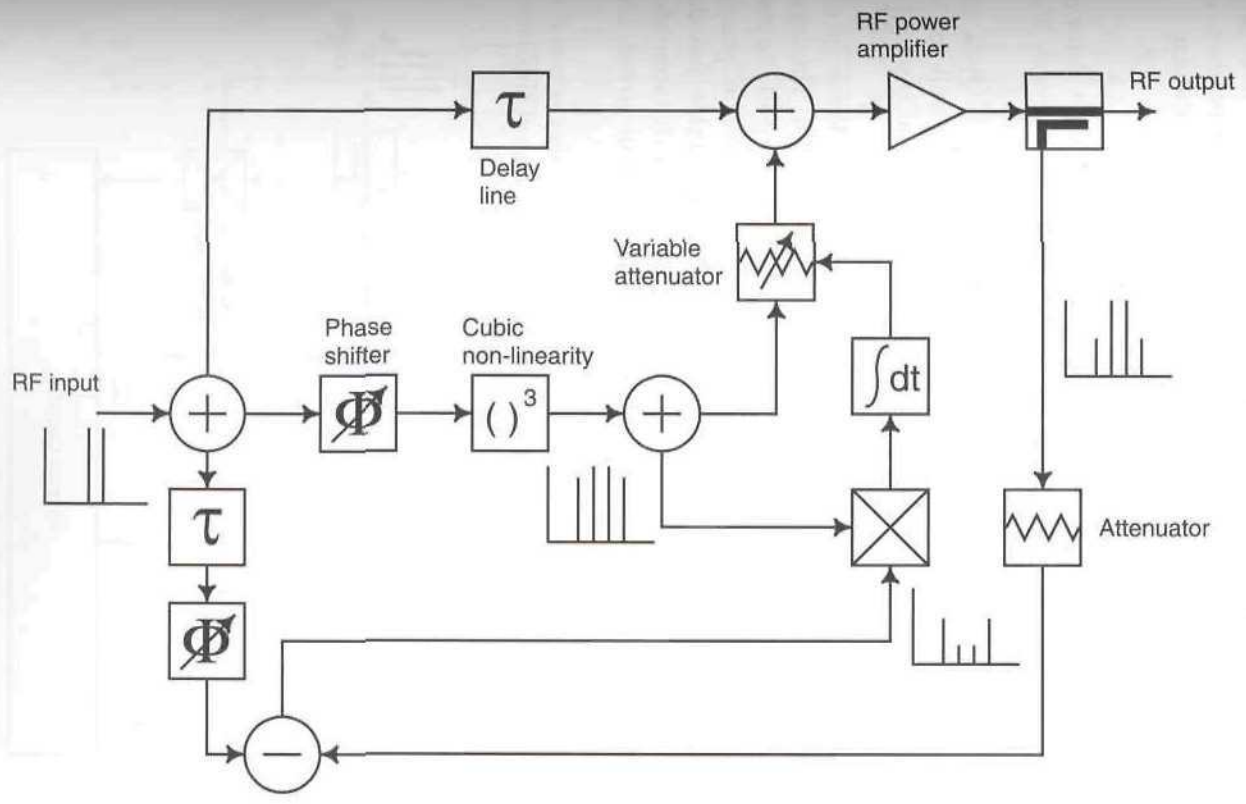


Figure 6.26 Correlation-based control of an RF predistortion lineariser.

manufacture, so long as the amplitude setting is controlled accurately, since relatively large errors in the phase setting can be tolerated whilst maintaining the relatively modest levels of correction expected of most third-order only predistorters. There is little point in suppressing the third-order products to much beyond the intrinsic level of the fifth- and higher-order products.

6.2.16.4 Feedback Control Employing Baseband Signals

The use of direct conversion to baseband, utilising either a local oscillator or envelope detection, may also be used to provide predistorter control. The basic schematic for either of these variants is shown in Figure 6.27.

The main disadvantages of the use of a local oscillator (LO) in the control system are first, that one must be available, which may well not be the case for a multicarrier input signal, for example, and, secondly, that the system is no longer 'stand-alone' or truly generic in nature.

A block diagram of a control scheme without LO-based downconversion, is shown in Figure 6.28 and its operation is as follows [39]. The input signal is sampled and downconverted as a reference for the control process, before being fed to the predistort subsystem. The degree of nonlinearity present in the predistort element is externally-controllable and this provides the required variation in order to compensate for the varying characteristics of the power amplifier with, for example, temperature, output power, and frequency.

The predistorted signal is then fed to the power amplifier where it is distorted in a roughly complementary manner to that of the predistortion.

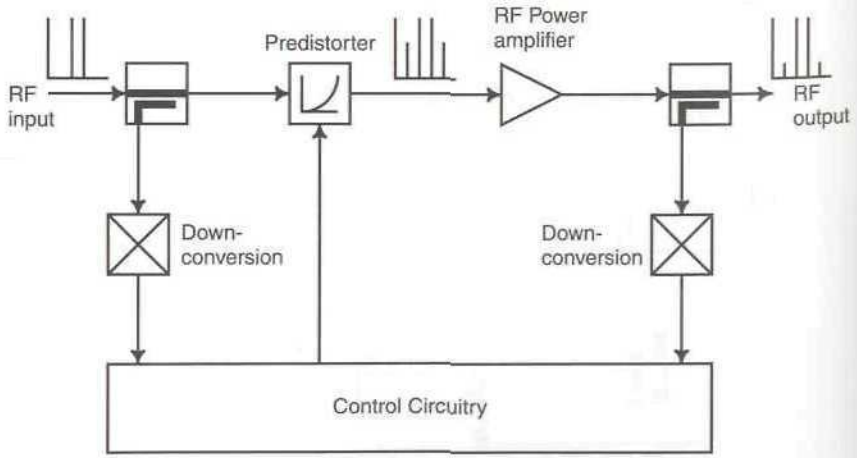


Figure 6.27 Predistorter control employing downconversion.

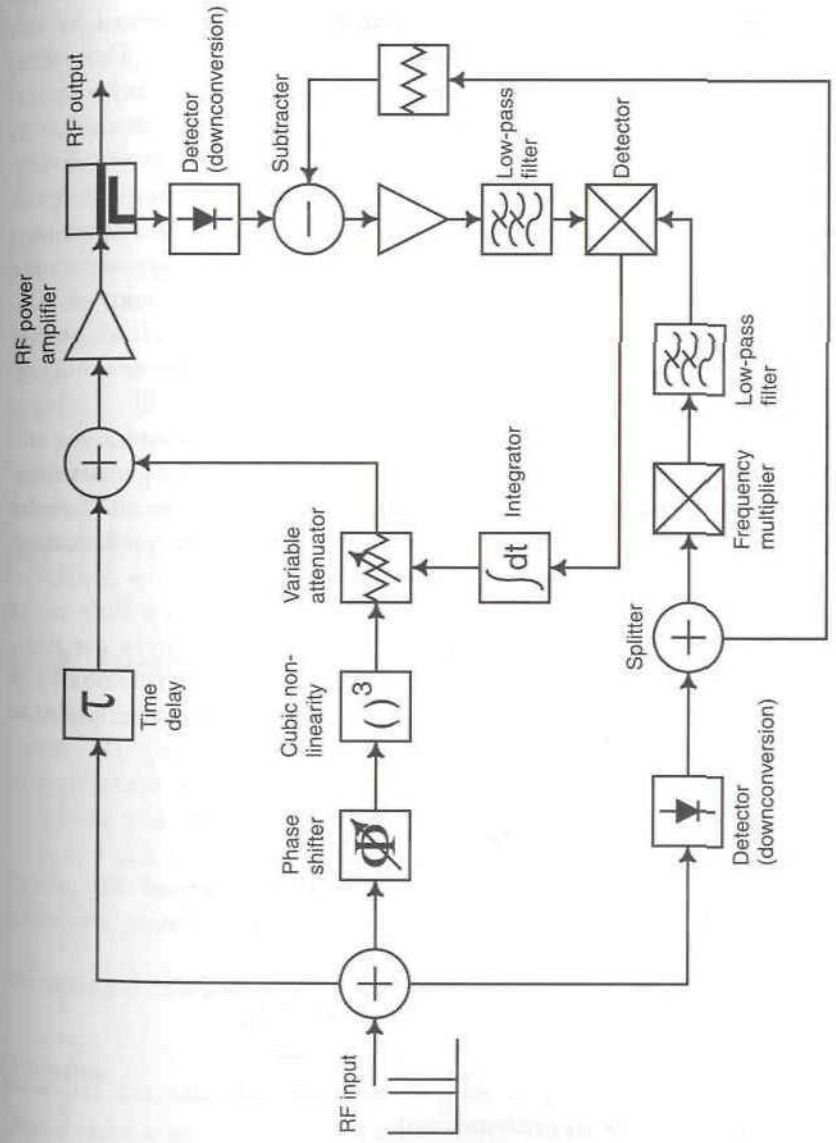


Figure 6.28 Predistorter control employing direct detection.

Finally, the output of the power amplifier is sampled prior to the signal being fed to the output. This output sample provides a measure of the distortion present in the output signal and this is used in conjunction with the input reference signal to determine the achieved linearity improvement. An error signal is generated, based on this improvement, and is minimised by the control system in order to optimise the linearisation process. This minimisation occurs dynamically in real-time and does not require any external input to indicate, for example, operating frequency, or power level.

Figure 6.28 also shows an outline of the functionality of the downconversion processing and control circuitry. The input and output signals are detected to extract the baseband envelope information together with a measure of the distortion present in the output spectrum. The dynamic range of the detected output signal is reduced by subtracting a copy of the downconverted reference (input) signal and this is fed to the output detector. The reference signal is frequency multiplied to generate a higher-order signal and this is fed to the other input of the detector.

The signal resulting from the detection process feeds an integrator and this in turn controls the degree of nonlinearity provided by the predistorter. This feedback process therefore attempts to minimise the third-order component in the output signal by controlling the third-order predistortion element.

Although a practical implementation of this scheme is a little more complex than that shown in Figure 6.28, this figure does convey the basic mechanism of operation. In addition, dual-variable control versions have been used to provide improved performance over a very wide temperature range.

6.2.17 Efficiency of Predistortion Techniques

Predicting the overall efficiency of a predistortion linearised RF power amplifier system is not a straightforward process. The system efficiency depends upon a number of factors:

1. The class of power amplifier to be linearised.
2. The power consumption of the predistorter.
3. The degree of output back-off necessary to meet the required specification (with predistortion).
4. The envelope characteristics of the signal to be amplified.

The efficiency of a nominally linear amplifier (class-A, -AB or -B) will decrease as the output power is reduced. The average efficiency of a

predistortion linearised linear amplifier, when amplifying an envelope varying signal, will therefore be significantly lower than its peak CW efficiency and almost certainly lower than its uncompensated efficiency (i.e., without predistortion), since the predistorter itself will consume some power. The degree of efficiency reduction will, however, depend primarily upon the envelope characteristics of the input signal. This apparent loss in efficiency must be traded against the requirement to meet a given linearity specification and the alternative methods of achieving that specification (e.g., back-off). The overall efficiency of the linearised system *for a given ACP or SVE requirement*, will therefore be higher (in most cases significantly so) than the unlinearised (backed-off) alternative (see Table 6.1 for an example).

In some systems, it may still be necessary to back-off the output power of the RF amplifier in order to meet the required linearity specification (despite the use of predistortion). In such cases, the degree of back-off required will also reduce the overall efficiency of the system.

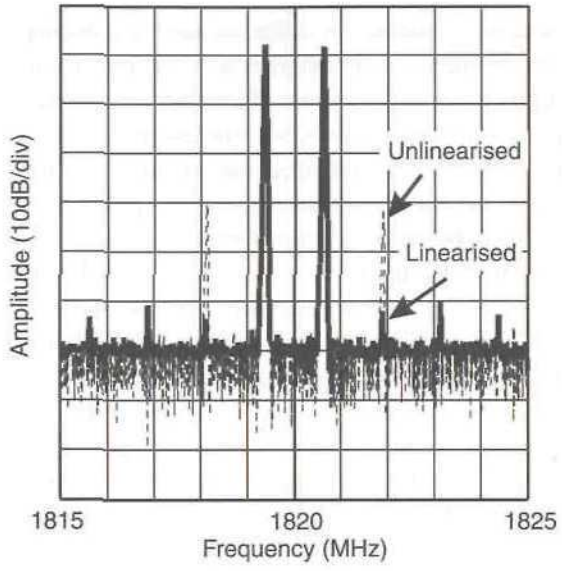
In all cases, the efficiency of the predistorted system should be higher (usually significantly so) than the efficiency of a backed-off amplifier producing the same level of intermodulation distortion (or signal vector error, if this is more important in a given application). Predistortion is therefore a useful efficiency enhancement technique where linearity is an issue in a system specification.

6.2.18 Example Performance of a Simple Predistorter

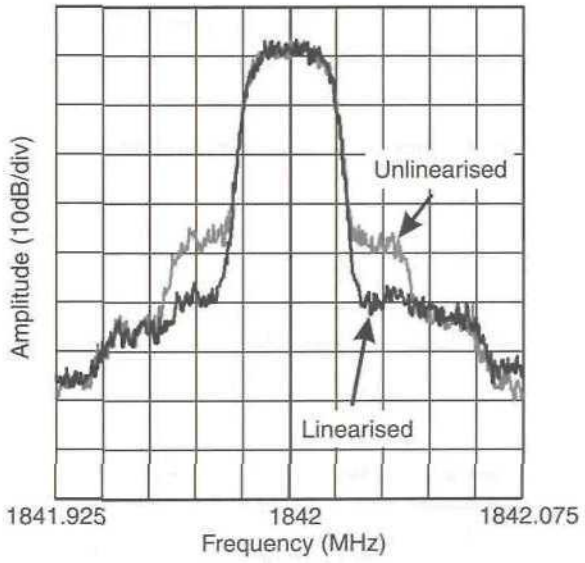
In order to illustrate some typical performance metrics for a simple (dioide-based) RF predistorter, this section presents some results for spectral improvement, efficiency gain, signal vector error improvement and mean power output improvement. In each case, the predistorter used is the same (i.e., not just the type of predistorter, the same physical circuit was used),

Table 6.1
Measured performance results from a simple broadband RF predistortion system for 16-QAM modulation (with a relaxed ACP requirement)

Parameter	Unlinearised	Linearised	Improvement
Signal vector error	8.2%	4.4%	46%
Efficiency	22%	29%	32%
Maximum mean output power whilst still meeting system mask	4.4W	8.8W	100%

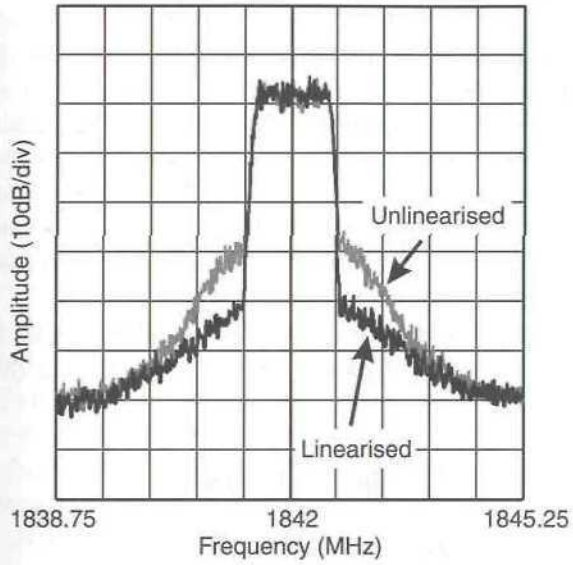


(a)

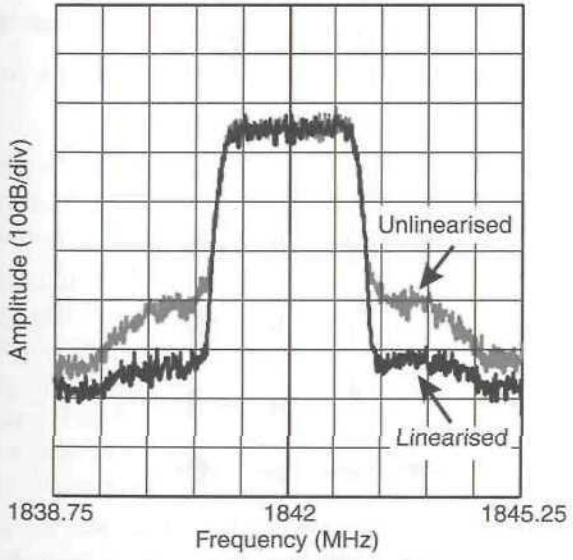


(b)

Figure 6.29 Results from a simple broadband RF predistortion system for a range of modulation formats and bandwidths, with 30W and 120W power amplifiers: (a) two-tone test (30W PA); (b) DAMPS IS-136 (120W PA); (c) IS-95 CDMA (120W PA); (d) Wideband CDMA (30W PA). © Wireless Systems International Ltd.



(c)

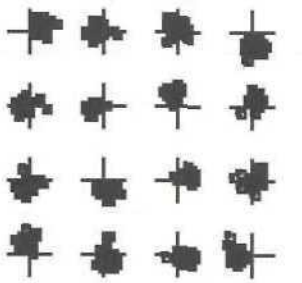


(d)

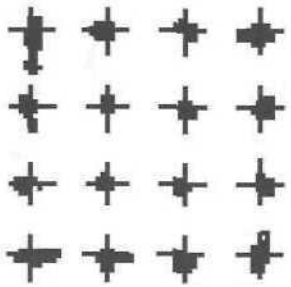
although a variety of power amplifier types and modulation formats are represented.

Some spectral results from the simple predistorter are provided in Figure 6.29 for a range of common modulation formats (including wideband CDMA). These results show the generic nature of the linearisation provided, both in terms of modulation format and RF occupied bandwidth, with channel bandwidths ranging from 25 kHz to 5 MHz being linearised using the same predistorter. Results from two different power amplifiers are illustrated (in this case high-power base station designs), to further underline the generic nature of the predistorter.

Adjacent channel performance is not the only system requirement which is degraded by PA nonlinearity. Signal vector error (SVE) is also



(a)



(b)

Figure 6.30 Results from a simple broadband RF predistortion system for 16-QAM modulation showing: (a) unlinearised constellation; (b) linearised constellation. © Wireless Systems International Ltd.

affected, although ACP is usually the more difficult requirement. Figure 6.30(a) illustrates the effect of severe nonlinearity on a 16-QAM constellation, with the outer points being constrained and the constellation being rotated due to the AM–PM conversion within the amplifier. The amplifier was operating in class-AB, but driven well into compression; the equivalent two-tone performance was around 15 dBc. Figure 6.30(b) illustrates the improvement afforded by the predistortion-based lineariser, with both of the above effects being corrected. The reduction in EVM in this case is close to 50%, as indicated in Table 6.1. This result is based around a satellite application, where ACP performance is relatively unimportant and SVE becomes the dominant requirement on linearity.

The use of a predistorter provides both an improvement in efficiency and also in output power capability, whilst still meeting a given ACP and SVE specification. In the case illustrated here, the efficiency improves by around one-third and the power output capability doubles. This clearly has a significant impact upon the cost and size of the product (whether base station or handset). Note that with a more stringent ACP requirement, efficiency improvements close to 50% have been observed.

6.2.19 Advantages and Disadvantages of RF/IF Predistortion

The principal advantages of RF/IF predistortion can be summarised as follows:

1. Simplicity of implementation. Predistortion systems are generally perceived to require few components and hence can be relatively low-cost.
2. Unconditionally stable. The basic predistortion technique is open loop and hence does not have the stability problems of feedback systems (e.g., RF feedback and Cartesian loop). Even where adaptive systems are used to adjust gain and/or phase values within the predistorter, their update rate is usually relatively slow (compared with the linearisation bandwidth) and hence stability is easy to achieve.
3. Ease of use at high frequencies. The relative simplicity of the predistortion technique and the absence of critical time-delay elements (in some implementations) means that it can be applied to microwave circuits without undue difficulty. Diode-based predistorters, in particular, can be used well into the microwave region.

4. Wide linearisation bandwidth. The instantaneous bandwidth of a predistortion lineariser is often very wide indeed and in this respect it is akin to a feedforward system.

The main disadvantages are:

1. Relatively modest linearity improvement. Simple predistortion techniques do not, in general, provide the same potential linearity improvements as, say, feedforward or Cartesian loop, although newer techniques, such as APL™, are capable of rivalling the more traditional alternatives in the area of linearity improvement.
2. Difficulty in dealing with many orders of distortion. Most simple predistortion systems can only reduce one or two orders of distortion (usually third- or third- and fifth-order) and indeed may increase the level of other, higher-order, products. Again, newer techniques are challenging this position and are demonstrating multi-order precorrection.
3. Poor performance with loading. In the case of relatively nonlinear amplifiers in particular, a predistorter configured for optimum performance at, say, full PEP will have reduced cancellation elsewhere in the input power range. Indeed, distortion enhancement, relative to that which might be anticipated from the amplifier IP3 specification, may be encountered at some points in the transfer characteristic. The system may, however, still meet the required emission mask, depending upon whether the mask is defined as a relative level or an absolute power. This disadvantage is generally eliminated by a well designed control technique.

Simple RF and IF predistortion techniques can therefore sometimes be seen as a useful 'helping hand' to aid other techniques (such as feedforward or Cartesian loop) in achieving a higher degree of absolute performance.

6.2.20 RF Predistorter Applications

Single-order polynomial predistortion is, in general, only suitable for use in systems with a predominantly single-order nonlinearity (usually third-order). For systems with more than one order of nonlinearity, a multiple-order polynomial predistorter would be required [40]. Polynomial predistortion, as demonstrated here, has application in satellite systems where the linearity improvement required is small, typically 5 dB to 15 dB [41]. The results show that this system can easily meet that type of specification.

The predistorter may also be of use in feedforward systems, to provide

initial linearity improvement for the main and/or error amplifiers, either resulting in an overall improvement in linearity, or a reduction in complexity of the feedforward amplifier, by allowing the use of only a single feedforward process where two might otherwise be required to meet a given specification.

6.3 Baseband Predistortion

The earlier sections in this chapter have dealt with the concept of providing predistortion at RF and/or IF, with the consequent compromises in attempting to model a complex nonlinear transfer characteristic with other nonlinear RF/IF elements. It is possible to provide these nonlinear modelling elements at baseband, where in many cases, they are simpler to realise. A simple block diagram illustrating the concept is shown in Figure 6.31.

It is much simpler, for example, to create the amplifier-based curve-fit predistorter of Figure 6.21, using a conventional audio operational amplifier. A complex version could also be constructed from two such units forming the I and Q nonlinearities in Figure 6.22, although the broadband audio phase-shift required to create the quadrature paths may prove difficult, without the use of a Hilbert transform filter in a digital signal processor (DSP).

If it is necessary to employ a DSP for the Hilbert transform, then it is a simple extension of the idea to employ the DSP in performing the predistorting function itself. This is an attractive concept since most modern radio transceivers employ some form of DSP in their baseband processing anyway.

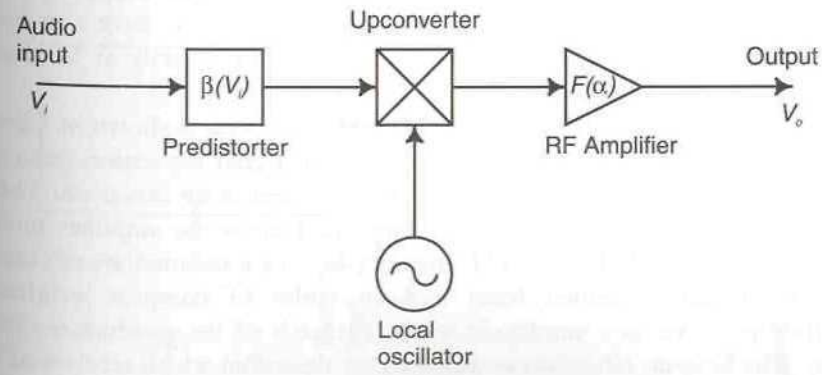


Figure 6.31 Schematic of a complete transmitter, employing simple baseband predistortion.

Having placed the predistortion process at baseband, and in a flexible processing device such as a DSP, it is then a further basic step to provide some form of dynamic adaption to the predistortion characteristic stored within the DSP. This then forms the full adaptive baseband predistortion system described below. The popularity of this form of predistortion is increasing as DSP processing power becomes progressively cheaper, smaller and, more importantly, more power efficient.

6.4 Adaptive (Baseband) Predistortion

6.4.1 Introduction

Adaptive predistortion has long been a promising technique for the narrowband linearisation of RF amplifiers. Like the Cartesian loop technique, it is a complete transmitter linearisation technique, as the input signal information is at baseband and the predistortion system incorporates the upconversion process. Another similarity to Cartesian loop lies in the common use of quadrature signals for both up- and downconversion, and for the actual predistortion process itself. There are thus two sets of coefficients, for the two quadrature channels, and both sets may be updated simultaneously to take account of gain and phase nonlinearities within the RF amplifier chain.

The use of digital signal processing to realise the predistorter element has been suggested by a number of the proposers of predistortion [42–44] and represents a useful solution to many of the updating and accuracy problems. A DSP device is capable of performing the rapid, complex multiplications and table look-up operations required if a predistortion system is to be employed successfully for significant IMD reduction. It is not, however, essential as adaptive analogue predistorters have also been constructed [34] and could be modified to operate directly at baseband frequencies.

A basic form of quadrature baseband predistorter is shown in Figure 6.32. The digital signal processor contains the signal separation (into in-phase and quadrature paths) and the complex weighting functions. These may be constructed in many ways, depending upon the amplifier model chosen (e.g. AM–AM, AM–PM, memory-less, or combinations of these), but are typically formed from look-up tables of complex weighting coefficients at various amplitude levels (for each of the quadrature channels). The look-up tables are accessed by an algorithm which receives as its input the fed-back downconverted RF output from the power amplifier. The coefficients in the look-up tables may then be up-dated in the light of the

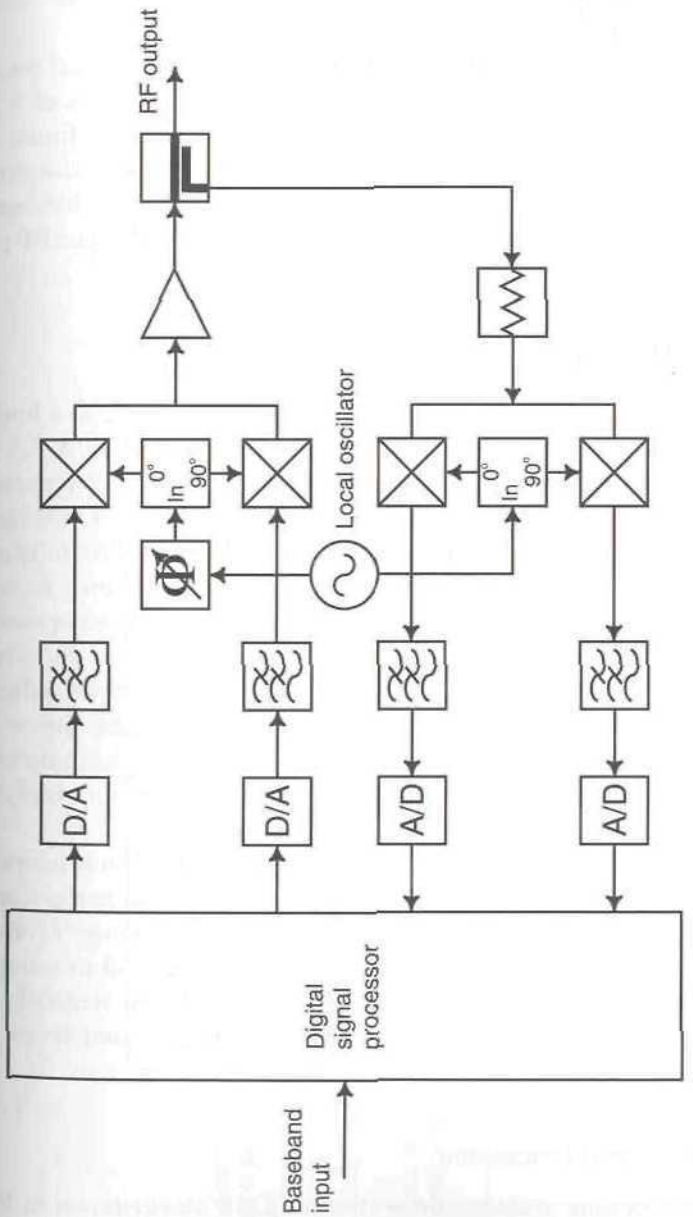


Figure 6.32 Hardware schematic of a complete transmitter, employing adaptive baseband predistortion using Cartesian components.

difference between this downconverted signal and the input signal. This *adaption algorithm* has been the basis of much research and is an obvious area where improvements in both speed and accuracy enhance the practicality of the adaptive predistortion technique.

Note that although an RF phase-shift is included in the local oscillator path in Figure 6.32, this function is often included in the form of a phase rotator at baseband (i.e., within the digital signal processing function) in many implementations. It performs a similar function to its counterpart in the Cartesian loop, namely that of compensating for the phase-shift between up- and downconversion caused by the finite time delay within the RF power amplifier.

6.4.2 Power-Efficiency

One of the major drawbacks of digital adaptive predistortion as a linearisation technique is that of the power consumption of the A/D, D/A, DSP and memory devices required in its implementation. Nagata [42] reports that 20Mbits of memory were required in his design, along with 13000 gates in dedicated DSP hardware and up to 2W in the ADCs and DACs. The total power consumption was therefore up to 4W which is significantly more than the power output of the RF amplifier in most handportable equipment.

It is therefore evident that the power-efficiency of an adaptive predistortion system will be poor until device technology is sufficiently advanced that the power consumption of the lineariser becomes a small fraction of that of the RF power amplifier. This situation is changing rapidly and a careful watch needs to be maintained to ensure that an optimum design methodology is chosen at any given point in time.

A recent advance in this area has been described by Sundstrom *et al.* [45], in which a dedicated predistorter ASIC is outlined. The performance of this device was shown to be very good over a broad range of channel bandwidths (up to 300 kHz) and the use of an ASIC helped to reduce the lineariser power consumption to sensible levels (roughly one-tenth of that of an equivalent clock-rate DSP device, whilst providing around seven *times* the channel bandwidth).

6.4.3 Generic Signal Processing

The signal processing architecture within the DSP block shown in Figure 6.32 may be summarised as shown in Figure 6.33. A baseband voice or data input signal is converted to a suitable sampled I/Q format of the desired modulation scheme by the voice or data coder. The I/Q signals then undergo

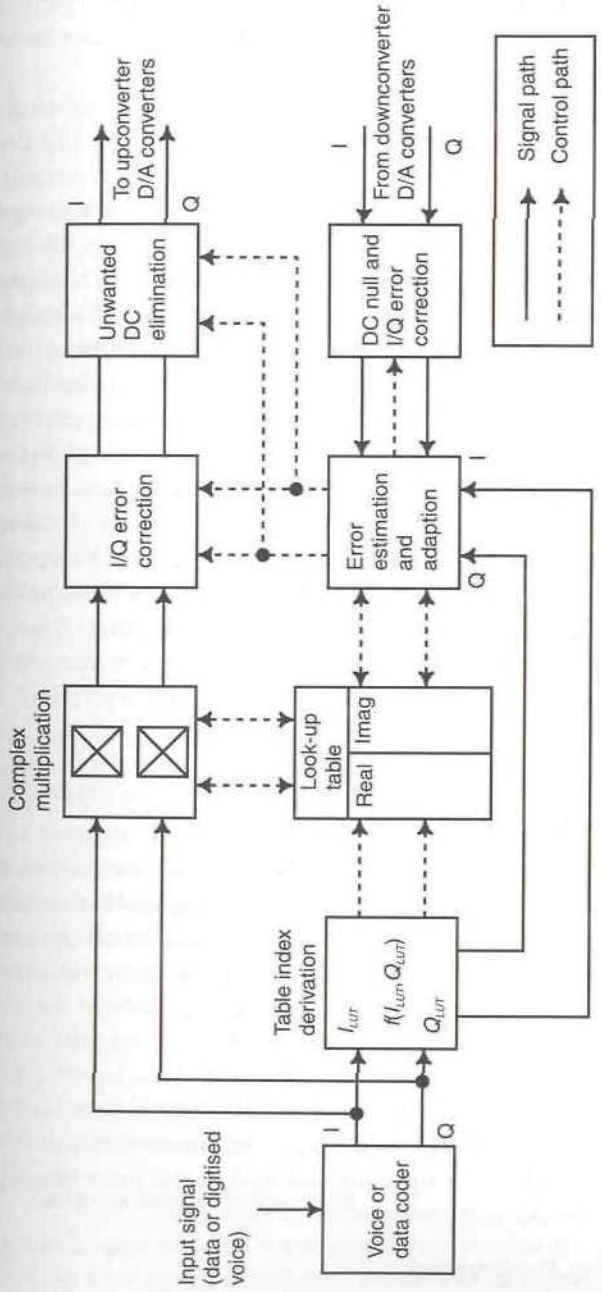


Figure 6.33 Signal processing architecture for an adaptive baseband predistortion system.

a complex multiplication with the relevant coefficients from the look-up table (or interpolated values derived from it) before DC elimination (if required) and adaptive error correction (based on the difference between the baseband I/Q and downconverted I/Q signals).

The look-up table is improved by knowledge of its performance derived from comparing the original I/Q sampled signals with the I/Q downconverted sample of the output signal. The accuracy and hence overall performance of the system is limited by the quality of this feedback signal. It is therefore essential to eliminate all known sources of error (e.g., DC offsets) or distortion from this signal; it is thus usually necessary to operate the downconverter mixers at a relatively low RF signal level. This signal level has been found experimentally to be around −20 dBm or lower for level seven diode-ring mixers (depending upon the desired final IMD specification).

The adaption block may also be used to aid in the removal of DC offsets in the main processing path, to eliminate carrier leakage for example, and also to optimise the sample timing. The use of discontinuous feedback (i.e., periodic rather than continuous updating of the look-up table coefficients) allows a higher level of correction gain to be applied than could be accommodated in, for example, Cartesian loop. This performance gain may be used either to allow a more nonlinear and hence more efficient RF power amplifier to be employed, or to provide a wider margin, for example, to accommodate production spreads for an equivalent level of performance.

6.4.4 Main Predistorter Elements

This section will outline the contents of the main blocks shown in Figure 6.33. There are a range of options in each case, in particular for the look-up table indexing mechanism and a complete description of all options would become unwieldy. A significant body of literature exists in this area and the reader is encouraged to refer to this for more detail.

6.4.4.1 Table Index Calculation

There are two main methods of look-up table indexing: Cartesian mapping and envelope mapping. These two methods result in markedly differing sizes and complexities for the look-up table and lead to the methods of: *mapping predistortion* and *complex gain predistortion*, respectively.

6.4.4.1.1 Mapping Predistorters

The mapping predistorter [42,46] is directly analogous to its continuous feedback equivalent, the Cartesian loop. In this case, two look-up tables are required, each of which is two-dimensional (i.e., a function of two variables: I_{IN} and Q_{IN}):

$$I_{OUT} = F_I(I_{IN}, Q_{IN}) \tag{6.15}$$

$$Q_{OUT} = F_Q(I_{IN}, Q_{IN}) \tag{6.16}$$

The advantage of this approach is that errors associated with the upconversion process can also be eliminated (e.g., DC offsets, I/Q imbalance) in a similar manner to Cartesian loop, with the limit of performance in this respect being determined by the downconversion (reference) path (i.e., DC offsets in the downconverter and its I/Q imbalance).

The disadvantages of this approach are in the potential size of the look-up table or the processing overhead required if interpolation is used and in a low speed of convergence. The low convergence speed results from the need to address all points in the I/Q complex plane before convergence can be completed.

The memory requirement (without interpolation) for a system with a table value for each potential pair of input samples (I, Q) is given by:

$$S_{LUT} = 2 \times (2^N)^2 \tag{6.17}$$

where S_{LUT} is the memory size required for the look-up table and N is the system resolution in bits. Thus, a system with 12-bit resolution would require 33554432 bits of memory, assuming that 12-bit RAM was used (the use of, for example, 16-bit RAM would result in 4-bits being unused in each location and hence require one-third more bits). Interpolation is therefore potentially extremely beneficial, however, this is at the expense of increased processing overhead and hence a slower convergence rate.

Practical two-tone test results of this type of predistorter have been presented by Mansell and Bateman [47]. They report an improvement in third-order distortion of 43 dB for a 25W VHF power amplifier and a 256 × 256 element look-up table. This excellent performance was achieved without full two-dimensional interpolation, but was at the expense of an extremely long adaption time.

6.4.4.1.2 Complex Gain Predistorters

This is a much more popular form of predistorter due to its greatly reduced look-up table size and the reduced processing power required for both interpolation and adaption [43]. These savings are achieved at the expense of lower predistortion accuracy and hence overall IMD suppression.

The principle of operation of a complex-gain predistorter is to attempt to force the predistorter and the RF PA to jointly ensure that a constant overall gain, G, is maintained at all power levels. This is performed using complex (I/Q) gain coefficients stored in the look-up table, which is now one-dimensional as it need only be indexed by the input signal envelope level.

A simple method of indexing the look-up table (reported in [47]) is:

$$I_{LUT} = \mathrm{INT}\left(K_{LUT} \times |x|^2\right) \tag{6.18}$$

where I_{LUT} is the look-up table index, K_{LUT} is a scaling factor and x is the complex baseband input signal. The 'INT' function returns the integer value of the product within the parentheses. The advantage of using $|x|^2$ as the input to the indexing process is that it provides a distribution of table entries which is linear with respect to the input signal power level. It can therefore provide a greater density of table entries close to PEP, which is where the greatest degree of nonlinearity will occur in a typical power amplifier, than can a distribution which is linear with input signal envelope. This situation is also achieved with a minimum amount of processing overhead as typical DSP devices have hardware multiplication functions built in.

Faulkner *et al.* [48] have used a one-dimensional look-up table based on polar co-ordinates and report a significant reduction in look-up table size and a simplification of the interpolation process. The tables employed were of the form:

$$R_{OUT} = F_R(R_{IN}) \tag{6.19}$$

$$\theta_{OUT} = F_\theta(R_{OUT}) \tag{6.20}$$

The phase look-up table was constructed as a function of R_{OUT} rather than R_{IN} since this allows the convergence of the two tables to be assessed separately.

The principal disadvantage of this approach is in the rectangular to polar conversion process which took 83% of the total processing time; this is obviously a significant processing overhead.

In the case of digital modulation schemes (without filtering), it is possible to eliminate interpolation altogether, as only the fixed constellation points need be considered. Even with high M-ary schemes, such as 256 QAM, a reasonable size of look-up table results (comparable with that of a general purpose system incorporating interpolation). Thus it is possible to provide good performance with a minimum of processing overhead.

Saleh and Salz [49] report on an early use of this type of system for a 64-QAM transmitter and on an earlier patent covering a similar operating principle [50]. Random access memory (RAM) is used to contain the predistorted values of each constellation point in I and Q components and an encoder accesses the relevant memory location based upon the input signal point at any instant in time. The corresponding predistorted voltage values are then accessed and converted to analog voltages via a D/A converter. These voltages are then upconverted, via a quadrature modulator, to provide the predistorted RF signal for the required symbol period.

The system is made adaptive by quadrature downconversion of a sample of the output signal and comparing the digitised version of this signal (in I and Q) with the input signal, resulting in an error signal which is used to compute the new predistortion value at that constellation point.

In the case of the predistorter ASIC mentioned above, the table index calculation was performed by means of a magnitude calculation on the I and Q input signals in a manner similar to that shown in Figure 6.34. This is an improvement on the (simpler) 'squared magnitude' method (illustrated by (6.18)) as it provides a more even distribution of table entries throughout the input signal range. The squared magnitude method, by comparison, provides a greater concentration of entries at high amplitudes and a relatively coarse representation at lower amplitudes. This in turn leads, in general, to a poorer adjacent channel performance for a given number of table entries.

An index calculation method of the form shown in Figure 6.34 is not an exact algorithm and will result in a small error relative to the correct square-root calculation. This error is reported to contribute less than 2 dB of adjacent channel degradation, relative to an exact (and far slower) calculation.

The basis of the algorithm is to index two small tables using the integer part of the $I_{LUT}^2 + Q_{LUT}^2$ calculation. These tables contain square-root values at the integer points and difference values between the points. The algorithm therefore effectively performs a linear interpolation of the square-root function by taking an exact square-root value at an integer point and adding to it a fractionally-weighted version of the difference between the integer value square-root and its neighbour. This difference is, however, stored in a ready-prepared table and therefore need not be calculated explicitly each iteration. The tables are reported to contain only eight entries each based on a 100 entry predistorter table.

6.4.4.2 Complex Multiplication and Look-Up Table Format

The heart of the predistortion system is the look-up table and complex multiplication function (Figure 6.35). The look-up table stores the set of

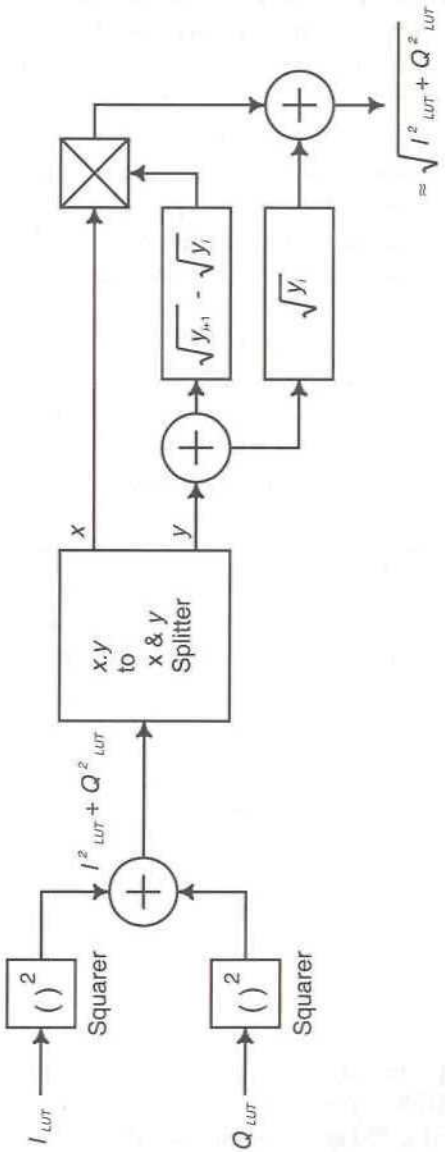


Figure 6.34 Magnitude-based table index calculation.

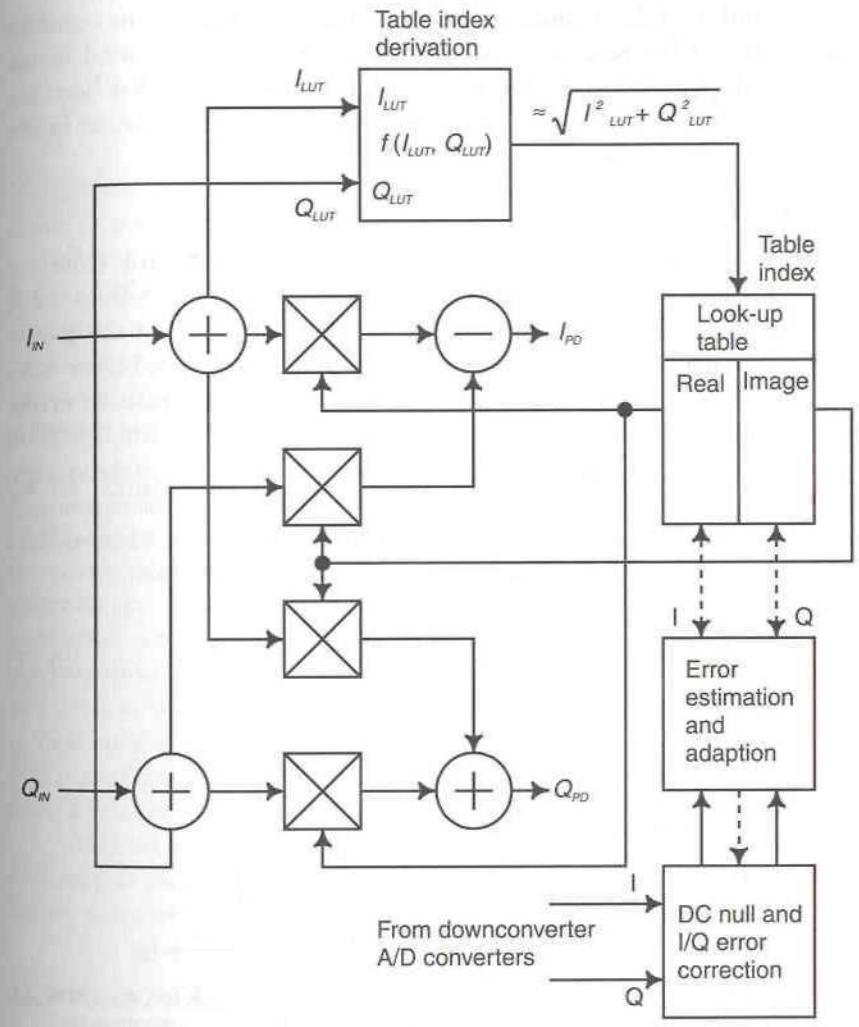


Figure 6.35 Complex multiplication within the predistorter.

weighting coefficients with which the input samples are multiplied in order to form the predistorted signal. The table itself will be updated or 'adapted' in order to minimise the level of adjacent channel interference (ACI) in the fed-back sample of the output spectrum. Note that any distortion of this feedback signal caused by, for example, nonlinearity in the downconverter, cannot be corrected and hence places a fundamental limit on the performance of the system. This is analogous to the situation also found with the Cartesian loop technique.

The look-up table is indexed based on the magnitude of the complex input signal and the resulting real and imaginary coefficients used in the complex multiplication process. The size of this look-up table has been the subject of much research [43], with typical numbers of entries being in the region of 64 to 128, based on ACI levels of −60 to −75 dBc.

6.4.4.3 DC Nulling and Error Correction

This section of the system corrects the complex output signals from the predistortion process itself (i.e., after weighting by the look-up table coefficients) for quadrature modulator errors and DC offsets in the modulator. A block diagram of the required functionality is shown in Figure 6.36.

The network shown in Figure 6.36 compensates for relative errors between the I and Q arms of the system (i.e., gain differences and deviation from perfect phase quadrature), but not for absolute gain errors in the system; these are more typically corrected within the look-up table. If the excess gain of the I signal over that of the Q signal is ΔG and the phase of the I signal is in error by $\Delta\phi$ (additional phase-shift over that of the Q arm), then:

$$G_{ERR} = \frac{\cos(\Delta\phi)}{1+\Delta G} \tag{6.21}$$

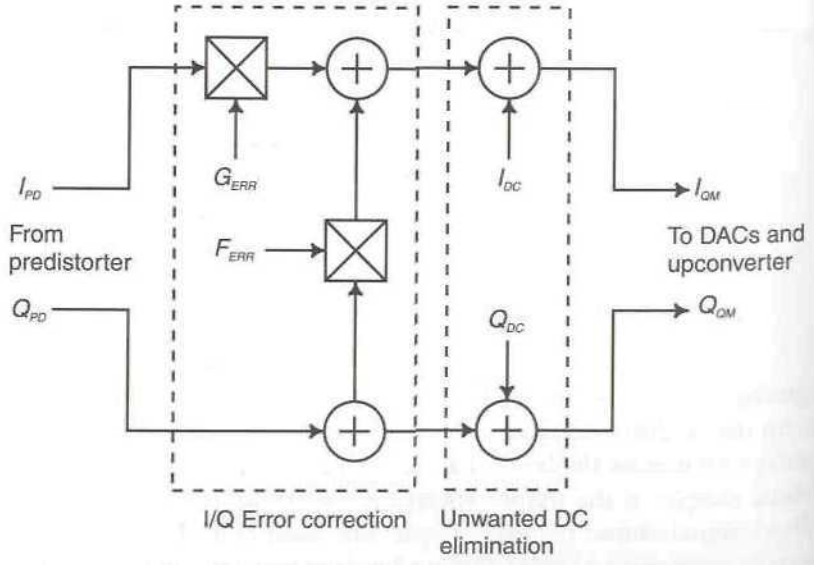


Figure 6.36 Correction of DC and gain and phase errors in the forward-path quadrature modulator.

and

$$F_{ERR} = \frac{\sin(\Delta\phi)}{1+\Delta G} \tag{6.22}$$

Using these equations, it is possible to calculate the required coefficients in order to compensate for the gain and phase errors in the modulator.

The I and Q channel DC injection coefficients are used to compensate for the DC offsets in the upconverter; these signals, if uncorrected, would lead to an unwanted carrier 'leakage' in the output spectrum.

Note that a similar mechanism to that of Figure 6.36 may also be employed in the feedback path as a post-correction system following the downconverter.

6.4.4.4 Error Estimation and Adaption

This is the process by which the look-up table entries are updated in the light of recent experience of the linearity performance of the transmitter. The reference on which this experience is based is obtained from the down-converted output signal sample and this is compared with the output of the predistortion function. Any differences are calculated, averaged (if required) and used to modify the table entries. The speed of this adaption process is critical to the acceptability of adaptive baseband predistortion systems in many applications and in particular the speed of adaption from switch-on (i.e., when the transceiver is cold) can be a problem in some applications.

This adaption process will undoubtedly improve in both speed and accuracy as more powerful signal processing devices become available and better adaption algorithms are developed.

6.4.5 Data Predistorters

This type of predistorter is a variant of those described above and attempts to restore the constellation of a data signal to that which the receiver would ideally like to see (in the absence of amplifier nonlinearity) by predistorting the input data constellation [49,51,52]. Such predistorters compensate for warping and clustering effects on the data constellation and therefore improve the eye openings at the maximum eye opening instants. They will thus improve the error vector magnitude of the amplifier output signal, but do not usually improve, intentionally, the adjacent channel performance or spectral purity of the transmitter output. They commonly employ look-up table based techniques to form the predistortion function and operate in a broadly similar manner to their conventional equivalents. A key disadvan-

tage with this form of predistorter is that it is generally modulation format specific. In a multimode terminal, therefore, a range of different look-up tables would be required, probably with differing numbers of entries.

6.4.6 Practical Issues

6.4.6.1 Sources of Error

Many practical considerations will affect the performance of a digital baseband predistortion system; these include:

1. Temperature changes: results in the literature [53] suggest that for π/4-QPSK, the increase in adjacent channel power is approximately proportional to the cube of the temperature change (without adaptive correction).
2. Drive level: the same simulation study suggests that the increase in adjacent channel power is approximately proportional to the change in the drive level (again for π/4-QPSK, without adaptive correction).
3. Carrier leakage: here, again without adaptive correction, the increase in adjacent channel power is approximately proportional to the carrier leakage level (π/4-QPSK) [54]. As in the case of Cartesian loop, this results from DC errors/leakage around the loop.
4. Gain error: in this case, the relationship between adjacent channel power and the error in the gain between the I and Q paths (for π/4-QPSK) is of the form:

$$S = k_1(\Delta G)^{1.3} \tag{6.23}$$

where S is the power of the adjacent channel leakage signal, k_1 is a constant of proportionality and ΔG is the gain difference between the two paths.

5. Phase error: in this case the relationship (for π/4-QPSK) is of the form:

$$S = k_2(20\log_{10}(\Delta\phi)) \tag{6.24}$$

where S is the power of the adjacent channel leakage signal, k_2 is a constant of proportionality and $\Delta\phi$ is the phase difference between the two paths.

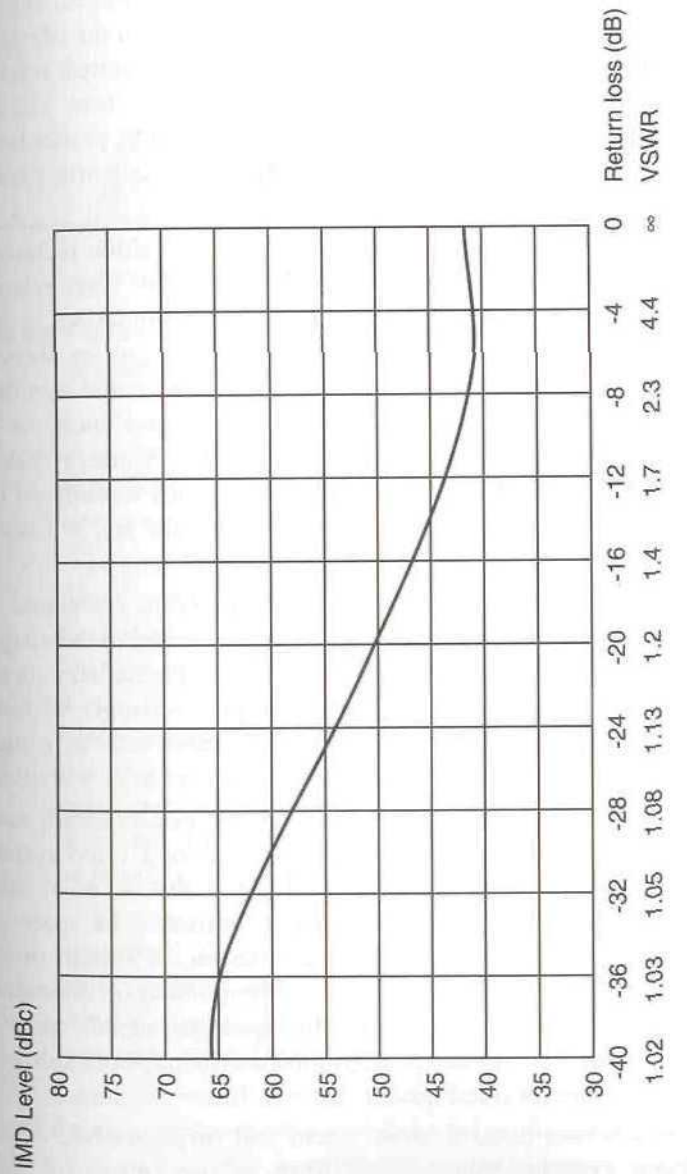


Figure 6.37 Simulated effect of output VSWR on an adaptive baseband predistortion transmitter [55]. © RF Design 1996. N.B. All system elements (bar the PA) were assumed ideal; the IMD suppression (from the predistorter) was 32 dB, with the unlinearised performance being 32 dBc.

6. Antenna load mismatch: a poor antenna VSWR is common in mobile communications applications, particularly for handportable transmitters. In the case of a Cartesian loop transmitter, it can cause instability due to the gross change in phase-shift around the loop which it causes. In the case of a predistortion system the effects will cause a gain and/or phase error in the downconverted reference vectors used in the adaption feedback loop. This in turn will affect the IMD suppression of which the system is capable, particularly as the VSWR changes are likely to occur more frequently than the update rate of the adaption system.

   An analysis of this issue, using computer simulation techniques, is presented by Gloeckler [55] and indicates that even relatively modest VSWR figures can cause a severe degradation in high-linearity transmitters. The results of this simulation are shown in Figure 6.37, and provide an indication of the potential severity of the problem. Clearly the actual results will depend upon the type and design of predistorter used; however, this figure provides an indication that the VSWR issue must be carefully considered in an adaptive predistortion design (more so than in, say, a Cartesian loop design).

7. Physical shock: this is a further practical problem associated with handheld transmitters in particular. The update rate of the adaption scheme must be sufficiently regular that any momentary physical disturbances (resulting in small gain and phase changes within the transmitter) can be adaptively neutralised before causing a significant degradation in the performance of the system as a whole.

8. Effect of external transmitters: the coupling of unwanted, nearby transmitter signals into the transmit antenna of a radio system is becoming increasingly common with the proliferation of mobile terminals and the increased use of base station mast space. The response of a predistortion transmitter to such a system must be such as not to cause undue additional intermodulation distortion at its output, nor to generate any other spurious signals, such as a 'mirror image' of the interfering signal on the opposite side of the *transmitter carrier (see Figure 6.38).*

9. Effect of power control: most recent and future mobile communications systems require some form of power control in the transmitter to reduce the interference in 'cellular'-style system scenarios. These requirements vary from a small number of relatively large steps (in PMR systems, for example) to a very large number of small steps for frequency hopped or direct-sequence CDMA systems (such systems do require linear transmitters in most cases).

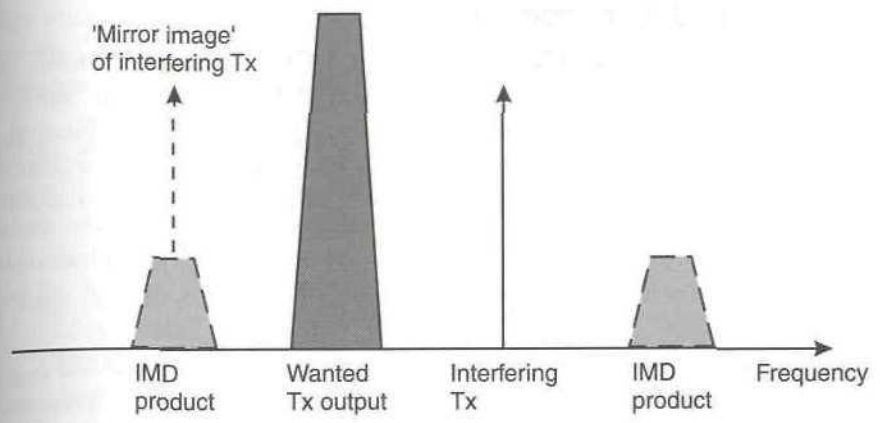


Figure 6.38 Possible effects of an external transmitter coupling into the antenna of a predistortion transmitter.

In general, the gain and phase characteristics of a notionally linear amplifier will change with power level and hence the predistortion system must be able to either adapt to these changes, or have a set of look-up tables, one for each power level. This latter option could, in many cases, prove extremely costly in memory provision. This would be particularly evident in the case of mapping predistorters which have large memory requirements anyway. These memory requirements would be increased by a factor equal to the number of power levels required, which could be six or more even in PMR systems.

The use of predistortion to linearise highly nonlinear power amplifiers (e.g., class-C) can prove very difficult even without power control and will certainly prove more difficult with power control, due to the very significant gain and phase changes with operating point (power level) in such systems.

It should be noted that the above relationships (used in 1–5) were derived for an amplifier with a well-behaved nonlinear characterisitc (i.e., a square-law region at very low powers, a linear region and a high-power saturation region). The device used was a BLU98 (Philips) power transistor. These results may not hold true for devices with highly nonlinear characteristics or those with nonlinearities in the (notionally) linear region.

6.4.6.2 Sampling-Rate Considerations

A baseband predistorter should be capable of eliminating all orders of distortion, provided that its baseband bandwidth is sufficient, and is adequately sampled (i.e., sampled at the Nyquist rate or greater). Since the processing in most baseband predistortion systems occurs in quadrature components, it is only necessary to sample each channel at half the full-bandwidth rate. In other words:

$$f_S = B_{IMD} \tag{6.25}$$

where f_S is the required sampling rate (the same for both channels) and B_{IMD} is the full bandwidth of the RF channel including all significant orders of distortion (see Figure 6.39).

Reduction of the sampling rate below this level will reduce the level of suppression of higher-order IMD products (possibly to zero) and hence may be undesirable. Recent work has, however, reported a reduction in sampling rate of 30% to 35% without a significant degradation in performance [47].

Compensation for the loop delay must be provided in order to allow the comparison of the correct input sample with its corresponding feedback sample. This may be provided partly within the software (for integer multiples of the sampling interval) and partly by skewing the A/D or D/A sampling clocks to provide a fractional sample delay.

The loop delay is dominated by the reconstruction filters required in the D/A processing and hence variations in the delay of these filters (e.g., with temperature) will lead to errors in the feedback comparison process (and possibly even instability in extreme circumstances). It is therefore necessary to remove the effects of this delay by calibration, although the use of oversampling converters can significantly reduce the analog filtering

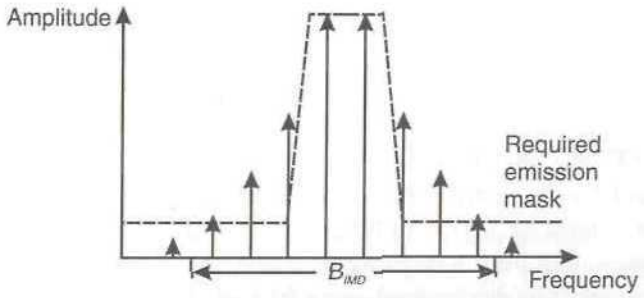


Figure 6.39 RF bandwidth (and hence required baseband bandwidth) for a predistortion system.

requirements of the D/A part of the system and they also have a fixed (known) delay. This calibration process may therefore be able to be eliminated in future systems.

6.4.6.3 Quantization and Resolution

In any digital system operating on essentially analog signals, the issue of quantization must be addressed. It is necessary to ensure that an appropriate resolution is used at all points in the system, from the sampling of the input signals, through the various internal calculations, to the output digital to analog conversion.

This issue has been investigated in detail by Sundstrom *et al.* [45], based on a complex gain predistorter and a class-AB power amplifier. Five areas in which quantization effects can have a bearing were studied:

1. source (input) signal;
2. calculation of the table index;
3. the look-up table coefficients;
4. the complex multiplications/additions performed in the predistortion process itself;
5. the modulator error correction process.

The results of this investigation for $\pi/4$-DQPSK with a filter roll-off (α) of 0.35 are summarised in Tables 6.2 and 6.3. Table 6.2 indicates the effect on adjacent channel interference of a range of standard word-lengths used in various parts of the predistorter system and Table 6.3 presents the minimum required word-length for an ACI of −70 dBc. Note that this latter table takes no account of the additional dynamic range required in a practical system in order to compensate for gain variations in the forward path, particularly the RF power amplifier chain. For example, a gain error of ±20% would require an extra 0.6 bits of resolution to be added to the look-up table and predistorter output. The two quantization effects in the output correction networks in the I and Q channels (quantization in the output signals and quantization of the error coefficients) require further additions in word-length to account for the gain error, DC offsets and modulator errors. Even with these additions, however, the limiting wordlength is still that of the upconverter error correction coefficients at 12.9 bits.

6.4.6.4 Calibration Issues

In addition to the loop delay effects mentioned above, there are a number of other parameters which must be removed or suppressed by a periodic calibration procedure; these are:

Table 6.2
Comparison of adjacent channel interference levels (simulated) from various parts of a predistortion system for π/4-DQPSK with $\alpha = 0.35$

Section of predistorter	Predistorter output word-length (bits)	Adjacent channel Interference ratio (dBc)
Look-up table	8	−54.0
	10	−65.4
	12	−77.2
Predistorter output	8	−52.0
	10	−64.1
	12	−76.2
Upconverter error correction coefficients	8	−48.3
	10	−60.4
	12	−72.5

Table 6.3
Wordlengths required in different parts of a predistortion system in order to obtain a −70 dBc adjacent channel interference level for π/4-DQPSK with $\alpha = 0.35$

Section of predistorter	Predistorter output word-length (bits)
Input	10.3
Look-up table	11.5
Predistorter output	11.0
Upconverter error correction coefficients	12.9

1. DC offsets.
2. Carrier feedthrough (partially caused by DC offsets, but also present in the upconverter).
3. Feedback loop phase-shift.
4. Cartesian errors (I/Q path gain and phase errors). These can be reduced to a level dictated by the quadrature feedback path. The errors in this path are usually dominated by the downconverter.

Measurement of the feedback path DC offsets can be achieved by powering-down the RF PA (to provide isolation between the forward and feedback paths) and reading the relevant values in the I and Q paths using the feedback A/D converters. Once these offsets are known, the PA can then be powered-up and the upconverter (including D/A) offsets can be removed. The degree of suppression which can be achieved by this process is generally limited by the difference in the DC level produced by the downconverter when operated with and without an input signal.

The feedback loop phase-shift can be measured (if required) by means of a low-level pilot-tone generated within the DSP. A similar mechanism could also be used to remove Cartesian errors, by monitoring the residual image level of a known (e.g., internally generated) signal.

6.5 Postdistortion Linearisation

In essence, the postdistortion approach to amplifier linearisation (Figure 6.40) is very similar to predistortion, with the obvious exception that the linearising element must be capable of handling the full power capability of the PA output stage [56]. It is therefore inherently less desirable as a linearisation technique due to the restriction this places on the range of available nonlinear elements which may be used in the postdistorter. In addition, the inevitable losses in this block have a significant effect on the overall efficiency of the amplifier system. Indeed in many cases it may be more efficient to provide linearisation by back-off rather than to use postdistortion.

An alternative, and rather more interesting, approach is to place the postdistorting element in the receiver rather than in the transmitter. The signal levels will then be significantly lower and the losses (other than if it is placed in the antenna path) much more acceptable. It also has the advantage in a system employing some form of base station, that the complexity of the mobile or handportable terminal is reduced, at the expense of the base station receiver. Complexity at the base station is usually much more acceptable due to the increased size and cost which is tolerable in that part of the system.

The approach proposed [57] to allow adjustment of the postdistorters at the base station involves measuring the level of distortion present in a

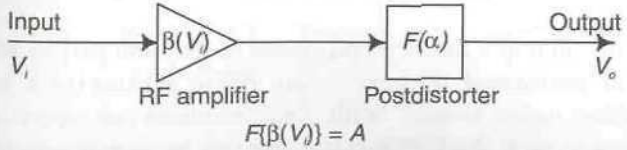


Figure 6.40 Schematic of an RF amplifier using postdistortion linearisation.

vacant channel and adjusting the parameters of the postdistorters in the two adjacent channels to eliminate the distortion present in the vacant channel. When this vacant channel is in use, it should then, theoretically, enjoy almost interference-free reception.

There are a number of inherent disadvantages with this approach, however, and these will severely limit the available performance. First, the degree of IMD reduction which can be achieved will be small (< 15 dB according to simulation results in the above paper), largely due to the lack of knowledge of the original signal in the transmitter (i.e., the input signal) which would be required for good adaption. A linearisation scheme would therefore still be required in the mobile transmitter, although its performance need not meet the full required mask (e.g., > 60 dBc IMD ratio). This largely removes the complexity reduction advantage in the mobile terminal.

The system also relies on vacant channels being available in order to adapt its postdistorters. This may be a problem in a heavily-loaded system and the non-regular frequency allocations of many systems may make the use of this technique difficult (the nearest channels may be far from adjacent).

Finally, the requirement for an adaptive postdistorter for each channel in the base station will result in significantly increased complexity in that element of the system. This is only worthwhile if the cost/size/power consumption of the mobile terminal can be significantly reduced.

Adaptive channel equalisation has also been suggested as a method of removing amplifier nonlinearities in the receiver [58,59].

A more worthwhile use for postdistortion is in receiver systems, and in particular receiver front-end amplifiers. Here nonlinearities in the front-end amplifier may be eliminated with little or no sacrifice in noise performance, which would be inherent in practice in the use of most other linearisation schemes in this application (e.g., predistortion). The postdistorter could also double as a predistorter for the first mixer, hence providing a dynamic range enhancement for the complete receiver front-end up to (and possibly including) the IF stages. In this way a relatively simple, broadband, high dynamic range receiver front-end could be created.

6.6 IMD Cancellation at the Antenna

An interesting technique has been proposed in [60] and [61] in which IMD cancellation is performed by appropriate phase shifting of a number of power amplifier input signals, with final combination occurring at the transmit antenna itself. In [60] it is shown that an amplifier consisting of two parallel (identical) combined amplifiers, with a phase difference of 45° at

the input and output results in a reduction of one of the third-order intermodulation products, as follows.

Consider the two amplifier outputs to be represented by current sources of the form:

$$I_1 = \sum_{m,n=-\infty}^{\infty} I_{m,n} \exp[j\{(m\omega_2 + n\omega_1)t + n\theta\}] \tag{6.26}$$

$$I_2 = \sum_{m,n=-\infty}^{\infty} I_{m,n} \exp[j\{(m\omega_2 + n\omega_1)t + m\theta\}] \tag{6.27}$$

where $\theta = \pi/4$, $I_{m,n} = I^*_{-m,-n}$ and ω_1 and ω_2 are the angular frequencies of the two tones.

The resulting third-order output can then be shown to be:

$$I_{IM3} = I_{\pm 2,\mp 1} \exp[\pm j\{(2\omega_2 - \omega_1)t + \phi\}] \tag{6.28}$$

where $\phi = 0, \pi/2$. Note that no $(2\omega_1 - \omega_2)$ term appears. Alternatively, a similar combination of two amplifiers, but with a phase difference of 90° at the input and output, can be shown to result in a single IMD of the form:

$$I_{IM3} = 2I_{\mp 1,\pm 2} \exp\left[\pm j\left\{(2\omega_1 - \omega_2)t + \frac{3\pi}{2}\right\}\right] \tag{6.29}$$

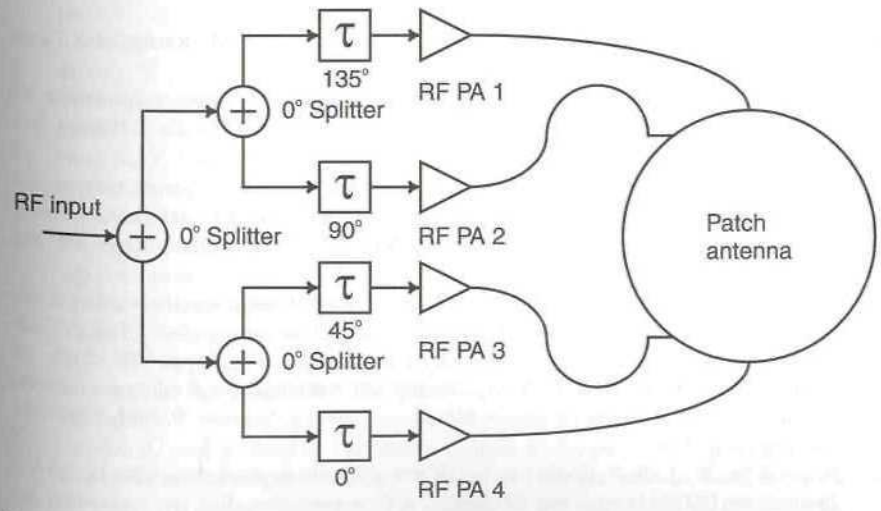


Figure 6.41 Cancellation-based linearisation of multiple identical power amplifiers by combination in a patch antenna (from [61] © IEE 1999).

Combining these two techniques results in both IM products being cancelled.

One method of providing the required combination is using a patch antenna, as shown in Figure 6.41. In this arrangement, the input splitting is performed using zero degree splitters (e.g., Wilkinson dividers), followed by delay lines of an appropriate length to provide the phase differences required. The combination process, after power amplification, is then performed by the patch antenna, with the feed points chosen to be 45° apart around the arc of the circular patch. This results in a planar structure with circular polarisation. Using this method, an IMD reduction of some 23 dB is reported for a two-tone test with a tone spacing of approximately 200 kHz.

References

1. Kenington, P. B., "Achieving High-efficiency in Multi-carrier Base-station Power Amplifiers," *Microwave Engineering Europe*, September 1999, pp. 83–90.
2. Bura, P., D. Gelerman, and P. Ntake, "Linear solid-state power amplifier for 64 QAM radios," in *Proc. of the International Conference on Communications*, Vol. 1, Seattle, Washington, 7–10 June, 1987, pp. 1.1.1–1.1.5.
3. Locatelli, G. P., L. Ricco, and M. Nannicini, "Microwave linear power amplifier with micromodule technology," in *Proc. of the International Conference on Communications*, Vol. 1, Seattle, Washington, 7–10 June, 1987, pp. 1.3.1–1.3.6.
4. Kenington, P. B., K. J. Parsons, and D. W. Bennett, "Broadband linearisation of high-efficiency power amplifiers," *Proc. of the Third International Mobile Satellite Conference*, Pasadena, California, June 1993, pp. 59–64.
5. "RF Power Amplifier with Signal Predistortion for Improved Linearity," US Patent No. 5,570,063, 29 October 1996.
6. Nojima, T., and T. Konno, "Cuber predistorter linearizer for relay equipment in 800 MHz band land mobile telephone system," *IEEE Trans. on Vehicular Technology*, Vol. VT-34, November 1985, pp. 169–177.
7. Aihara, S., T. Nishiumi, Y. Fujiki, and S. Fukuda, "GaAs FET power amplifiers as substitutes for TWT amplifiers in a multi-level QAM digital radio system," in *Proc. of the International Conference on Communications*, Vol. 1, Seattle, Washington, 7–10 June 1987, pp. 1.2.1–1.2.5.
8. Kumar, M., J. Whartenby, and H. Wolkstein, "Predistortion linearizer using GaAs dual-gate MESFET for TWTA and SSPA used in satellite transponders," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-33, December 1985, pp. 1479–1488.
9. Egger, A., M. Horn, and T. Vien, "Broadband linearisation of microwave power amplifiers," *Proc. of the 10th European Microwave Conference*, Warsaw, Poland, September 1980, pp. 490–494.
10. Schwarz Jr., W. J., R. P. Slade, and J. J. Kenny, "Radio repeater design for 16 QAM," *Proc. of the IEEE International Conference on Communications*, Vol. 1, June 1981, pp. 13.5.1–13.5.7.
11. Sun, J., B. Li, and Y. W. M. Chia, "A novel CDMA power amplifier for high efficiency and linearity," *Proc. of the 50th IEEE Vehicular Technology Conference, Fall 99*, Vol. 4, September 1999, pp. 2044–2047.
12. Sun, J., B. Li, and M. Y. W. Chia, "Linearised and highly-efficient CDMA power amplifier," *IEE Electronics Letters*, Vol. 35, No. 10, 13 May 1999, pp. 786–787.
13. Yamauchi, K., *et al.*, "A novel series diode linearizer for mobile radio power amplifier," *Proc. of the IEEE Symposium on Microwave Theory and Techniques*, 1996, pp. 831–834.
14. Yu, C. S., W. S. Chan, and W. L. Chan, "1.9GHz low loss varactor diode predistorter," *IEE Electronics Letters*, Vol. 35, No. 20, 30 September 1999, pp. 1681–1682.
15. Maeda, M., H. Masato, H. Takehara, M. Nakamura, S. Morimoto, and H. Fujimoto, "Source second harmonic control for high efficiency power amplifiers," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-43, December 1995, pp. 2952–2957.
16. Katz A., and S. S. Moochalla, "Nonlinearity generator using FET source-to-drain conductive path," US Patent no. 5,038,113, 6 August 1991.
17. Dorval, R., "MMIC linearizers for C and Ku-band satellite applications," *IEEE MTT-S Workshop on Advances in Amplifier Linearization*, Paper 7, 8 June 1998.
18. Nazarathy, M., *et al.*, "Predistorter for high frequency optical communication devices," US Patent No. 5424680, 1995.
19. Blauvelt, H., *et al.*, "Predistorter for linearization of electronic and optical signals," US Patent No. 4992754, 1991.
20. Grebliumas, J., *et al.*, "Microwave predistortion linearizer," US Patent No. 5523716, 1996.
21. Namiki, J., "An automatically controlled predistorter for multilevel quadrature amplitude modulation," *IEEE Trans. on Communications*, Vol. COM-31, May 1983, pp. 707–712.
22. Nojima, T., T. Murase, and N. Imai, "The design of a predistortion linearization circuit for high-level modulation radio systems," *Proc. of GLOBECOM '85*, 1985, pp. 1466–1471.
23. Bernardini, A., and S. DeFina, "Analysis of different optimization criteria for IF predistortion in digital radio links with nonlinear amplifiers," *IEEE Trans. on Communications*, Vol. 45, No. 4, April 1997, pp. 421–428.
24. Huang, W., and R. E. Saad, "Residual second order intermodulation suppression in third order distortion generators," *IEEE MTT-S Conference Digest*, Vol. 3, June 1998, pp. 737–740.
25. Imai, N., T. Nojima, and T. Murase, "Novel linearizer using balanced circulators and its application to multilevel digital radio systems," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 37, No. 8, August 1989, pp. 1237–1243.
26. Huang, W., and R. E. Saad, "Novel third-order distortion generator with residual IM2 suppression capabilities," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 46, No. 12, December 1998, pp. 2372–2382.
27. Maas, S., *Nonlinear Microwave Circuits*, Artech House, Norwood, MA, 1996.
28. Li, S. M., D. Jing, and W. S. Chan, "Verification of practicality of using the second-harmonic for reducing IMD," *IEE Electronics Letters*, Vol. 34, No. 11, 28 May 1998, pp. 1097–1098.
29. Charas, P., and J. Rogers, "Improvements in on-board TWTA performance using amplitude and phase predistortion," *Proc. of the IEE International Conference on Satellite Communication System Techniques*, Birmingham, UK, April 1975, Vol. CP 126, pp. 270–280.
30. Abuelma'atti, M. T., "Synthesis of non-monotonic single-valued function generators

without using operational amplifiers," *International Journal of Electronics*, Vol. 51, No. 6, 1981, pp. 803–809.

31. Chua, L. O., and F. Ayrom, "Designing non-linear single op-amp circuits: a cookbook approach," *International Journal on Circuit Theory and Applications*, Vol. 13, No. 3, July 1985, pp. 235–268.
32. Horn, M., and A. Egger, "Design and performance of microwave predistortion networks using digital circuits," *Proc. of the 14th European Microwave Conference*, Belgium, September 1984, pp. 549–554.
33. Kumar, M., J. Whartenby, and H. Wolkstein, "Predistortion linearizer using GaAs dual-gate MESFET for TWTA and SSPA used in satellite transponders," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-33, December 1985, pp. 1479–1488.
34. Stapleton, S. P., and F. C. Costescu, "An adaptive predistorter for a power amplifier based on adjacent channel emissions," *IEEE Trans. on Vehicular Technology*, Vol. 41, No. 1, February 1992, pp. 49–56.
35. Stapleton, S. P., and J. K. Cavers, "A New Technique for Adaption of Linearizing Predistorters," *Proc. of the 41st IEEE Vehicular Technology Conference*, St. Louis, Missouri, 19–22 May 1991, pp. 753–758.
36. Namiki, J., "An automatically controlled predistorter for multi-level quadrature amplitude modulation," *IEEE Trans. on Communications*, Vol. COM-31, May 1983, pp. 707–712.
37. Ashida, H., Y. Suzuki, I. Umino, and N. Tozawa, "C-band 100 Watts GaAs FET amplifier for digital radio," *Proc. of the IEEE International Conference on Communications*, Seattle, Washington, June 1987, pp. 1/23–1/27.
38. Nannicini, M., P. Magni, and F. Oggionni, "Temperature controlled predistortion circuits for 64 QAM microwave power amplifiers," *Proc. of the IEEE Symposium on Microwave Theory and Techniques*, 1985, pp. 99–102.
39. Kenington, P. B., S. J. Gillard, and A. E. New, "An Ultra-Broadband Power Amplifier using Dynamically-Controlled Linearisation," *Proc. of IEEE International Symposium on Microwave Theory and Techniques*, Anaheim, California, USA, Vol. 1, 13–19 June 1999, pp. 355–358.
40. Bernardini, A., and S. DeFina, "Analysis of different optimization criteria for IF predistortion in digital radio links with nonlinear amplifiers," *IEEE Trans. on Communications*, Vol. 45, No. 4, April 1997, pp. 421–428.
41. Katz, A., "TWTA Linearization," *Microwave Journal*, April 1996.
42. Nagata, Y., "Linear amplification technique for digital mobile communications," *IEEE Vehicular Technology Conference*, San Fransisco, USA, May 1989, pp. 159–164.
43. Cavers, J. K., "A linearizing predistorter with fast adaption," in *Proc. of the 40th IEEE Vehicular Technology Conference*, Orlando, Florida, 6–9 May 1990, pp. 41–47.
44. Wright, A. S., and W. G. Durtler, "Experimental performance of an adaptive digital linearized power amplifier," *IEEE Trans. on Vehicular Technology*, Vol. 41, No. 4, November 1992, pp. 395–400.
45. Sundstrom, L., M. Faulkner, and M. Johansson, "Quantization analysis and design of a digital predistortion lineariser for RF power amplifiers," *IEEE Trans. on Vehicular Technology*, Vol. 45, No. 4, November 1996, pp. 707–719.
46. Minowa, M., M. Onoda, E. Fukuda, and Y. Daido, "Backoff improvement of an 800 MHz GaAs FET amplifier for a QPSK transmitter using an adaptive nonlinear distortion canceller," in *Proc. of the 40th IEEE Vehicular Technology Conference*, Orlando, Florida, 6–9 May 1990, pp. 542–546.

47. Mansell, A., and A. Bateman, "Practical implementation issues for adaptive predistortion transmitter linearisation," *IEE Colloquium Digest no. 1994/089: Linear RF Amplifiers and Transmitters*, April 1994, pp. 5/1–5/7.
48. Faulkner, M., T. Mattsson, and W. Yates, "Adaptive linearisation using predistortion," *Proc. of the 40th IEEE Vehicular Technology Conference, Orlando, Florida*, 6–9 May 1990, pp. 35–40.
49. Saleh, A. A. M., and J. Salz, "Adaptive linearization of power amplifiers in digital radio systems," *The Bell System Technical Journal*, Vol. 62, No. 4, April 1983, pp. 1019–1033.
50. Davis, R. C., and W. Boyd, "Adaprive predistortion technique for linearizing a power amplifier for digital data systems," U.S. Patent No. 4,291,277, Issued 22 September 1981.
51. Karam, G., and H. Sari, "A data predistortion technique with memory for QAM radio systems," *IEEE Trans. on Communications*, Vol. 39, February 1991, pp. 336–343.
52. Jeon, W. G., K. H. Chang, and Y. S. Cho, "An adaptive data predistorter for compensation of nonlinear distortion in OFDM systems," *IEEE Trans. on Communications*, Vol. 45, No. 10, October 1997, pp. 1167–1171.
53. Faulkner, M., M. Johansson, and W. Yates, "Error sensitivity of power amplifiers using predistortion," *Proc. of the 41st IEEE Vehicular Technology Conference*, St. Louis, Missouri, 19–22 May 1991, pp. 451–456.
54. Ren, Q., and I. Wolff, "Effect of demodulator errors on predistortion linearisation," *IEEE Trans. on Broadcasting*, Vol. 45, No. 2, June 1999, pp. 153–161.
55. Gloeckler, R., "Use adaptive digital predistortion to simulate a linearization system," *RF Design*, October 1996, pp. 24–40.
56. Prochazka, A., P. Lancaster, and R. Neumann, "Amplifier linearisation by complementary pre- or post-distortion," *IEEE Trans. on Cable Television*, Vol. CATV–1, October 1976, pp. 31–39.
57. Stapleton, S., and L. Quach, "Reduction of adjacent channel interference using postdistortion," in *Proc. of the 42nd IEEE Vehicular Technology Conference*, Vol. 2, Denver, Colorado, May 1992, pp. 915–918.
58. Falconer, D. D., "Adaptive equalization of channel nonlinearities in QAM data transmission systems," *Bell System Technical Journal*, Vol. 57, No. 7, September 1978, pp. 2589–2611.
59. Benedetto, S., and E. Biglieri, "Nonlinear equalization of digital satellite channels," *IEEE Journal of Selected Areas in Communications: Special Issue on Digital Satellite Communications*, Vol. SAC–1, No. 1, January 1983, pp. 57–62.
60. Hayashi, H., and M. Muraguchi, "A low distortion technique for reducing transmitter intermodulation," *IEICE Trans. Electron.*, Vol. E80-C, No. 6, 1997, pp. 768–774.
61. Yang, C. C., W. L. Chan, and W. S. Chan, "Low IMD, antenna combined, circularly polarised, integrated antenna," *IEE Electronics Letters*, Vol. 35, No. 17, 19 August 1999, pp. 1401–1403.

7

Linear Transmitters Employing Signal Processing

7.1 Introduction

The linearisation systems to be considered in this chapter all involve some form of signal processing to radically alter the original signal, whether it be a baseband input signal or an already modulated carrier. They attempt to synthesise an envelope and phase modulated signal at high power, with highly efficient (and usually nonlinear) processing throughout the remainder of the system. In their basic configurations, no real-time control is utilised and hence the systems may be considered to be open-loop, although high-performance, practical embodiments require some form of adaption to maintain their performance. Feedback may be added to the basic techniques described with a resulting improvement in performance. Polar feedback, for example, maps well to the Envelope Elimination and Restoration (EE&R) technique, whilst Cartesian feedback may be more appropriate for LINC (*LI*near amplification using *N*onlinear *C*omponents) and LIST (*LI*near amplification using *S*ampling *T*echniques). Further details of these feedback mechanisms may be found in Chapter 4.

All of the techniques have the potential for very high efficiencies, theoretically approaching 100% in most cases. This is their principal attraction and the reason that much research effort has been directed in this area in recent years. Practical problems in their realisation and their perceived complexity have, however, resulted in the appearance of few practical, commercial systems to date.

The advent of digital signal processing techniques has proved very beneficial for the systems considered below. It alleviates many of the problems associated with generating the required functions utilising analog techniques (e.g., the $\cos^{-1}$ function required for LINC) and hence enables the techniques to become a practical proposition. The two major drawbacks of digital signal processors, namely processing speed and power consumption, limit the near ideal performance of the schemes in which they are utilised. As technology continues to advance the techniques described in this chapter will become increasingly attractive and will ultimately be more widely adopted.

7.2 Envelope Elimination and Restoration

7.2.1 Introduction

The Envelope Elimination and Restoration (EE&R) technique was first proposed by Kahn in 1952 [1] (and hence is sometimes referred to as the Kahn Technique in the literature) and has received further attention since [e.g., 2–4]. It is essentially a high-level modulation technique, and as such may be implemented as either a complete linear transmitter, or as an RF linear amplifier. It is in this latter guise that it is most commonly found in the literature and, indeed, as it was originally proposed, although both forms will be discussed in this section.

The EE&R technique was originally employed for linear HF amplification of SSB signals, but is now also employed in high power television and radio broadcast transmitters due to its potential for highly-efficient operation. It has also been proposed for mobile radio applications [4], again due to its potential for high efficiency and also its relative simplicity of implementation.

It does, however, suffer from inherent device limitations in the final RF power stage when operating at low signal levels and these will be discussed below.

7.2.2 Operation of an EE&R Amplifier

The configuration of a basic EE&R *amplifier* is shown in Figure 7.1. The input signal, which may contain both amplitude and phase modulation, is split to form a baseband path containing the envelope of the input signal and an RF path containing a constant-envelope phase modulated carrier signal. The latter signal may be created simply by limiting the RF input, to remove

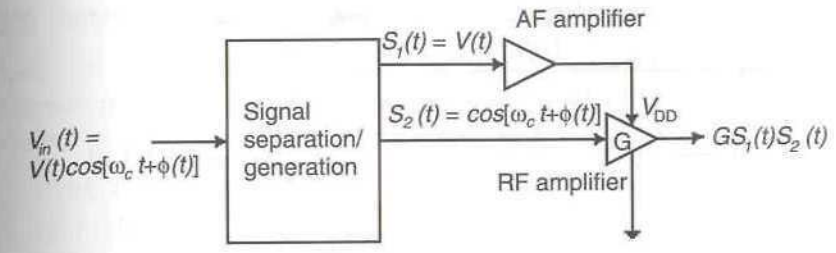


Figure 7.1 Schematic of an envelope elimination and restoration amplifier.

the amplitude modulation components, which will then leave only the phase (or frequency) modulation on the original input signal. The baseband signal corresponding to the amplitude variations of the input envelope may be created either by diode detection of the input signal, or by coherent detection utilising the carrier signal after the above-mentioned limiter. The latter method will produce more accurate (lower distortion) results, but will lead to increased complexity for the overall system. For this reason, diode-based envelope detection has been employed in many of the EE&R transmitters mentioned in the literature (including [4]).

The constant-envelope, phase-modulated carrier signal ($S_2(t)$ in Figure 7.1) is then amplified by a high-efficiency RF amplifier, for example, of class-C, -D, or -E. This will preserve the phase modulation information and transmit this to the output of the system. The baseband AM signal ($S_1(t)$ in Figure 7.1) is amplified by a suitably efficient audio amplifier, or is used to feed a pulse-width modulator (see Chapter 3), with subsequent class-D power amplification. Finally, the resulting high-power audio signal is used to modulate the collector or power supply of the final RF power stage. This high-level modulation process restores the signal envelope and, assuming that the relevant delays between the two paths are suitably equalised, results in a high power replica of the input signal being produced at the output.

7.2.3 Operation of an EE&R Transmitter

The operation of a complete EE&R transmitter is, in many respects, simpler than that of the corresponding EE&R amplifier. In the complete transmitter, shown in Figure 7.2, the input signals are generated at baseband as separate amplitude and phase modulating signals. This is comparable to the generation of I and Q signals required by the Cartesian loop transmitter (see Chapter 4) and may be performed by a digital signal processor in a similar manner.

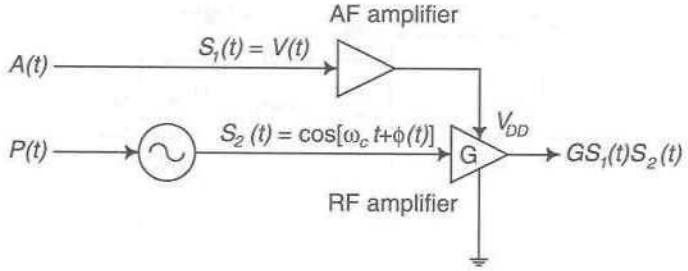


Figure 7.2 Schematic of an envelope elimination and restoration transmitter.

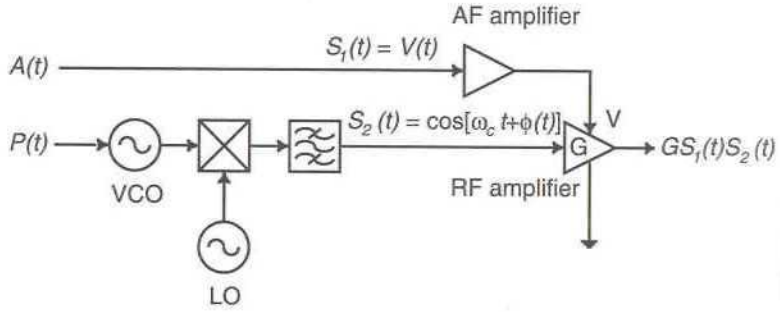


Figure 7.3 Schematic of an envelope elimination and restoration transmitter employing upconversion.

Generating the amplitude and phase (polar) signals by this technique removes the need to modulate the carrier elsewhere in the transmitter architecture and also eliminates the requirement for a limiter and amplitude detector to perform the component separation process at the input.

An alternative form of EE&R transmitter is shown in Figure 7.3. This configuration removes the requirement for the VCO generating the phase-modulated carrier signal, to operate at the carrier frequency (operation of a VCO at the carrier frequency is often undesirable). The VCO may now operate at any convenient frequency and the operating channel frequency may be determined by a separate synthesiser. This arrangement is considerably more convenient, particularly for channelised systems.

The nonlinearities present in the upconversion mixer are unimportant, since the signals it is processing are constant-envelope. The only new concern is in ensuring that the unwanted mixer products fall outside the bandwidth of the RF amplifier, or are suitably attenuated by the bandpass filter. This filter may be of any suitable high-Q design (including ceramic or crystal), since it is only required to process low-power signals.

7.2.4 Intermodulation Distortion in EE&R Transmitters

There are a large number of potential sources of IMD in EE&R-type transmitters, due to the range of types of signal processing which are required in the system. Some of the principal sources include:

1. bandwidth of the envelope modulator (e.g., class-S amplifier);
2. differential delay between the envelope and phase signals;
3. nonlinearity in the envelope detector;
4. AM–PM conversion in the limiter;
5. AM–PM conversion in the final power amplifier stage (when being high-level amplitude modulated);
6. 'encoding' error in the class-S modulator (i.e., errors between the actual and ideal response of the modulator at a given input amplitude level). These errors are usually minimised by amplitude feedback employed around the modulator;
7. cut-off occurring in the RF power amplifier at low envelope levels.

The first two effects are arguably the most significant potential source of IMD in the majority of EE&R transmitters and have been analysed by Raab [5].

7.2.4.1 Envelope and Phase Functions for a Two-tone Test

In order to examine the linearity of an EE&R system, it is necessary to consider an envelope varying input signal. The most commonly used test signal of this type is the two-tone test, which may be described by:

$$\begin{aligned} v_i(t) &= \frac{1}{2}[\cos(\omega_c + \omega_m)t + \cos(\omega_c - \omega_m)t] \\ &= M_i(\theta)\cos[\omega_c t + \phi_i(\theta)] \end{aligned} \tag{7.1}$$

where $\theta = \omega_m t$.

The resulting output signal from the EE&R system will be a delayed version of the input signal, with the envelope delay potentially different to that of the RF signal (since both are processed separately):

$$v_o(t) = M_o(\theta - \tau)\cos[\omega_c t + \phi_o(\theta)] \tag{7.2}$$

where τ is the envelope phase delay in radians. The bandwidth of the signal is given by:

$$B_{RF} = 2\omega_m \tag{7.3}$$

The input signal is separately envelope detected and limited resulting in envelope and phase functions, given by:

$$M_i(\theta) = |\cos\theta| \tag{7.4}$$

and

$$\phi_i(\theta) = \frac{\pi}{2}[1 - c(\theta)] \tag{7.5}$$

where $c(\theta)$ is a squarewave with a cosinusoidal phase characteristic and an amplitude level of ± 1.

The envelope function may be approximated by a Fourier series of finite length:

$$M_o(\theta) = \left[a_0 + \sum_{m=2,4,6,\cdots} a_m \cos m\theta\right] \tag{7.6}$$

with a_m given by:

$$a_m = \begin{cases} \dfrac{2}{\pi} & m = 0 \\ \dfrac{4}{\pi}\dfrac{(-1)^{(m-2)/2}}{m^2 - 1} & m \geq 2 \end{cases} \tag{7.7}$$

A Fourier expansion of the squarewave switching function gives:

$$c(\theta) = \left[\sum_{n=1,3,5,\cdots} c_n \cos n\theta\right] \tag{7.8}$$

where c_n is given by:

$$c_n = \frac{4}{\pi}\frac{(-1)^{(n-1)/2}}{n} \tag{7.9}$$

7.2.4.2 Theoretical Effect of Finite Envelope Bandwidth on IMD

For a two-tone signal, (7.2) may be re-written based on the envelope and phase-switching functions ($M_o(\theta)$ and $c(\theta)$, respectively) as:

$$\begin{aligned} v_o(t) &= M_o(\theta)c(\theta)\cos(\omega_c t) \\ &= f(\theta)\cos(\omega_c t) \end{aligned} \tag{7.10}$$

assuming, for the present, a negligible time delay error.

For the output envelope to be a precise replica of the input envelope, the modulation function, $f(\theta)$, must be cosinusoidal, and may be represented by a Fourier series expansion as:

$$\begin{aligned} f(\theta) &= \cos(\theta) \\ &= b_1\cos(\theta) + b_3\cos(3\theta) + b_5\cos(5\theta) + \cdots \end{aligned} \tag{7.11}$$

where b_1 represents the amplitude of the wanted signal and, for example, b_3 and b_5 represent the amplitudes of the various IMD products.

The finite bandwidth of the class-S modulator may be modelled by truncating the Fourier series representation of the envelope function after the Nth component, giving:

$$M_o(\theta) = \left[a_0 + \sum_{m=2,4,6,\cdots,N} a_m \cos m\theta\right] \tag{7.12}$$

Relating this to the envelope and RF bandwidths, gives:

$$N = \mathrm{INT}\left(\frac{2B_E}{B_{RF}}\right) \tag{7.13}$$

where 'INT' is the integer function, which truncates a decimal to only its integer part (rounding down).

The output spectrum of the composite signal from the EE&R transmitter will be formed by mixing each of the spectral components of the envelope function $M_o(\theta)$ with each of the spectral components of the phase switching function $c(\theta)$. This mixing will result in terms of the form:

$$2\cos n\theta\cos m\theta = \cos(m+n)\theta + \cos(m-n)\theta \tag{7.14}$$

The resulting coefficients of the modulation function (7.11) are then:

$$\begin{aligned} b_1 &= \frac{1}{2}[2a_0c_1 + a_2(c_1 + c_3) + a_4(c_3 + c_5) + \cdots] \\ b_3 &= \frac{1}{2}[2a_0c_3 + a_2(c_1 + c_5) + a_4(c_3 + c_7) + \cdots] \\ b_5 &= \frac{1}{2}[2a_0c_5 + a_2(c_3 + c_7) + a_4(c_1 + c_9) + a_6(c_1 + c_{11}) + a_8(c_3 + c_{13}) + \cdots] \end{aligned} \tag{7.15}$$

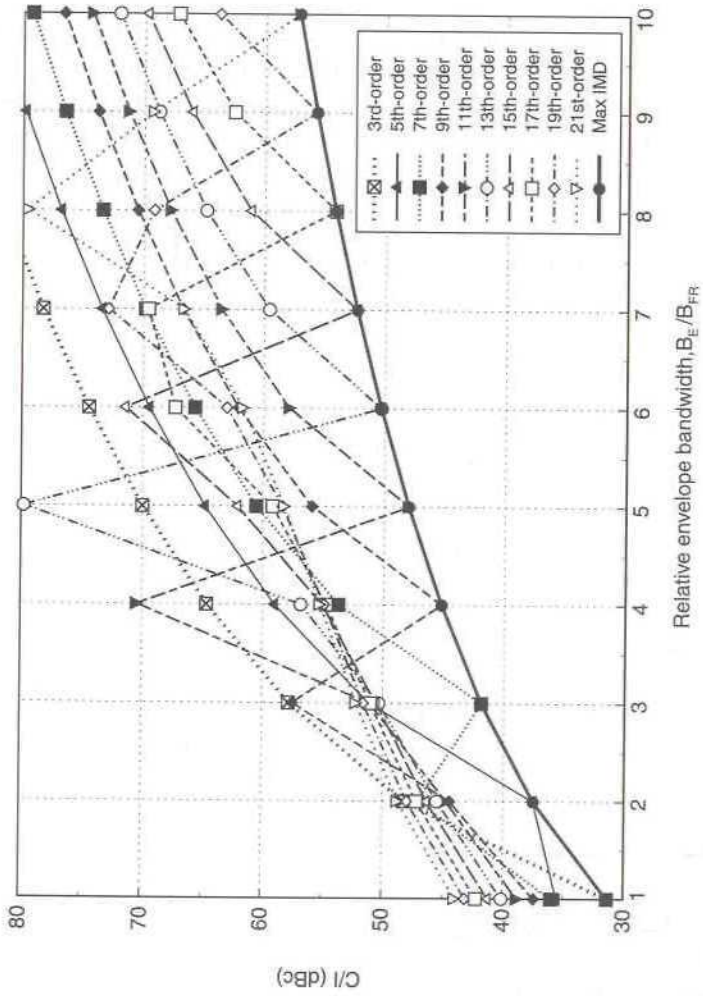


Figure 7.4 Theoretical IMD levels for an EE&R system with a finite envelope bandwidth (after [5]).

Since modulation of the carrier by the modulation function will produce two components of amplitude $b_k/2$, the C/I ratio (defined with respect to an unmodulated carrier) is given by:

$$C/I = 20\log\left(\frac{2}{|b_k|}\right) \quad \text{dB} \tag{7.16}$$

Using the above equations, it is possible to predict the C/I ratio for a given ratio of envelope, B_E, and RF, B_{RF}, bandwidths, where the envelope bandwidth is, in practice, typically limited by the bandwidth of the class-S modulator. These characteristics are shown in Figure 7.4 for a range of typical values.

The order of the largest IMD product increases with the ratio of envelope to RF bandwidths, consequently it is not typically the third-order product (as is the case with most other forms of linear power amplifier or transmitter). This order will be given by the next odd number greater than N (from (7.13)). Thus, for $B_E = 3B_{RF}$, the largest IMD product will be seventh-order.

The practical results reported in [5] indicate that the third-order products were almost always the largest (although not by a large margin), despite the theoretical predictions to the contrary. This, it was suggested, was due to the finite roll-off of the envelope filter (a 'brick-wall' filter is assumed by (7.12)) and the limits imposed by DC linearity within the system.

7.2.4.3 Theoretical Effects of Differential Delay in an EE&R System

A differential delay between the envelope and switching functions results in an unintentional inversion of the polarity of the modulation function, $f(\theta)$ immediately prior to or immediately following the waveform zero-crossing points (depending upon which path has the greater delay). This is shown in Figure 7.5.

The resultant IMD from this process is then derived from the added signal, $p(\theta)$, which is simply an appropriate portion of a cosinusoidal waveform:

$$f(\theta) = |\cos(\theta)|c(\theta + \tau) = \cos(\theta) + p(\theta) \tag{7.17}$$

where

$$p(\theta) = \begin{cases} -2\cos\theta, & \pi/2 - \tau \leq \theta < \pi/2 \\ -2\cos\theta, & 3\pi/2 - \tau \leq \theta < 3\pi/2 \\ 0, & \text{otherwise} \end{cases} \tag{7.18}$$

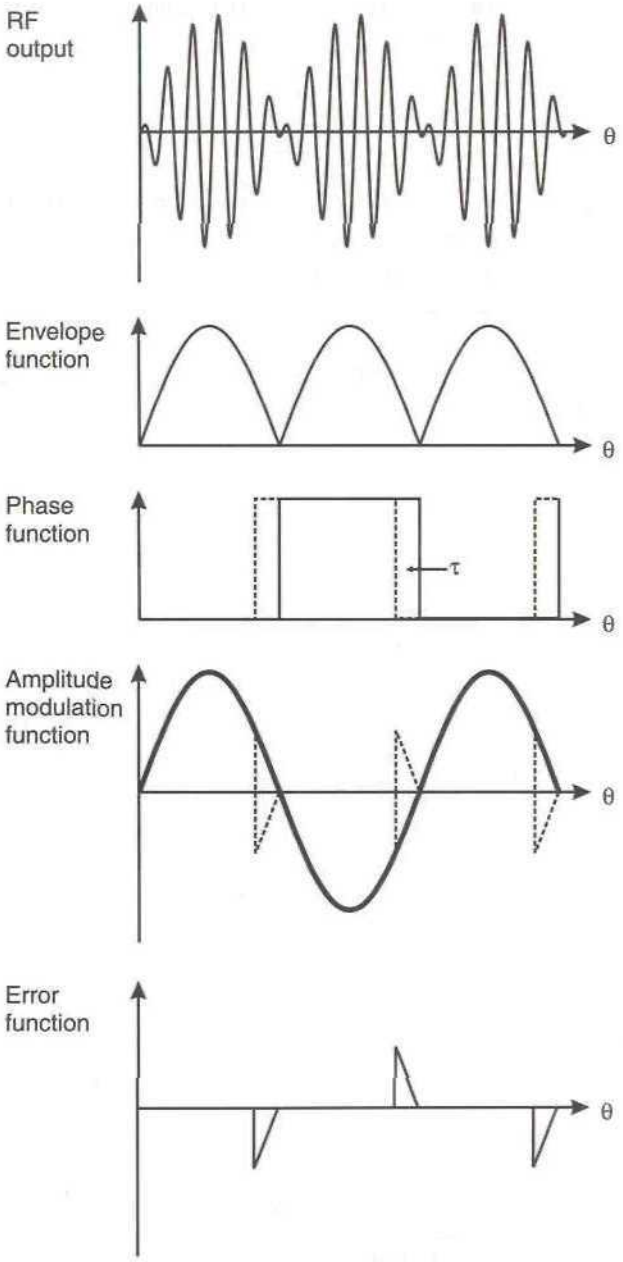


Figure 7.5 Waveforms in an EE&R transmitter, showing the effect of differential delay between the envelope and switching functions (after [5]).

The differential delay, τ, (in radians) is related to the differential delay, Δt, in seconds by:

$$\Delta t = \frac{\tau}{2\pi f_m} = \frac{\tau}{\pi B_{RF}} \tag{7.19}$$

since $f_m = B_{RF}/2$. For convenience, (7.18) may be re-written (neglecting phase information since only the magnitude of the IMD is of interest):

$$p'(\theta) = \begin{cases} -2\sin\theta, & 0 \le \theta < \tau \\ -2\sin\theta, & \pi \le \theta < \pi + \tau \\ 0, & \text{otherwise} \end{cases} \tag{7.20}$$

A Fourier expansion of (7.20) yields:

$$p'(\theta) = \sum_{k=1}^{\infty} (q_k \cos k\theta + r_k \sin k\theta) \tag{7.21}$$

The even order coefficients will be zero due to the symmetry of the function, hence the odd-order coefficients are:

$$\begin{aligned} q_k &= -\frac{4}{\pi}\int_0^{\tau} \sin\theta \cos k\theta \, d\theta \\ &= -\frac{2}{\pi}\left[\frac{1-\cos(k+1)\tau}{k+1} + \frac{\cos(k-1)\tau - 1}{k-1}\right] \end{aligned} \tag{7.22}$$

and

$$\begin{aligned} r_k &= -\frac{4}{\pi}\int_0^{\tau} \sin\theta \sin k\theta \, d\theta \\ &= -\frac{2}{\pi}\left[\frac{\sin(k-1)\tau}{k-1} - \frac{\sin(k+1)\tau}{k+1}\right] \end{aligned} \tag{7.23}$$

Equations (7.22) and (7.23) may be simplified using the approximation:

$$\cos\varphi \approx 1 - \varphi^2/2 \quad \forall\, |\varphi| << 1 \tag{7.24}$$

where $\varphi = k\tau$ in this case. The resultant expressions are then, respectively:

$$\begin{aligned} q_k &\cong -\frac{2}{\pi}\left[\frac{(k+1)^2\tau^2}{2(k+1)} - \frac{(k-1)^2\tau^2}{2(k-1)}\right] \\ &= -\frac{2\tau^2}{\pi} \end{aligned} \tag{7.25}$$

and

$$r_k \cong -\frac{2}{\pi}\left[\frac{(k-1)\tau}{(k-1)} - \frac{(k+1)\tau}{(k+1)}\right] = 0 \tag{7.26}$$

For small values of differential delay, the magnitude of the intermodulation product is therefore:

$$s_k = \sqrt{q_k^2 + r_k^2} \approx |q_k| = \frac{2\tau^2}{\pi} \tag{7.27}$$

Combining with (7.19) and accounting for the fact that the distortion function, S_k, produces two IMD sidebands of value $S_k/2$, gives:

$$s_{IMD} = \pi(\Delta t\, B_{RF})^2 \tag{7.28}$$

This function is plotted for a range of delay values (relative to the RF bandwidth) in Figure 7.6. As an example, consider a 25 kHz channel bandwidth (e.g., for TETRA), in order to achieve a C/I ratio of 60 dBc,

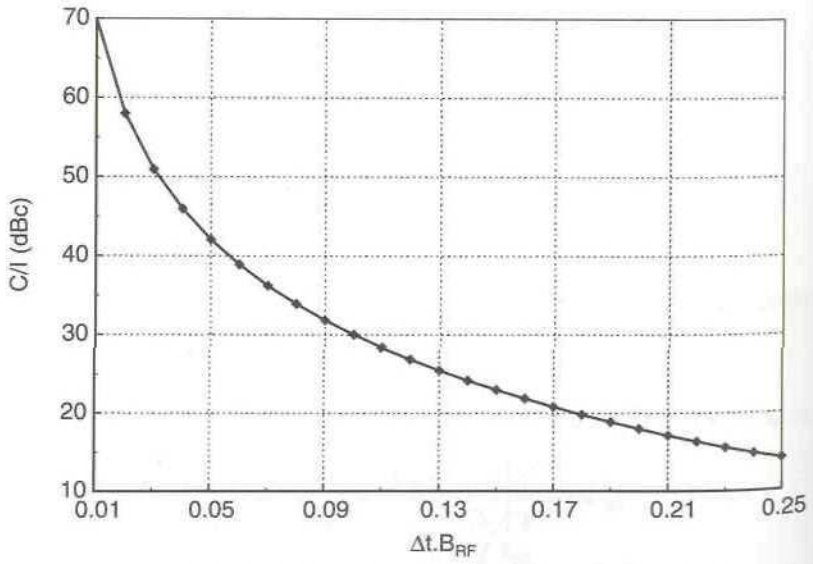


Figure 7.6 Theoretical IMD levels for an EE&R transmitter with differential delay between the envelope and phase paths (after [5]).

the maximum allowable time delay error is 0.72 μs. Note that in a typical system, the envelope suffers the greater delay and hence the modulator signal processing must add delay to the phase-modulation signal, typically in the order of 5 μs to 20 μs. The minimum envelope bandwidth required in order to meet this specification would be 325 kHz. It should also be noted that the remainder of the system would need to be capable of meeting these IMD figures (e.g., the envelope detector, class-S modulator) and that this would be extremely difficult, based on the results reported to date (e.g., [6,7]). More realistic linearity figures are currently in the region of 30 dBc to 40 dBc.

Practical results presented in [5] indicate that the odd-order products do dominate (assumed in the expansion of (7.21), due to the symmetry of the function) and that the roll-off of the products with increasing order is slow (although not flat as predicted by (7.27)).

7.2.5 Practical Delay Considerations With EE&R Transmitters

The delay between the RF and AF paths must be closely matched in order to minimise IMD at the transmitter output, as outlined in Section 7.2.4.3. The delay in the class-S modulator output filter is likely to be large in many implementations, due to the desire to choose as low a switching frequency as possible (to maximise efficiency and ease implementation). This delay is likely to be very much larger than the delay in the RF (or IF) path containing the constant-envelope, phase-modulated signal. Since these two delays must be matched in order to optimise IMD performance, a compensating delay must be inserted in the RF/IF path.

In a typical narrowband implementation, this compensating delay may be many microseconds, hence precluding the use of traditional transmission-line based delay lines. In an HF transmitter described in [6–8] with a class-S modulator operating at 250 kHz, the compensating delay required was 9 μs (corresponding to around 2 miles of coax for a transmission-line based delay). It was implemented using four 4-pole LC Butterworth filters (each contributing 2 μs) and a tapped delay line (2 μs in 100 ns steps).

In order to minimise broadband accuracy of the compensating delay, it is important to place the limiter *after* the delay element, since limiting prior to this element would increase the bandwidth over which the delay element must operate.

Increasing the instantaneous bandwidth of operation of the transmitter will result in an increased class-S modulator switching frequency and hence a wider output filter bandwidth. This will lower the absolute value of the compensating delay required, but will increase the delay matching accuracy needed for a given IMD performance.

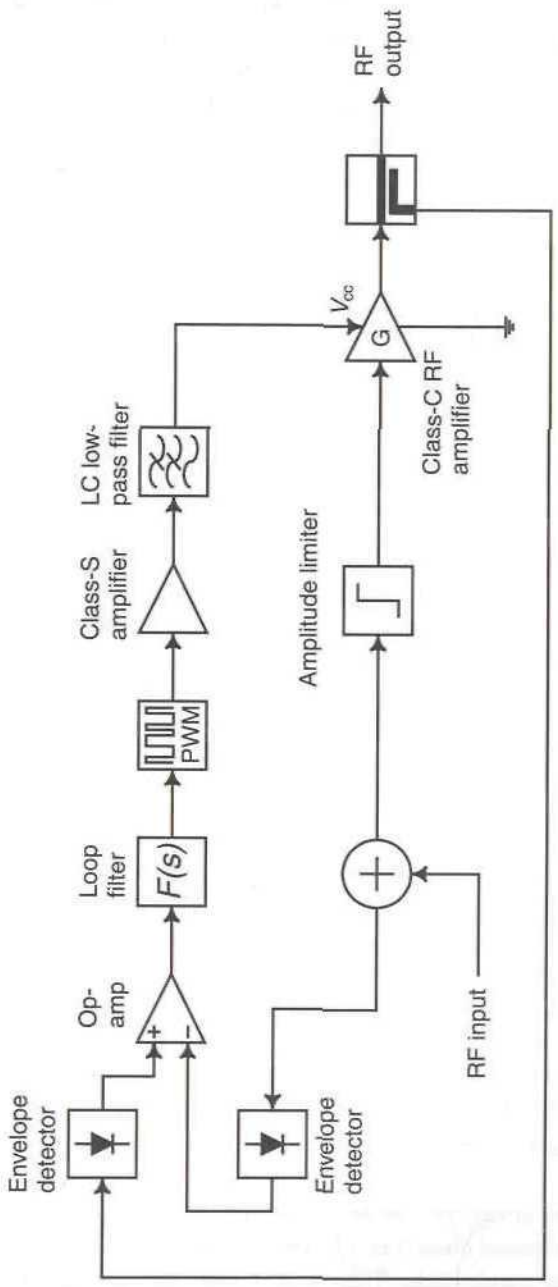


Figure 7.7 EE&R amplifier employing envelope feedback.

7.2.6 Applying Feedback Around EE&R Systems

It is possible to reduce the effect of nonlinearities in the system by means of feedback and this feedback can take a number of forms [9]:

1. envelope feedback;
2. polar feedback;
4. Cartesian feedback.

Indeed, it was the idea of applying feedback correction around an EE&R transmitter which first produced the concept of the polar-loop transmitter discussed in Chapter 4. The practical implementation of the first of these schemes is discussed below.

7.2.7 An EE&R Amplifier With Envelope Feedback

The application of simple envelope feedback around an EE&R amplifier is relatively straightforward, as shown in Figure 7.7 and its operation is as follows. The RF input signal is split to form two equal paths; the upper path feeds an amplitude limiter to remove all envelope variations (whilst preserving phase-modulation information) and the lower path is envelope-detected to extract the amplitude information. The constant-envelope signal at the output of the limiter feeds the RF input to the class-C amplifier; the amplitude information from the envelope detector is used to feed the PWM generator or class-S amplifier which in turn modulates the supply voltage to the class-C PA.

The feedback around the system is performed by taking a sample of the RF output signal, performing envelope detection and comparing the result with the detected envelope of the input signal. It is this difference signal which drives the PWM generator, hence providing improved amplitude linearity. This simple system will, of course, have no effect on the phase response of the system or on any AM–PM conversion taking place in the RF power amplifier.

A system of this type has been discussed by Koch and Fisher [4], exploiting the relative independence of efficiency and power supply voltage in a saturated class-C amplifier stage (shown in Figure 7.8). High levels of efficiency are therefore theoretically possible at all envelope levels.

7.2.8 Efficiency of an EE&R System

The efficiency of an EE&R system depends primarily on the efficiencies of the high-power audio amplifier and the nonlinear RF amplifier, assuming that the output power is sufficiently high to ensure that the power consumption of the signal processing devices is negligible.

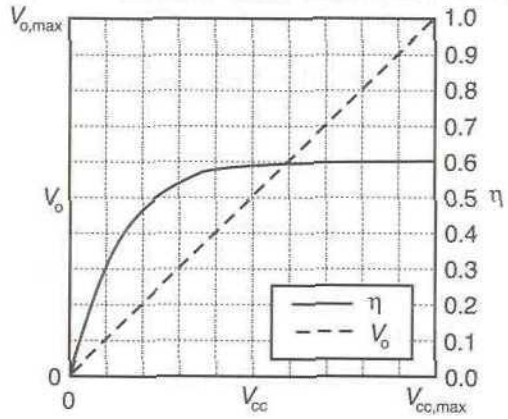


Figure 7.8 Efficiency and output voltage of a class-C PA as a function of DC collector voltage.

If this assumption is made, then the efficiency of an EE&R system is simply the product of the efficiencies of the high-power audio amplifier stage and the nonlinear RF power amplifier:

$$\eta_{tot} = \eta_{AF}\eta_{RF} \tag{7.29}$$

A typical system employing a class-C RF power amplifier ($\eta = 0.6$) and a class-S (pulse-width modulation) audio amplifier ($\eta = 0.9$) will yield an overall efficiency of around 54%. If a switching RF power amplifier is employed, with a basic efficiency of, say, 80%, then the overall efficiency will increase to around 72%.

Note that this efficiency will be maintained at *all* envelope levels in a linearly modulated signal and does not merely represent the efficiency when operating the transmitter at its full peak power rating. This is an important point as it demonstrates that the benefits of an analog linear modulation scheme, in terms of the low speech duty-cycle, may be fully realised with this form of transmitter; during gaps in the user's conversation (e.g., between words), the transmitter will (theoretically at least) consume no power.

7.2.9 An EE&R Transmitter Employing Envelope and Modulator Feedback

Nonlinearities in the pulse-width modulation process may also be reduced by means of a feedback loop, leading to a consequent improvement in overall linearity performance. Raab *et al.* [10] describe an experimental transmitter of this type intended for mobile satellite applications. A block diagram of the transmitter is shown in Figure 7.9.

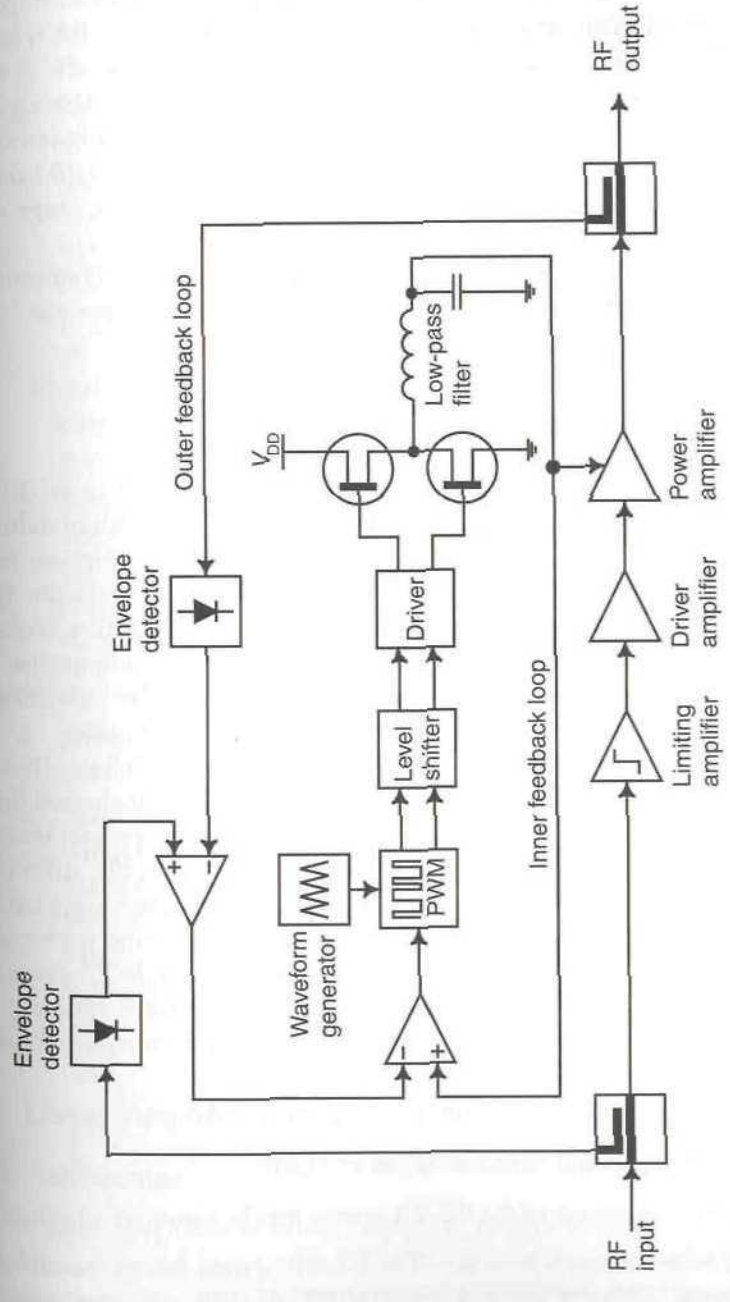


Figure 7.9 Block diagram of an L-band EE&R transmitter (based on [10]).

A class-B PA was employed as the output stage and this was driven 3 dB into saturation for all supply voltage levels (hard limiting was employed earlier in the RF chain to ensure this). The CW efficiency of the PA was 65% at PEP (20W) and this degraded by less than 5% over an 18 dB dynamic range. The class-S modulator employed a PWM with a 3.3 MHz triangle-wave input and a pair of GaAs FETs in a totem-pole configuration. It achieved an efficiency of 90% at PEP and over 80% at up to 10 dB back-off. It is reported to have a linear response over a 26 dB dynamic range and a virtually flat frequency response to 150 kHz.

Using a two-tone test, the results indicated a 57% overall efficiency at PEP and a 35% efficiency at 18 dB back-off; linearity performance was 30 dBc and 39 dBc, respectively.

7.2.10 Integrated Circuit Implementation of EE&R

It is possible to integrate many of the functions required in an EE&R transmitter, and such a device is reported in [11]. A delta-modulated switching power supply, limiter and envelope detectors (for use in the envelope feedback linearisation process) were integrated using 0.8 μm digital CMOS technology and combined with a separate power amplifier, to form a complete EE&R amplifier. The integrated circuit (excluding the PA) occupied less than 4 mm^2 on the above process, with the vast majority being occupied by the switching power supply and output buffer.

The IC was designed for a North American Digital Cellular (NADC) application and reportedly achieved first and second adjacent channel figures of better than 30 dBc and 48 dBc, respectively. These figures were achieved at a power-added efficiency of 50% and an output power of +29.5 dBm (4.8V supply). This figure compares to an efficiency of only 36% when using back-off (with the same amplifier) to achieve the required adjacent channel performance and with a maximum efficiency (in AMPS mode) of 58% for the PA alone. The EVM figure recorded was 3.4% r.m.s. when using the EE&R system.

A more recent example of an EE&R IC has also been reported for use in the 800 MHz band [12].

7.2.11 Advantages and Disadvantages of EE&R

The principle advantages of the EE&R system may be summarised as follows:

1. Potentially high linearity. The EE&R system has the potential for good linearity over a wide range of low envelope variation modulation schemes. Examples are filtered digital schemes, such as π/4-DQPSK. The available linearity will degrade for systems requiring a full envelope variation, such as SSB and 16-QAM, due to problems with the RF power output device when it is required to operate at very low levels of collector-emitter voltage.
2. Potentially high efficiency. The use of a class-C (or even a class-D, or -E) amplifier in the RF power stage will ensure that a high DC–RF conversion efficiency is maintained at this stage. The use of class-S PWM (pulse-width modulation) techniques for the high-level modulator and class-D for the modulator amplifier will also ensure that this part of the system operates at a high efficiency. Together these techniques can lead to a highly efficient RF amplifier or transmitter. Furthermore, this high level of efficiency can be maintained over a wide range of power levels and not just at full output. This is a significant benefit of the EE&R technique over many conventional linearisation schemes.
3. Simplicity of implementation. The basic EE&R technique does not involve any undue complexity other than, perhaps, the design of the high-efficiency amplifiers. The addition of feedback, predistortion or other forms linearisation can, however, greatly add to complexity.

The major disadvantages of the technique are:

1. Difficulties involving signals with a large envelope variation (high peak-to-mean ratio). These are detailed in 1 above.
2. Inability of the basic technique to monitor its own output and correct for nonlinearities present in the system components, for example, the high-level modulator or the high-level modulation process itself (taking place in the RF output stage). The techniques suggested by Koch and Fisher [4] and Raab [10] may partially overcome these problems.

7.3 Linear Amplification Using Nonlinear Components

7.3.1 Introduction

The *LI*near amplification using *N*onlinear *C*omponents (LINC) technique was first proposed by Cox in 1974 [13] as a method of achieving linear amplification at microwave frequencies—a feat which was virtually impossible at the time due to the lack of suitable linear devices at microwave

frequencies. It followed on from work by Chireix [14], who proposed a similar form of outphasing modulator in the 1930s. The intention of the LINC technique was to create a complete linear amplifier, that is, an amplifier with a linear input-output relationship, where the intermediate stages of RF power amplification could employ highly nonlinear devices.

There are many potential advantages with the LINC technique:

1. The ability to use nonlinear devices permits the technique to be used at high microwave and millimetric-wave frequencies.
2. High-power phase-locked signal sources can also be employed by the technique. These are often easier to construct at high frequencies and efficiencies than amplifiers of equivalent power.
4. The use of highly nonlinear RF amplifiers (e.g., class-C, -D or -E) results in the potential for very high efficiencies indeed to be realised.
4. The technique is capable (theoretically) of an ideal 100% efficiency at *all* envelope levels of the output RF signal. Any degradation due to, for example, non-ideal components, and power amplifier efficiencies will therefore be a degradation from an ideal 100% efficiency. This contrasts with, for example, class-A amplification (with a maximum theoretical efficiency of 50% *at full output*) or class-B amplification, with a maximum theoretical efficiency of 78%.
5. The technique is (conceptually at least) straightforward to understand and implement. There are, however, a number of practical difficulties which have led to the relatively few applications of the technique to date.

7.3.2 Operation of a LINC Transmitter

The basic schematic of a LINC transmitter is shown in Figure 7.10, where the two RF amplifiers are assumed to be high-efficiency and highly nonlinear. The RF input signal, $v_{in}(t)$, is split into two constant-envelope, phase-modulated signals by the signal separation or generation process, and each is fed to its own nonlinear RF power amplifier. The power amplifiers separately increase the power of each signal by an identical amount, before feeding them to an ideal summing junction for recombination. The resulting output signal from the summing junction is then an amplified version of the original input signal with (ideally) no added distortion.

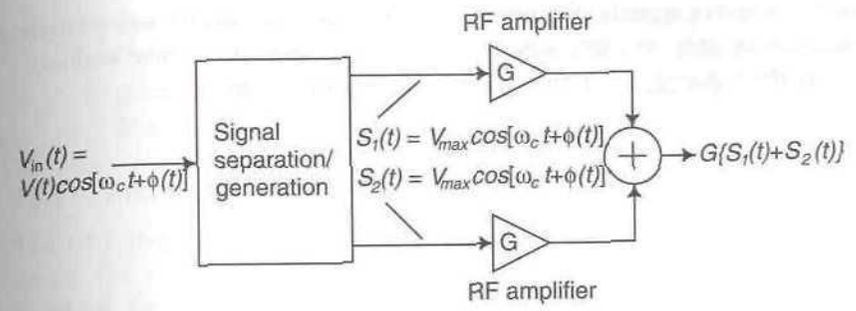


Figure 7.10 Schematic of a LINC transmitter.

The input signal, $S(t)$, is given by:

$$S(t) = V(t)\cos[\omega_c t + \phi(t)] \tag{7.30}$$

where $V(t)$ is the amplitude modulation present on the signal, ω_C is the carrier frequency and $\phi(t)$ is the phase-modulation component of the signal. The input signal is split into two constant-envelope phase modulated signals, $S_1(t)$ and $S_2(t)$, where:

$$S_1(t) = V_{\max}\cos[\omega_c t + \varphi(t)] \tag{7.31}$$

and

$$S_2(t) = V_{\max}\cos[\omega_c t + \theta(t)] \tag{7.32}$$

where:

$$\varphi(t) = \phi(t) + \alpha(t) \tag{7.33}$$

and

$$\theta(t) = \phi(t) - \alpha(t) \tag{7.34}$$

For these signals to recombine and produce the correct linearly amplified version of the input signal, the following relationships must also hold:

$$2S(t) = S_1(t) + S_2(t) \tag{7.35}$$

and

$$\alpha(t) = \cos^{-1}[V(t)/V_{\max}] \tag{7.36}$$

Thus, the above signals, $S_1(t)$ and $S_2(t)$, must be successfully and accurately generated in order for the benefits of the LINC technique to be realised.

If the input signal is provided in quadrature form as:

$$S(t) = s_I(t) + j\,s_Q(t) \tag{7.37}$$

Then the two LINC component signals may be defined as:

$$\begin{aligned} s_1(t) &= s(t) + e(t) \\ s_2(t) &= s(t) - e(t) \end{aligned} \tag{7.38}$$

Where

$$e(t) = -s_Q(t)\sqrt{\frac{1}{\left(s_I^2(t) + s_Q^2(t)\right)} - 1} + j\,s_I(t)\sqrt{\frac{1}{\left(s_I^2(t) + s_Q^2(t)\right)} - 1} \tag{7.39}$$

Quadrature signal components are commonly provided by digital modulation formats and hence the above signal mapping may be preferable in many implementations, in particular those based around a DSP.

7.3.3 Signal Separation/Generation

There are a number of techniques by which the LINC signals may be generated:

1. Analogue signal processing of a modulated RF input signal. This was the method originally envisaged by Cox, however, it is somewhat complex in implementation. It is also difficult to realise the LINC signals accurately by this technique and hence the overall linearity performance of LINC is limited by this process.
2. Digital signal processor generation of the LINC signals at baseband. The two LINC signals may be generated from a baseband input signal by means of a digital signal processor. The resulting baseband output signals can then be upconverted to RF before amplification by the nonlinear amplifiers. This method has a number of advantages over its analogue counterpart, corresponding in general to the advantages of digital over analogue signal processing. This method and its advantages are discussed in more detail in Section 7.3.6.

3. Generation of a version of the baseband input modulation in Cartesian format, followed by the CALLUM [15] modulator to generate and correct the resulting LINC signals. This method is described in more detail in Section 7.5.

7.3.3.1 Analogue Generation of LINC Signals

The first stage in generating the required LINC component signals is to derive the envelope and phase modulation information separately. The envelope signal will be a baseband signal and the phase modulated signal can either be derived in the form of a baseband phase modulating signal or as an RF carrier at the required output frequency, complete with the required phase modulation imposed upon it. This therefore yields three possible methods of realising the envelope signal and phase-modulated carrier:

1. Baseband input signals. If the amplifier forms part of a complete modulator and transmitter, the baseband input signal information will be available and hence can be processed at baseband to yield the amplitude and phase modulation (baseband) signals. This is most easily performed by a DSP or ASIC (as shown in Figure 7.11), but other (analogue) implementations are possible.

   Once these signals have been generated, the phase modulating signal may be modulated onto a carrier at the required frequency (e.g., using a voltage controlled oscillator or frequency synthesiser). Both signals (the baseband envelope and the RF phase-modulated carrier) may then be processed by the signal generator/modulator shown in Figure 7.14. Note that the modulation bandwidth required for the VCO (or PLL synthesiser) will be at least ten times that of the desired RF channel bandwidth (see Section 7.3.8 for further details on LINC signal bandwidths).

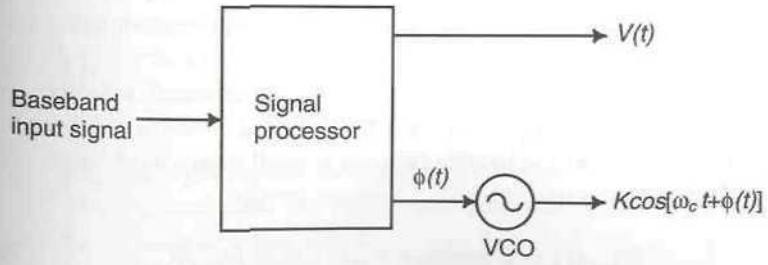


Figure 7.11 Generation of the signal envelope and phase-modulated carrier from a baseband input signal.

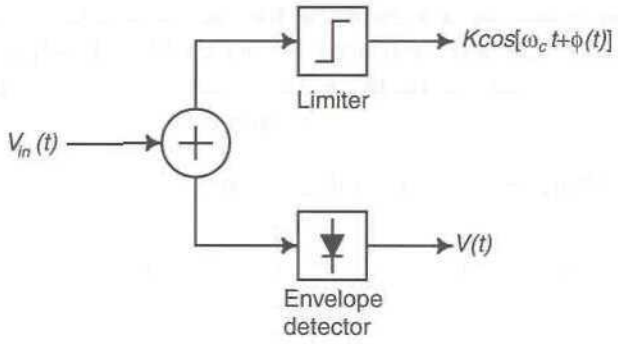


Figure 7.12 Generation of the signal envelope and phase-modulated carrier utilising a limiter and an envelope detector.

This method is potentially a little cumbersome and is particularly inefficient bearing in mind that a suitably powerful DSP can generate the LINC signals at baseband, thus rendering the use of the hardware in Figure 7.14 unnecessary.

2. Limiter and envelope detector. This is the simplest method of realising the required signals and is shown in Figure 7.12. Limiters, whether based around diode circuitry or active high-gain limiting amplifiers, are commonly available and easy to use. Likewise, envelope detectors are straightforward to implement, although they do generally suffer from poor linearity and hence will degrade the achievable linearity from the LINC system.
3. Limiter and synchronous detection. This is still a relatively simple method, but benefits from the more linear transfer characteristic of the coherent detection process. A schematic of the system is shown in Figure 7.13.

Having separated the input signal into its envelope and RF phase-modulated carrier components, these can then be utilised in a baseband feedback system to generate the required LINC component signals (at RF), as shown in Figure 7.14 [13].

Consider the feedback loop around the operational amplifier with voltage gain, $G_A = V_O(t)/V_i(t)$. A phase-shifted version of the phase-modulated carrier (from the limiter output) is itself phase modulated by the operational amplifier output signal, $V_O(t)$, to produce:

$$V_{out1}(t) = K\sin[\omega_C t + \phi(t) + k_P V_O(t)] \tag{7.40}$$

where k_P is the gain constant of the phase modulator.

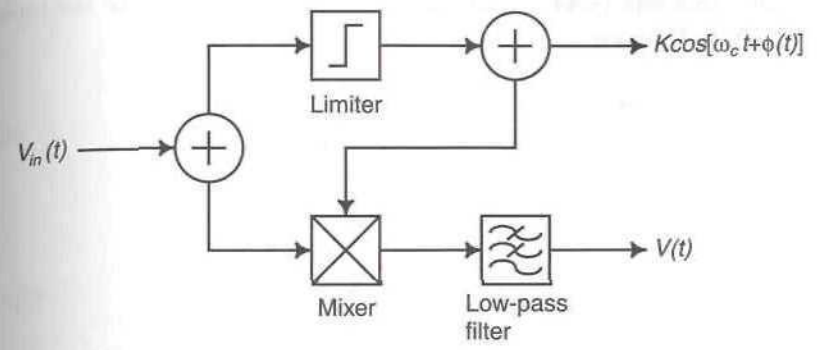


Figure 7.13 Generation of the signal envelope and phase-modulated carrier utilising a limiter and coherent detection.

The output of the mixer is therefore:

$$M_{out}(t) = K^2\sin[\omega_C t + \phi(t) + k_P V_O(t)]\cos[\omega_C t + \phi(t)] \tag{7.41}$$

The output of the baseband low-pass filter, $V_{fb}(t)$, is then:

$$V_{fb}(t) = \frac{K^2}{2}\sin[k_P V_O(t)] \tag{7.42}$$

To ensure stability of the feedback loop, the argument of the sine term must remain within the first quadrant, that is, $|k_P V_O(t)|$ $\pi/2$.

The operational amplifier acts as a summing amplifier, thus:

$$V_i(t) = \frac{R_1 V(t) + R_2 V_{fb}(t)}{R_1 + R_2} \tag{7.43}$$

Completing the feedback loop to form the output of the operational amplifier, $V_O(t) = G_A V_i(t)$, gives:

$$V_O(t) = -G_A V_i(t) = -\frac{G_A R_1 V(t)}{R_1 + R_2} - \frac{G_A R_2 (K^2/2)\sin(k_P V_O(t))}{R_1 + R_2} \tag{7.44}$$

With the restriction detailed above (to ensure stability of the feedback loop), it is possible to determine the maximum value of the sine term in the above expression. This will occur when:

$$K_P V_O(t) = \frac{\pi}{2} \tag{7.45}$$

Substituting this into (7.44) yields the maximum value, V_{max}, of the input signal, $V(t)$, as follows:

$$\frac{\pi}{2k_P} = -\frac{G_A R_1 V(t)}{R_1 + R_2} - \frac{G_A R_2 (K^2/2)}{R_1 + R_2} \tag{7.46}$$

hence:

$$\frac{\pi(R_1 + R_2)}{K^2 k_P R_2} = -G_A[1 + 2R_1 V_{max}/K^2 R_2] \tag{7.47}$$

Assuming that:

$$G_A >> \frac{\pi(R_1 + R_2)}{K^2 k_P R_2} \tag{7.48}$$

then (7.47) reduces to:

$$V_{max} = -\frac{K^2 R_2}{2R_1} \tag{7.49}$$

Assuming that K, R_2 and R_1 are chosen as outlined in (7.49) above, then (7.44) becomes:

$$V(t) = -\frac{K^2 R_2}{2R_1} \sin[k_P V_O(t)] \tag{7.50}$$

and hence:

$$\alpha(t) = -k_P V_O(t) \tag{7.51}$$

Thus, the outputs from the component separator are the desired constant amplitude phase-modulated signals:

$$\begin{aligned} V_{out1}(t) &= K \sin[\omega_0 t + \phi(t) - \alpha(t)] \\ V_{out2}(t) &= K \sin[\omega_0 t + \phi(t) + \alpha(t)] \end{aligned} \tag{7.52}$$

The phase modulator is an important component in this form of signal separator, and in particular, the matching between the two phase modulators shown in Figure 7.14 must be accurately maintained in order to ensure a high

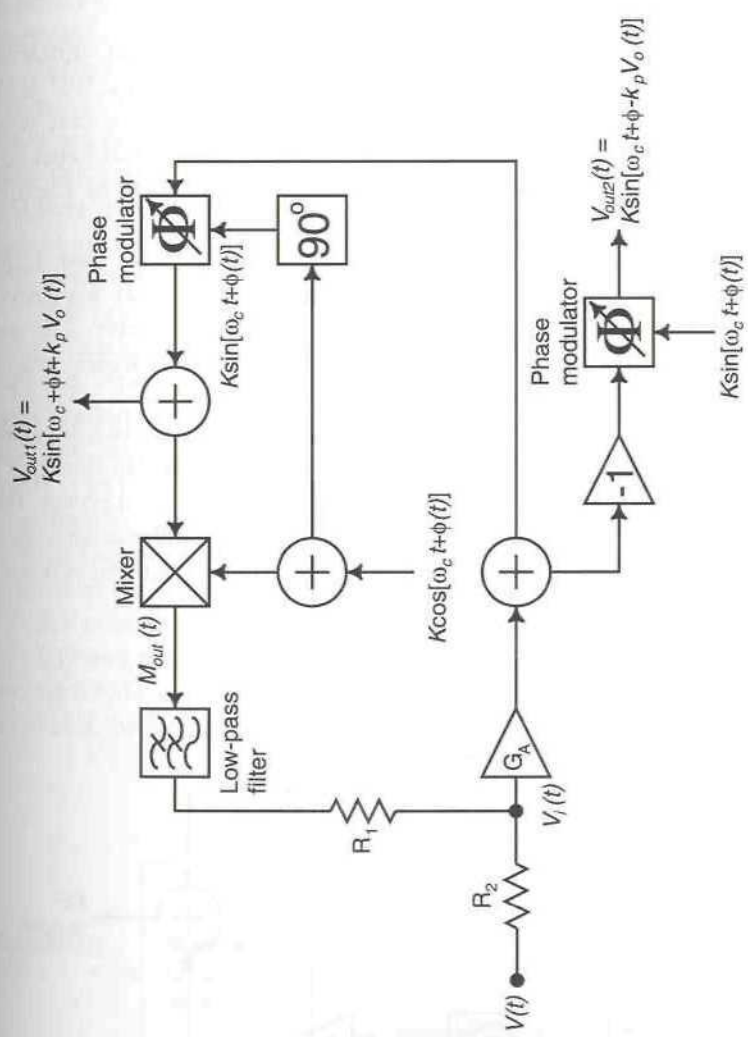


Figure 7.14 LINC component signal generation utilising analogue signal processing techniques (from [13] © IEEE 1974).

degree of linearity for the overall LINC amplifier. This requirement, together with that of accurate matching between the RF amplifiers and subsequent combiner inputs, limits the achievable performance of this analog signal generation technique.

7.3.4 Frequency Translation Within the LINC Technique

In many situations it is desirable to generate complex component signals at a convenient frequency, for example, a standard IF frequency (e.g., 10.7 MHz, or 21.4 MHz) or at baseband (see Section 7.3.6). This is also the case when considering the generation of the LINC component signals and hence it is desirable to be able to implement the upconversion process to the final output frequency within the LINC technique.

This can be achieved as shown in Figure 7.15—the two LINC component signals, $V_{out1}(t)$ and $V_{out2}(t)$, may be generated at any convenient frequency before feeding into the dual-channel upconverter. The only additional component requirements are the upconversion mixers and the associated band-selection filters.

Since the LINC component signals are phase-modulated, nonlinearities in the upconversion mixers have no effect upon the achievable linearity of the technique. Hence the mixers may be driven at their full rated power, thus minimising the gain required in the RF or microwave nonlinear amplifiers. This is an important advantage, since gain at high frequencies and powers is difficult and expensive to achieve.

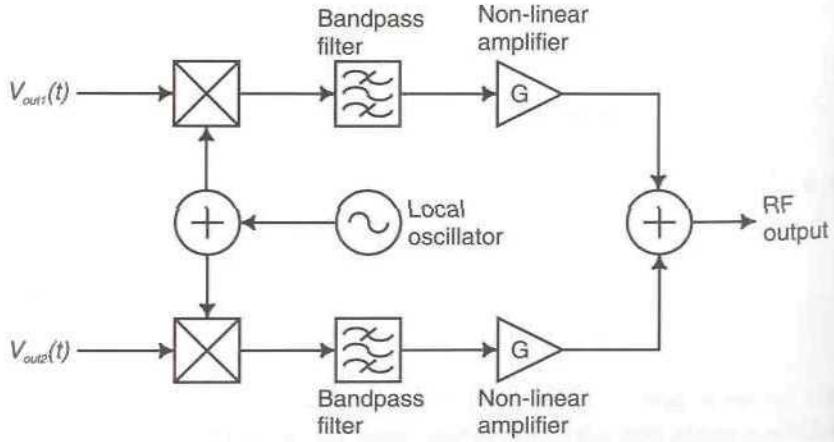


Figure 7.15 Upconversion within the LINC technique.

7.3.5 Use of Voltage-Controlled Oscillators in the LINC Technique

It is clear from the above discussions that the primary attribute of the LINC signals lies in their broadband phase-modulated nature. This type of signal lends itself well to the use of voltage-controlled oscillators (VCOs) as a simple means of generation, including upconversion to the final carrier frequency (by the simple expedient of using VCOs operating at this final frequency).

Figure 7.16 shows one possible implementation of the LINC technique utilising VCOs operating at the final carrier frequency. In this instance, the VCOs are conventional low-power devices, with the final output power being generated in the two (identical) power amplifiers. As before, these amplifiers should be nonlinear, high-efficiency types (e.g., switching amplifiers) in order to realise the optimum overall efficiency for the complete LINC transmitter.

The principal advantage of the use of voltage-controlled oscillators in this application lies in the reduced complexity of the overall system. The baseband input signal can be processed at baseband (i.e., without any upconversion) by, for example, the use of digital signal processing techniques (see Section 7.3.6) to produce the necessary two LINC (baseband) signals. The frequency translation/modulation process takes place directly in the voltage-controlled oscillators, rather than requiring additional mixing, filtering and amplification (to overcome mixer and filter losses) as was required in the system illustrated in Figure 7.15.

An even simpler system is illustrated in Figure 7.17. In this instance, the VCO and nonlinear amplifier in each arm of the LINC transmitter have been replaced by a high-power VCO operating at the final output frequency and at half the final output power. It is relatively straightforward to create a

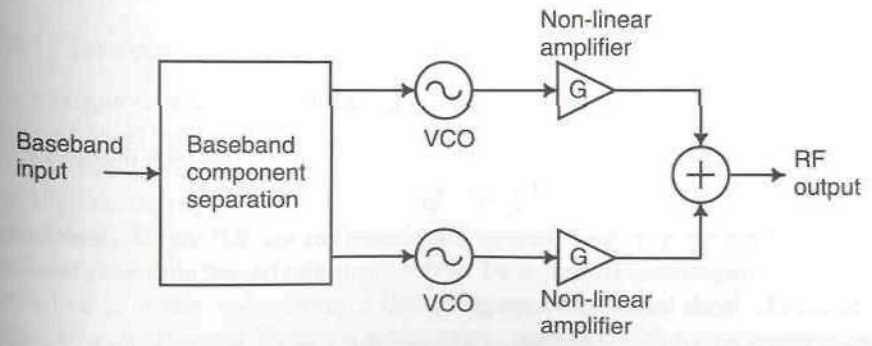


Figure 7.16 Use of voltage-controlled oscillators in the LINC technique.

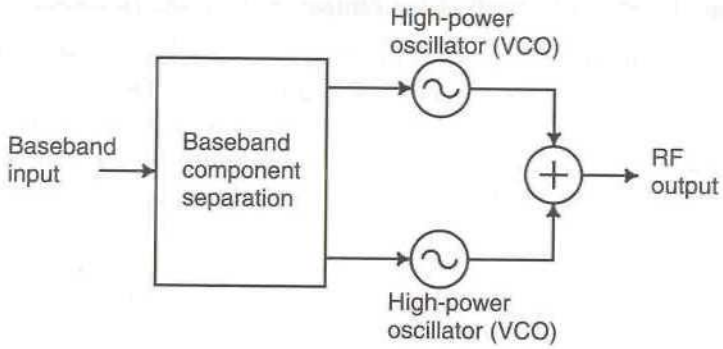


Figure 7.17 Use of high-power oscillators in the LINC technique.

medium or even high-power VCO at VHF, UHF, and some SHF frequencies, since a number of applications require such devices and hence suitable semiconductors are available.

It is, of course, essential that the two oscillators are operating on precisely the same center frequency, although small differences may be removed by the addition of a DC level to the relevant output of the baseband component separator (or just a constant in the case of a DSP implementation of this element). At microwave frequencies, oscillator components (e.g., Gunn diodes) are often available before their transistor (or GaAsFET) counterparts, and hence the use of high-power oscillators may be the only option initially. 'High-power' in this instance may be only a few milliwatts.

The use of high-power oscillators is also sensible from the point of view of efficiency, since oscillators are inherently constant-envelope devices and may be designed to be very efficient. They are also relatively simple devices in construction and hence the overall LINC transmitter, built in this manner, is perhaps the simplest implementation of all.

7.3.6 DSP Implementation of a LINC Transmitter

The generation of LINC signals by the analogue techniques described in Section 7.3.3 has a number of obvious problems:

1. Hardware complexity: a large number of RF signal processing components are required in the implementation, and these tend to be both bulky and expensive.
2. Cost: as was mentioned above, the use of RF signal processing components, particularly in significant quantities, is often expensive.

3. Power consumption: this is particularly a problem when looking to integrate the RF components, such as mixers, but is also of concern in a discrete solution. The large number of components involved on its own is a strong indicator of potential problems in this area.
4. Difficulty of integration: this is a major drawback since one of the main potential applications of the LINC technique is in hand-portable radio transmitters.

The major part of the hardware complexity in the analogue implementation arises from the requirement to accurately generate the $\cos^{-1}$ term in (7.36). The advent of digital signal processing has made this task relatively straightforward (at baseband) and hence the use of DSP techniques within a LINC transmitter has many advantages. The use of DSP techniques moves the limitation on linearity performance from the signal separator to the gain and phase match which may be achieved in the RF path. If this is maintained to a high degree, very good linearity performance may be achieved. Results reported in the literature [16,17] indicate that two-tone IMD levels of −60 dBc (max) may be achieved in a practical implementation.

A major issue with DSP implementation is, however, the modulation signal bandwidths which must be generated. These are typically 10 or more times the bandwidth of the modulation to be transmitted (e.g., over 2 MHz for a single GSM-EDGE carrier). This issue is explored in more detail below.

An example of the baseband signal processing required for a DSP implementation of the LINC system, employing quadrature upconversion, is given in Figure 7.18. The equations defined by the look-up table have been outlined earlier (7.39).

7.3.7 Example LINC Signals

An example of a LINC component signal (both would look identical on this type of plot) and the resulting output spectrum of a LINC transmitter is shown in Figure 7.19. The modulation format which is ultimately generated by the transmitter is $\pi/4$-DQPSK. It is obvious from this result that the bandwidth of the FM component signals extends to many times the wanted channel bandwidth and that relatively broadband cancellation must be achieved in order to provide a linear signal output for even a relatively narrowband channel. The gain and phase matching of the two paths must therefore be maintained accurately over this bandwidth.

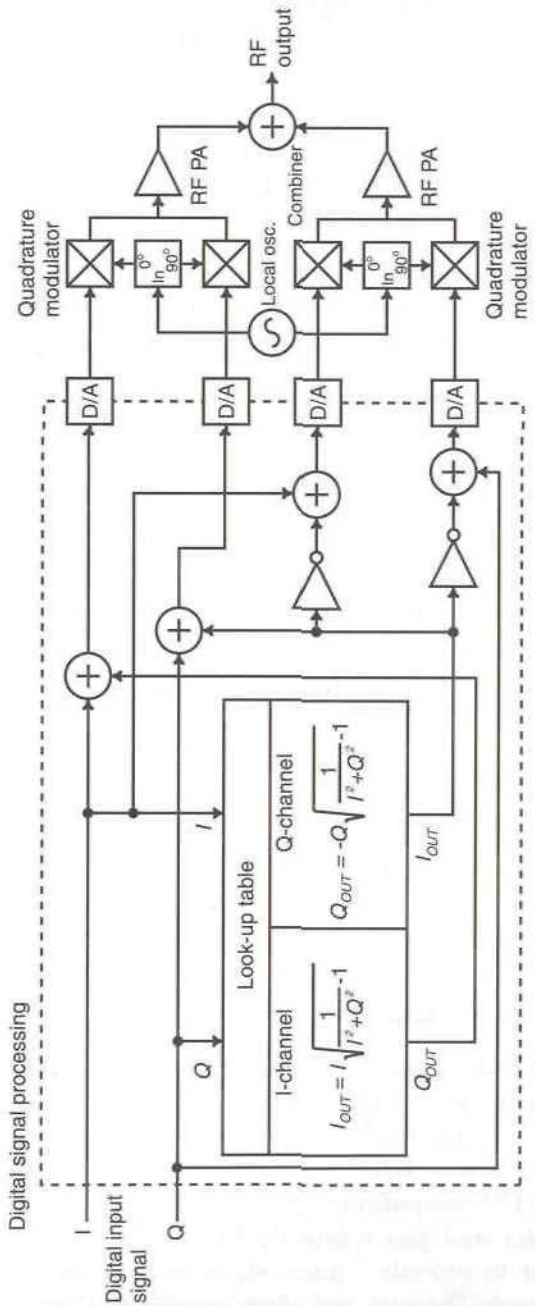


Figure 7.18 Use of DSP in the generation of LINC signals.

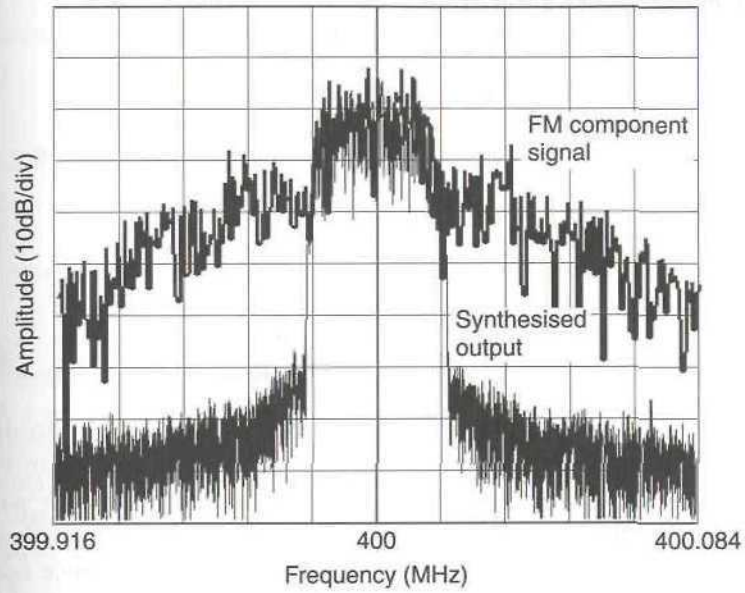


Figure 7.19 Experimental results from a LINC transmitter using π/4-DQPSK.

7.3.8 Gain and Phase Matching of the Two RF Paths in a LINC Transmitter System

The difficulty of achieving the accurate gain and phase matching required between the two paths in a LINC transmitter is one of the reasons why this technique has seen relatively little application to date. Errors in the gain and/or phase matching will result in incomplete cancellation of the unwanted elements of the wideband phase modulated signals and hence result in a large number of unwanted spurious products appearing in the output spectrum (i.e., a much larger number of products than those produced by conventional intermodulation distortion).

The effect of gain and phase imbalances between the two paths in a LINC transmitter may be analysed as follows [18]. Equation (7.30) may be re-written in complex envelope form as:

$$S(t) = r(t)\, e^{j\rho(t)} \tag{7.53}$$

where: $0 < r(t) \leq r_{\max}$

Similarly, the two LINC signals ((7.31) and (7.32)) may be rewritten in terms of the wanted signal information and the additional (unwanted) signals

which will ultimately cancel in the combining process:

$$S_1(t) = S(t) + E(t)$$
$$S_2(t) = S(t) - E(t) \tag{7.54}$$

where:

$$E(t) = jS(t)\sqrt{[r_{\max}^2/r^2(t) - 1]} \tag{7.55}$$

In the frequency domain, (7.54) becomes:

$$S_1(f) = S(f) + E(f)$$
$$S_2(f) = S(f) - E(f) \tag{7.56}$$

A comparison between the bandwidths of the original signal information and that required for the additional (unwanted) signals is shown in Figure 7.20, for various modulation schemes. A root-raised cosine filter with an alpha (roll-off factor) of 0.35 was used for the pulse shaping. It is evident that the spectra for $E(f)$ extend far into adjacent channels and hence must be eliminated in the cancellation process at the output of the LINC transmitter (see also Figure 7.19). For this to be performed correctly, a very high degree of gain and phase matching is required between the two paths over this (very wide) bandwidth.

To illustrate the effect of gain and phase imbalances, a gain imbalance,

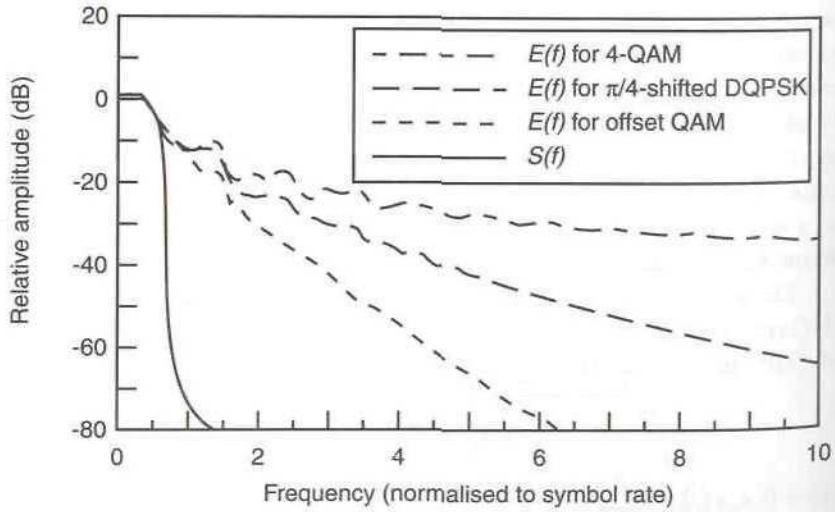


Figure 7.20 Simulated $S(f)$ and $E(f)$ for various modulation schemes (from [18] © IEE 1995).

g_i and a phase imbalance ρ_i are introduced in the second branch of a LINC transmitter. The corresponding signal component is then:

$$S_{2i}(f) = [S(f) + E(f)]\,[1 + g_i]\,e^{j\rho_i} \tag{7.57}$$

The LINC transmitter output then becomes (neglecting the power amplifier gains):

$$S_{out}(f) = S_1(f) + S_{2i}(f)$$
$$= S(f)\left[1 + (1 + g_i)\,e^{j\rho_i}\right] + E(f)\left[1 - (1 + g_i)\,e^{j\rho_i}\right] \tag{7.58}$$

The first term in (7.58) indicates that the source signal is scaled by a factor $1 + (1 + g_i)\,e^{j\rho_i}$, which is approximately 2 for small imbalances. The second term shows that the residual unwanted signal due to imperfect cancellation is $E(f)$ scaled by a factor $1 - (1 + g_i)\,e^{j\rho_i}$. This will only become zero upon perfection being achieved in both gain *and* phase matching.

Since the cancellation process is essentially vector cancellation (i.e., the same as that occurring in feedforward), the degree of gain and phase imbalance which may be permitted for a given level of cancellation is the same as that illustrated in Chapter 5. Taking as an example the π/4-DQPSK spectrum shown in Figure 7.20 it is evident that some 50 dB of cancellation is required in order to meet a 60 dBc first adjacent channel specification and 40 dB required to meet a 70 dBc second adjacent channel specification (such as would be required for a TETRA mobile station). These figures assume that the summation process occurring within the wanted channel (hence providing additional ACPR over that illustrated in Figure 7.20) is absorbed in the design 'margin' for the transmitter. The above example specification would require a gain and phase balance of around 0.02 dB and 0.1° respectively, for the first adjacent channel; an exacting requirement, even over the narrow (3 × 25 kHz = 75 kHz) bandwidth in this illustration. In addition, greater than about 10 dB of cancellation would be required in the fifth adjacent channel (250 kHz bandwidth). It is likely, therefore, that the practical use of the LINC technique will be restricted to narrowband, probably single-channel applications with relatively undemanding adjacent channel requirements, for the foreseeable future.

7.3.8.1 Methods of Achieving Gain and Phase Matching

There are three potential methods of achieving the required gain and phase matching between the paths:

1. Utilise a control scheme, in a similar manner to that described for a

feedforward amplifier in Chapter 5, in order to actively control the gain or phase matching. This can be either a real time system, in which the gain and phase are continuously adjusted whilst transmitting (e.g., as described in [18]), or a periodic update technique occurring immediately before the modulation is transmitted or during periods when the transmitter is not required.

2. Integrate all of the essential RF components of both power amplifier (or power oscillator) paths onto a single piece of silicon. The high degree of matching which should be possible using this technique will reduce or eliminate the need for a control scheme. This approach relies on a reasonable IMD and spurious response specification having been set for the transmitter and incorporates some inherent risk as the transmitter cannot monitor its own performance during service and hence may exceed its permitted spurious levels without being aware of it. This situation would also not be obvious to the user, as the transmitter would operate quite normally until the situation became severe.
3. Utilise a continuous modulation feedback technique, such as the CALLUM technique described below. Such techniques 'force' the output modulation to conform to the input signals and as a consequence ensure that the matching between the two paths is maintained continuously in real time. The principal disadvantage of this approach is that the large feedback bandwidth required, relative to the baseband bandwidth of the input signals, restricts its use to relatively narrow-band transmitters.

7.3.9 Influence of PA Output Match on a LINC Transmitter

The imperfect output match of a power amplifier (or high-power oscillator) has been assessed in terms of its effect on LINC transmitter performance [19]. Imperfections in this match result in a reduction of the theoretical adjacent channel suppression over and above that due to imperfections in the gain and phase balance between the two paths (discussed above).

Taken alone (i.e., assuming perfect gain and phase matching), the results indicate that an output return loss of better than 20 dB is required in order for the ACI degradation to be small (2 dB or 3 dB). A return loss of only 10 dB results in an ACI degradation approaching 10 dB. These results are based on an antenna return loss of 10 dB and an output power combiner isolation of 25 dB and show that a good PA output return loss is essential for good linearity performance from the LINC technique.

7.4 Vector Locked Loop

The vector locked loop technique [20,21] is a mechanism for generation of the signals required in an outphasing modulator, incorporating feedback control to eliminate the effects of gain and phase errors in the two paths. The system consists of two cross-coupled phase-locked loops employing phase and magnitude detection so that both phase and magnitude may be employed as feedback signals. A block diagram of the approach is shown in Figure 7.21 and its operation is as follows.

The high-power RF output signal is generated by the summation of the two RF power amplifier output signals in a similar manner to that of the conventional LINC technique. A sample of the output signal is fed back and split to feed a phase detector and a magnitude detector. The outputs of these detectors are then a phase error signal and a magnitude error signal. The difference between the phase error signal and the magnitude error signal is used to modulate the first (upper path) VCO via the first loop filter. The sum of the phase and magnitude error signals is used to modulate the second VCO, via the second loop filter.

When the loop is locked, both the phase and magnitude errors are minimised and the output signal will then track the phase, frequency and amplitude of the input signal. A change to either the amplitude or phase response of either path, resulting from a change in gain of one of the power amplifiers, for example, will be compensated for by the action of the feedback loop.

The main drawback of this technique, relative to the CALLUM technique described below, is in the difficulty of realising appropriately low-distortion magnitude and phase detectors with a suitable broadband response and operating at high carrier frequencies.

Note also, that if the two power amplifiers are unequal in output power, which will always be the case, to some degree, in a practical implementation, then it is not possible for the technique to generate a zero output. This is true despite the use of feedback to compensate for differences between the two paths. Generation of a zero output would be necessary, for example, to generate the envelope minimum in a two-tone test. For many practical modulation formats, however, it is not necessary to generate a zero output and hence this limitation is not of great concern. It is also a limitation common to other LINC implementations and is not specific to the vector locked loop.

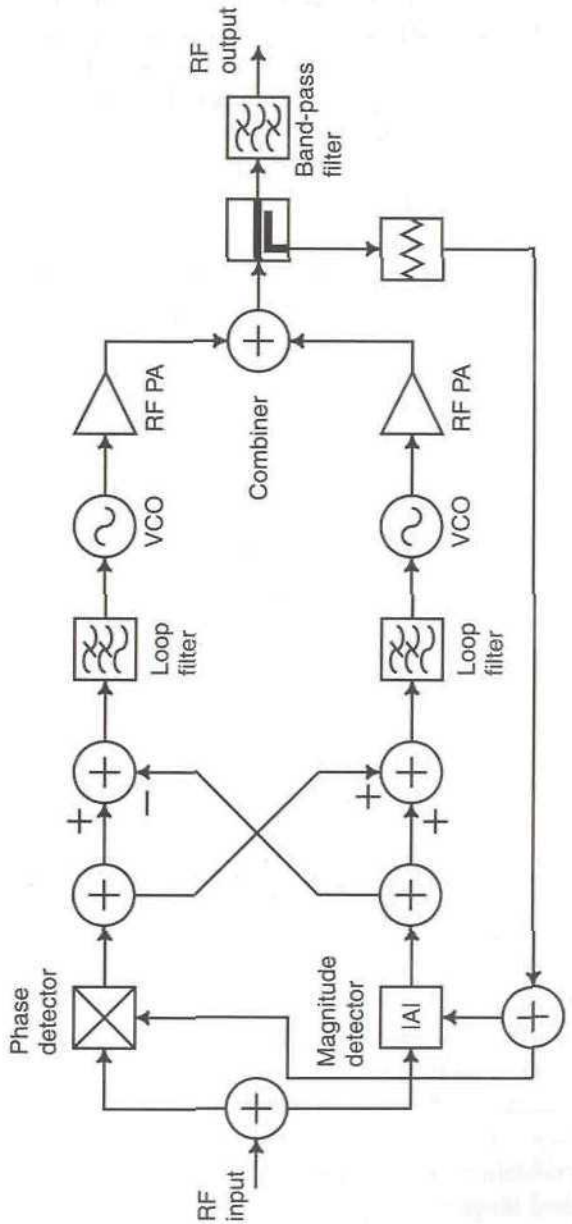


Figure 7.21 Vector locked loop.

7.5 Combined Analogue-Locked Loop Universal Modulator—CALLUM

7.5.1 Introduction

The CALLUM technique is similar in some respects to the vector locked loop described above. The main difference is in the use of quadrature detection, rather than magnitude and phase detection of the VLL technique. The two main aims of the CALLUM [15,22] modulator are:

1. To generate the required signals for the LINC technique in a simple, cheap and integrable form;

and

2. To provide a control mechanism to automatically compensate for gain and phase errors in the RF amplifier chains and hence to compensate for thermal drift and component aging, as well as variations due to sudden disturbances (such as would occur if the transmitter was dropped).

This latter aim, in particular, is key to the potential success of the LINC technique, as it relieves the uncertainty as to what performance could be achieved with good thermal tracking. As is evident from the discussions covered earlier in this chapter, the most attractive application of the LINC technique is in linear transmitters for use in handportables. The transmitter section is thus likely to be integrated and the good matching and thermal tracking which would be obtained by integrating both amplifier chains as part of the same module (or even the same piece of silicon) has long been thought to be the key to the successful implementation of the LINC system. Unfortunately, there is a considerable financial risk in taking a system to this stage of integration, with no certainty that the required matching and thermal tracking can be achieved. It is not surprising, therefore, that no products of this type have, to date, appeared on the market.

Implementation of the CALLUM technique is therefore a potential solution to this problem, as the matching and tracking performance of the LINC system implemented in this way can be assessed without the need to resort to costly integration techniques.

A block diagram of the basic CALLUM modulator is shown in Figure 7.22. The baseband input signal is fed to the modulator in I & Q component form, for example from a DSP, in a similar manner to that of the Cartesian loop transmitter discussed in Chapter 4. The upconversion and LINC signal generation process is then performed by two voltage-controlled oscillators (as suggested by Cox [13]) with subsequent power amplification being provided by two highly nonlinear amplifiers (G_I and G_Q in Figure 7.22).

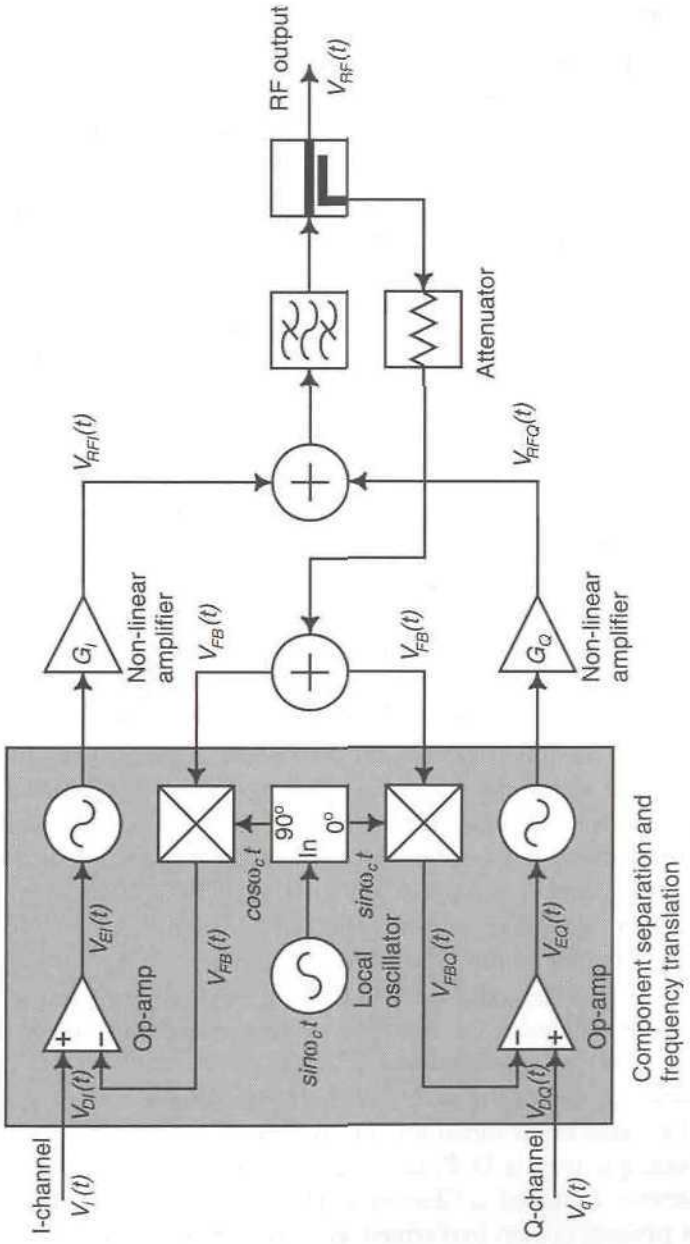


Figure 7.22 Schematic of a basic CALLUM modulator.

The desired VCO control voltage is generated by an error signal formed from the subtraction of the relevant component (I or Q) of the input signal from a downconverted version of the output signal (from the summation hybrid). The local oscillator of the downconverter section is provided in quadrature form to feed both loops and consequently is the sole determinant of the channel frequency of the transmitter. The two loops are therefore acting in a manner akin to quadrature phase-locked loops (PLLs) and hence the stable region of operation of the transmitter may be deduced from PLL theory.

The action of the feedback control system will be twofold: first, it will cause the free-running VCOs to lock to the channel center frequency as defined by the local oscillator signal and secondly, it will cause the combined LINC output signal to closely match the input signal, thus implying that the LINC transmitter output is a perfect linear, amplified representation of the input signal. This, of course, assumes that the loop gain is infinite and that the demodulation process has perfect gain and phase balance, which will not be the case in practice. The performance achieved from an experimental system [15,23] is, however, very impressive, thus demonstrating the practicality of the technique.

7.5.2 Analysis of the CALLUM Modulator

It is possible to demonstrate that the CALLUM technique provides the required LINC signal components outlined in (7.30) to (7.36) and hence will operate correctly as a LINC transmitter. As a consequence of the derivation, it is also possible to demonstrate that the technique will compensate for differential gain errors between the amplifier chains.

Consider, for simplicity, a CALLUM system with only an I-channel input signal, that is:

$$v_i(t) = V_I(t) \tag{7.59}$$

and

$$v_q(t) = 0 \tag{7.60}$$

The resulting RF output signal is given by:

$$V_{RF}(t) = V_{RFI}(t) + V_{RFQ}(t) \tag{7.61}$$

Hence the RF feedback signal from the splitter, which feeds the RF port of

each of the mixers is (assuming a perfect amplitude balance in the splitter):

$$V_{FB}(t) = \frac{V_{RFI}(t) + V_{RFQ}(t)}{K_S} \tag{7.62}$$

where K_S includes both the sampling ratio of the feedback coupler (or hybrid) and the feedback path splitter.

The downconverted feedback signals are now:

$$\begin{aligned} V_{FBI}(t) &= V_{FB}(t)\cos(\omega_C t) \\ &= \frac{V_{RFI}(t) + V_{RFQ}(t)}{K_S}\cos(\omega_C t) \end{aligned} \tag{7.63}$$

and

$$\begin{aligned} V_{FBQ}(t) &= V_{FB}(t)\sin(\omega_C t) \\ &= \frac{V_{RFI}(t) + V_{RFQ}(t)}{K_S}\sin(\omega_C t) \end{aligned} \tag{7.64}$$

Now, consider the I-channel (upper loop) error signal:

$$V_{EI}(t) = K_{AI}V_{DI}(t) \tag{7.65}$$

where $V_{DI}(t)$ is the differential voltage between the noninverting and inverting inputs to the operational amplifier and K_{AI} is the gain of the operational amplifier.

Hence:

$$\begin{aligned} V_{EI}(t) &= K_{AI}[V_I(t) - V_{FBI}(t)] \\ &= \frac{K_{AI}}{K_S}[K_S V_I(t) - (V_{RFI}(t) + V_{RFQ}(t))\cos(\omega_C t)] \end{aligned} \tag{7.66}$$

When the loop gain is large (e.g., $K_{AI} \to \infty$), the action of the feedback forming the upper loop of Figure 7.22 will be to cause $V_{DI}(t) \to 0$; similarly, the feedback action of the lower loop will cause $V_{DQ}(t) \to 0$. Hence:

$$V_I(t) - \frac{V_{RF}(t)}{K_S}\cos(\omega_C t) = 0 \tag{7.67}$$

The RF output in the I and Q channels is formed by amplifying the outputs

of the two VCOs; these are given by:

$$\begin{aligned} V_{VCOI}(t) &= K_I\cos(\omega_{OI}t + \theta_{OI}) \\ V_{VCOQ}(t) &= K_Q\cos(\omega_{OQ}t + \theta_{OQ}) \end{aligned} \tag{7.68}$$

Combining (7.61), (7.69) and the RF amplifier gains gives:

$$\begin{aligned} V_{RF}(t) &= V_{RFI}(t) + V_{RFQ}(t) \\ &= G_I K_I\cos(\omega_{OI}t + \theta_{OI}) + G_Q K_Q\sin(\omega_{OQ}t + \theta_{OQ}) \end{aligned} \tag{7.69}$$

And hence (combining (7.67) and (7.69)):

$$\begin{aligned} V_I(t) = {} & \frac{G_I K_I}{2K_S}\{\cos[(\omega_C - \omega_{OI})t - \theta_{OI}] + \cos[(\omega_C + \omega_{OI})t + \theta_{OI}]\} \\ & + \frac{G_Q K_Q}{2K_S}\{\sin[(\omega_C - \omega_{OQ})t - \theta_{OQ}] + \sin[(\omega_C + \omega_{OQ})t + \theta_{OQ}]\} \end{aligned} \tag{7.70}$$

If it is assumed that the two phase-locked loops, formed by the upper and lower feedback paths, are locked, then $\omega_c = \omega_{OI} = \omega_{OQ}$. The double-frequency (second harmonic) terms are removed by the natural low-pass filtering action of the operational amplifiers (or can be explicitly removed by filtering if necessary). Equation (7.70) now reduces to:

$$V_I(t) = \frac{G_I K_I}{2K_S}\cos\theta_{OI} - \frac{G_Q K_Q}{2K_S}\sin\theta_{OQ} \tag{7.71}$$

Similarly, for $V_{DQ}(t) \to 0$, and bearing in mind that $v_q(t) = 0$ (from (7.66)):

$$0 - \frac{V_{RF}(t)}{K_S}\sin(\omega_C t) = 0 \tag{7.72}$$

Substituting for $V_{RF}(t)$ from (7.69) gives:

$$[G_I K_I\cos(\omega_{OI}t + \theta_{OI}) + G_Q K_Q\sin(\omega_{OQ}t + \theta_{OQ})]\,\frac{\sin(\omega_C t)}{K_S} = 0 \tag{7.73}$$

Again, this can be expanded to yield the baseband and double-frequency terms:

$$0 = \frac{G_I K_I}{2K_S}\{\sin[(\omega_C - \omega_{OI})t - \theta_{OI}] + \sin[(\omega_C + \omega_{OI})t + \theta_{OI}]\} + \frac{G_Q K_Q}{2K_S}\{\cos[(\omega_C - \omega_{OQ})t - \theta_{OQ}] - \cos[(\omega_C + \omega_{OQ})t + \theta_{OQ}]\} \quad (7.74)$$

Making the same assumptions as before regarding the locking of the (Q-channel) phase-locked loop and the filtering of the double-frequency terms by the (Q-channel) operational amplifier, gives:

$$\frac{G_I K_I}{2K_S}\sin\theta_{OI} = \frac{G_Q K_Q}{2K_S}\cos\theta_{OQ} \quad (7.75)$$

If it is assumed, for the present, that:

$$G_I K_I = G_Q K_Q \quad (7.76)$$

in other words, that the RF gains in the two arms of the system are balanced, then it follows from (7.75) that:

$$\sin\theta_{OI} = \cos\theta_{OQ} \quad (7.77)$$

It can therefore be deduced that θ_{OI} and θ_{OQ} are in quadrature. Hence allowing:

$$\sin(\theta_{OI} + \pi/2) = \cos(\theta_{OQ} + \pi/2) \quad (7.78)$$

or

$$\cos\theta_{OI} = -\sin\theta_{OQ} \quad (7.79)$$

Substituting this into (7.71) gives:

$$V_I(t) = \frac{G_I K_I}{K_S}\cos\theta_{OI} \quad (7.80)$$

Finally, re-arranging the above gives:

$$\theta_{OI} = \cos^{-1}\frac{K_S V_I(t)}{G_I K_I} \quad (7.81)$$

which is of the required form for a LINC signal:

$$S_I(t) = V_{\max}\cos[\omega_C t + \phi(t) + \theta_{OI}(t)] \quad (7.82)$$

and

$$\theta_{OI} = \cos^{-1}\frac{V_I(t)}{V_{\max}} \quad (7.83)$$

Since θ_{OI} and θ_{OQ} are in quadrature, it follows that:

$$\theta_{OQ} = -\frac{\pi}{2} + \cos^{-1}\frac{K_S V_I(t)}{G_I K_I} \quad (7.84)$$

It was previously assumed that $G_I K_I = G_Q K_Q$ (from (7.76)). If this assumption is no longer made, in other words, if it is now assumed that a gain imbalance exists between the upper and lower paths of the system, it is possible to analyse the effect this has on the phase relationship of the two phase-locked loops. For example, if $\theta_{OI} = \pi/2$, then (7.75) becomes:

$$\frac{G_I K_I}{2K_S} = \frac{G_Q K_Q}{2K_S}\cos\theta_{OQ} \quad (7.85)$$

or

$$\cos\theta_{OQ} = \frac{G_I K_I}{G_Q K_Q} \quad (7.86)$$

The phase angle, θ_{OQ}, now no longer equals zero (i.e., the two vectors are no longer in quadrature) and consequently the two LINC signals appearing at the inputs to the RF summing junction are no longer in perfect anti-phase. The result of this is that partial (rather than complete) cancellation will occur in this coupler and this will compensate for the gain imbalance present between the two halves of the system.

Clearly a similar analysis to that shown above could be performed for a Q-channel input signal and this would yield a similar result. It is therefore evident that the CALLUM system is capable of both generation of the required LINC signals and of compensating for gain errors between the two RF paths. By modifying, slightly, the above analysis, it is also possible to show that phase errors between the upper and lower paths are also compensated by the CALLUM technique. Two of the fundamental difficul-

ties associated with the LINC technique are therefore addressed, for narrowband applications, with this configuration.

The above system analysis has been performed on a CALLUM system without loop integrators. It is possible to add loop integration to both of the loops and this provides an improvement in tracking performance [23], although care must be taken to avoid overshoot and ringing when faced with step changes in input phase.

7.5.3 Stability of the Basic CALLUM Modulator

The CALLUM modulator consists of two phase-locked loops, loop A and loop B. Loop A will only maintain lock whilst the phase of the RF output signal is within ±90° of the sine of the local oscillator vector, whilst loop B will only maintain lock whilst the phase of the RF output signal is within

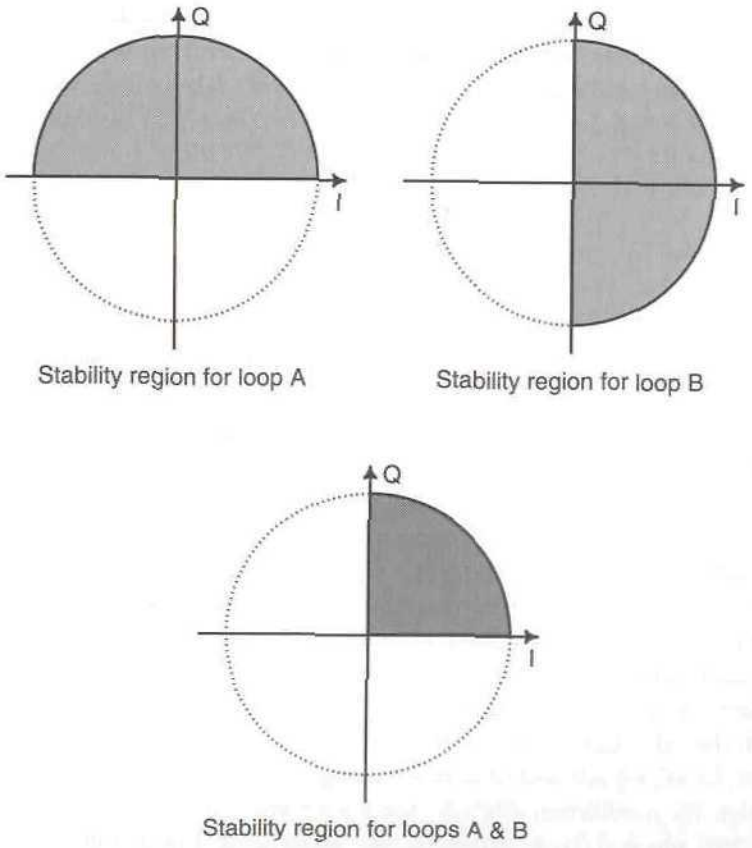


Figure 7.23 Stability regions for the basic CALLUM modulator.

±90° of the cosine of the local oscillator vector. Consequently, there exists a stability region within which both loops can achieve and maintain lock, and this is shown in Figure 7.23. Loop A is stable in the first and second quadrants, whilst loop B is stable in the first and fourth quadrants. Neither loop is stable in the third quadrant.

The only region, therefore, in which the complete modulator is stable is that of the first quadrant. This therefore limits the use of the basic modulator to signals which remain wholly within the first quadrant; relevant modulation schemes include amplitude shift keying (ASK) and full-carrier AM.

Whilst these schemes are useful in some circumstances, for the technique to be truly useful, a method must be found to extend the stable region of operation to all quadrants. Work has been performed to solve this problem by including additional signal processing within the basic CALLUM modulator [22,23]. One option is to control the sign of the VCO input signals in order to ensure that the loop is always operating in its stable region. This can be achieved by means of a switching matrix inserted between the output of the differential amplifier and the input to the VCO in each channel, as shown in Figure 7.24. A high loop gain must be used in this configuration in order to minimise the impulse noise created when the input signal crosses from one quadrant to another.

7.6 Linear Amplification Employing Sampling Techniques (LIST)

7.6.1 Introduction

The use of pulse-width modulation (PWM) in RF linear amplification has traditionally been restricted to low-frequency operation, due to the high sampling frequencies which would otherwise be required. LIST [24] is an attempt to bring the advantages of PWM techniques to RF amplification at sensible carrier frequencies and hence may be viewed as a 'rival' to LINC, in as much as it makes use of the combined output of two highly nonlinear amplifiers.

The LIST technique differs from LINC in a number of important aspects. It does not rely on phase cancellation in the output combiner in order to generate a linearly amplified signal; reconstruction of the linear signal is achieved purely by the output bandpass filter which simulates the action of the low-pass filter used to perform 'detection' of baseband PWM signals. Subtraction does occur at the output of the LIST transmitter,

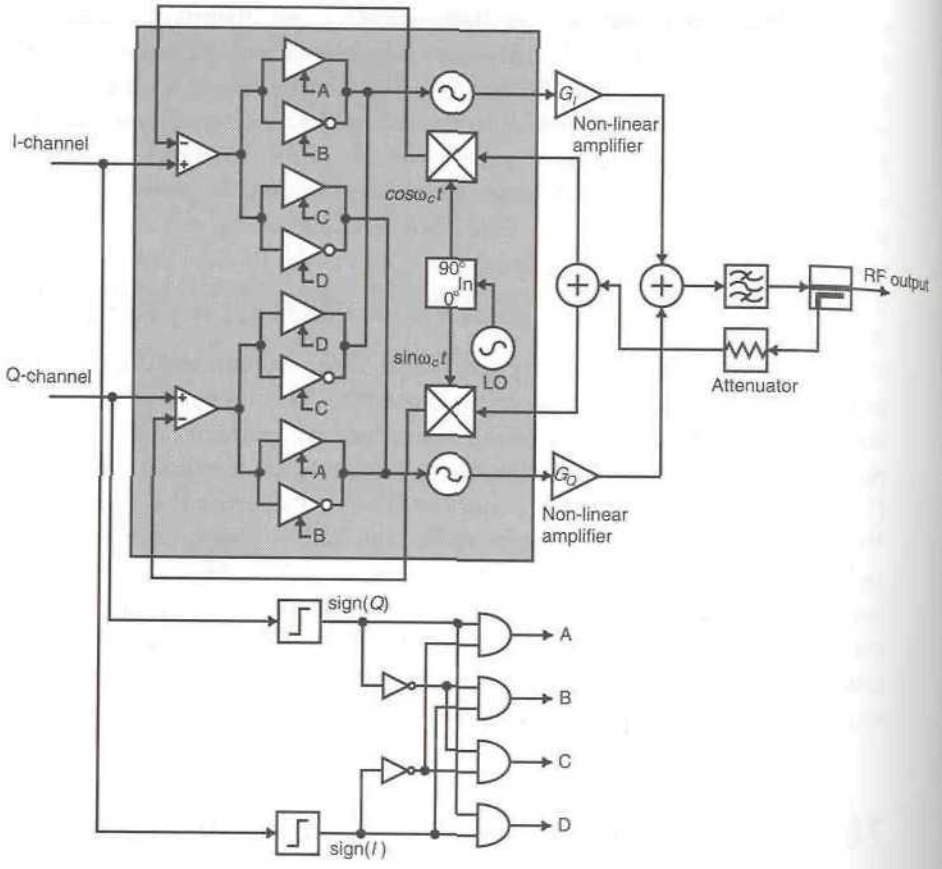


Figure 7.24 CALLUM modulator employing a switching matrix to allow full four-quadrant operation.

however, this is used to remove the image signal and its performance is therefore usually less critical than that of the combiner in LINC (e.g., in the case of an in-band image created by Weaver upconversion). The modulation method used to generate the LIST signals is also essentially digital in nature, whereas the LINC methods are essentially analogue (although they may be implemented using digital techniques, for example, a DSP).

7.6.2 Operation of a LIST Transmitter

Like Cartesian loop and CALLUM (among others), a LIST transmitter takes baseband input signals in I and Q format and performs all of the upconversion and modulation processes, prior to amplification and transmission. It is

therefore a genuine linear transmitter and not merely a linear amplifier.

The basic structure of a LIST transmitter is shown in Figure 7.25. The I and Q signals are fed directly into a delta coder [25] (one per channel) in which the original analog information is converted to a data-stream of value ±K. The resulting output signals are given by (7.87) and (7.88):

$$\Delta I(t) = K\Delta[i(t)] \tag{7.87}$$

$$\Delta Q(t) = K\Delta[q(t)] \tag{7.88}$$

where $\Delta[\bullet]$ represents the delta coding function. In the frequency domain, these signals include alias products centered on a frequency, f_A, as shown in Figure 7.25.

The delta-coded signals (including alias products) are then quadrature upconverted and fed to the two nonlinear power amplifiers. The binary (±K) output from the delta coders results in a carrier phase shift of ±180° during the upconversion process; this is more commonly known as Phase Reversal Keying (PRK) [26]. The resulting upconverted signals are given by:

$$\Delta I_{RF}(t) = AK_{M1}\Delta I(t)\cos(\omega_0 t) \tag{7.89}$$

$$\Delta Q_{RF}(t) = AK_{M2}\Delta Q(t)\sin(\omega_0 t) \tag{7.90}$$

In an optimum system, the power amplifiers would be constructed from, for example, class-E power stages, with no loss of signal fidelity (since the signals are constant-envelope at this point in the system) and with an excellent potential efficiency. Fabrication on a common substrate would, of course, greatly aid tracking of the two paths and hence image (or SVE) performance.

Following the nonlinear amplification, the two paths are combined and fed to a bandpass filter. The combination process should, ideally, be lossless, although this is unlikely to be the case in practice. The combiner problem is, however, different to that of the LINC technique, since the two signal paths do not need to subtract to generate the wanted linear modulation, but only to remove the unwanted image.

A final benefit results from the construction of the signals in quadrature component form. Any imbalance between the two paths should not result in a degradation in linearity performance, but will result in an image signal. This image signal can be arranged to be in-band (using Weaver upconversion [27], illustrated by the spectra in Figure 7.25) and hence the suppression required can be relatively modest. As a result, a gain and phase error between

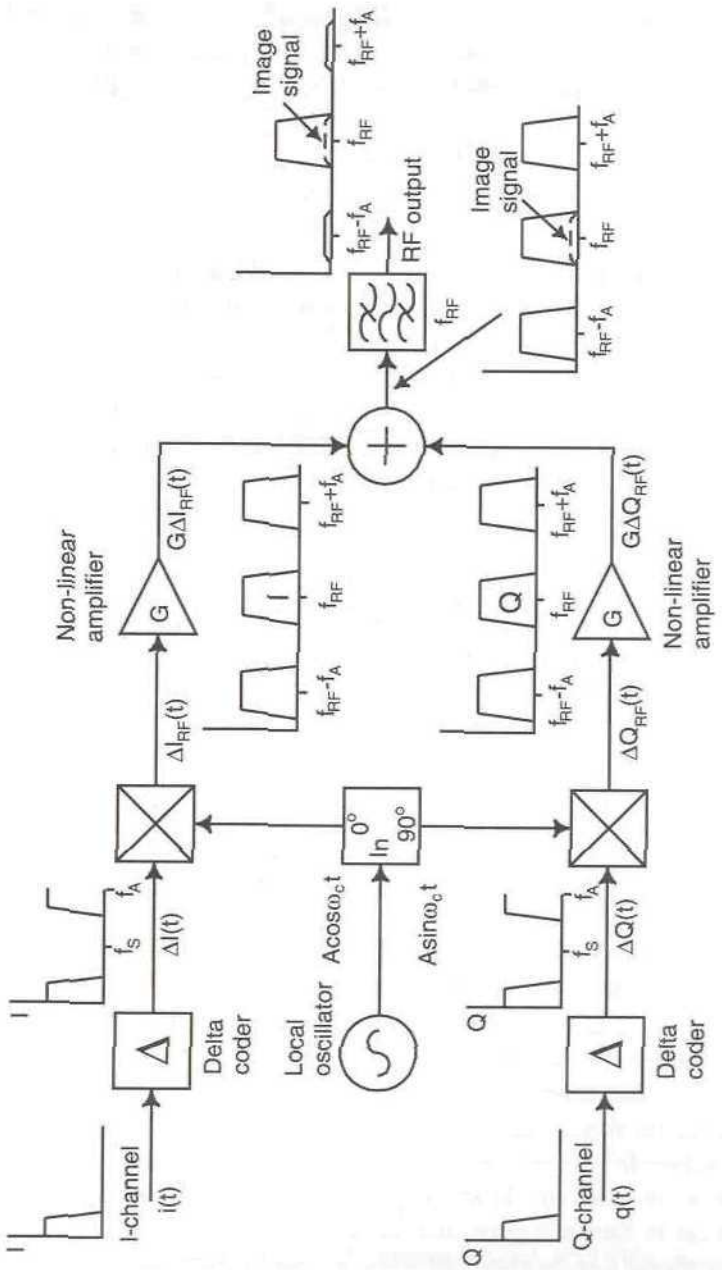


Figure 7.25 Basic architecture of a LIST transmitter.

the two paths of 0.3 dB and 3° may well be adequate for most applications. This specification can be achieved by current commercial I/Q upconverter components and hence external feedback (for this reason) may well not be required, so long as the tracking between the amplifiers is of a similar order. Whilst this latter requirement may be difficult to achieve with discrete amplifier designs over a wide range of temperatures, it should be possible with a matched pair of amplifiers fabricated on the same substrate.

The bandpass filter at the output is required to perform reconstruction of the delta-coded signals; this is equivalent to using a low-pass filter at baseband. A low-pass filter is a suboptimal solution (with the optimal solution being an integrator), however, it is adequate in most cases. The bandpass filter must be sufficiently good to remove the many out of band products generated by the LIST modulator and hence is one of the main drawbacks of this technique.

7.6.3 Bandpass Filter Specification

The bandpass filter is required to remove the out-of-band (or out-of-channel) products generated by upconversion of the wideband delta coded signals. If the sampling frequency of the delta coders is set to be just adequate to cover the wanted channel bandwidth (plus a suitable guard band to allow for the bandpass filter roll-off), then this bandpass filter must either be tunable, or switchable to cover all of the required channels. This is obviously a complex and potentially inflexible solution.

An alternative is to make use of the recent advances in digital logic and comparator technologies and hence to use very high-speed devices in both of these parts of the delta coders (comparators with switching times of 2.5 ns and less have been available for a number of years [28]). The unwanted products should then be forced to appear well out-of-band, thus making the job of the bandpass filter somewhat easier (i.e., it could be combined with the conventional bandpass transmit filter and could even be part of the diplexer filter in a full-duplex radio). This overcomes one of the fundamental drawbacks to the practical use of LIST in multi-channel systems.

The unwanted products mentioned above arise due to the aliasing from what is essentially a sampling process occurring in the delta coders. It is therefore useful to ensure that these products appear well out of band, and this may be achieved by effectively grossly over-sampling the required channel. It is necessary to bear in mind that this over-sampling must ensure a sufficient separation between the wanted channel and the alias products, such that even at the band extremes, the alias products occurring close to the opposite band extreme are capable of removal by the bandpass

filter. This represents the 'worst-case' scenario, with the simplest case occurring when the wanted channel appears in the center of the operational band.

It should be noted here that the heavy reliance of the LIST technique on the performance (and loss) of this output filter is a key disadvantage of this system. The creation of a suitably low-loss, narrow-band bandpass filter is a formidable challenge, particularly at high carrier frequencies, and this may well be the deciding factor in assessing the practicability of a LIST solution in a given application.

7.6.4 Delta Coding

The performance of the delta-coding operation within the LIST transmitter is crucial in obtaining good linear performance from the overall design. In particular, the reconstruction filter must be chosen appropriately in order to take advantage of the use of a wideband bandpass filter as outlined above. This will then result in a transmitter with a bandpass performance characteristic suitable for use over a complete system bandwidth, rather than merely a single channel bandwidth. In addition, the specification of the bandpass filter becomes more realistic, since obtaining a good bandpass filter with a pass-band bandwidth of (say) 25 kHz at (say) 900 MHz would be extremely challenging.

A standard delta-coder schematic is shown in Figure 7.26. The coder produces an output binary value corresponding to the sign of the difference between the analog input signal, $i(t)$, and the predicted value for this signal, based on the recent past history of the signal. In the case of a linear delta modulation system, this predicted signal, $i_o(t)$, is simply generated by filtering the binary output stream using (ideally) an integrator.

When a varying input signal is fed to the delta coder, the comparator will yield a value of $+1$, if the input signal voltage is greater than that predicted by the reconstruction integrator and -1, if it is less. This signal is then sampled by the D-type flip-flop, resulting in a sampled data stream of value $\pm K$ at the output of the coder (with relevant level-shifting if necessary). The predicted value is derived by integrating this output stream, so that, for example, a continuous stream of value $+K$ would result in the integrator ramping up, which is the direction in which the signal could be predicted to continue, based on the recent past history of $+K$ values.

The 'receiver' for a delta coded signal can therefore be deduced as an integrator of the type used as the reconstruction filter (Figure 7.26(b)). This equivalence is exploited later when considering reconstruction utilising a

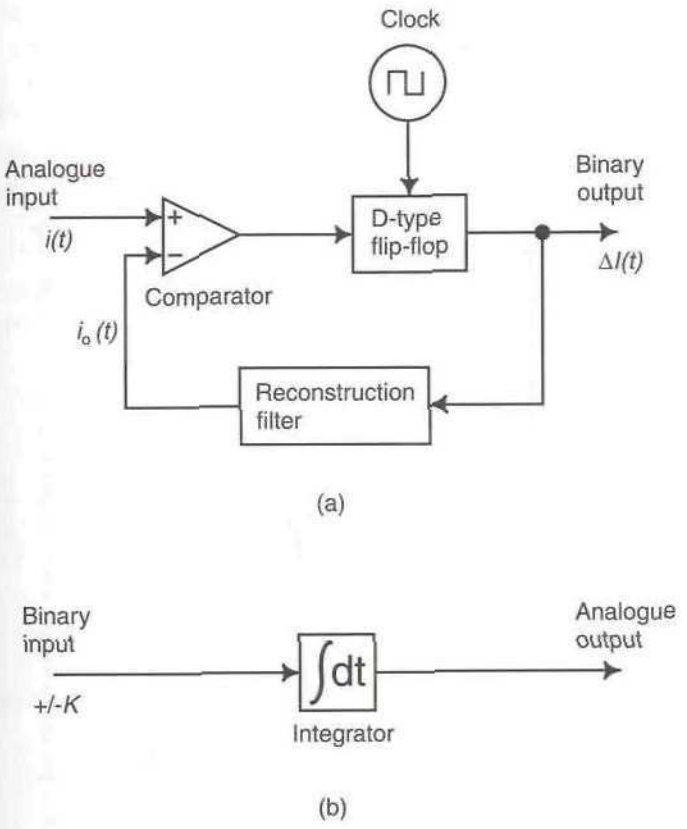


Figure 7.26 Schematic of a delta coder system: (a) basic structure; (b) decoder.

bandpass filter at the output of the LIST transmitter.

Although simple to construct, the delta coder is difficult to analyse, and a number of texts have attempted to do this (e.g., [29,30]). Some of the resulting practical errors and limitations are outlined below, in the context of a LIST transmitter.

7.6.5 Errors in Delta Coding

There are two main sources of error in delta coding: *granular noise* and *slope overload*. It is necessary to minimise both of these effects in order to achieve a good level of in-band and out-of-band performance for the complete LIST transmitter. Both of these effects are illustrated in Figure 7.27 for a delta coder employing a digital integrator. The use of a digital integrator allows the origins of both effects to be seen more clearly, however, these effects will

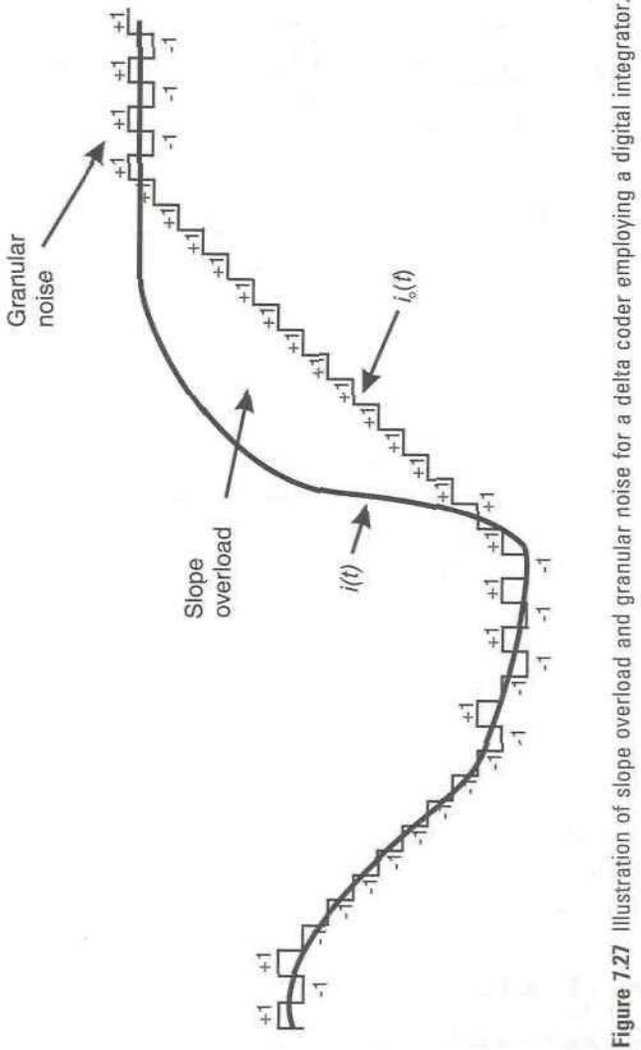


Figure 7.27 Illustration of slope overload and granular noise for a delta coder employing a digital integrator.

also be experienced with an analog integrator.

As can be seen from Figure 7.27, *granular noise* is effectively the quantisation noise of the sampled system and is controlled by the step size, κ. Therefore a reduction in the step-size, and hence an increase in the clock frequency, of the coder will result in a reduction in the amount of granular noise experienced in the reconstructed signal. In the case of a LIST transmitter, this will result in a lower output noise floor for the system and will ease the specification placed upon the output bandpass filter.

For a system employing an analog integrator, the granular noise level will be given by [31]:

$$N_q = \frac{\kappa^2}{3} \tag{7.91}$$

Thus, the signal to (granular) noise ratio for a sampled sinewave, with amplitude A, is:

$$SNR_q = \frac{3A^2}{2\kappa^2} \tag{7.92}$$

The step-size, κ, is given by:

$$\kappa = \frac{V_{in}}{RC} \cdot \frac{1}{f_s} \tag{7.93}$$

where RC is the integrator timeconstant, V_{in} is the input signal level to the integrator and f_s is the sampling (clock) frequency.

Slope overload occurs when the maximum slope of the reconstructed signal is exceeded by the input signal. The integrator is no longer able to accurately track the input signal and the resulting reconstructed signal is distorted (Figure 7.27). The effect of this upon the output spectrum of a LIST transmitter is to create sidebands around the wanted signal at harmonics of the baseband input signal frequency. Thus, for a 25 kHz sinewave input signal, sidebands at, for example, 50, 75, and 100 kHz from the channel centre frequency would be created. This will result in adjacent channel interference (ACI) being created by the transmitter and may cause the adjacent channel power (ACP) mask to be broken. These cannot be removed by the proposed bandpass filtering at the output and hence slope overload must be avoided at all costs during the design of the transmitter.

The maximum slope of the reconstructed signal is limited by κf_s; if the rate of change of the input signal is greater than this value, then the

reconstructed signal will be unable to track the input signal accurately. Consider a sinusoidal input signal given by:

$$x(t) = A \sin \omega_m t \tag{7.94}$$

This may be viewed as the maximum signal frequency in a Fourier analysis of a more complex modulating signal, allowing the delta coder to be designed appropriately. The maximum rate of change of the input signal can be found using differentiation and this must be less than the maximum allowable slope, hence:

$$\kappa f_s \geq A\omega_m \tag{7.95}$$

or

$$\alpha \equiv \frac{\kappa f_s}{A\omega_m} \equiv \frac{\kappa f_s}{2\pi AB} \geq 1 \tag{7.96}$$

where B is the bandwidth of the baseband input signal.

When an analog integrator is used as the reconstruction filter, the step size is given by (7.93). Substituting this expression into (7.96) gives:

$$\alpha = \frac{V_{in}}{2\pi ABRC} \tag{7.97}$$

which is independent of the clock frequency employed in the coder. Therefore, in this case, it may be concluded that an increase in the clock frequency will result in a reduction of the granular noise, but will not affect the slope overload characteristics.

If the inequality expressed in (7.96) is not satisfied, then slope overload appears, and the resulting signal to noise ratio is given by [29]:

$$\frac{S_o}{N_o} = \frac{A^2}{2N_o} = 0.56(1-\alpha)^{-5/2} \tag{7.98}$$

The range of validity for (7.98) is given by: $0.7 \leq \alpha \leq 1$.

In this case, the optimum choice of the step size is therefore a compromise between avoiding slope overload and achieving a low granular noise level. Note that this compromise only appears if the inequality given in (7.96) is not satisfied. Otherwise, the choice of step size depends only on the granular noise, because slope overload no longer occurs.

Finally, the relationship between the signal to (granular) noise ratio and clock frequency results from (7.91) and (7.93). Combining these equations gives:

$$N_q = \frac{V_{in}^2}{3R^2C^2f_s^2} \tag{7.99}$$

which shows that the granular noise is inversely proportional to the square of the clock frequency. Thus for a factor of 10 increase in clock frequency, a 40 dB reduction in noise should be anticipated.

7.6.6 Choice of Reconstruction Filter

The choice of the reconstruction filter in the delta coder is important in ensuring that the 'predicted' signal generated within the coder is an accurate representation of what is being achieved at the output of the transmitter. Whilst the ideal reconstruction filter for the coder alone is an integrator, this is not the case when a complete LIST transmitter is considered.

The reconstruction filter at the output of the LIST transmitter is formed using a bandpass filter. The characteristics of this filter do not map well to a baseband integrator due to the relatively flat passband characteristic of the bandpass filter. It is not possible to build an 'integrating' bandpass filter with continuous 20 dB/decade roll-off both above and below the center frequency, as a wideband characteristic is necessary to cover an entire band of operation, without switching or re-tuning.

If an integrator were to be used as the reconstruction filter in the delta coder, and a (relatively) broadband bandpass filter as the transmitter output reconstruction filter, the resulting mismatch between the two characteristics would lead to distortion of the transmitter output spectrum.

It is therefore necessary to use a reconstruction low-pass filter which matches, as closely as possible, the pass-band and roll-off characteristics of the output bandpass filter. In addition, the sampling rate used by the delta coders must ensure that the reconstruction process is taking place in the 20 dB/decade roll-off of the low-pass filter, that is, the sampling rate must be greater than the system (not channel) bandwidth of the required transmitter. This is necessary anyway, to ensure that the sampling clock frequency and granular noise spectrum are both able to be removed by the output bandpass filter.

If an active low-pass filter of the form shown in Figure 7.28 is used in the delta coder, then $R = R_1$ in (7.93) and (7.97). Thus the relevant

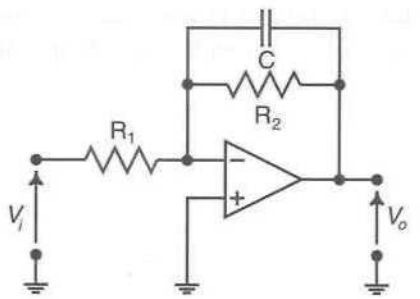


Figure 7.28 Active low-pass filter.

parameters may be calculated to avoid slope overload.

7.6.7 Gain and Phase Imbalances in a LIST Transmitter

Gain and phase imbalances can occur at any point in the baseband, upconversion, power amplification and combining stages of the transmitter, in addition to imperfections in the quadrature generation of the local oscillator. Of these, the baseband sections should be inherently the best matched, as the signal levels are based on digital levels (from a D-type flip-flop). Of the remaining elements, the power amplifier gains will, in general, be the poorest matched. The imbalance mechanisms described here, and their spectral implications, are similar to those described for the Cartesian loop in Chapter 4 (although they are modified slightly here for the particular case of a LIST transmitter).

The effects of gain and phase imbalance may be analysed as follows. With reference to Figure 7.29, the magnitude of the signal error vector may be determined using the cosine rule as:

$$E_V = \sqrt{[R^2 + M^2] - 2RM\cos(\phi_\epsilon)} \tag{7.100}$$

where R is the magnitude of the reference ('ideal') vector, M is the magnitude of the 'actual' vector, and ϕ_ϵ is the phase error between them. The measured ('actual') vector magnitude, M, is composed of the reference vector magnitude, R, plus a component resulting from the gain error present in the system, G_ϵ:

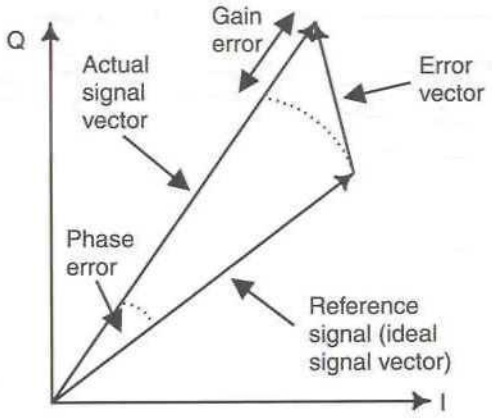


Figure 7.29 Illustration of vector error due to imperfect gain and phase accuracy in the I/Q plane.

$$M = R + G_\epsilon \tag{7.101}$$

If the reference vector magnitude is set to unity, then the magnitude of the resultant error vector, E_{VM}, may be found from:

$$E_{VM} = \sqrt{[1 + M^2] - 2M\cos(\phi_\epsilon)} \tag{7.102}$$

where $M = 1 + G_\epsilon$.

The image suppression resulting from the imperfect gain and phase balance is therefore:

$$S_{dB} = 20\log\left\{\sqrt{[1 + (1 + G_\epsilon)^2] - 2(1 + G_\epsilon)\cos(\phi_\epsilon)}\right\} - 6 \text{ dB} \tag{7.103}$$

where the additional factor of 6 dB arises from the fact that the components forming the wanted sideband sum in phase, hence resulting in a wanted signal to image ratio (usually referred to as an 'image suppression') 6 dB greater than the cancellation of the unwanted sideband alone.

A typical specification, for example, is a gain error of 0.3 dB (i.e., $1 + G_\epsilon = 1.035$) and a phase error of 3°; using (7.103), this results in an image suppression, relative to the wanted sideband, of 30 dB, which is acceptable for many applications. These values are well within the capability of current quadrature modulators, but would require careful balancing of the two power amplifiers. This could be achieved by fabrication of both in the same

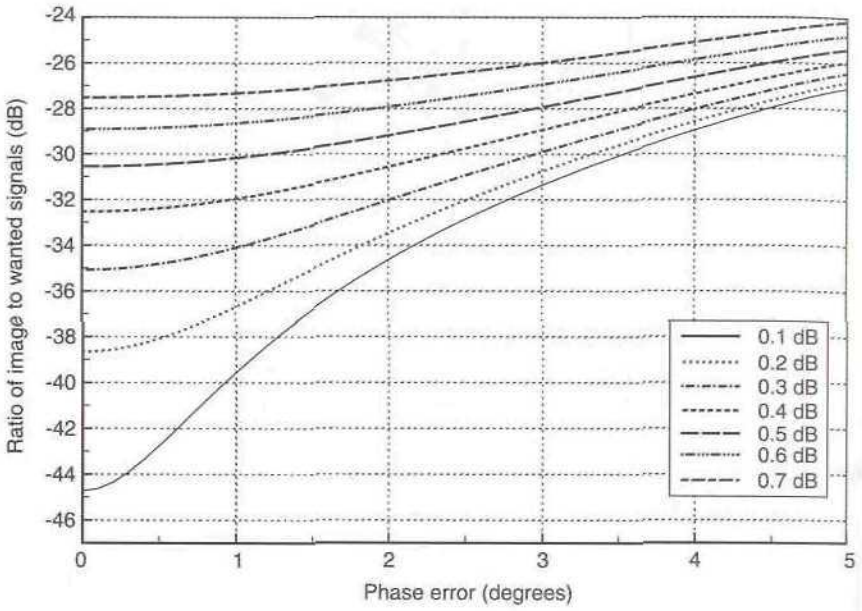


Figure 7.30 Image suppression in a LIST transmitter.

substrate (in the case of a handset transmitter) or by means of an automatic control technique for larger, higher-power systems. Such control techniques are used in feedforward systems for maintaining a similar 'balance' (although in that case for IMD rather than image cancellation).

An illustration of a range of gain and phase error values and their effect on a LIST transmitter is shown in Figure 7.30. Note that the gain error values in this figure ($G_{err,dB}$) indicate the gain between one branch in the LIST transmitter and the other. They are related to the gain error, G_e, by the following relationship:

$$G_{err,dB} = 20\log(1 + G_e) \tag{7.104}$$

It is possible to 'predistort' the input signal vectors in order to compensate for the errors present in an upconverter (and hence a LIST transmitter) [32], however, this usually requires some form of calibration on a 'per-unit' basis and this is generally undesirable in a production environment, unless absolutely necessary.

7.6.8 Sources of Noise and Distortion

There are two potential sources of noise and distortion in a LIST system (assuming that no slope overload can occur):

1. Signal-to-noise ratio (SNR) degradation produced by the delta coding itself.
2. Quadrature distortion caused by phase and/or amplitude imbalances in the quadrature paths. Careful (possibly automatic) control of these parameters is required in order to maintain a 'clean' spectrum (cf. image rejection in a Weaver upconverter). Some methods of control are discussed in Chapter 4.

   The use of LIST for narrowband (single carrier) applications results in these quadrature constraints being somewhat relaxed, since most systems can tolerate quite a high level of co-channel interference (e.g., −30 dB) without any noticeable degradation in SNR or BER performance. However, this problem becomes very significant if the LIST technique is considered for use in multicarrier systems with a non-symmetric distribution of carriers.

The SNR degradation introduced by the delta modulation PWM coding may be reduced to an arbitrarily low level by suitable choice of the sampling frequency and corresponding step size, κ. Results are presented in the literature [24] which indicate that for a SNR of 55 dB (often considered a minimum for mobile radio applications), a sampling rate of around 200 times the baseband bandwidth is required. This corresponds to a sampling rate of 5 MHz for a 25 kHz channel, which is easily achievable with most logic systems. The sampling rates become less achievable for multicarrier systems, but may still be realisable for relatively narrowband systems.

7.6.9 Feedback Correction

It is possible to employ Cartesian feedback around a LIST transmitter in order to overcome some of the shortcomings mentioned above. This arrangement is shown in Figure 7.31 and its analysis is similar to that detailed in Chapter 4 for the Cartesian loop.

The use of feedback will force the output signal to closely resemble the input signal, but the gain/bandwidth/delay constraint it introduces will inherently limit the system's achievable bandwidth. The forward path delay can be minimised by correct choice of the feedback low-pass filter bandwidths; they are only required to remove mixer products and should not perform any additional reconstruction filtering of the delta modulated signal (which

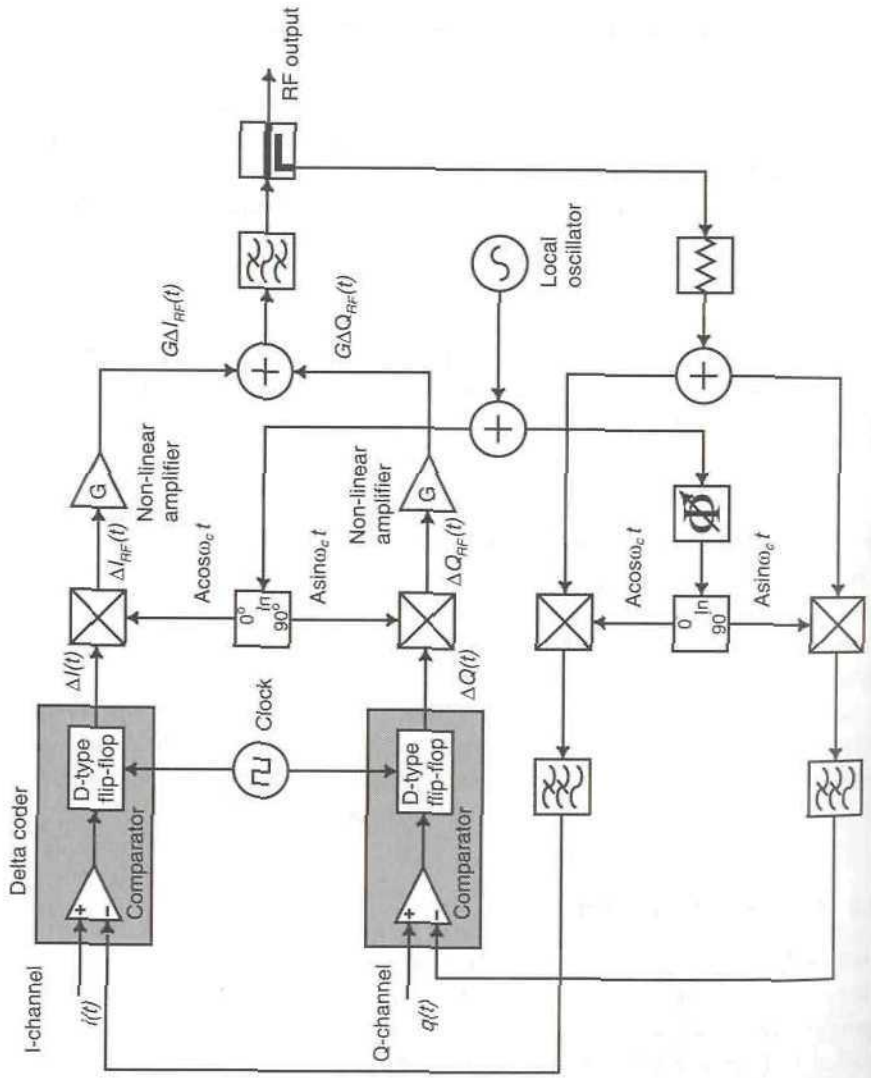


Figure 7.31 Quadrature feedback applied around a LIST transmitter (after [24]).

should be fully reconstructed at this point by the output bandpass filter).

7.6.10 Efficiency and Combining Issues With LINC and LIST Transmitters

It is evident that a key benefit of both the LINC and LIST techniques is in their potential for very high levels of power efficiency. In both cases the realisation of this high level of efficiency is critically dependent on the output combiner and, in particular, on what happens to the signal energy when subtraction of the two signal vectors takes place. Ideally, this should either result in no energy being consumed, or in the energy being recovered in some way and hence not wasted as heat.

7.6.10.1 Efficiency of the LINC or LIST Techniques

Assuming that the combining process at the output of the two high-efficiency power amplifiers (or oscillators) is 100% efficient, and that the amplifiers themselves are 100% efficient, then the overall LINC (or LIST) amplifier efficiency is also 100% *at all envelope levels* (i.e., not only at PEP, as with many other high-efficiency linear amplifier designs). This statement does, of course, assume that the signal generation mechanism consumes no power, although in practice this can be made negligibly small in a high power amplifier. The situation becomes somewhat less clear for a lower power (e.g., handportable) power amplifier, when the signal separation and generation mechanism becomes significant in terms of relative power consumption [33].

7.6.10.2 Output Combiner Issues

A key problem with the LINC or LIST techniques is in the design of the output combiner. If a conventional 3 dB hybrid (or equivalent) combiner is used, 50% of the power (at least) is wasted in the combiner. This is due to the subtracted signal energy being 'dumped' in the load of the hybrid and hence wasted as heat. Alternative passive combiner designs are possible which may be optimised for a particular modulation type, but these are still sub-optimal.

There are two alternative, theoretically optimal, solutions:

1. Design the amplifiers and combining process as one element and configure the amplifier output stages such that a *voltage* summation occurs at the summing point. This is akin to, for example, the process occurring at the input of a summing operational amplifier stage.
2. Utilise an energy recovery process to 're-use' the energy which

would otherwise be dissipated in the load, when using a conventional hybrid (or equivalent) combiner. This method is covered in more detail below.

Of these two alternatives, the former has the greater potential for achieving near 100% efficiency, but is also the greater design challenge.

7.6.10.3 Simple Combiner Architecture Employing Energy Recovery

One solution to the 'energy recovery' method of improving efficiency in the LINC or LIST techniques is outlined below. It has been used experimentally at 220 MHz, by the author, and has given some promising results in that frequency band. More recently, a similar scheme has been proposed at 1.9 GHz [34].

The basis of the technique is shown in Figure 7.32. The outputs of the RF power amplifiers are combined using a hybrid combiner, with the 'sum' port forming the wanted RF output and the 'difference' port being connected to an energy recovery network. For correct operation of the hybrid, the difference port must be terminated in a resistive 50Ω load and this would normally be a high-power resistor designed to absorb all of the unwanted signal energy. This is clearly a wasteful technique and hence a network which can recover this energy as efficiently as possible is highly desirable.

The network consists of an impedance matching circuit, an RF rectifier, a smoothing circuit and a dc to dc converter. The purpose of this network is to replenish some of the energy consumed from the battery and hence overcome some of the energy wastage. Theoretically, all of the energy could be recovered in this way and hence the ideal of a 100% efficient transmitter could be realised in this manner. In practice, however, there are a number of losses in this system and hence only a proportion of the energy is recovered.

An experimental system, constructed for use in the 220 MHz band, shows that although inefficiencies do exist, the technique is still useful. A

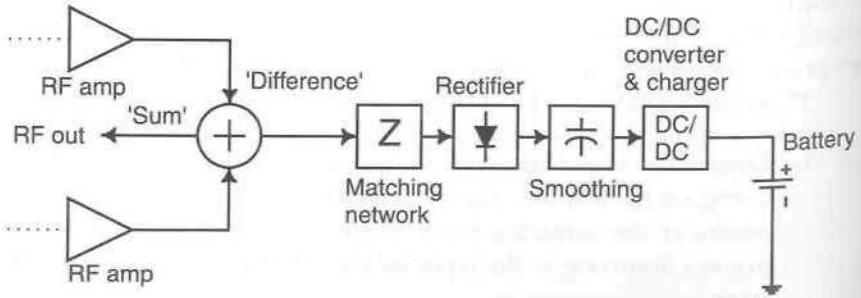


Figure 7.32 Simple energy recovery combiner for use in LIST (and LINC) systems.

Table 7.1
Results from the energy recovery combiner at 220 MHz

Frequency range	220–225 MHz
Combiner/rectifier efficiency	71%
Switching converter efficiency	90%
Battery charging efficiency	70%
Overall efficiency	45%

number of different rectifier types were tried experimentally and the results are based on the most efficient of those used. The efficiency range of the different devices studied was 52% to 71% (including losses in the combiner itself, although these were small). These results are summarised in Table 7.1.

The overall recovery efficiency of around 45%, whilst not spectacular is still a useful improvement over a standard hybrid combiner. Thus, for example, a system employing this technique along with a pair of class-E amplifiers at 80% efficiency would achieve an overall efficiency of 58%. This is similar to many class-C amplifiers and hence is quite a reasonable figure, particularly bearing in mind that this is a linear transmitter and that the efficiency should remain largely unchanged at all envelope levels.

The methodology employed in this technique may not be appropriate for all transceiver designs or battery types, but it is still useful in the areas where it can be employed successfully. Similarly, it is not necessary to actually 'charge' the battery, since a greater amount of power is drawn from it in supplying the transmitter; the purpose of the circuit, in most designs, would therefore be to supply some of the power for the transmitter. This could increase the 'battery charging efficiency' figure in Table 7.1 to perhaps 90% or more. This would result in an overall efficiency for the transmitter of > 63%.

A similar scheme has been proposed for use at 1.9 GHz [34] and has been shown to achieve a peak reuse efficiency of over 60%. This paper also highlights the fundamental problem with this approach, namely that the power recovery circuit can only be ideally matched at *one* input power level, hence the conversion efficiency will degrade at higher or lower power levels. The power conversion efficiency also critically depends upon the power supply voltage for the system into which the energy is being returned, the series resistance of the rectifier diodes (usually Schottky) and their cut-off frequency and built-in voltage.

An approximate analysis for the power returned may be derived as follows (based on [34]), assuming a simple resistive model for the diode. The power returned to the supply is approximately:

$$P_r = V_S I_{D,dc} \tag{7.105}$$

where V_S is the power supply voltage and $I_{D,dc}$ is the DC current from the recovery power converter.

The power available for return (assuming perfect conversion from RF to DC) is given by:

$$P_{RF-DC} = \frac{V_{S,rms}^2}{4R_S} \tag{7.106}$$

Where $V_{S,rms}$ is the output voltage of the output combiner difference port and R_S is its source impedance (usually 50Ω).

The overall efficiency of the power recovery circuit is then:

$$\eta = \frac{P_r}{P_{RF-DC}} \tag{7.107}$$

The instantaneous current through the converter is given by (approximately):

$$I_D(t) = \begin{cases} \left(V_{pk}\sin\omega_c t - (V_S + V_D)\right)/R_D & \forall\ V_{pk}\sin\omega_c t \geq (V_S + V_D) \\ \left(-V_{pk}\sin\omega_c t + (V_S + V_D)\right)/R_D & \forall\ V_{pk}\sin\omega_c t \leq (-V_S - V_D) \\ 0 & \forall\ \left|V_{pk}\sin\omega_c t\right| < (V_S + V_D) \end{cases} \tag{7.108}$$

where V_{pk} is the magnitude of the diode network input voltage, V_D is the diode built-in voltage (forward voltage drop), R_D is the diode series 'on' resistance and ω_c is the carrier frequency. The DC value of the diode current can therefore be approximated by:

$$I_{D,dc} \approx \frac{2\left[\sqrt{V_{pk}^2 - (V_S + V_D)^2} - \theta(V_S + V_D)\right]}{\pi R_D} \tag{7.109}$$

where $\theta = \cos^{-1}\left\{(V_S + V_D)/V_{pk}\right\}$ for $V_{pk} > V_S + V_D$.

The impedance of the converter at the carrier frequency is given by:

$$Z_D \approx \frac{\pi R_D}{2}\left[\theta - \frac{(V_S + V_D)\sqrt{V_{pk}^2 - (V_S + V_D)^2}}{V_{pk}^2}\right]^{-1} \tag{7.110}$$

As V_{pk} changes (with input power) the impedance given by (7.110) will also change and hence an ideal match is only possible at one input power level. At other power levels, the impedance mismatch will cause the power recovery process to have additional losses. For an impedance transformation of n^2, then the source voltage will be given by:

$$V_{S,rms}^2 \approx \frac{n^2 V_{pk}^2}{2}\left[\frac{R_S + Z_{in}}{Z_{in}}\right]^2 \tag{7.111}$$

and the power delivered to the detector is then:

$$P_{del} \approx \frac{V_S^2 Z_{in}}{(R_S + Z_{in})^2} \tag{7.112}$$

where $Z_{in} = n^2 Z_D$.

Combining (7.111), (7.109), (7.106) and (7.107), allows the overall power reuse efficiency to be calculated.

References

1. Kahn, L. R., "Single-sideband transmission by envelope elimination and restoration," *Proc. of the IRE*, Vol. 40, July 1952, pp. 803–806.
2. Kahn, L. R., "Comparison of linear SSB transmitters with envelope restoration transmitters," *Proc. of the IRE*, December 1956, pp. 1706–1712.
3. Raab, F. H., "Envelope elimination and restoration concepts," *Proc. of RF Expo East*, Boston, USA, November 1987, pp. 167–177.
4. Koch, M. J., and R. E. Fisher, "A high efficiency 835 MHz linear power amplifier for digital cellular telephony," *Proc. of the 39th IEEE Vehicular Technology Conference*, San Fransisco, California, USA, Vol. 1, May 1989, pp. 17–18.
5. Raab, F. H., "Intermodulation distortion in Kahn-technique transmitters," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 44, No. 12, December 1996, pp. 2273–2278.
6. Raab, F. H., and D. J. Rupp, "Class-S high-efficiency amplitude modulator," *RF Design*, May 1994, pp. 70–74.
7. Raab, F. H., and D. J. Rupp, "High-efficiency single-sideband HF/VHF transmitter based upon envelope elimination and restoration," *Proc. of the 6th International Conference on HF Radio Systems and Techniques*, IEE Conference Publication 392, York, UK, 4–7 July 1994, pp. 21–25.
8. Raab, F. H., and D. J. Rupp, "High-efficiency single-sideband HF/VHF transmitter

based upon envelope elimination and restoration," *IEE Conference on HF Radio Systems and Techniques*, Conference Publication No. 392, July 1994, pp. 21–25.
9. Raab, F. H., "Envelope elimination and restoration concepts," *Proc. of RF Expo East*, Boston, USA, November 1987, pp. 167–177.
10. Raab, F. H., B. E. Sigmon, R. G. Myers, and R. M. Jackson, "High-efficiency L-band Kahn-technique transmitter," *Proc. of IEEE MTT-S*, Baltimore, USA, Vol. 2, June 1998, pp. 585–588.
11. Su, D., and W. McFarland, "An IC for linearizing RF power amplifiers using envelope elimination and restoration," *IEEE International Solid State Circuits Conference*, paper TP 3.6, February 1998, pp. 54–55.
12. Staudinger, J., *et al.*, "800 MHz power amplifier using envelope following techniques," *Proc. of the IEEE Radio and Wireless Conference (RAWCON) 1999*, Denver, Colorado, August 1999.
13. Cox, D. C., "Linear amplification with nonlinear components," *IEEE Trans. on Communications*, Vol. COM–22, December 1974, pp. 1942–1945.
14. Chireix, H., "High power outphasing modulation," *Proc. of the Institute of Radio Engineers*, Vol. 23, No. 11, November 1935, pp. 1370–1392.
15. Bateman, A., "The combined analogue locked loop universal modulator (CALLUM)," *IEEE Vehicular Technology Conference*, May 1992, pp. 759–763.
16. Hetzel, S. A., A. Bateman, and J. P. McGeehan, "LINC transmitter," *IEE Electronics Letters*, Vol. 27, No. 10, 9 May 1991, pp. 844–846.
17. Hetzel, S. A., A. Bateman, and J. P. McGeehan, "A LINC transmitter," *IEEE Vehicular Technology Conference*, St. Louis, Missouri, USA, 19–22 May 1991, pp. 133–137.
18. Sundstrom, L., "Automatic adjustment of gain and phase imbalances in LINC transmitters," *IEE Electronics Letters*, Vol. 31, No. 3, 2 February 1995, pp. 155–156.
19. Ren, Q., and I. Wolff, "Influence of some imperfect system performances on linearizers," *MTT-S IEEE International Microwave Symposium Digest*, 1998, Vol. 2, pp. 973–976.
20. Senderowicz, D *et al.*, "An NMOS integrated vector-locked loop," *IEEE International Symposium on Circuits and Systems*, vol. III, 1982, pp. 1164–1167.
21. Vector Locked Loop, US patent number 5,105,168, 14 April 1992.
22. Chan, K. Y., and A. Bateman, "Linear modulators based on RF synthesis: realization and analysis," *IEEE Trans. on Circuits and Systems—I: Fundamental Theory and Applications*, Vol. 42, No. 6, June 1995, pp. 321–333.
23. Bateman, A., and K. Y. Chan, "Analytical and measured performance of the combined analogue locked loop universal modulator (CALLUM)," *IEE Proc. on Communications*, Vol. 42, No. 5, October 1995, pp. 297–306.
24. Cox, D. C., "Linear amplification by sampling techniques: a new application for delta coders," *IEEE Trans. on Communications*, Vol. COM–23, No. 8, August 1975.
25. Black, H. S., *Modulation Theory*, Van Nostrand, Princeton, N.J., 1955.
26. Stremler, F. G., *Introduction to Communication Systems*, Second Edition, Addison-Wesley, 1982.
27. Weaver, D. K., "A third method of generation and detection of SSB signals," *Proc. of the Institute of Radio Engineers*, No. 44, 1956, pp. 1703–1705.
28. Analogue Devices, *1992 Amplifier Reference Manual*, pp. 3–21.
29. Steele, R., *Delta Modulation Systems*, Pentech Press, 1975.
30. Schindler, H. R., "Linear, non-linear and adaptive delta modulation," *IEEE Trans. on Communications*, Vol. COM–22, No. 11, November 1974, pp. 1807–1823.
31. Schwartz, M., *Information, Transmission, Modulation and Noise*, 3rd Edition, McGraw-

8

Efficiency Boosting Systems

8.1 Introduction

A number of power amplification techniques are often described as performing amplifier linearisation, when in actual fact they are merely boosting the efficiency of already linear amplifiers. Whereas amplifier linearisation schemes will often attempt to make use of the least linear (and hence most efficient) amplifier possible and then linearise it to provide the necessary low distortion, efficiency-boosting schemes will attempt to increase the efficiency of class-A, -AB or -B amplifiers.

The levels of distortion resulting from such systems can obviously be no better than the basic performance of the original (e.g., class-A) amplifier and hence the range of applications for such techniques is necessarily limited. However, they do provide a useful alternative for applications where efficiency is the primary concern and linearity is a necessary, but secondary consideration. Satellite systems are an obvious example, with broadcast transmitters being another.

All of the techniques are applicable to RF applications, although the applicable frequency range of some is fundamentally limited.

8.2 Doherty

The Doherty technique, first proposed in 1936 for use in high-power broadcast transmitters [1], has continued to find application in this area due to its relative simplicity of implementation and ease of application to

such high-power systems [2–7]. As was suggested in Section 8.1, it is not a linearisation technique as such, it is more an efficient method of utilising conventional linear amplifiers and overcoming some of their inherent inefficiency when operated with envelope-varying signals.

The efficiency gains which are possible with this technique can be as high as two or three times that of a conventional class-B PA when operating below PEP (the efficiency being approximately the same at PEP). The use of this system alone does, however, assume that the linearity of a conventional class-B PA is acceptable for the required application, since the linearity achieved by the Doherty system is largely reliant on the linearity of the two (or more) class-B amplifiers used in its implementation. This does not, of course, preclude the application of a linearisation technique in addition to the Doherty technique, but such systems will not be considered here.

The implementation of a two-stage Doherty system is shown in Figure 8.1. The impedance inverting network at the output of $A1$ is shown as a quarter-wave transmission line although it could, of course, be replaced by a suitable lumped-element equivalent with no loss of functionality.

The drive control is shown as operating on the input signals to the two power amplifiers, although it could equally be applied to the bias on the amplifiers themselves.

There are a number of stipulations on the design of the power amplifiers $A1$ and $A2$. They must not generate signals at the harmonics of the carrier frequency and hence are often implemented utilising class-B push-pull amplifiers, which inherently suppress even-order harmonics (see Chapter 3). Alternatively, other linear amplifier configurations can be used, such as single-ended designs, with the addition of harmonic suppression filters at their outputs. Additionally, the active devices used in the two amplifiers must be adequately rated (particularly those in $A1$) to allow them to operate in compression without damage.

Correct phasing of the outputs from $A1$ and $A2$ is essential to the operation of the amplifier and errors result in severe distortion of the modulation. The quarter-wave delay (or lumped-element equivalent) in the input path to $A2$ compensates for the $\lambda/4$ delay in the output path of $A1$ and hence ensures that the correct phase relationship is restored.

8.2.1 Operation of the Doherty Technique

The basic concept behind the Doherty technique is to allow one or more amplifiers to operate at their PEP level, and hence at maximum efficiency, whilst allowing a 'final' linear amplifier to deal with the modulation peaks. At lower envelope levels the lower-order amplifier(s) take over the linear

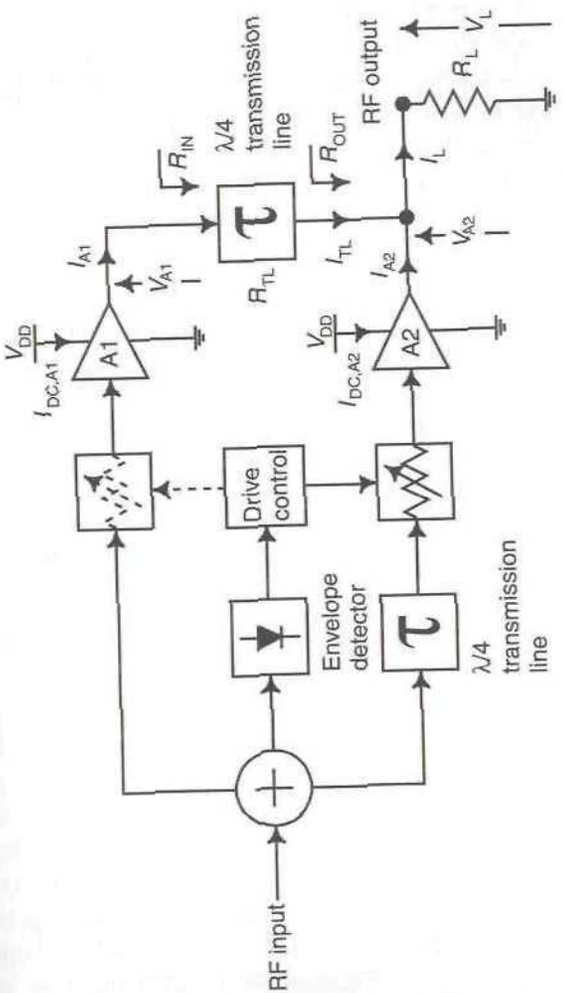


Figure 8.1 Schematic of a Doherty-type amplifier.

amplification function, with the higher-order amplifiers being shut off by the drive control circuitry.

In the specific case of a two-stage Doherty amplifier, its operation may be described as follows. At low power output levels, $A2$ is shut down, either by removal of its drive signal or by a suitable alteration of its bias level. $A1$ receives all of the input signal and operates in a conventional linear mode. The impedance present at its output, due to the impedance transformation performed by the $\lambda/4$ transmission-line, ensures that $A1$ saturates at a level well below the desired system PEP. This level is known as the *transition point*. In a standard two-stage Doherty system, the transition point occurs at half of the maximum output voltage. At this transition point, the first amplifier ($A1$) is operating at maximum efficiency, with the second ($A2$) shut off; the overall system is therefore operating at maximum efficiency.

At all output levels above the transition point, $A2$ commences activity and takes over operation as a linear amplifier, performing as a controlled current source. $A1$ remains saturated and therefore operates as a controlled voltage source, the action of the $\lambda/4$ transmission-line converting this to appear as a current source at the point where the outputs from $A1$ and $A2$ combine. The form of the output voltage from $A1$ and current from $A2$, together with the combined signal, are shown in Figure 8.2 for a full-carrier AM signal with a modulation index of approximately 0.7.

The portion of the load current from $A2$ increases the RF output voltage from the complete amplifier above that due to $A1$ alone and hence the impedance seen by the $\lambda/4$ transmission-line at its output is greater than the actual load impedance. Furthermore, as the contribution of $A2$ increases, the impedance seen by the $\lambda/4$ transmission line increases and therefore the load impedance seen by $A1$ decreases. The effect of this decreasing impedance seen by $A1$ is to increase its output contribution as well. Therefore the output power from both $A1$ and $A2$ will increase with increasing signal level until full PEP is reached.

Within the region between the transition point and full PEP, the efficiency of $A1$ will remain at its maximum value and the efficiency of $A2$ will vary between half its maximum level (at the transition point) and its maximum level (at its rated PEP). When both amplifiers are operating at their rated PEP levels, the overall system is operating at full PEP and hence the efficiency of a two-stage Doherty system is a maximum at both the transition point and at full PEP.

The variable attenuator prior to $A2$ has two functions. First, it serves to shut down the input to $A2$ at low envelope levels (below the transition point), thus ensuring that $A2$ makes no contribution to the overall RF output. This could also be achieved by an appropriate change to the biasing

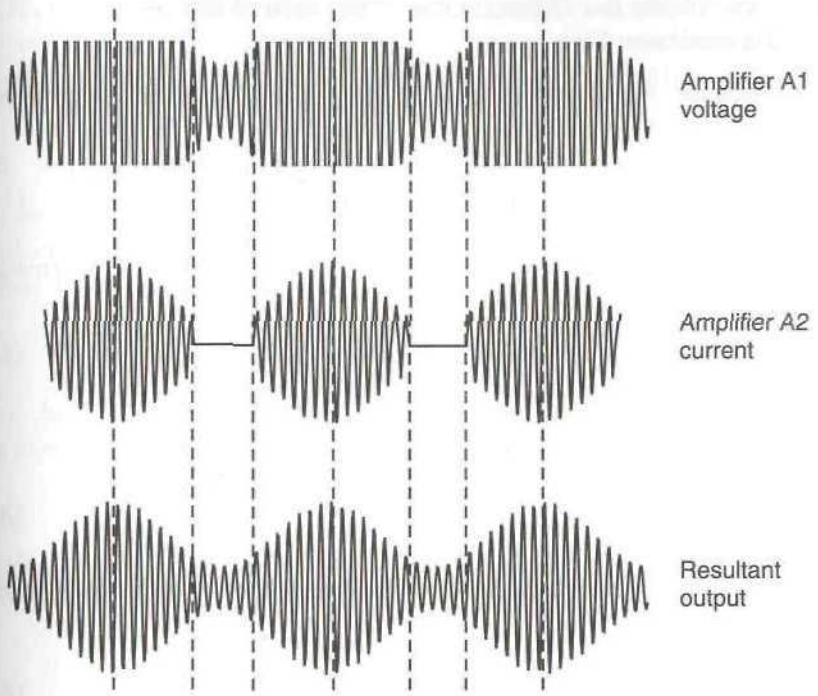


Figure 8.2 Envelopes of the voltage and current outputs from $A1$ and $A2$ respectively, together with the resultant combined output of the Doherty amplifier.

of $A2$ or the use of a PIN diode switch. Secondly, it is required to provide appropriate gain adjustment with drive level, based upon the transconductance characteristic for $A2$. This is a highly nonlinear function, for example, for an FET and this may be best achieved by means of a look-up table within a DSP device, driven by the output from the envelope detector. This is the 'drive control' function illustrated in Figure 8.1.

It is not strictly necessary to control the input level to $A1$ for correct operation of an ideal Doherty system, however, it may be advantageous in overcoming non-ideal behaviour of $A1$ with changing drive level, particularly beyond the transition point. An attenuator may therefore be used in this position also, as shown in Figure 8.1.

The following is an analysis of the operation and efficiency characteristics of a Doherty system and is based on that performed by Raab [8]. It assumes that the amplifiers employed are ideal class-B devices and that the complete system is operating into a nominally resistive load. The various symbols are shown on Figure 8.1.

8.2.2 Determining the Characteristic Impedance of the λ/4 Transmission Line

The impedance transformation of a quarter-wave transmission line is given by:

$$R_{TL} = \sqrt{R_{IN}R_{OUT}} \tag{8.1}$$

The ratio of the power amplifier supply voltages sets the transformation ratio, T, of the transmission line:

$$V_{DD,A1} = TV_{DD,A2} \tag{8.2}$$

If this line is assumed lossless, then the current delivered to the load, from $A1$, is related to its output voltage by:

$$I_{TL} = \frac{V_{A1}}{R_{TL}} \tag{8.3}$$

The λ/4 transmission line acts as a voltage transformer, with a transformation factor, T, given by:

$$T = \frac{V_{A1}}{V_{OUT}} \tag{8.4}$$

And, from (8.3):

$$I_{TL} = \frac{TV_{OUT}}{R_{TL}} \tag{8.5}$$

also:

$$V_{OUT} = I_{TL}R_{OUT} \tag{8.6}$$

Combining (8.5) and (8.6), gives:

$$T = \frac{R_{TL}}{R_{OUT}} \tag{8.7}$$

or

$$R_{TL} = TR_{OUT} \tag{8.8}$$

It is usually most convenient if both amplifiers are run from the same supply voltage and hence, $T = 1$. Therefore:

$$V_{DD,A1} = V_{DD,A2} = V_{DD} \tag{8.9}$$

and $R_{TL} = R_{OUT}$. Thus for a Doherty system utilising equal supply voltages for the two class-B amplifiers, operating into a 50Ω load, the required impedance of the quarter-wave transmission line is also 50Ω.

8.2.3 Impedances Seen by the Amplifiers

The impedances seen by the two power amplifiers will depend upon the division of the output (load) current between them. This *division ratio* is defined as:

$$S = \frac{I_{TL}}{I_L} \tag{8.10}$$

The power into the load is therefore the sum of:

$$P_{A1} = \frac{V_L I_{TL}}{2} = SP_L \tag{8.11}$$

and

$$P_{A2} = \frac{V_L I_{A2}}{2} = (1 - S)P_L \tag{8.12}$$

The impedance seen by the output end of the λ/4 transmission-line is therefore:

$$R_{OUT} = \frac{V_L}{I_{TL}} = \frac{V_L}{SI_L} = \frac{R_L}{S} \tag{8.13}$$

and the impedance seen by the output of the second PA ($A2$) is:

$$R_{A2} = \frac{R_L}{1 - S} \tag{8.14}$$

8.2.4 Efficiency of a Doherty Amplifier

The derivation of overall efficiency for a Doherty system breaks down into three parts:

1. Operation below the transition point. In this region, $A2$ is shut down and $A1$ is operating as a conventional linear (class-B) amplifier.

2. Operation above the transition point. $A1$ is now saturated and acting as a voltage source; $A2$ takes over linear operation and acts as a controlled current-source.
3. Operation at full PEP. Both amplifiers are saturated, with the peak output voltage of the complete amplifier being the power supply voltage, V_{DD} (assuming ideal amplifiers).

8.2.4.1 Below the Transition Point

When the amplifier is operating below its transition point, only $A1$ is supplying power. Since it is assumed to be an ideal class-B amplifier, it will achieve a peak efficiency at full output power (i.e., at the transition point) of:

$$\eta = \frac{\pi}{4} \tag{8.15}$$

Saturation of $A1$ occurs when $A1$ is supplying its maximum share of the system output power, as determined by the division ratio (usually 0.5 for a two-stage Doherty system). At this point the voltage supplied to the load, V_L, is the required proportion of the supply voltage, that is,

$$V_L = \alpha V_{DD} \tag{8.16}$$

where α is the power division ratio at PEP (i.e., $S = \alpha$ at PEP).

Below this level, $V_L < \alpha V_{DD}$, and hence the overall efficiency is given by:

$$\eta = \frac{P_L}{P_{DC}} = \frac{\pi V_L}{4\alpha V_{DD}} \tag{8.17}$$

8.2.4.2 Above the Transition Point

Above the transition point, $A1$ is saturated and hence operating as a voltage source and $A2$ is operating as a linear current source. The current delivered to the load from the saturated $A1$ (via the $\lambda/4$ transmission line) is essentially constant and is the relevant proportion of the output current set by the transition point, hence:

$$I_{OUT,M} = \frac{\alpha V_{DD}}{R_L} \tag{8.18}$$

The remainder of the output current must be delivered by $A2$ in order to

produce the required output voltage across the load, V_L. Thus:

$$I_{A2} = I_L - I_{OUT,M} = \frac{V_L - \alpha V_{DD}}{R_L} \tag{8.19}$$

The voltage transformation performed by the coupler in this region is therefore (from (8.4)):

$$T = \frac{V_{DD}}{V_L} \tag{8.20}$$

For $T = 1$, $V_{DD} = V_L$, hence the RF output current of $A1$ in this region is:

$$I_{A1} = I_{OUT,M} = \frac{\alpha V_L}{R_L} \tag{8.21}$$

For an ideal class-B PA, the DC input current drawn by $A1$ is:

$$I_{DC,A1} = \frac{2I_{A1}}{\pi} = \frac{2\alpha V_L}{\pi R_L} \tag{8.22}$$

Similarly, the DC input current to $A2$ is (from (8.19)):

$$I_{DC,A2} = \frac{2I_{A2}}{\pi} = \frac{2(V_L - \alpha V_{DD})}{\pi R_L} \tag{8.23}$$

The total DC input current to the Doherty amplifier in the medium power region is therefore:

$$I_{DC} = I_{DC,A1} + I_{DC,A2} = \frac{2V_L(1+\alpha) - 2\alpha V_{DD}}{\pi R_L} \tag{8.24}$$

The efficiency in the medium power region is therefore:

$$\eta = \frac{P_L}{P_{DC}} = \frac{V_L^2/2R_L}{V_{DD}[2V_L(1+\alpha) - 2\alpha V_{DD}]/\pi R_L} = \frac{\pi V_L^2}{4V_{DD}(V_L + \alpha V_L - \alpha V_{DD})} \tag{8.25}$$

8.2.4.3 At Full Power

At peak power, $V_L = V_{DD}$, and hence (from (8.25)):

$$\eta = \frac{P_L}{P_{DC}} = \frac{\pi V_{DD}^2}{4V_{DD}(V_{DD} + \alpha V_{DD} - \alpha V_{DD})} = \frac{\pi}{4} \tag{8.26}$$

which is the same efficiency as would be obtained from an ideal class-B PA used alone.

8.2.5 Overall Efficiency of a Doherty Amplifier

Figure 8.3 shows the overall efficiency of a Doherty amplifier, using (8.17) and (8.25). For low-power operation, the efficiency increases linearly with output power level. Above the transition point, the efficiency initially degrades, but then increases again, yielding the same efficiency at PEP as at the transition point. The minimum efficiency in this medium power region is around 70% (for the ideal class-B PAs assumed here and $\alpha = 0.5$), which is

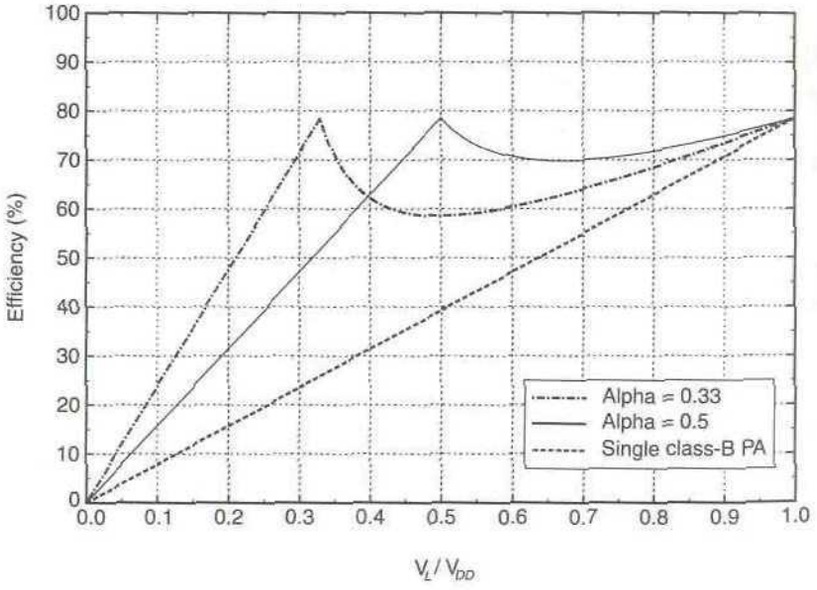


Figure 8.3 Efficiency of a two-stage Doherty amplifier system; solid-line: two-stage Doherty system, utilising ideal class-B amplifiers, $\alpha = 0.5$; dash-dot line: two-stage Doherty system, utilising ideal class-B amplifiers, $\alpha = 0.33$; dashed-line: Single (ideal) class-B amplifier.

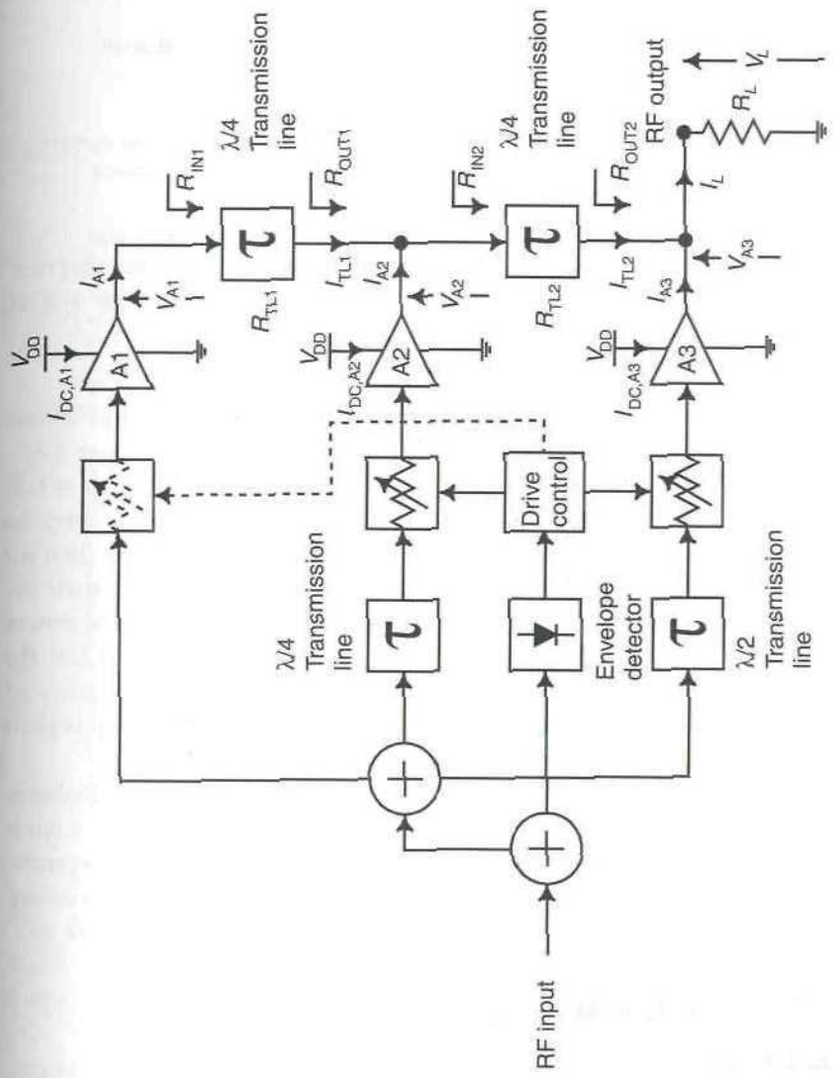


Figure 8.4 Three-stage Doherty amplifier.

Table 8.1
Regions of operation of a three-stage Doherty amplifier.

Region of Operation	*A1* Function	*A2* Function	*A3* Function
Low-power	Linear current source	Shut-off	Shut-off
Medium-power	Saturated (operating as a voltage source)	Linear current source	Shut-off
High-power	Saturated (operating as a voltage source)	Saturated (operating as a voltage source)	Linear current source
Full-output	Saturated (operating as a voltage source)	Saturated (operating as a voltage source)	Saturated (operating as a voltage source)

markedly better than a conventional single class-B amplifier's efficiency would be at an equivalent output power level (dashed line in Figure 8.3).

It is, of course, possible to utilise a transition point other than $\alpha = 0.5$. This is also illustrated in Figure 8.3, for $\alpha = 0.33$. Note that in this case, the minimum efficiency in the medium power region is markedly lower than for the optimum value of α. It is still, however, significantly better than the single class-B stage alternative. Note also that the efficiency in the low power region, up to 33% of the peak output voltage, is markedly *better* for the system with the lower transition voltage. Note also that an efficiency of nearly 60% or better is maintained up to 12 dB of back-off; the equivalent figure for the 'optimum' transition point is only 8 dB.

It is therefore evident that the optimum transition point for a Doherty system depends upon the characteristics of the modulation scheme which it is required to amplify and the degree of power control present in the system. Some examples of different types of modulation scheme and corresponding values for the transition point are given in the literature [8].

8.2.6 Three- (or More) Stage Doherty Systems

Figure 8.4 below shows the configuration of a three-stage Doherty system; the principle can, of course, be extended to as many stages as desired, although there are obviously diminishing returns and an increase in complexity as the number of stages grows.

The operation of a three-stage system is identical to that of the two-stage version, with the exception that there are now two transition points and hence a more complex drive control arrangement is required. There are now three regions of operation and these are summarised in Table 8.1.

In order to set the two transition points, the impedances of the two transmission lines must be:

$$R_{TL1} = \frac{R_L}{\alpha_1 \alpha_2} \tag{8.27}$$

and

$$R_{TL2} = \frac{R_L}{\alpha_2} \tag{8.28}$$

The resulting transition points are then α_1, for the transition between *A*1 and *A*2 and α_2, for the transition between *A*2 and *A*3.

The efficiency of a three-stage Doherty system is given by (from (8.17) and (8.25)):

$$\eta = \begin{cases} \dfrac{\pi V_L}{4\alpha_1 V_{DD}} & 0 < V_L < \alpha_1 V_{DD} \\ \dfrac{\pi V_L^2}{4V_{DD}(\alpha_1 V_L + \alpha_2 V_L - \alpha_1\alpha_2 V_{DD})} & \alpha_1 V_{DD} < V_L < \alpha_2 V_{DD} \\ \dfrac{\pi V_L^2}{4V_{DD}(V_L + \alpha_2 V_L - \alpha_2 V_{DD})} & \alpha_2 V_{DD} < V_L < V_{DD} \\ \dfrac{\pi}{4} & V_L = V_{DD} \end{cases} \tag{8.29}$$

The theoretical optimum transition points in this case are: $\alpha_1 = 0.33$ and $\alpha_2 = 0.66$, although again, this will depend upon the modulation scheme employed.

The efficiency characteristics for Doherty amplifiers of two- to five-stages are shown in Figure 8.5, where in each case the optimum transition points are used. It is clear from Figure 8.5 that an increasing number of stages brings the overall Doherty system closer to the ideal, limiting case of an almost constant 78% efficiency at all envelope levels.

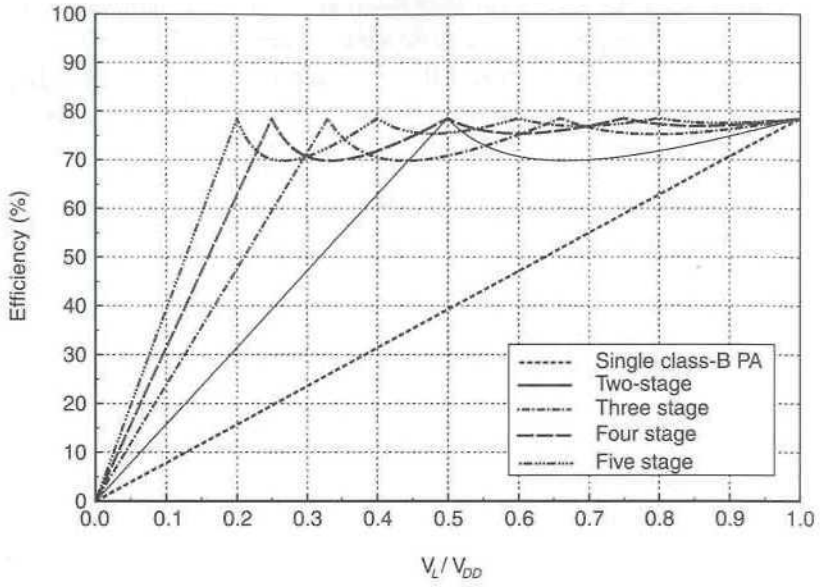


Figure 8.5 Efficiency characteristics for multi-stage Doherty amplifiers when utilising optimum transition points.

8.2.7 Advantages and Disadvantages of the Doherty Technique

The Doherty system has traditionally found application in broadcast systems and other very high-power applications, but is worth considering for use in a wider variety of areas. In particular, its use in a multicarrier power amplifier, for example, as the main amplifier in a feedforward system, could lead to a significant improvement in the efficiency of such systems.

8.2.7.1 Potential Advantages

1. Relatively low complexity. It is also potentially easier to implement than standard high-efficiency amplifier classes (-D, or -E, for example) at very high frequencies, due to it being relatively undemanding of RF power device characteristics.
2. It does not require a separate high-power modulator or modulation amplifier (such as is used in the EE&R technique).
3. Its modulation bandwidth is also not restricted by a requirement for a high-power high-efficiency modulator.
4. Its efficiency rivals that of many alternative techniques, such as envelope tracking.

5. The degree of control required is relatively small and undemanding (unlike, for example, feedforward systems).
6. Additional linearisation can be achieved by conventional means (e.g., envelope feedback) if required.

8.2.7.2 Possible Disadvantages

1. The use of $\lambda/4$ transmission-lines and the requirements for accurate phase matching between the two paths generally restricts Doherty systems to fixed (single) frequency operation (or operation over a very narrow band of frequencies).
2. Operation into antennas with a poor VSWR (such as typical mobile antennas) will upset the operation of the Doherty system and hence its achievable efficiency. This problem may be overcome in some frequency bands by the use of isolators.
3. The IMD performance of a Doherty system alone is relatively poor and the addition of a linearisation scheme will increase system complexity.
4. The class-B PAs used in a Doherty system must be designed to cope with variable (but still resistive) load impedances, without damage or loss of performance.

8.3 Adaptive Bias

8.3.1 Operation of Adaptive Bias

The adaptive bias technique is similar in its methodology for improving efficiency to the envelope tracking technique described below. The main difference is that in this case, it is the standing DC bias on a class-A amplifier which is varied with the envelope level, rather than the supply voltage (of a class-B stage), which is used in envelope tracking.

The adaptive bias technique was primarily proposed to overcome the poor power-added efficiencies found in linear class-A amplification, caused by the use of back-off as a linearisation technique. The efficiency of a class-A stage will typically be lowered in proportion to the square of the back-off used and hence can very quickly fall to only a few percent (for, say, 10 dB back-off).

The principle of operation of an adaptive bias scheme is shown in Figure 8.6. A directional coupler takes a sample of the input signal and feeds it to an envelope detector. The resulting detected signal is used to vary the

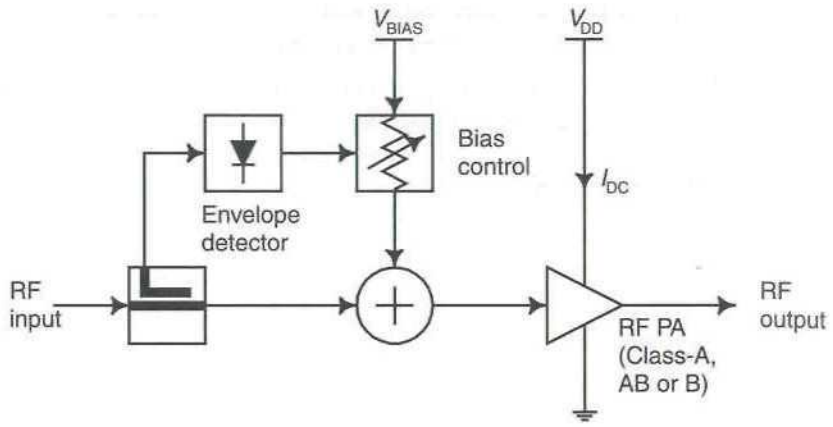


Figure 8.6 Adaptive bias applied to an RF amplifier.

bias on, for example, the gate of an FET, in order to vary the standing bias current drawn from the supply. The result of this is to ensure that the amplifier always has sufficient bias current to operate in its linear region, and hence maintain relatively low levels of distortion, without drawing an excessive supply current at low envelope levels (where this is unnecessary).

The additional circuitry required in an adaptive bias scheme is relatively simple and, due to the high resistance of an FET gate, should require little power from the supply. Thus almost none of the efficiency gained by using the technique is lost in its control circuitry.

The only major restriction placed on the active power device is that its gain should remain the same at (ideally) all levels of gate bias voltage. If this is not the case, significant AM–AM distortion will result and hence the operation of an adaptive bias scheme will significantly degrade the linearity of the amplifier, compared with conventional class-A operation.

8.3.2 Efficiency of an Adaptive Bias Class-A Amplifier

8.3.2.1 Simple Analysis

Consider a class-A amplifier with a supply voltage, V_{CC}, an output voltage, V_O, supplied to a load R_L, and operating at a quiescent bias, I_{DC}:

$$I_{DC} = \frac{V_{CC}}{R_L} \tag{8.30}$$

Hence:

$$\eta_A = \frac{V_O^2}{2V_{CC}^2} \tag{8.31}$$

If an optimum bias level is provided adaptively, such that:

$$I_{DC} = \frac{V_O}{R_L} \tag{8.32}$$

Then the overall efficiency of the stage becomes:

$$\eta_{AB} = \frac{V_O}{2V_{CC}} \tag{8.33}$$

This efficiency characteristic varies linearly with output power in contrast to that of a conventional class-A amplifier (from (8.31)) in which it varies with the square of the output power. Efficiency is therefore improved over that of a conventional class-A stage at all envelope levels, with the exception of PEP, at which the efficiencies are equal.

8.3.2.2 FET Amplifier With Envelope-Varying Signals

It is shown in [9] that the drain efficiencies of a conventional class-A amplifier, and the same amplifier with an adaptive bias scheme employed, are given by (8.34) and (8.35) respectively:

$$\eta_A = \frac{\overline{r^2}}{2} \tag{8.34}$$

$$\eta_{AB} = \frac{\overline{r^2}}{2\overline{r}} \tag{8.35}$$

where $\overline{r^2}$, the mean-square value of the normalised output envelope, $r(t)$, is related to the output power back-off level, B_o, by:

$$1/\overline{r^2} = B_o = \frac{\text{max. single-carrier linear output power}}{\text{average multi-carrier output power}} \tag{8.36}$$

and, assuming that the output signal has a Rayleigh distribution, the mean

value of the normalised output envelope is given by [10]:

$$\bar{r} = \sqrt{\frac{\pi \overline{r^2}}{4}} \tag{8.37}$$

where $r(t)$ is defined as:

$$r(t) = \frac{V_O(t)}{V_{O,\max}} \tag{8.38}$$

$V_O(t)$ being the time varying output envelope and $V_{O,\max}$ defining the maximum output voltage whilst still remaining in the linear region of operation of the FET.

Using the above equations and some typical mean-square values for the output envelope, when utilising some common signals [9], results in the typical comparative efficiencies given in Table 8.2.

It is evident from these figures, that the theoretical improvements in efficiency which result from the use of adaptive bias can be dramatic, particularly where high degrees of back-off are employed. In such cases, the raw efficiency can be more than doubled, making this a potentially very useful technique at frequencies where high-efficiency amplification and conventional linearisation schemes cannot be employed. This is currently true, for example, at higher microwave frequencies where active devices do not have sufficient gain to be operated satisfactorily in class-B or -C modes.

Table 8.2
Theoretical comparison between the efficiency of a conventional class-A FET amplifier and a similar amplifier employing an adaptive bias scheme

Modulation	$\overline{r^2}$	$\bar{r}$	Class-A efficiency, η_A	Adaptive bias efficiency, η_{AB}
Single FM carrier	1	1	0.5	0.5
Two-tone test	0.5	0.637	0.25	0.392
6 dB back-off (e.g., for a large number of constant envelope carriers)	0.25	0.443	0.125	0.282
16-QAM	0.556	0.706	0.278	0.394
64-QAM	0.429	0.615	0.215	0.350
256-QAM	0.378	0.576	0.189	0.328

8.4 Envelope Tracking

The format of an envelope tracking system [11] is similar to that of the envelope elimination and restoration system discussed in Chapter 7. The main difference is that in the case of an envelope tracking system, the RF amplifier is a class-A, -AB or -B design and not a high-efficiency nonlinear design. A schematic of the system is shown in Figure 8.7.

An envelope detector extracts the envelope information from a sample of the modulated RF input signal and feeds it to a high-efficiency audio amplifier (e.g., class-S) or a switching regulator. The purpose of this amplifier is not to amplitude modulate the supply voltage of the final RF amplifier, as in the case of EE&R, but merely to provide the RF amplifier with just sufficient supply voltage to permit linear amplification at the instantaneous envelope level present at that particular point in time.

This has the effect of significantly reducing the power consumption at low power levels, hence enhancing the average efficiency of the linear amplifier without compromising its linearity in any way (ideally).

A recent example of this technique is described by Hanington *et al.* [12] (with earlier work in this area having been conducted by Raab [13]), using heterojunction bipolar transistor (HBT) technology in a DC–DC converter supplying the RF amplifier. The converter operated at a switching frequency of 10 MHz, supplying the 300mW MESFET RF power amplifier. The converter itself operated within an envelope feedback loop to ensure that the high-power output envelope closely followed the envelope of the RF input signal (after envelope detection).

The efficiency of the system was based on a 74% DC–DC converter efficiency at full output (1W) and a MESFET amplifier peak efficiency of

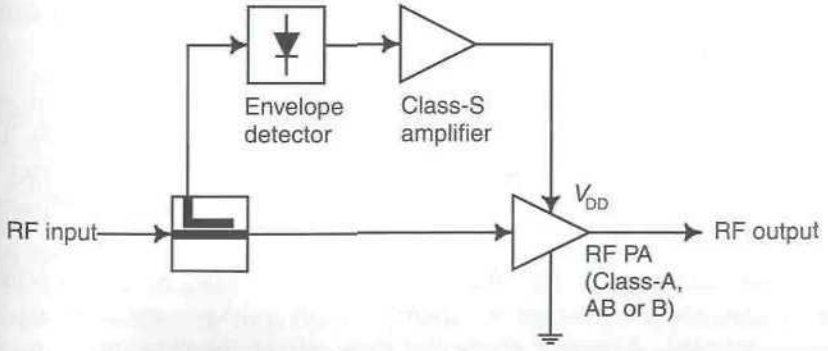


Figure 8.7 Schematic of an RF amplifier employing envelope tracking.

35% when run in class-A. Clearly at full output power, the efficiency of the system is poorer than that of the RF amplifier alone, however, this is not a realistic assessment of system performance. In a practical mobile communications system, the use of envelope-varying modulation formats and system power control (e.g., the closed-loop control of IS-95 CDMA), results in the RF amplifier rarely operating at PEP. The efficiency of the amplifier therefore depends critically on the 'typical' useage profile (and modulation format) assumed. In this case, a power useage profile was assumed which placed the most probable RF output power at roughly half the peak power capability of the transmitter. Based on this assumption, the reported improvement in battery life was 40%, which is an attractive benefit. Linearity was also reasonable at 30 dBc for a two-tone test.

8.5 Class-H Amplification

A class-H amplifier [14] is very similar to the envelope tracking amplifier described (Section 8.4). The main difference is that the supply voltage adjustment operates on the instantaneous signal amplitude, rather than on the envelope amplitude. The class-S (PWM) amplifier or switching regulator must therefore operate at a much higher frequency, hence restricting the technique to lower operating frequencies.

8.6 Dual-Bias Control

An obvious extension to the envelope tracking and adaptive bias schemes discussed above is to control both [15]. For the base bias control, efficiency of the controller is relatively unimportant and hence an analog control may be used (similar to that of adaptive bias). For the collector bias controller, two options exist.

1. Continuous bias control. This employs a DC–DC converter in a manner similar to the envelope tracking scheme (Section 8.4). The efficiency of this converter is then a key design parameter in ensuring that the overall system displays a useful efficiency improvement.
2. Discrete (switched) bias control. In this case, a series of FET switches, connected to different supply voltages, are connected in parallel. A switch controller then selects the appropriate power supply voltage for the instantaneous envelope level. If these voltages are derived using DC–DC converters, then the same issues as in (1) above will apply; if, however, they are derived from a series of individual cells (in a battery-powered application, for example) then converter efficiency is not an issue.

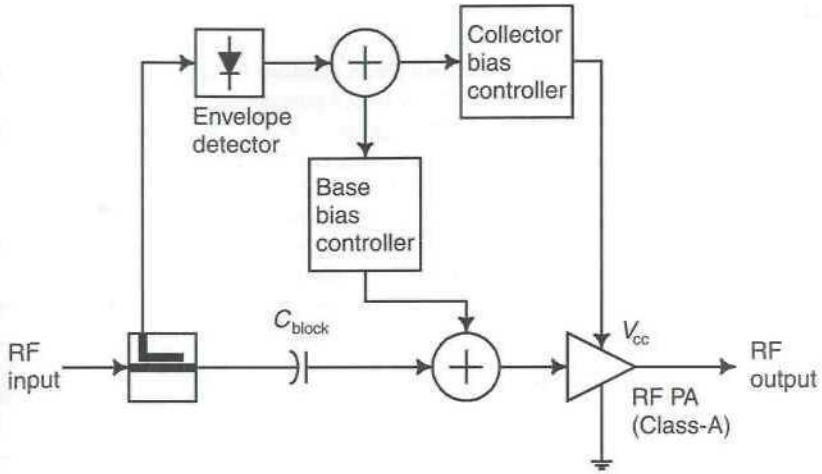


Figure 8.8 Dual bias control scheme.

Both of these schemes are represented by Figure 8.8.

Simulation results from this technique (detailed in [15]) indicate that significant improvements in efficiency are possible over a conventional class-A amplifier. Based on a starting point of 3.3% PAE for the class-A amplifier (assuming a multicarrier signal with a Rayleigh distribution and 10 dB of output back-off), this is improved to 9% or 10% with continuous bias or envelope control (individually) and to 24% with both controlled.

Although these efficiency improvements are impressive (assuming that they may be approached in a practical implementation), it is worth comparing them to the approach of linearising a class-C amplifier to produce an equivalent adjacent-channel power level. Cartesian loop, for example, should be capable of achieving at least this efficiency, although it is fundamentally band-limited, (unlike the envelope tracking or adaptive bias schemes discussed here). Alternatively, predistortion of a class-B amplifier (for example) will allow it to be driven at a higher mean power level and hence achieve higher efficiency, whilst still achieving broadband operation. Clearly the application of *both* linearisation *and* efficiency boosting techniques will lead to the ultimate in efficiency performance (outside of RF synthesis techniques).

References

1. Doherty, W. H., "A new high efficiency power amplifier for modulated waves," *Proc. of the Institute of Radio Engineers*, Vol. 24, No. 9, September 1936, pp. 1163–1182.
2. Terman, F. E., and J. R. Woodyard, "A high efficiency grid modulated amplifier," *Proc. Institute of Radio Engineers*, Vol. 26, No. 8, August 1938, pp. 929–945.
3. Smith, C. E., J. R. Hall, and J. O. Weldon, "Very high power long-wave broadcast station," *Proc. Institute of Radio Engineers*, Vol. 42, No. 8, August 1954, pp. 1222–1235.
4. Sainton, J. B., "A 500kW medium-wave standard broadcast transmitter," *Cathode Press (Machlett Company)*, Vol. 22, No. 4, 1965, pp. 22–29.
5. Bowers, D. F., "HEAD—a high efficiency amplitude modulation system for broadcasting transmitters," *Communications and Broadcasting*, Vol. 7, No. 2, February 1982, pp. 15–23.
6. Vinogradov, P. Y., N. I. Vorobyev, E. P. Sokolov, and N. S. Fuzik, "Amplification of a modulated signal by the Doherty method in a transistorised power amplifier," *Telecommunications and Radio Engineering*, Pt. 1, Vol. 31, No. 10, October 1977, pp. 38–41.
7. Upton, D. M., and P. R. Massey, "A new circuit topology to realize high efficiency, high linearity, and high power microwave amplifiers," *Proc. Radio and Wireless Conference (RAWCON)*, Colorado Springs, pp. 317–320, August 1998.
8. Raab, F. H., "Efficiency of Doherty RF power-amplifier systems," *IEEE Trans. on Broadcasting*, Vol. BC-33, No. 3, September 1987, pp. 77–83.
9. Saleh, A. A. M., and D. C. Cox, "Improving the power-added efficiency of FET amplifiers operating with varying envelope signals," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 31, No. 1 January 1983, pp. 51–56.
10. Papoulis, A., *Probability, Random Variables and Stochastic Processes*, New York: McGraw-Hill, 1965. p. 148.
11. Raab, F. H., "High efficiency amplification techniques," *IEEE Circuits and Systems Journal*, No. 7, December 1975, pp. 3–11.
12. Hannington, G., P. F. Chen, V. Radisic, T. Itoh, and P. M. Asbeck, "Microwave power amplifier efficiency improvement with a 10 MHz HBT DC–DC converter," *Proc. of IEEE MTT-S*, Baltimore, USA, June 1998, pp. 589–592.
13. Raab, F. H., "Efficiency of envelope-tracking RF power-amplifier systems," *Proc. of RF Expo East*, Boston, USA, November 1986, pp. 303–311.
14. Feldman, L., "Class-H variproportional amplifier," *Radio Electronics*, No. 48, October 1977, pp. 53–56.
15. Yang, K., G. I. Haddad, and J. R. East, "High-efficiency class-A power amplifiers with a dual-bias control scheme," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 47, No. 8, August 1999, pp. 1426–1432.

Biography

Dr. Kenington received B.Eng. (with first-class honours) and Ph.D. degrees from the University of Bristol, U.K. in 1986 and 1989, respectively. His Ph.D. work concerned the design of novel receiver systems for satellite earth stations.

From 1989 to 1990 he was engaged as a research assistant at the University of Bristol, performing research into novel transmitter and receiver architectures, as well as the EMC issues surrounding mobile radio technology. In 1990 he was appointed as a lecturer and continued research into high-linearity RF amplifier and transmitter systems and software radio techniques. During this time he undertook a wide range of industrial contract work encompassing linear power amplifier designs and complete software radio systems for commercial realisation by a number of major companies. In 1995, he helped to found Wireless Systems International Ltd., and joined them full-time in 1997. He is currently the Head of Advanced Development and supervises teams concerned with all aspects of technology and systems design within the company.

Dr Kenington received the Institution of Electrical Engineers (IEE) prize for 'Outstanding Academic Achievement' in 1986, the IEE Mountbatten Premium in 1989 and the IEE Engineering Science and Education Journal Premium in 1999. He has served on IEE Professional Group E9 (Satellite Communications) and on the IEE Science, Education and Technology Divisional Board. He has also served on CISPR (International Special Committee on Radio Interference) and on the editorial board of a number of scientific journals. He is a chartered engineer and a corporate

member of the IEE and is the author of over 90 published papers and 25 patents in the radio technology field.

He may be contacted by e-mail using: PBKenington@wsi.co.uk or PBKenington@iee.org

Index