Transmitter Linearization for Portable Wireless Communication Systems

Luns Tee

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-42
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-42.html

April 12, 2007

Acknowledgement

This work was funded in part by: the California MICRO program; NSF grant #MIP 9412940; and C2S2, the MARCO Focus Centre for Circuit & System Solutions, under MARCO contract 2003-CT-888.

Transmitter Linearization for Portable Wireless Communication Systems

by

Luns Tee

B.A.Sc. (University of Toronto) 1995

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy

in

Engineering-Electrical Engineering
and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Paul R. Gray, Chair
Professor Robert G. Meyer
Professor Paul K. Wright

Spring 2007

The dissertation of Luns Tee is approved:

Professor Paul R. Gray, Chair Date

Professor Robert G. Meyer Date

Professor Paul K. Wright Date

University of California, Berkeley

Spring 2007

Transmitter Linearization for Portable Wireless Communication Systems

by

Luns Tee

Abstract

Transmitter Linearization for
Portable Wireless Communication Systems
by
Luns Tee

Doctor of Philosophy in Engineering-
Electrical Engineering and Computer Sciences

University of California, Berkeley
Professor Paul R. Gray, Chair

Recent years have seen much progress in the integration of RF transceivers in low-cost CMOS technology, with many commercial transceivers on the market now being CMOS designs. However, it is still common in applications requiring high power output and high linearity to find discrete power amplifiers (PAs) implemented in other technologies. One of the obstacles to integrating the PA in CMOS is the linearity requirements of the nonconstant-envelope modulation schemes used in high data-rate systems. Linear class A or AB PAs have poor power efficiency compared to other topologies, but more power-efficient amplifiers such as class C, E, or Doherty configurations can only be used with constant-envelope modulation unless some form of linearization is utilized.

Cartesian Feedback is a well known linearization technique; however, its use in integrated transceivers has primarily been limited to either low output power or non-CMOS amplifiers. Existing analyses of Cartesian Feedback assume a linear amplifier as found in these designs, and do not offer useful

intuition for application to highly nonlinear PAs as would be found in CMOS implementations.

This thesis investigates applying Cartesian Feedback to enable the use of a CMOS PA, allowing integration of the PA together with other radio components already available in CMOS. Contributions of this work include: development of analytical techniques applicable to highly nonlinear amplifiers leading to a stability criterion for the design of a Cartesian-Feedback loop; introduction of circuit techniques for CMOS implementations of PA, mixers and loop filter appropriate for the needs of Cartesian-Feedback; and application of these techniques to a monolithic transmitter for cellular telephone applications.

To demonstrate the validity of the developed analysis and circuits, an integrated CMOS transmitter, including an on-chip PA, was designed to produce an EDGE modulated signal. This prototype, implemented in a standard 0.18µm CMOS technology, meets GSM spectral mask and EVM requirements, producing an 18dBm output with 18% drain efficiency. The linearized prototype has distortion 21dB lower at 400kHz offset compared with open-loop operation, demonstrating the effectiveness of Cartesian Feedback for this application.

_____

Paul R. Gray, Chairman of Committee

Acknowledgments

The road to a Ph.D. is an arduous one, but while the degree itself is ultimately a personal goal, the journey to it fortunately does not happen in isolation. If not for the contributions of many people, I would never have made it through to the end, and to these people, I am grateful.

First, it has been an honour to have had the opportunity to work under the guidance and support of my advisor, Professor Paul Gray. While he is well known for his vast technical knowledge and accomplishments, what has really impressed upon me has been his insight for future directions in technology, and even more, his high standards for research. Several times, I wanted to cut one corner or another and he would always discourage doing so; while some of the times left me feeling very deflated with thoughts of how much longer it would take for me to finish everything, looking back in retrospect, the extra time to do things right has been well spent. I am grateful for his keeping me on track, and hope to carry on the tradition of excellence.

I would also like to thank Professor Robert Meyer not only for serving as chair of my Qualifying Exam and reading my dissertation, but for keeping alive the sense of discovery and enjoyment in circuit design. Listening to his talks has always been inspiring as he always manages to frame technical discussions in a way that somehow avoids the feeling of being enslaved to the demands of industry or pressures of other interests and doesn't lose sight of learning little things that are interesting for their own sake. Of all the classes I've taken at Berkeley, his EE242 remains my favourite not just because of what I learned from it technically, but because he made learning it fun!

I also thank Professors Seth Sanders, Philip Stark, and Paul Wright for serving on my Qualifying Exam and dissertation committees. I'm also grateful for Professors Robert Broderson and Jan Rabaey for establishing outstanding facilities at the Berkeley wireless Research Centre, and also to Professor Borivoje Nikolic for his support and advice.

I am also greatly indebted to visiting industrial fellow Enrico Sacchi, and fellow students Ryan Bocock, and Tim Wongkomet who have saved me from the otherwise impossible task of implementing my prototype on my own. Enrico's contribution is particularly noteworthy; even after his VIF appointment was over, he continued to help, laying out a significant part of the chip from Pavia, and even coming back to Berkeley on his own vacation time - twice - to help with getting the chip taped out. Even more than their help with the chip though, I am grateful to have these folks as great friends.

Less directly involved with my chip, but still very helpful have been the more senior graduate students in the group. George Chien, Martin Tsai, and Li Lin were always ready available for interesting discussions, technical or otherwise, and sitting around the Mahjongg table with them was always fun. Seeing Chris Rudell and Jeff Weldon lead the group in designing 'the big chip' was a great inspiration. Andy Abo and Sekhar Narayanaswami, shared in administering the group's computing resources, and to this day, I still don't know whether to thank or curse them for bringing me into the fray. Newer students in the group, Yun Chiu and Cheol-Woong Lee were also very helpful when designing my test boards, and have also been interesting to interact with.

Outside of Professor Gray's group, Mike Shuo-Wei Chen and Ian O'Donnell were a great help as we navigated the maze of board fabrication, assembly and testing together.

Many staff members in the department have also been very helpful in my time here. I would like to thank Ruth Gjerde for clearing the way through all the administrative tangles that keep popping up. Elise Mills, Tom Boot, Carol Sitea, Diane Chang, Judy Fong, and Carol Zalon, have come to the rescue at various times for purchase orders and other things. Kevin Zimmerman, Brian Richards and Brad Krebs have also been there when I've needed help with computing or other facilities. Kevin and Brian always have interesting stories to share along the way with perceptive, pithy, or just plain mind-boggling comments handy to distract me from worrying about whatever problems may bring me to them.

I would like to thank STMicroelectronics for providing chip fabrication, and Bhusan Gupta and Benjamin Coates were instrumental to getting my chips down that pipe. June Sun and Rachel Lim at Marvell Semiconductor were very helpful in getting parts samples and arranging access to test equipment that weren't available at Berkeley.

Outside of school, Jennifer Tsoi has been practically a sister to me and I'm lucky to have had her to grow up together with in my time in California. Grace Yeung also shared in my getting settled in at Berkeley as we both figured out how things work around the Bay Area, and I'll never forget our impromptu trips to SF to stuff ourselves silly at Coriya.

In more recent years, Diane Kong has been a welcome dinner buddy, sharing opportunities for me to get away from Berkeley, and helping displace my worries and frustrations about school. For those worries and frustrations that refuse to stay displaced, Fei-Ling Woo has had a bizarre knack for making me feel more at ease with them by somehow describing to me exactly how I feel, the bizarre part being that she would do so thinking she was describing her own Ph.D.

Finally, I would never have made it without the patient support, love, and encouragement of my parents, Tiam-Tuan Tee and Shi-Ling Tong. They've done everything in their power to minimize the things I need to worry about so that I may focus on what I have to do progress in life.

This work was funded in part by: the California MICRO program; NSF grant #MIP 9412940; and C2S2, the MARCO Focus Centre for Circuit & System Solutions, under MARCO contract 2003-CT-888.

Table Of Contents

List of Figures

Chapter 1
Introduction

1.1 Motivation

The last two decades have brought a massive proliferation of personal wireless devices. Nowhere is this more apparent than in the mobile phone market, where phones have gone from being expensive equipment used by businesses to a low-cost everyday personal item. Other applications have also emerged in the meanwhile, with wireless LAN (Wi-Fi) and PAN (Bluetooth) devices now being affordable and commonplace. The market for wireless transceivers has grown tremendously, and with emerging standards and applications for third/fourth generation (3G/4G) mobile technology that blur the distinction between wide-area voice communication and local-area data, this growth can be expected to continue for years to come.

This growth in the market has come in conjunction with advances in transceiver design. Bag phones the size of a lunch box have given way to

handsets small enough to be lost in one's pocket. As the race to produce smaller handsets has levelled off, a trend of increasing functionality has taken its place, with many handsets now also including GPS receivers, cameras, PDA and MP3 player functionality. These trends have been made possible by advances in integrated-circuit technology.

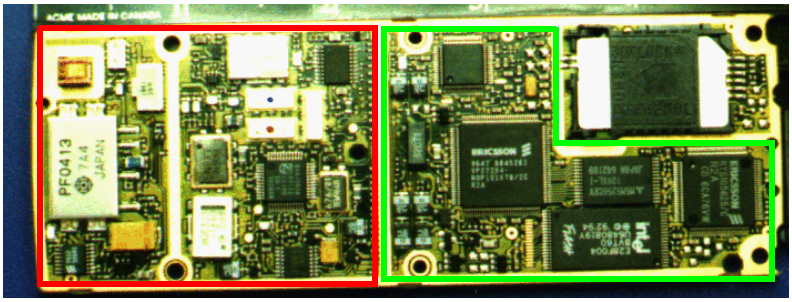


Fig. 1.1: Mobile Phone Circa 1995

Figure 1.1 shows a board photo of a GSM phone released in 1995. The radio section of the phone consists of many discrete components in various technologies, while the digital back-end is composed mainly of a handful of large, complicated, highly integrated chips. CMOS is the natural technology for such complex digital designs, and economies of scale from the large volume of digital chips shipped makes CMOS the least-expensive modern integrated-circuit technology available. This low cost, and the dream of integrating the radio front-end together with the digital back-end into a single-chip radio

transceiver, led to much research in the design of high-performance radios in CMOS technology.

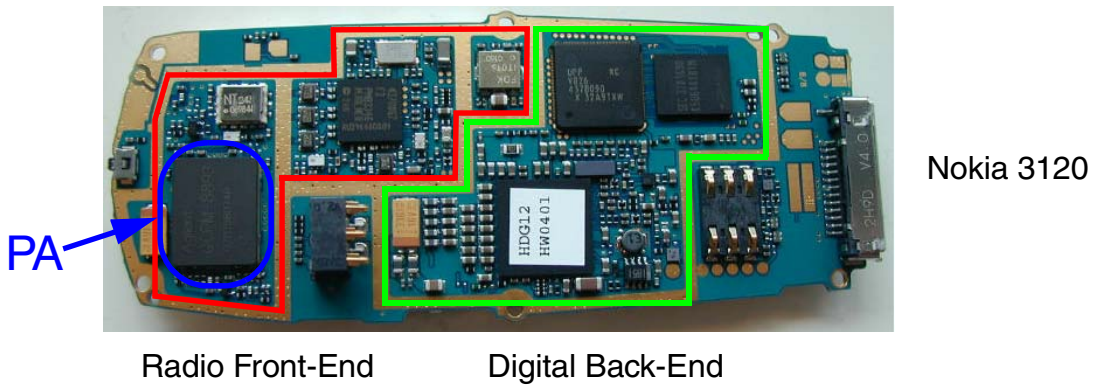


Fig. 1.2: Mobile Phone Circa 2004

Figure 1.2 shows a newer phone and illustrates the progress made since the mid-90's. The number of components in the radio section is substantially reduced with most radio functions being integrated together, however the power amplifier (PA) still remains as a separate module, in this case a Gallium-Arsenide multi-chip module. This is typical of the state of the art today: although there are now several so-called single-chip CMOS radio solutions for the mobile phone market, these designs still depend on an external PA. While transceiver integration has come a long way, integration of the PA remains an unsolved problem.

The PA is the last active component in a transmit chain and needs to produce an output strong enough to travel the distance to the intended receiver. This need to produce a large output signal means the PA can be a significant

portion of a transceiver's power budget, thus it is important for a PA to have good power efficiency. At the same time, the PA must not introduce distortion in the final output signal sent to the antenna. Older radio standards such as AMPS, DECT and GSM, used a class of modulation known as constant-envelope modulation schemes, whose signals are insensitive to distortion from PA nonlinearity. However, newer high data-rate standards have moved to nonconstant-envelope modulation schemes for better spectral efficiency - more bits/second per Hz of RF bandwidth occupied - and these schemes require a linear transmitter.

This need for both power efficiency and linearity is the main reason that PAs remain external. The most power-efficient PAs are nonlinear, and conversely, linear designs have poorer power efficiency. This tradeoff is particular bad for CMOS designs, hence the continued use of PAs in other technologies such as Silicon Bipolar, LDMOS, or GaAs, which offer better performance. While there are some CMOS PAs now emerging on the market, these target constant-envelope radio standards, while modern nonconstant-envelope applications are still without a CMOS solution.

It is important to realize however, that the requirement of good PA linearity is an artificial one - what ultimately matters is the linearity of the transmitter as a whole. While traditional radio designs achieve this by specifying adequate linearity of all its blocks, a transmitter built around

nonlinear components that still meets overall system linearity requirements is also legitimate.

Various architectures have been proposed in the past for linearizing transmitters built around nonlinear components, and may hold the key to achieving a fully-integrated transmitter. By relaxing linearity requirements for the PA, more power-efficient PA designs can be used. If linearization allows a CMOS PA to achieve performance competitive with non-CMOS PAs in traditional architectures, then this barrier to integration can be eliminated.

1.2 Research Goals

The main objective of this research is to demonstrate linearization as a suitable method to enable the integration of a CMOS PA. This work surveys several linearization architectures, identifying Cartesian Feedback as a promising approach. From this point, the work divides into two parts - system-level analysis, and circuit implementation.

Applications of Cartesian Feedback in the past have been for enhancing the linearity of inherently linear amplifiers, and the existing analyses for the architecture assume this and are inadequate for the levels of distortion found in a CMOS PA. One goal of this work is to provide a more thorough analytical

framework for evaluating the performance and stability of a Cartesian-Feedback transmitter, and apply it in designing a transmitter around an integrated PA.

With the analysis taken care of, practical demonstration of linearization is the other main goal. A prototype Cartesian Feedback transmitter was designed, integrating all active circuit blocks, including the PA, onto a single CMOS die. The prototype is designed targeting a real radio standard, GSM EDGE, operating in the DCS 1800 band.

Contributions of this research include:

- Development of an intuition for understanding the stability of the Cartesian Feedback architecture, and applying this to evaluate the impact of several typical nonidealities including PA distortion and other RF signal path effects
- Examined loop gain requirements, and the trade-off between gain needed for distortion suppression and stability, and proposing a high-order loop transfer function that offers a better compromise between these two than the traditional single-pole loop filter
- Identified a downconversion mixer architecture that achieves good flicker noise performance and high linearity as needed by the Cartesian Feedback architecture

- A prototype design was fabricated and measured. While the PA output power was less than designed, the functioning of the linearization is clearly demonstrated - the integrated PA operating in a traditional open-loop is unable to meet spectral mask and EVM requirements at any power level, but with the feedback loop closed, both requirements are met at the maximum power deliverable from the PA

1.3 Thesis Organization

This dissertation is organized as follows:

Chapter 2 gives an overview of the system level design of linearized radio transmitters, describing several different architectures. Some analysis techniques for nonlinear systems are presented along the way. The origins and effects of nonlinearity and the specifications they affect, are visited.

Chapter 3 looks into issues involved in the Cartesian Feedback architecture. Being a feedback system, stability is a concern, and an intuition for multivariate feedback is developed. Practical implementation problems are visited, and possible solutions presented.

Chapter 4 discusses the system-level design of a Cartesian-Feedback transmitter targeting the GSM EDGE standard. Specifications are examined and some performance requirements are found for various transmitter blocks. Some

system-level simulations are presented for estimating requirements on the loop transfer function, and to verify closed-loop performance of a transmitter.

Chapter 5 describes the practical implementation of a prototype that was designed and fabricated, and the results of measurements performed on it are given in Chapter 6.

Lastly, Chapter 7 contains concluding remarks, and suggestions for future work.

Chapter 2
Transmitter Fundamentals

The function of a transmitter in any radio system is to take information that it has, whether it be telegraph, audio, video, or arbitrary digital data, and produce a signal representing it that can propagate through the air to a remote receiver where it can be recovered and used. While earlier radio systems had fixed high-power transmitters that would broadcast over great distances to portable receivers, the dawn of modern wireless communication has come with two-way communication made possible by having not only portable receivers, but also portable transmitters. This chapter introduces the basics of radio signals and how transmitters synthesize them. Operation of the power amplifier, and the effects of distortion that it can introduce are discussed, and then some architectures for accommodating PA nonlinearity are introduced.

2.1 Transmitter Basics

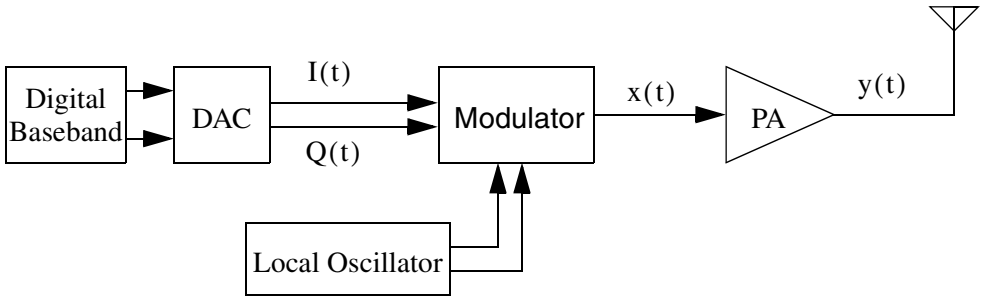


Fig. 2.1: Radio Transmitter Block Diagram

Figure 2.1 shows a simplified block diagram of a typical modern radio transmitter. Data to be transmitted is processed by digital baseband circuitry and converted to continuous-time analog baseband signals by digital-to-analog converters (DACs). A local oscillator (LO) generates a reference signal for the carrier. A modulator takes this LO signal and the baseband modulated signals to generate the desired modulated signal. The signal coming out of the modulator is usually too low in power to be transmitted very far, so it is amplified by a Power Amplifier (PA) to bring it up to useful power levels.

The baseband modulated signals and the local oscillator signals will be presumed available and their generation will not be discussed in this work. Many variations of the modulator exist, but a common trait among many of them is that the output of the modulator is fed to the PA with the understanding that the PA does not introduce significant distortion. Different architectures will be

discussed later, but variations of this architecture will not be discussed exhaustively: any architecture that generates the same modulated signal to give to the PA based on equivalent baseband inputs, will be considered equivalent.

2.1.1 Performance Metrics

Radio standards specify certain performance requirements for transmitters. Output power level requirements are specified to ensure the transmitter output can travel distances needed for the application. Of more direct interest to the user is battery life, and this relates to the power efficiency of the transceiver. While power efficiency requirements are not typically specified in radio standards, they are still of interest and commonly evaluated.

As will be seen later, power efficiency is typically achieved at the expense of linearity, and transmitter nonlinearity can introduce distortion. Noise and distortion of the output signal can cause problems of two sorts: they can corrupt the signal being transmitted and introduce errors when the signal is received; and they can interfere with other users trying to communicate on other frequencies. Other specifications given in radio standards relate to the accuracy of the output signal.

Three tests are commonly used to quantify noise and distortion performance: Spectral Mask, Adjacent Channel Power-Ratio, and Error-Vector

Magnitude. Effects of nonlinearity of the transmitter show up in these tests. These tests and some basics of power efficiency will be discussed next.

2.1.1.1 Spectral Mask

The most direct way to observe a transmitted signal is to observe the frequency spectrum of its output. The radio channel being transmitted on is defined by a centre frequency (carrier frequency) and a channel width, and the output spectrum should be contained within that channel. In practice, the modulation being used, as well as noise and distortion will introduce emissions at nearby frequencies as well. The spectral mask sets limits on how strong these emissions can be as a function of frequency relative to the carrier, and set a ceiling on the amount of interference than is allowed to spill onto adjacent channels. Fig. 2.2 shows the measured spectrum for a GSM EDGE signal in the PCS band together with the spectral mask from GSM specifications for handset emissions.

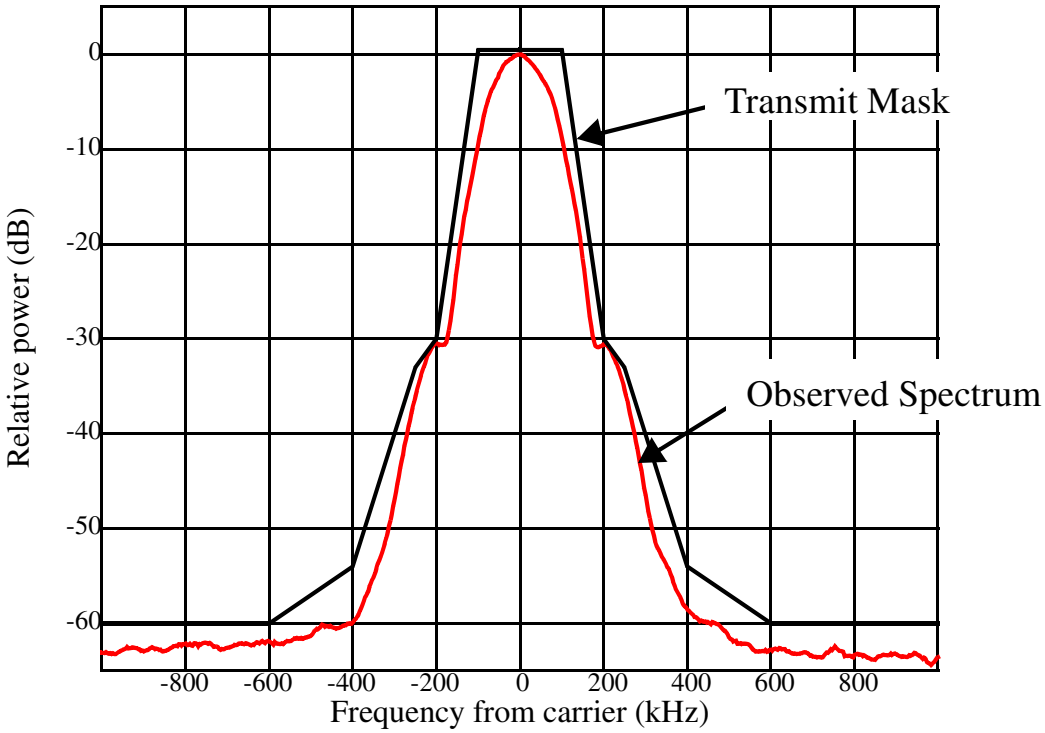


Fig. 2.2: GSM EDGE Spectral Mask for PCS band handsets

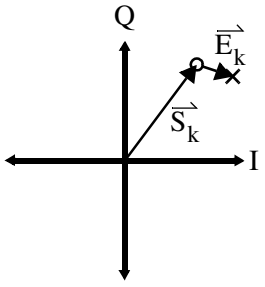
2.1.1.2 Adjacent Channel Power Ratio (ACPR)

Another metric for looking at the effect on nearby users is to consider the signal that the other users receive. Receivers will filter their input signal to isolate the channel they are operating on, so another test is to consider the amount of interference seen after this filtering. The output of a transmitter will generate a certain amount of power at the output of receive filters tuned to a neighbouring channel: the magnitude of this interfering power, relative to the intended carrier power is called the Adjacent Channel Power Ratio, also known as the Adjacent Channel Leakage Ratio.

Several radio standards specify ACPR for the first and second adjacent channels: for instance, IS-54 specifies that the ACPR for the first and second channels away from carrier be at most -30dB and -48dB of the carrier power. Other standards, such as GSM, do not specify ACPR performance, relying instead on the spectral mask to limit interference.

2.1.1.3 Error Vector Magnitude (EVM)

Spectral Mask and ACPR relate mainly to the output's effect on users of other channels, but does not always reflect how accurate the signal represents what was intended: a signal can meet Spectral Mask and/or ACPR specifications, yet still be too corrupt to use. The Error-Vector-Magnitude test looks at the transmitted signal as it is intended to be used: the transmitter's output signal is converted to a series of symbols in the IQ-plane (the meaning of this plane will be described in Section 2.2.3) just as it would be in a receiver. These symbols are compared to what they would be for an ideal transmitter. This is illustrated in Figure 2.3.



$$EVM_i = \sqrt{\frac{\left|\vec{E}_i\right|^2}{\sum_{k=1}^{N}\left|\vec{S}_k\right|^2/N}}$$

$$EVM_{RMS} = \sqrt{\sum_{k=1}^{N}\left|\vec{E}_k\right|^2/N}$$

Fig. 2.3: Error Vector Magnitude (EVM)

GSM specifications for EDGE modulation require the RMS EVM to be under 9%, with 95% of symbols having under 15% EVM, and with a peak EVM under 30%.

2.1.1.4 Power Efficiency

Radio standards specify output power level requirements for transmitters, but of more importance to a handset user is the battery life. This depends on battery capacity and the power drawn. Power for the transmitter output is of course drawn from the battery, so the power of the transmitter's output represents a limit to how low battery drain can be. In practice, significant amounts of power are consumed in creating the intended output, and more power yet is used for other transceiver functions, thus this limit is not one that is approached in reality.

A metric for how close power consumption comes to this limit is power efficiency, typically expressed as a percentage measure, which is defined as:

$$\eta = \frac{P_{RFout}}{P_{DC}} \qquad \text{(Eq 2-1)}$$

where P_{RFout} and P_{DC} are the transmitter's output power, and the power drawn from the DC supply (battery) respectively. There are different variations of this metric depending on where one measures the output power, and what component of power drawn from the supply is considered. It is common practice

however, to look at power efficiency of the PA alone as it is the PA that has to deliver this final output power. Power consumed by other transmitter blocks, and signal power lost in post-PA switches or filters are considered separately. For applications with significant output power, the PA dominates the power budget, making PA efficiency an important metric.

One common PA efficiency metric takes the PA's final output power as P_{RFout}, and power for the PA's final output stage as P_{DC}. For FET-based PAs, this is known as *drain efficiency* (*collector efficiency* for bipolar PAs), and is typically the most optimistic of the efficiency metrics as it neglects power consumed in generating the input to this final output stage.

The input to the final PA stage is often another amplifier, and several consecutive stages of amplification can collectively referred to as being a PA. A more meaningful efficiency measure would include power drawn from the supply for all PA stages. This is called the *overall efficiency*, denoted as $\eta_{overall}$, where

$$\eta_{overall} = \frac{P_{RFout}}{P_{DC,\,total}} \qquad \text{(Eq 2-2)}$$

Overall efficiency is a reasonable measure when the multiple stages of the PA provides a large gain. For a single-stage PA, the overall PA efficiency is just the drain efficiency.

This definition is still somewhat lacking: it only considers power drawn from the supply, but not power accepted at the amplifier input. If we consider a 'unity gain amplifier' consisting of just a wire from input to output, it can produce signal power at the output - all provided by its input - while drawing no power from the supply. By the preceding definitions, this would be an infinite efficiency, which is quite impressive for a circuit that does not do anything! Thus, another common efficiency measure is *Power-Added-Efficiency* (PAE), where instead of the output power from the amplifier, the signal power *added by the amplifier* is considered instead, or:

$$\text{PAE} = \frac{P_{\text{RFout}} - P_{\text{RFin}}}{P_{\text{DC, total}}} \qquad \text{(Eq 2-3)}$$

PAE is generally accepted as the most meaningful PA efficiency metric. For PAs that provide reasonably large gain, the overall efficiency is very close to the PAE, and is a reasonable measure as well.

2.2 Radio Signals and Linear Systems

Most wireless radio systems use signals that are based on simple sinusoids. This has been the practice right from the first demonstration of radio waves by Heinrich Hertz in the late 1880s and continues to this day for good reason. Electromagnetic waves had been previously predicted by James Clerk Maxwell[1], whose solution of the wave equation is satisfied by sinusoidal

waves, and the behaviour of any circuit or radio channel, is simple to characterize in terms of such signals. Also, it is reasonably simple to design circuits that are sensitive to sinewaves of a desired frequency, but insensitive to others, thus frequency selectivity of such sinewaves has long been a basis for sharing of the electromagnetic spectrum between different uses.

2.2.1 Pure sinewaves - phasor notation

A real sinewave $x(t)$ at a given frequency f_c can be characterized by two quantities: its amplitude E (for envelope) and phase ϕ. Denoting $\omega_c = 2\pi f_c$, the sinusoid $x(t)$ can be written as:

$$x(t) = E\cos(\omega_c t + \phi) \qquad \text{(Eq 2-4)}$$

This can also be written in an alternate form:

$$x(t) = \mathrm{Re}[\tilde{x}e^{j\omega_c t}] \qquad \text{(Eq 2-5)}$$

where $\tilde{x}$ is called the phasor representation of $x(t)$, and is given by:

$$\tilde{x} = Ee^{j\phi} \qquad \text{(Eq 2-6)}$$

A graph of (Eq 2-5) can be visualised by thinking of $e^{j\omega_c t}$ as an infinitely long corkscrew wrapped around the time axis, which is scaled and rotated by the magnitude and argument of $\tilde{x}$. Taking the Real component of this term is like taking the shadow of the corkscrew.

However, this representation of (Eq 2-4) is not unique: the same $x(t)$ could just as legitimately be written as:

$$x(t) = \mathrm{Re}[\tilde{x}^* e^{-j\omega_c t}] \qquad \text{(Eq 2-7)}$$

where the * in $\tilde{x}^*$ denotes it as the complex conjugate of $\tilde{x}$. This would be a corkscrew turning with a reverse spiral, and rotated opposite to what was considered earlier. The differences are lost when taking the shadow.

Rather than consider which of these two forms is more meaningful to use, it's convenient to use a notation that renders both of them symmetric, and at the same time eliminate the need to take a Real component. $x(t)$ can be written as.

$$x(t) = \frac{\tilde{x}}{2} e^{j\omega_c t} + \frac{\tilde{x}^*}{2} e^{-j\omega_c t} \qquad \text{(Eq 2-8)}$$

This can be described as being two complex sinusoids, of frequencies $+\omega_c$ and $-\omega_c$, scaled by phasors of $\frac{\tilde{x}}{2}$ and $\frac{\tilde{x}^*}{2}$ respectively. By the symmetry, it is common when speaking of real sinusoids, to speak only of ω_c and $\tilde{x}$, and neglect mention of the $-\omega_c$ term since everything that happens to it is the complex conjugate of what happens at ω_c; both terms are generally understood to exist even when only one is mentioned.

2.2.2 Linear Transfer Function - H(s) as an eigenvalue

Part of what makes a sinusoidal signal so convenient is the ease of characterizing a circuit or radio channel's response to it. If we consider feeding a sinewave $\tilde{x}$ at frequency f_c into a passive linear channel, its output will also be a sinewave of the same frequency, but of possibly different amplitude or phase.

The channel being linear means that feeding in a different input amplitude will result in a proportionally different output amplitude. Changing the phase of the input signal is equivalent to a time shift of some fraction of a cycle, and the response of the circuit, including its output, will experience the same time shift. The input to output response of the channel can be characterized simply by the ratio of output amplitude to input amplitude, and the difference between output phase and input phase.

Denote the output phasor as $\tilde{y}$, and consider the ratio of the output phasor to the input

$$\tilde{H} = \frac{\tilde{y}}{\tilde{x}} = \frac{E_{out}e^{j\phi_{out}}}{E_{in}e^{j\phi_{in}}} = \left(\frac{E_{out}}{E_{in}}\right)e^{j(\phi_{out}-\phi_{in})} \qquad \text{(Eq 2-9)}$$

This quantity is itself a phasor: its magnitude represents gain, while its argument represents the phase shift for a sinewave at a particular frequency going through the channel. This phasor exists for other frequencies as well, and as a function of the sinewave frequency f, is commonly written as $H(s)$, where

$s = j\omega$ and $\omega = 2\pi f$. $H(s)$ is referred to as the *Frequency Domain Transfer Function* of the channel, also referred to as the *Frequency Response* of the channel.

From symmetry of the definitions of $\tilde{x}$ and $\tilde{y}$ in terms of positive and negative frequency, it can be shown that $H(-j\omega_c) = H^*(j\omega_c)$. Noting that $\tilde{y} = \tilde{x}H(j\omega_c)$, we can write $y(t)$ as:

$$y(t) = \frac{\tilde{x}H(j\omega_c)}{2}e^{j\omega_c t} + \frac{\tilde{x}^*H(-j\omega_c)}{2}e^{-j\omega_c t} \qquad \text{(Eq 2-10)}$$

This consists of the same two terms containing the same $e^{j\omega_c t}$ and $e^{-j\omega_c t}$ that $x(t)$ had, except that these are multiplied by $H(j\omega_c)$ and $H(-j\omega_c)$ respectively. The linear system can be said to have *eigenvalues* of $H(j\omega_c)$ and $H(-j\omega_c)$ corresponding to *eigenvectors* of $e^{j\omega_c t}$ and $e^{-j\omega_c t}$ respectively. This eigenvector/eigenvalue perspective will be useful later on.

2.2.3 Modulated Signals

In practice, the pure sinusoid of (Eq 2-4) is not very useful - it exists over infinite time, and can only convey at most two values represented by the phasor. However, it is important to realize that for practical purposes, the signal need not span infinite time. Observing just several cycles - or even just one cycle - of the sinusoid is enough to identify the describing phasor. Also, the

output of a circuit is typically a function of only its present input and what the input has been in the recent past, with influence of earlier inputs falling off exponentially with time: the circuit can be said to have a memory only for its recent input, and is *memoryless* on a longer time scale.

If a linear circuit is given a sinusoid at its input for a sufficiently long but finite time, its output will approach what it would have been had that sinusoid been present for all of eternity. Knowing the circuit's transfer function, the input's phasor can be recovered. We can consider *modulating* a sinusoid - changing the amplitude and phase gradually over time - slowly enough that inputs recent enough for the circuit to remember, the amplitude and phase can be treated as constant. Then the output of the circuit at any time will behave as though the current amplitude and phase had existed for all of eternity.

This modulated signal takes a form quite similar to the pure sinusoid, except the describing phasor is replaced with a time-dependant quantity. The modulated sinewave can be expressed as

$$x(t) = \frac{\tilde{x}(t)}{2} e^{j\omega_c t} + \frac{\tilde{x}^*(t)}{2} e^{-j\omega_c t} \qquad \text{(Eq 2-11)}$$

$e^{j\omega_c t}$ is said to be the *carrier* for this modulation, with a *carrier frequency* of f_c or equivalently ω_c. $\tilde{x}(t)$ is referred to as the *complex envelope* of the modulated signal. The complex envelope $\tilde{x}(t)$ is often referred to

interchangeably with the modulated signal $x(t)$ that it represents, the distinction being understood by context.

The complex envelope $\tilde{x}(t)$ can be described in polar form as seen before:

$$\tilde{x}(t) = E(t)e^{j\phi(t)} \qquad \text{(Eq 2-12)}$$

or rectangular (Cartesian) form:

$$\tilde{x}(t) = I(t) + jQ(t) \qquad \text{(Eq 2-13)}$$

Substituting these back into the expression for $x(t)$, gives:

$$x(t) = E(t)\cos(\omega_c t + \phi(t)) \qquad \text{(Eq 2-14)}$$

or:

$$x(t) = I(t)\cos(\omega_c t) - Q(t)\sin(\omega_c t) \qquad \text{(Eq 2-15)}$$

These expressions offer two different ways of viewing the modulated signal. First, it can be thought of as a sinusoid, with time varying amplitude and phase as described earlier, with $E(t)$ and $\phi(t)$ said to represent Amplitude Modulation (AM) and Phase Modulation (PM) respectively. Or, the signal can be thought of as the sum of two sinusoids in quadrature, with their amplitudes being modulated by $I(t)$ (**I**n-phase component) and $Q(t)$ (**Q**uadrature component). Regrettably, this notation can be confusing in circuit discussions,

where the symbol I normally refers to current, however the distinction is normally understood from context.

The trajectory of $\tilde{x}(t)$ in the *Argand (complex) Plane* is what carries information. The real and imaginary parts of $\tilde{x}(t)$ correspond to $I(t)$ and $Q(t)$, so it is also common to refer to this as the *I-Q Plane*, or the *Fresnel Plane*. How data is encoded into this trajectory will not be discussed in depth, but the effects of distortion on a signal will be considered later in this plane. The rectangular form of (Eq 2-15) is widely used in practice, as the signals $I(t)$ and $Q(t)$ relate linearly to $x(t)$, and are straightforward baseband signals to synthesize. This approach to synthesis is often referred to as *quadrature modulation*.

Although the polar form in (Eq 2-14) is useful for visualizing the form of $x(t)$, it is of more limited use in the synthesis of $x(t)$. The main problem is that $\phi(t)$ is not bounded as $\tilde{x}(t)$ circles around the origin, and thus cannot be readily represented by a single voltage. Its derivative $\omega(t) = \frac{d\phi(t)}{dt}$ is better bounded and is sometimes used in less stringent frequency-modulation schemes which are designed to not care about the actual phase and only use the frequency $\omega(t)$ to convey information. Transmitter architectures that do use a polar representation of the signal, often start with the sine and cosine of $\phi(t)$ - which are essentially I and Q normalized to eliminate E - thus, even polar transmitters

typically include a quadrature modulator to create the desired carrier phase, to which the amplitude modulation of $E(t)$ may then be applied.

2.2.4 Frequency Spectrum - Narrowband Assumption

How well the approximation of treating the system as being memoryless to the modulation relates to how slowly $\tilde{x}(t)$ moves. By the *Fourier Transform*, $\tilde{x}(t)$ can be thought of as being composed of low frequency sinewaves, which are combined together by the integral expression:

$$\tilde{x}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{X}(j\omega)e^{j\omega t}d\omega \qquad \text{(Eq 2-16)}$$

where $\tilde{X}(j\omega)$ is referred to as the *frequency spectrum* of the signal $\tilde{x}(t)$. By this representation, $x(t)$ is a combination of complex sinusoids $e^{j\omega t}$ that are combined with a weighting of $\tilde{X}(j\omega)$. For a slow-moving signal, $\tilde{X}(j\omega)$ is nonzero for small values of ω, and can be considered to be zero for $|\omega| > \omega_b$ for a certain value of ω_b, referred to as the *bandwidth* of the signal.

Substituting this expression into (Eq 2-11), a little manipulation yields:

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\frac{\tilde{X}(j(\omega-\omega_c))}{2} + \frac{\tilde{X}^*(j(\omega+\omega_c))}{2}\right]e^{j\omega t}d\omega \qquad \text{(Eq 2-17)}$$

Thus, just as with the complex envelope, the modulated signal is also composed of complex sinewaves $e^{j\omega t}$, weighted by a frequency spectrum of:

$$X(\omega) = \frac{\tilde{X}(j(\omega - \omega_c))}{2} + \frac{\tilde{X}^*(j(\omega + \omega_c))}{2} \qquad \text{(Eq 2-18)}$$

Feeding this signal through a channel with frequency response of $H(j\omega)$, the eigenvector $e^{j\omega t}$ is simply scaled by eigenvalue $H(j\omega)$, giving an output $y(t)$ of:

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)H(j\omega)e^{j\omega t}d\omega \qquad \text{(Eq 2-19)}$$

For $\tilde{X}(j\omega)$ with bandwidth of ω_b, $X(\omega)$ and hence the whole integrand is nonzero only for values of ω in a band from $\omega_c - \omega_b$ to $\omega_c + \omega_b$, and another from $-\omega_c - \omega_b$ to $-\omega_c + \omega_b$. If the dependence of $H(j\omega)$ on ω is weak enough that it can be considered a constant value within each band - $H(j\omega_c)$ and its conjugate $H(-j\omega_c)$ respectively - then this integral collapses back into:

$$y(t) \cong \frac{\tilde{x}(t)}{2}H(j\omega_c)e^{j\omega_c t} + \frac{\tilde{x}^*(t)}{2}H(-j\omega_c)e^{-j\omega_c t} \qquad \text{(Eq 2-20)}$$

This is simply $x(t)$ with its amplitude and the phase of the carrier modified according to $H(j\omega_c)$. Although the frequency response is a function of frequency, the effect of a channel being memoryless is that the frequency

response can be boiled down to this single constant, called the *complex gain*. It is very convenient to be able to model a radio channel with just this complex gain, thus systems are commonly designed to keep ω_b small enough to allow this approximation. Systems where the bandwidth is small relative to the carrier frequency and allowing the memoryless assumption are referred to as *narrowband* systems.

Channels having significant delay are not strictly memoryless: a delay of τ has a frequency response of $H(j\omega) = e^{-j\omega\tau}$, and for a long delay, the phase of this may vary significantly even within a relatively narrow frequency band, thus the frequency response is not functionally constant. However, if a channel's response can be approximated by a constant complex gain together with a pure delay factor, then it is still common to refer to it as being in memoryless: delay is treated as a separate effect from what other memory a channel does or does not have.

In recent years, there has been interest in *UltraWideBand* radio systems - commonly defined as systems where ω_b exceeds 20% of ω_c. However, this work will look only at narrowband applications.

2.3 Modulators

2.3.1 Quadrature modulator

Figure 2.4 shows an ideal canonical implementation of (Eq 2-15)

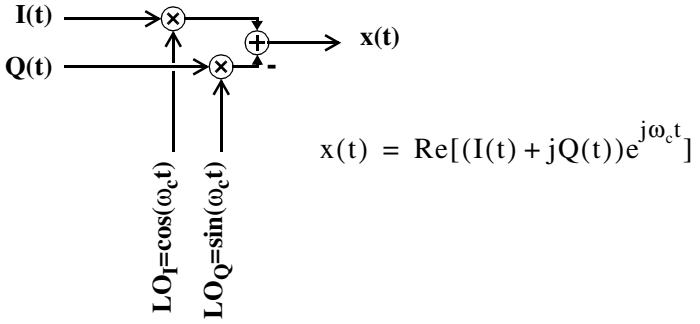


Fig. 2.4: Idealized direct conversion modulator

This modulator takes a baseband input and produces a frequency translated version of it at high frequency for its output, and thus is commonly referred to as an *upconverter*. When this upconverter is used to create the desired radio signal with a single upconversion, the architecture is referred to as the *direct conversion modulator*.

Many transmitters use direct conversion, but real implementations differ somewhat from the ideal diagram. It is difficult to implement an ideal multiplier that is linear for both its inputs, so it is common to use circuits that produce the desired product term along with products at other frequencies that can be filtered out. Some 'multiplier' implementations actually add the two signals to be multiplied, and then feed them through a nonlinear device. The nonlinear behaviour 'mixes' the two components of its input, creating the desired product,

and other by-products that are then filtered off or otherwise cancelled. Thus it is common to refer to any circuit intended to produce this multiplication as a *mixer*.

In CMOS designs, individual mixers are commonly implemented as *switching mixers*, in which a differential input signal is given to a set of switch transistors. These switch transistors are driven by the local oscillator, and with every half cycle of the oscillator signal, alternately pass the input through straight, or inverted by exchanging the differential signals. This switching amounts to multiplying the input by a +1/-1 squarewave rather than a pure sinusoid, but this squarewave consists of the desired fundamental sinewave, and odd harmonics. By-products from these harmonics can be filtered off.

The signal being switched can be either a voltage, or a current. Traditionally, mixers that switch voltage are referred to as *passiv*e mixers, with the switch devices appearing as passive switches, while current-mode mixers are usually referred to as *active mixers*.

For quadrature upconversion modulators, since the outputs of the two mixers are to be added, it is common to use current-mode active mixers. A typical circuit diagram of such a modulator is shown in Figure 2.5.

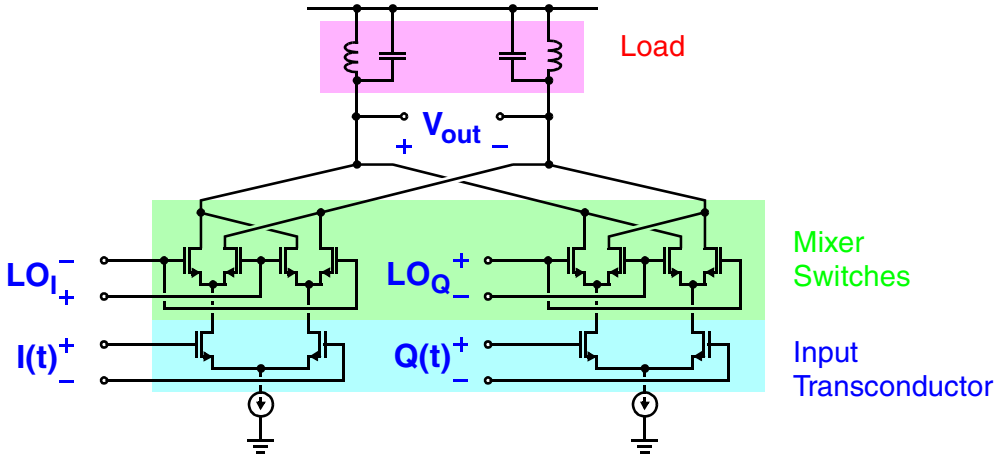


Fig. 2.5: CMOS Direct-conversion modulator

Input transconductors at the bottom convert the baseband signals to currents, which are commutated by the mixer switches. The output currents of the I and Q mixers are combined at the load to produce an output voltage. The resonant load provides filtering to attenuate products at harmonics of the carrier frequency.

2.3.2 Linear Modulator Impairments

Since a quadrature mixer combines the outputs of two signal paths, it is subject to nonidealities due to mismatch. The gains of the two signal paths, if not matched, result in errors that can be seen in the I-Q plane. If a gain

mismatch of $1 \pm \alpha$ is introduced between the I and Q channels, (Eq 2-15) becomes

$$\begin{aligned} x_g(t) &= (1+\alpha)I(t)\cos(\omega_c t) - (1-\alpha)Q(t)\sin(\omega_c t) \\ &= \mathrm{Re}[(\tilde{x}(t) + \alpha\tilde{x}^*(t))e^{j\omega_c t}] \end{aligned} \qquad \text{(Eq 2-21)}$$

Similarly, the phases of the two paths may differ from the ideal 90° quadrature between sine and cosine. If there is an error in the quadrature of $\pm\beta$, then

$$\begin{aligned} x_p(t) &= I(t)\cos(\omega_c t + \beta) - Q(t)\sin(\omega_c t - \beta) \\ &= \mathrm{Re}[(\tilde{x}(t)\cos(\beta) + j\tilde{x}^*(t)\sin(\beta))e^{j\omega_c t}] \end{aligned} \qquad \text{(Eq 2-22)}$$

The effects of gain and quadrature mismatch are very similar, in that both add an *image* of $\tilde{x}^*(t)$ to the intended complex envelope of $\tilde{x}(t)$.

Lastly, the $I(t)$ and $Q(t)$ inputs to the modulator, being baseband signals, can contain DC offsets.

$$\begin{aligned} x_{ofs}(t) &= (I(t) + I_{ofs})\cos(\omega_c t) - (Q(t) + Q_{ofs})\sin(\omega_c t) \\ &= \mathrm{Re}[(\tilde{x}(t) + (I_{ofs} + jQ_{ofs}))e^{j\omega_c t}] \end{aligned} \qquad \text{(Eq 2-23)}$$

The effect of such an offset is often called *carrier leak*, as the extra signal at the output is simply an unmodulated tone at the carrier frequency.

The effects of these impairments in the IQ plane are shown below in Figure 2.6

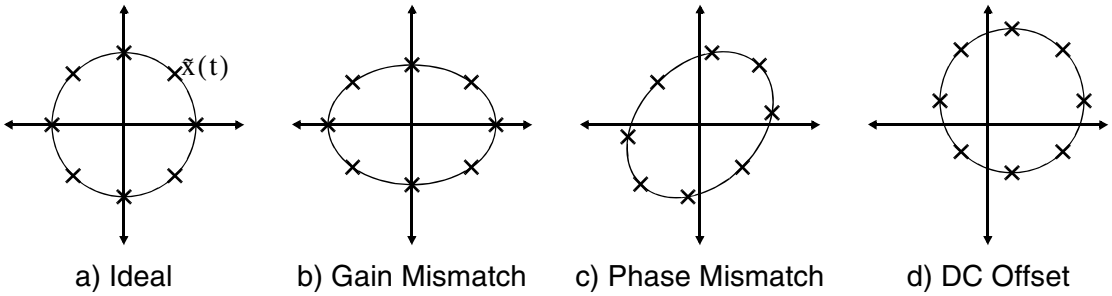


Fig. 2.6: Effect of Linear Impairments in IQ plane

These nonidealities are not necessarily the product of the modulator: the DC offset and gain mismatch can just as easily be a product of the DACs or other baseband circuitry before the modulator. Phase error is typically introduced where the local oscillator signal is split into the sine and cosine reference phases, rather than the at the mixer core. These errors all add linearly to the intended $\tilde{x}(t)$, and as long as the respective coefficients are known, can be corrected for by applying appropriate linear transforms to the baseband input signal before the modulator. The coefficients can be determined adaptively by monitoring the amplitude of the output as the modulation moves around the I-Q plane: this is described in more detail in [2][3], but the need to do this can be minimized with careful design.

These effects are typically not visible in the frequency domain when generating a normal modulated signal by direct conversion: the image and carrier leak are hidden by the spectrum of the desired signal. These nonidealities

can be exposed by using a modulation consisting of a single test tone: $\tilde{x}(t) = E_{test}e^{j\omega_{test}}$, in which case the modulated signal, its image, and carrier leak, end up spaced ω_{test} from each other. This is commonly referred to as the *single-sideband (SSB) test*. The magnitude of the image and carrier leak relative to the desired tone are common performance metrics for modulator performance evaluated by this test.

The SSB test will often also produce other tones in the vicinity of the carrier. These are distortion products produced by nonlinearity, either in the mixers or in the baseband circuitry preceding them. Nonlinearity will be discussed in more depth in the context of the power amplifier, however for characterizing mixer nonlinearity, it is common to look at the magnitude of these tones.

2.4 Nonlinearity

The treatment of radio signals in Section 2.2 assumed that circuits respond linearly to their inputs, and memoryless channels can be characterized by a simple constant complex gain. If we denote this gain as simply $\bar{A}$ (for **A**mplifier gain), then the output of the system, $\tilde{y}(t)$ for a given input $\tilde{x}$ is simply

$$\tilde{y}(t) = \bar{A}\tilde{x}(t) \qquad \text{(Eq 2-24)}$$

For a linear system, $\bar{A}$ is simply a constant, however for a real circuit, this is not necessarily so. As the signal through an amplifier varies in amplitude, the output does not necessarily follow proportionally, nor does the phase through the amplifier necessarily remain constant. We can still use this expression however, substituting an appropriate function for $\bar{A}$.

While $\bar{A}$ can be a function of the input amplitude, it does not depend on the phase of the input signal. A phase change of the input is equivalent to a time shift, and an amplifier has no basis for knowing the absolute time: its response to the input would follow the same time shift, but otherwise remain identical. Thus, $\bar{A}$ can be considered to be a function of the input amplitude only. Rather than the amplitude, the gain can be expressed as a function of the square of the amplitude instead - reasons to prefer this will be seen later. The input-output relationship of an amplifier can thus be expressed in terms of its gain as:

$$\tilde{y}(t) = \bar{A}(|\tilde{x}(t)|^2)\tilde{x}(t) \qquad \text{(Eq 2-25)}$$

Rather than express the relationship in terms of a large signal gain, we can also keep it in terms of the amplifier's input and output phasors, in which case the relationship can be written as:

$$\tilde{y}(t) = \tilde{F}(\tilde{x}(t)) \qquad \text{(Eq 2-26)}$$

where $\tilde{F}(\tilde{x})$ is the envelope transfer function given by:

$$\tilde{F}(\tilde{x}) = \bar{A}(|\tilde{x}|^2)\tilde{x} \qquad \text{(Eq 2-27)}$$

2.4.1 AM/AM, AM/PM

This input-output characteristic for an amplifier can be measured empirically: the output power of the amplifier and phase shift through it can be measured for inputs $\tilde{x}$ of various amplitudes, and $\tilde{F}(\tilde{x})$ can be determined from the results. The output amplitude of the amplifier as a function of its input amplitude is referred to as the amplifier's AM/AM characteristic and is just $|\tilde{F}(\tilde{x})|$. The phase shift as a function of input amplitude $\angle\bar{A}(|\tilde{x}|^2)$, or equivalently $\angle\tilde{F}(|\tilde{x}|)$, is referred to as the AM/PM characteristic. These measurements are a standard power-sweep procedure for vector network analysers. Examples of AM/AM and AM/PM curves are shown in Figure 2.7.

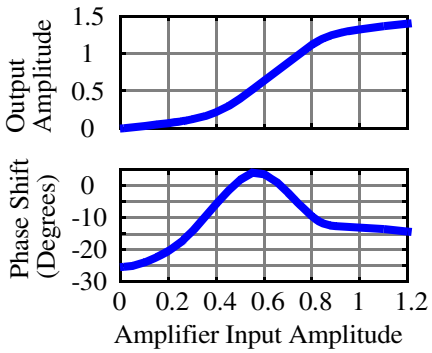


Fig. 2.7: AM/AM and AM/PM curves

For a perfectly linear amplifier, the AM/AM curve would be a perfectly straight line, while the AM/PM curve would be constant. If the AM/AM curve increases more slowly than it would for a linear amplifier, the amplifier is said to be in *gain compression*, while an AM/AM curve that increases faster than linearly is said to be in *gain expansion*. Nonlinearities of device I-V characteristics will generally introduce AM/AM effects. AM/PM effects can

arise from voltage-sensitive capacitances that change the apparent reactance seen at a node as signal amplitude changes. These mechanisms are interrelated though - it would be wrong to say that AM/AM and AM/PM result exclusively from one effect or the other, but both effects be seen empirically without regard for the underlying mechanisms.

The effects of AM/AM and AM/PM are easily seen in the IQ plane. Every input phasor $\tilde{x}$ gets mapped to an output phasor $\tilde{F}(\tilde{x})$. This mapping is easily seen to be rotationally symmetric:

$$\tilde{F}(\tilde{x}e^{j\theta}) = \bar{A}(\left|\tilde{x}e^{j\theta}\right|^2)\tilde{x}e^{j\theta} = \bar{A}(|\tilde{x}|^2)\tilde{x}e^{j\theta} = \tilde{F}(\tilde{x})e^{j\theta} \qquad \text{(Eq 2-28)}$$

That is, $\tilde{x}$ can be rotated by an arbitrary angle θ, and the resulting output is equivalent to taking the system's behaviour around $\tilde{x}$ and rotating its output by the same amount. Circles around the origin in the $\tilde{x}$ domain represent signals of a particular amplitudes, and map to $\tilde{F}(\tilde{x})$ of particular amplitudes - again circles around the origin. The AM/AM characteristic of the amplifier is captured in the spacing of these circles, while the AM/PM is captured by how much the output twists as a function of amplitude. This is illustrated in Figure 2.8.

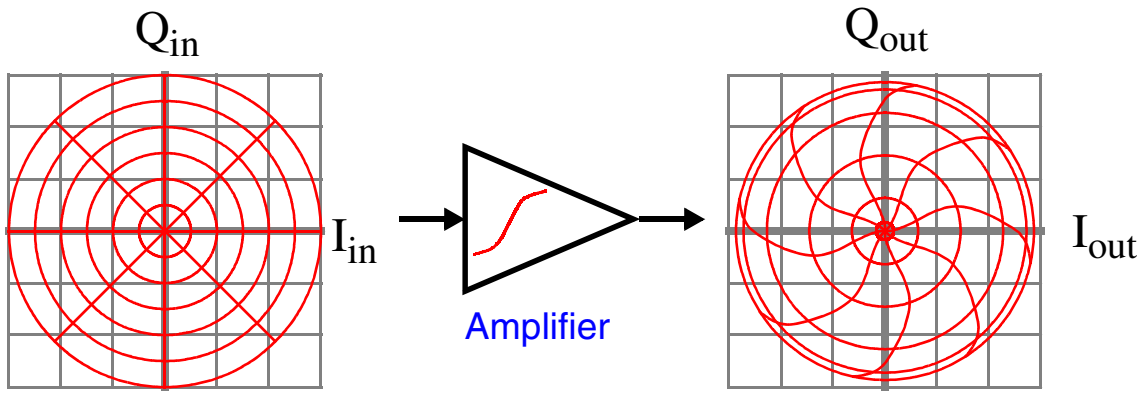


Fig. 2.8: Effect of Amplifier Nonlinearity in IQ plane

Knowing $\bar{F}(\bar{x})$ is extremely useful for system simulations. A transient circuit simulation involving modulated signals can take a very long time to execute as the simulator needs to follow every cycle of the carrier, over the long time scales of the modulation. By characterizing the amplifier in terms of $\bar{F}(\bar{x})$, the effects of the amplifier can be considered in IQ space, where there are no fast-moving carrier frequency effects that need be considered, and a transient simulation can be performed in terms of baseband modulated signals alone.

EVM introduced by amplifier nonlinearity is easily evaluated with this mapping: a modulated signal fed through the amplifier goes through this mapping when considered in the IQ plane. The distortion of the output in the IQ plane compared to what it should be relates directly to the EVM measurement.

Spectral mask performance for an amplifier can also be evaluated using this mapping: the IQ plane representation of the amplifier's output is found

using this mapping, and the fourier transform of this baseband representation gives the spectrum of the output signal. Spectral mask performance is also often considered conceptually using volterra series concepts, which are discussed next.

2.4.2 Volterra Series

For weakly nonlinear systems, it is common to consider Volterra-series representations of the system[4][5]. The Volterra series is an extension of the linear frequency response from Section 2.2.2 to nonlinear behaviour. The functions that characterize distortion are known as Volterra-series kernels

An amplifier's linear transfer function is its first-order Volterra Kernel, and its contribution to the output as given in (Eq 2-19) can be written in a shorthand notation [6]:

$$y(t) = H_1(j\omega) \lozenge x(t) \qquad \text{(Eq 2-29)}$$

This is read as $H_1(j\omega)$ acting on $x(t)$, where $H_1(j\omega)$ is the linear $H(j\omega)$ seen earlier. The input signal $x(t)$ is understood to consist of complex sinusoids $e^{j\omega t}$, and the Volterra kernel is a scalar multiplier on each sinusoid.

The Volterra series extends this concept to higher powers of $x(t)$. The second-order response of a system can be written as:

$$H_2(j\omega_1, j\omega_2) \Diamond (x(t)x(t)) \qquad \text{(Eq 2-30)}$$

The operand of $x(t)x(t)$ consists of complex sinusoids which are the product of $e^{j\omega_1 t}$ and $e^{j\omega_2 t}$ coming from the first and second $x(t)$ factor respectively, and these products are multiplied by $H_2(j\omega_1, j\omega_2)$. For an input signal with components only around ω_c and $-\omega_c$, even-order distortion is of limited interest, producing distortion around even multiples of ω_c, which are far enough away from the carrier to be easily filtered off. However, the same is not true of odd-order distortion.

The third-order volterra kernel can be considered in the frequency domain. For a modulated signal with frequency spectrum of $X(\omega)$, $x(t)^3$ has a spectrum that consists of $X(\omega)$ convolved with itself twice, that is: $Y_3(\omega) = X(\omega) \otimes X(\omega) \otimes X(\omega)$. If $X(\omega)$ is contained within a certain bandwidth around ω_c, then $Y_3(\omega)$ has a product around the carrier that comes from convolving $X(\omega)$ near ω_c, ω_c, and $-\omega_c$ and is contained in a bandwidth three times as wide as the original. The volterra kernel $H_3(j\omega_1, j\omega_2, j\omega_3)$ is involved in the convolution, and can affect the shape of the resulting spectrum somewhat, but does not impact its bandwidth.

Higher-order odd-order kernels introduce distortion products with accordingly wider bandwidths around the carrier. The frequency spectrum of a

hypothetical modulated signal with third and fifth order distortion products is shown in Figure 2.9.

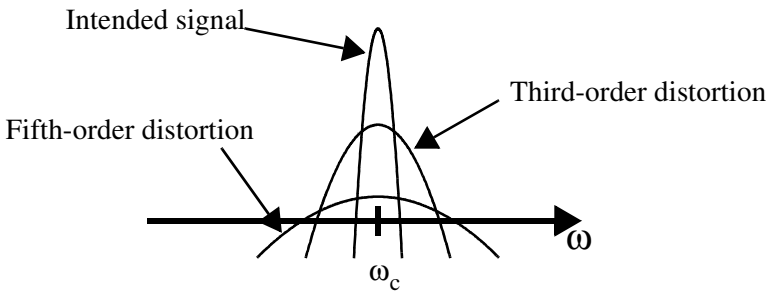


Fig. 2.9: Effect of Amplifier Nonlinearity in frequency domain

The distortion adds 'skirts' to the spectrum of the modulated signal, and this spreading in the frequency domain is referred to as *spectral regrowth*. This spectral regrowth is what spectral mask and ACPR measurements aim to characterize.

For weak nonlinearities, the distortion is dominated by the third-order term, and tests such as the two-tone test estimate the magnitude of this H_3, using metrics such as the IP_3 intercept point or the 1dB compression point. However, for stronger nonlinearities as can be found in a power amplifier, the higher-order terms are also significant.

2.4.3 AM/AM, AM/PM and Volterra Series equivalence

It is common in textbooks [7][8][9] trying to analyse AM/AM behaviour, to model amplifier nonlinearity with a completely memoryless

power-series, where the instantaneous output voltage is a power-series function of only the amplifier's input voltage at that instant, e.g.

$$y(t) = a_1 x(t) + a_2 x^2(t) + a_3 x^3(t) + \ldots \qquad \text{(Eq 2-31)}$$

Substituting in $x(t) = E(t)\cos(\omega_c t + \phi(t))$, the fundamental component of the output can then be shown to be:

$$y(t) = \left[a_1 + \frac{3}{4}a_3 E^2(t) + \frac{5}{8}a_5 E^4(t) + \ldots\right] E(t)\cos(\omega_c t + \phi(t)) \qquad \text{(Eq 2-32)}$$

$$+ \text{higher frequency terms}$$

While the polynomial found does give an expression for AM/AM, this approach is unsatisfying. Most importantly, no mechanism for AM/PM is offered. Phase modulation is a memory effect, but by starting with a completely memoryless power series, any hope of capturing AM/PM is lost.

The memoryless power-series is used since Volterra-series analysis is usually deemed to be too complicated for meaningful manual analysis. Even when it is used, it is typically only carried out to third-order kernels. While this generally works well for the analysis of linear circuits such as found in receivers, for a circuit as nonlinear as a power amplifier, this is inadequate for describing observed distortion, and higher-order kernels need to be considered.

Volterra-series analysis need not be unwieldy, however. The same memoryless assumption used in looking at the linear response of a system can be

applied to the nonlinear Volterra-series analysis as well. The amplifier may have memory on the time scale of the carrier frequency, but no longer-term memory that responds on the time scale of the modulation: indeed, these are the assumptions necessary for defining AM/AM and AM/PM in the first place.

Consider how AM/AM and AM/PM are measured: an input sinusoid of some amplitude is given to an amplifier, and the phase shift and output amplitude of the amplifier are measured, and the measurement is performed at different amplitudes. The results of this are easily predicted for a system whose volterra kernels are known.

Feeding in an input of

$$x(t) = \frac{\tilde{x}}{2} e^{j\omega_c t} + \frac{\tilde{x}^*}{2} e^{-j\omega_c t} \qquad \text{(Eq 2-33)}$$

the first-order volterra kernel H_1 gives an output of simply:

$$y_1(t) \cong \frac{\tilde{x}}{2} H_1(j\omega_c) e^{j\omega_c t} + \frac{\tilde{x}^*}{2} H_1(-j\omega_c) e^{-j\omega_c t} \qquad \text{(Eq 2-34)}$$

or using phasor notation,

$$\tilde{y}_1 = H_1(j\omega_c)\tilde{x} \qquad \text{(Eq 2-35)}$$

Now consider the output of the third-order kernel:

$$y_3(t) = H_3(j\omega_1, j\omega_2, j\omega_3) \lozenge (x(t))^3 \qquad \text{(Eq 2-36)}$$

The kernel acts on the operand of $x(t)^3$, which can be expanded as:

$$\begin{aligned} x(t)^3 &= \left(\frac{\tilde{x}}{2}e^{j\omega_c t} + \frac{\tilde{x}^*}{2}e^{-j\omega_c t}\right)^3 \\ &= \frac{\tilde{x}^3}{8}e^{j(\omega_c+\omega_c+\omega_c)t} + \frac{\tilde{x}^{*3}}{8}e^{j(-\omega_c-\omega_c-\omega_c)t} \\ &+ \frac{3\tilde{x}\tilde{x}^*}{4}\left(\frac{\tilde{x}}{2}e^{j(\omega_c+\omega_c-\omega_c)t} + \frac{\tilde{x}^*}{2}e^{j(\omega_c-\omega_c-\omega_c)t}\right) \end{aligned} \qquad \text{(Eq 2-37)}$$

The first two terms are at $\pm 3\omega_c$ and are readily filtered off being far away in frequency from the intended signal, however, the other two terms lie at $\pm\omega_c$, which is right at the frequency of interest. $H_3(j\omega_c, j\omega_c, -j\omega_c)$ and $H_3(j\omega_c, -j\omega_c, -j\omega_c)$ will act on these two terms respectively. Using phasor notation, the output at ω_c is just:

$$\tilde{y}_3 = \frac{3\tilde{x}\tilde{x}^*}{4}H_3(j\omega_c, j\omega_c, -j\omega_c)\tilde{x} \qquad \text{(Eq 2-38)}$$

Similar products get created for higher odd-order kernels as well. The net contribution of these to the output at $\pm\omega_c$ can be summed up as

$$\tilde{y} = \left[H_1(j\omega_c) + \frac{3}{4}\tilde{x}\tilde{x}^*H_3(j\omega_c, j\omega_c, -\omega_c) + \frac{5}{8}(\tilde{x}\tilde{x}^*)^2H_5(j\omega_c, j\omega_c, j\omega_c, -j\omega_c, -j\omega_c) + \ldots\right]\tilde{x} \qquad \text{(Eq 2-39)}$$

This expression is for the amplifier being fed a pure sinusoid. As with the memoryless assumption for the linear system, it can be argued that under certain conditions, these higher-order volterra kernels can also be treated as

constants, in which case the same expression can be used for narrowband modulated signals:

$$\tilde{y}(t) = \bar{A}(|\tilde{x}(t)|^2)\tilde{x}(t) \qquad \text{(Eq 2-40)}$$

where:

$$\bar{A}(|\tilde{x}|^2) = H_1 + \frac{3|\tilde{x}|^2}{4}H_3 + \frac{5|\tilde{x}|^4}{8}H_5 + \frac{35|\tilde{x}|^4}{64}H_7 + \ldots \qquad \text{(Eq 2-41)}$$

or more concisely

$$\bar{A}(|\tilde{x}|^2) = \sum_{i=0}^{\infty} A_{2i+1}|\tilde{x}|^{2i} \qquad \text{(Eq 2-42)}$$

where

$$A_{2i+1} = \frac{(2i+1)!}{4^i(i+1)!i!}H_{2i+1} \qquad \text{(Eq 2-43)}$$

This complex polynomial is the amplifier's large-signal gain, and is simply the result of the memoryless power-series analysis with complex constants from the Volterra kernels substituting for the real coefficients of the power series. As with the power-series analysis, the AM/AM curve is easily extracted, and is simply $\left|\bar{A}(|\tilde{x}|^2)\tilde{x}\right|$. However, this polynomial also gives the amplifier's AM/PM characteristic, which is $\angle A(|\tilde{x}|^2)$.

It should be noted that the Volterra kernels functions may contain factors of the form $\frac{1-(\omega_1-\omega_2)/\omega_z}{1-(\omega_1-\omega_2)/\omega_p}$, which would behave as 1 for a pure sinusoidal input, but for a modulated input with nonzero bandwidth, would appear as $\frac{\omega_p}{\omega_z}$ for the products of input signals from sufficiently different frequencies. Factors like this relate to low-frequency nodes in the circuit: bias nodes or slow thermal effects being affected by signal power. Although a narrowband modulation is assumed to move slowly relative to the carrier frequency, it still moves quickly relative to these low-frequency effects, thus the appropriate values to take for the Volterra kernels are not necessarily the coefficients in (Eq 2-39). Any measurements performed to extract AM/AM and AM/PM curves need to be done in a manner such that these slow states get set appropriately to what they would be for the modulated signals being modelled. Pulsed power-sweep measurements of the PA aim to keep thermal effects more faithful to what they would be in actual use.

Knowing the Volterra kernels, it is easy to extract the AM/AM and AM/PM curves for the amplifier. Conversely, given AM/AM and AM/PM curves, one could fit a polynomial to $\bar{A}(|\bar{x}|^2)$, and from the coefficients get representative values for the Volterra kernels. In a sense, traditional Volterra-series analysis does exactly that: normal hand-analysis techniques look at the bias point and first and higher-order derivatives of device characteristics there to derive the

kernels. This amounts to taking the Maclaurin series of $\bar{A}(|\bar{x}|^2)$. Other methods of fitting a polynomial - polynomial regression for instance - are just as legitimate as the Mclaurin series, and the coefficients found by other fitting methods may actually turn out to be more meaningful than volterra kernels found with traditional techniques. Regression fits would capture operation across different device operating regions better than arbitrarily many derivatives of a device model within one operating region. What fitting metric should be used for the regression is an open question, but for modelling the response to modulated signals, a mean-square fit weighted by the modulation's amplitude probability density function would seem appropriate.

Older works [10][11][12][13] recognize the correct relationship between Volterra series and AM/AM and AM/PM curves, but the derivation comes about as a special case of complex multitone analyses and the simplicity of the single-carrier result is obscured. Examples given in these works only give linear and third-order terms, and extending to higher degrees, although alluded to, is not pursued.

The idea of using a complex polynomial to represent $\bar{A}(|\bar{x}|^2)$ is one which has been proposed as an empirical model: Kenington [9] gives an example of fitting a polynomial in $|\bar{x}|$ to AM/AM and AM/PM curves for a class A and class C amplifier. This empirical fit uses all orders - both even and odd - of the

input amplitude up to a certain degree, and while coefficients of the fit are listed, they do not relate to anything.

Cripps, in section 3.3 of [14], recognizes that a polynomial fit of AM/AM and AM/PM curves relates to the Volterra kernels. The need for both magnitude and phase coefficients for the polynomial is acknowledged, but the relationship between these coefficients, the AM/AM and AM/PM curves, and volterra kernels is not clearly given, although the descriptions imply the correct model. Cripps offers a comparison between the results of polynomials fitted to power-sweep measurements and two-tone test results, noting some challenges in getting accurate coefficients from the power-sweep results.

2.4.4 Constant-Envelope Modulation

Nonlinearity of the amplifier means that the large-signal gain is nonconstant with respect to its input amplitude, and this variation can introduce distortion. Conversely, if the amplifier can be made to have a constant large-signal gain, then it functions as a linear amplifier would and does not introduce distortion. Since variation of the large-signal gain comes from variations of the signal amplitude, one approach to eliminating the variation in gain is to hold the signal amplitude constant. As long as the signal amplitude does not change, the amplifier's gain does not either, and the amplifier is indistinguishable from a linear amplifier with the same gain.

Some radio systems take this approach to help mitigate the effects of amplifier nonlinearity. The modulation schemes they use, which keep signal amplitude constant are known as *Constant-Envelope Modulation*.

It is easy to see in the IQ domain that constant-envelope signals are not distorted: the modulated signal remains on a circle in the IQ plane, and as seen in Section 2.4.1, this maps to a circle in the output IQ plane.

The picture in frequency domain is slightly more difficult to see, but it still holds that no distortion is created. The spectrum for third order distortion is $Y_3(\omega) = X(\omega) \otimes X(\omega) \otimes X(\omega)$. For convenience, define an intermediate product of $U(\omega) = X(\omega) \otimes X(\omega)$ - the distortion is $Y_3(\omega) = X(\omega) \otimes U(\omega)$. In time domain, the intermediate product is simply

$$u(t) = x(t)^2 = \tilde{x}(t)\tilde{x}^*(t) + \frac{\tilde{x}(t)^2}{4} e^{2j\omega_c t} + \frac{\tilde{x}^*(t)^2}{4} e^{-2j\omega_c t} \qquad \text{(Eq 2-44)}$$

For constant $|\tilde{x}(t)|$, the first term is a constant at DC or $\omega = (\omega_c - \omega_c) = 0$. The spectrum $U(\omega)$ is thus a delta at $\omega = 0$ (from the DC constant) and some modulated products around $\pm 2\omega_c$.

Convolving this with $X(\omega)$ to get $Y_3(\omega)$, the delta simply reproduces the input after convolution. The other two terms of $u(t)$ after convolving, produce the same products around $\pm\omega_c$; to arrive at the same frequency, they are the same

convolution of $X(\omega)$ around ω_c, ω_c and $-\omega_c$, just performed in different orders. Thus, all the products of $Y_3(\omega)$ around ω_c are just the spectrum of $X(\omega)$.

Constant-envelope modulation is used in older radio standards such as AMPS, DECT and GSM, to avoid linearity issues of the PA. These modulation schemes represent data only in the signal phase, but not in the amplitude. The signal amplitude is a degree of freedom which can carry information, and by not using the amplitude, these systems carry less data over their assigned radio bandwidths than would otherwise be possible. As radio spectrum is a finite resource, newer radio applications have adopted nonconstant-envelope modulation schemes to deliver higher data rates within the available spectrum, thus PA nonlinearity can no longer be ignored like it used to be.

2.4.5 Power Backoff, Peak to Average Ratio (PAR)

PA designs are commonly rated for a maximum power they can deliver at their output with a constant-envelope signal. The actual power being produced at the PA output relative to this maximum power is referred to as the *power backoff*.

The need for some backoff is easily seen by considering what happens when a clipping but otherwise linear amplifier is driven to its maximum output power. Further increasing the input, no additional output power, so any nonconstant-envelope signal being given to the amplifier like this will

experience distortion. Reducing signal levels brings the PA out of clipping, allowing for reasonable modulation of its output amplitude.

A nonconstant-envelope modulated signal will have a peak signal power that is larger than its average. The ratio of these - the *peak-average ratio (PAR)*, also known as the *crest factor* - is a common metric for modulation schemes, and represents the amount of backoff required for a clipping but otherwise perfectly linear amplifier to reproduce the modulated signal without clipping.

The large-signal gain given by (Eq 2-41) highlights this relationship between backoff and linearity: if the amplitude of the signal is kept small enough, the higher-order distortion terms are small, and the amplifier's gain effectively looks like the constant H_1. Reducing the amplitude of the signal (increasing backoff) reduces these distortion products faster than the intended signal, thus improving linearity.

While increasing backoff seems like an easy fix to linearity concerns, it comes at a price: as will be seen in the next section, reducing output power reduces power efficiency, so there is an inherent tradeoff between linearity and power efficiency.

2.5 Power Amplifiers

Figure 2.10 shows a simplified schematic of a typical power amplifier

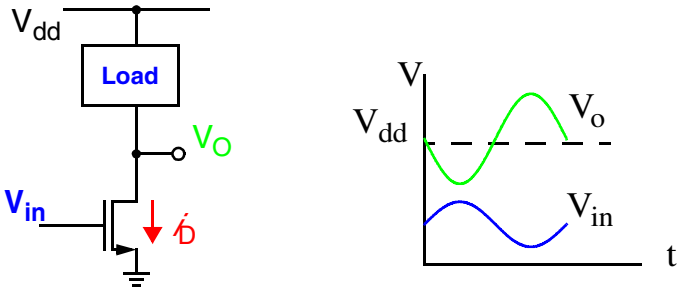


Fig. 2.10: Simplified PA output stage

together with waveforms for its input and output voltages. The circuit consists of an active device and a load. The active device may be a bipolar junction or field-effect device, and may be in silicon, gallium arsenide, or other technology. The load typically consists of the output to be driven, and an impedance matching network to scale signal voltages to a level appropriate for the amplifier. Also included in the load is an inductor to provide the DC current for the active device, as well as capacitances from the device itself or the matching network.

A sinusoidal input signal is given to the active device, which operates in inverting mode. The device pulls more current to bring the output voltage low when the input voltage is high, and lets the output voltage go high when the input is low.

The device passes current through some or all of the cycle, with the amount of time turned on being referred to as the *conduction angle*. A PA whose device is turned on all the time is said to have a conduction angle of 360°, and PAs whose devices are turned on less of the time have conduction angles proportionally less.

In the simplified schematic, all current drawn by the device comes from the supply. The instantaneous power dissipated in the device is $(V_{DD} - V_o)i_D$, while power drawn from the supply is $V_{DD}i_D$. The instantaneous drain efficiency is $1 - \frac{V_o}{V_{DD}}$, thus the best efficiency is achieved by drawing current when V_o is low. This in turn is achieved with the output amplitude as large as achievable.

2.5.1 PA Classes

The drain efficiency of $1 - \frac{V_o}{V_{DD}}$ given above is an instantaneous efficiency, but the average efficiency through a full cycle of the carrier depends on the drain voltage and current waveform through the cycle. How much current is drawn through each cycle of the input varies with PA design, but designs generally fall into one of several classes. The major classes will be summarized next. Less common classes such as class F or J are omitted, but descriptions for

these as well as more comprehensive discussion of the standard classes, can be found in Cripps [8], Kenington [9] and other texts.

2.5.1.1 Class A

A class A amplifier is conceptually the simplest to understand. The device is biased so that it conducts a current at all times. If the active device is assumed to be a linear transconductor, then a sinusoidal input voltage causes the device to draw a sinusoidal current, which causes a sinusoidal output voltage, and the relationship between the input signal's amplitude will be reflected linearly at the output. The bias at the input causes the transistor to draw an average current $\hat{I_D}$, which the output sinewave is then superposed on to.

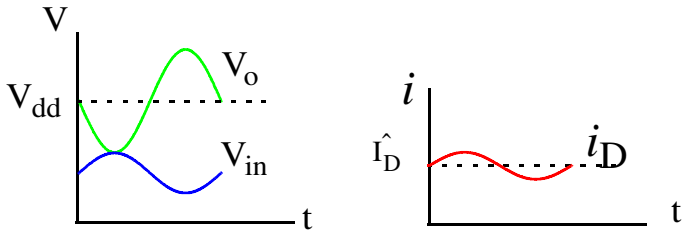


Fig. 2.11: Class A waveforms

Maximum power efficiency for Class A is achieved when the output voltage has an amplitude of V_{DD} and the device is drawing a sinusoidal current with amplitude of $\hat{I_D}$. Power delivered to the load in this case is $(V_{DD}\hat{I_D})/2$, while power drawn from the supply is $V_{DD}\hat{I_D}$. The best achievable drain efficiency of a Class-A amplifier with a linear transconductor is thus:

$$\eta_{max} = \frac{P_{RFout}}{P_{DC}} = \frac{(V_{DD}\hat{I}_D)/2}{V_{DD}\hat{I}_D} = 50\% \qquad \text{(Eq 2-45)}$$

The class A amplifier is considered to have good linearity, but the ideal maximum power efficiency is only 50%.

With a linear transconductor, the average current drawn is not a direct function of the output amplitude - the average of $\hat{I}_D$ is independent of the zero-average sinusoidal signal current superimposed on top of it. Thus, ideal class A amplifiers can be considered to have constant power consumption, independent of the output signal, and power efficiency proportional to the output power.

2.5.1.2 Class B

One approach to improving power efficiency is to recognize that the bias current of the class A amplifier is a significant waste of power for low signal amplitudes. Reducing the bias current reduces power consumption, but introduces clipping of the drain current when the signal amplitude exceeds the bias.

There is a continuum of what one can choose for the bias point. Setting the bias point such that with zero input, the device is on the threshold of turning on results in a conduction angle of 180°. This is referred to as class B operation.

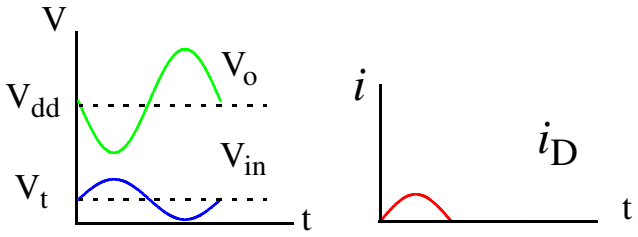


Fig. 2.12: Ideal Class B waveforms

Resonance of the load makes the drain voltage appear roughly sinusoidal despite non-sinusoidal current: the ideal class B analysis assumes the voltage is to be a pure sinusoid. With a linear transconductor, drain current is a pure sinewave with the negative half-cycles truncated. Under these assumptions, the class B PA has a maximum drain efficiency of:

$$\eta_{max} = \frac{P_{RFout}}{P_{DC}} = \frac{\int (V_{DD} - V_o(t)) i_D(t) dt}{\int V_{DD} i_D(t) dt} = \frac{\int_0^{\pi} (\sin\theta)^2 d\theta}{\int_0^{\pi} \sin\theta d\theta} = \frac{\pi/2}{2} \cong 78.5\% \qquad \text{(Eq 2-46)}$$

With a linear transconductor, the average current drawn is proportional to the signal amplitude, while the output power is proportional to the square of the amplitude. Thus, class B has a drain efficiency proportional to the output amplitude.

2.5.1.3 Class A,B nonidealities - Class AB

Class A and Class B both have the property that their output signals are directly proportional to their input signals if the device behaves as an ideal linear transconductor when turned on. However, real devices do not behave according to this ideal. For real devices, the transition from being turned-off to operating as a transconductor is a gradual one, and the effective transconductance varies through each cycle. A class B amplifier has gain expansion: as the signal amplitude becomes larger, the device gets pushed into having a larger average transconductance. Class A has an overall gain that is less sensitive to signal amplitude: there is a similar gain expansion for the half-cycle where a class B amplifier is turned on, but the device is also turned on in the other half of the cycle where it experiences a complementary gain compression. For square-law devices, these two effects cancel and the Class-A amplifier is linear with respect to the fundamental.

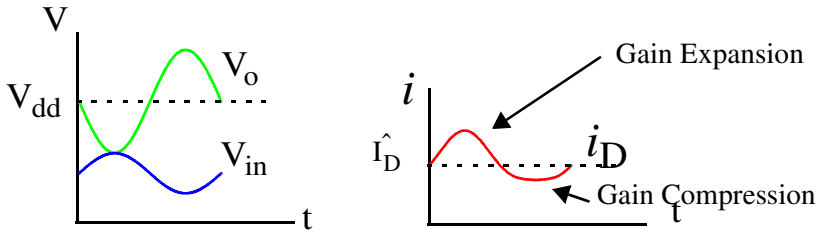


Fig. 2.13: Class A with square-law device

The assumption so far has been that the drain current is a function of only the input voltage. When the output voltage swings low enough, the current becomes dependent on not just the input voltage but the output as well, and passes less current than what would otherwise be expected given the input

signal. This is referred to as the knee-effect, causing gain compression in the amplifier which can be seen as saturation in the AM/AM curve. The voltage below which this happens is called the knee voltage.

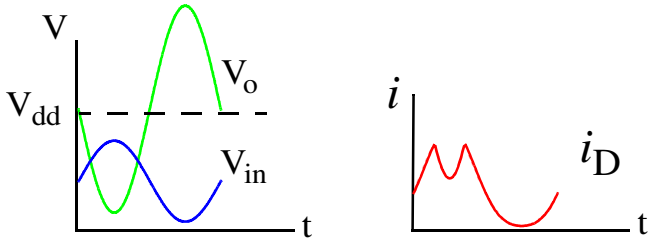


Fig. 2.14: Knee Effect (output saturation)

This knee effect is one of the factors that makes it difficult to design linear CMOS PAs. The low transconductance of a CMOS transistor relative to other technologies means that the input amplitude must be large. Square-law CMOS transistors go into triode operation when the drain voltage falls a threshold voltage below the drain voltage, and with the large input swing, this knee voltage is relatively high. In contrast to this, bipolar devices appear essentially as current sources for collector voltages down to collector voltages on the order of 0.3V.

It is common for PA designers seeking good linearity to try and balance gain expansion against knee-effect saturation by biasing the device in between class A and class B operation. For small signals, the device behaves as class A, giving good linearity. For larger signals where the output begins to saturate, the device also starts to turn off for part of the cycle, allowing more of the gain expansion of the rest of the cycle to dominate. This gain expansion can roughly

cancel the gain compression from output saturation, thus extending the amplifier's linear operation to larger amplitudes than would have been achievable for either class A or class B.

This operation is referred to as class AB, and is characterized by the device operating with a conduction angle between 360° and 180°. While class A and class B refer to specific conduction angles, class AB operation covers the range in between. The power efficiency of class AB lies somewhere between class A and B. It is sometimes stated [9] that distortion of class AB also lies between class A and B, but this is relevant only looking at harmonics of the carrier: when considering the AM/AM characteristic of the fundamental, class AB linearity can actually exceed either class A or B.

2.5.1.4 Class C

Going from class A to class B, the conduction angle reducing from 360° to 180° improves the drain efficiency. It stands to reason that further reducing the conduction angle should improve drain efficiency further. Amplifiers with a conduction angle less than 180° are referred to as class C.

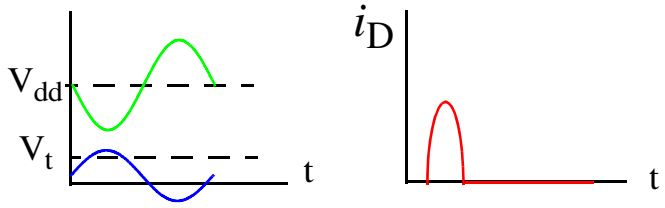


Fig. 2.15: Class C Waveforms

Shorter conduction angles offer better drain efficiency by drawing current only when the drain voltage is at its lowest: the efficiency with ideal components approaches 100% as the conduction angle approaches zero. This limit is not a meaningful one, however; as the conduction angle is reduced, power to the load is delivered over a shorter time, and to keep the output power reasonable, the current supplied by the device, and hence device sizes must be increased inversely to the conduction angle. As the conduction angle tends to zero, device sizes needed tend to infinity, and parasitic effects of the large devices make such an approach impractical.

While the limit of zero conduction angle is not a realistic one, it is still reasonable to design amplifiers with conduction angles more modestly less than than 180°.

As with class A and B, the amplitude of the input signal affects the amplitude of the output signal in class C, however there is no longer an ideally linear design. Reduced conduction angles are achieved by biasing the device below threshold, and input signals that are small enough in amplitude are insufficient to turn the device on. The AM/AM curve has a dead zone for small

input signals, and doesn't produce an output until the input is large enough to cross the device turn-on threshold.

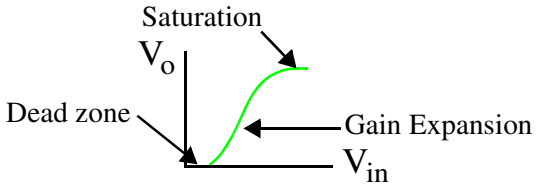


Fig. 2.16: Class C AM/AM curve

Once the device is turned on, the conduction angle is a function of the input amplitude. Together with the average device transconductance increasing, both effects cause gain expansion, until knee-effect saturation comes into play. Only in the narrow region in the transition from gain expansion to saturation does the amplifier appear linear. Being so narrow, this transition region is not a useful design point, so class C is generally considered to be nonlinear.

2.5.1.5 Switch-Mode Class D/Class E

The ideal 100% efficiency of class C is not practically achievable, as the efficiency comes from drawing current while the drain voltage is near ground, and with a sinusoidal drain voltage normally assumed for classes A, B and C, this occurs only for an arbitrarily narrow time window at the sinewave's bottom. One approach to keep device sizes reasonable compared to class C is to abandon the sinusoidal voltages at the drain, and use a drain voltage that stays around zero for a more significant fraction of the carrier cycle. The non-sinusoidal voltage is then filtered to attenuate harmonics delivered to the load.

Ideal efficiency is achieved by the device drawing current only when the drain voltage is zero, and passing no current when the voltage is higher. This is achieved with an ideal switch. A class D [15] amplifier alternately switches its output voltage between ground and a bypassed supply voltage, creating a squarewave which is then filtered. In principle, this approach should give good power efficiency, however fCV^2 losses for any capacitances directly driven by this squarewave limit efficiency.

Class E [16] remedies the fCV^2 losses by letting resonance of the load provide the high-swing of the waveform, while using a switch only when the voltage is to be held low. The centre frequency and damping of the load are designed so that the voltage across the switch are brought to zero by the resonance before the switch turns on - this is referred to as soft-switching.

With both class D and E, the device functions as a switch and is sensitive to input signal amplitude only as a parasitic effect. The amplitude of the output signal depends primarily on the supply voltage rather than the magnitude of the input signal. Decreasing the amplitude of the input signal will increase switch resistance, but by the time the output amplitude decreases significantly, the device no longer behaves as a switch, with the amplifier behaving more like a class B or class C amplifier.

Thus, switch-mode amplifiers are essentially constant-envelope amplifiers, unable to modulate their output signal via the RF input.

2.6 Overview of Linearization Schemes

As has been seen, the best PA power efficiency is generally achieved by PA classes with the worst linearity. Nonconstant-envelope modulation schemes require good linearity to avoid both in-channel distortion and spectral regrowth, so PAs for these applications have typically been class AB rather than more power-efficient alternatives. However, what matters is not the linearity of the amplifier but rather the accuracy of the modulated signal coming out, so while a traditional transmitter depends on PA linearity to accurately reproduce the signal coming out of a modulator, architecture, other approaches are possible. Several linearizing architectures exist to produce nonconstant-envelope modulated signals despite PA nonlinearities, thus allowing more the use of power-efficient PAs. This section will briefly review the main categories, but more extensive descriptions can be found in Kenington [9] as well as other sources.

2.6.1 Polar approaches: Envelope-Elimination and Restoration (EE&R)

The best PA power efficiencies are achieved with switch-mode PAs whose output amplitudes are independent of their input amplitudes, and are set

basically by the PA's supply voltage. While nonconstant-envelope modulation cannot be achieved by modulating the PA input, by varying the PA supply voltage, the output amplitude can be affected. Thus it is still possible to achieve amplitude modulation with switch-mode PAs. The output phase is still controlled by the PA's signal input. This is often referred to as polar modulation.

A simplified block diagram of a polar modulator is given in Figure 2.17

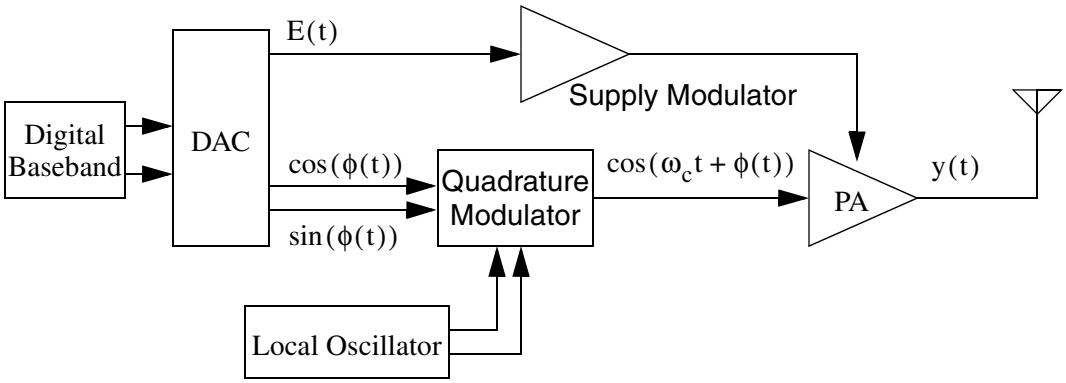


Fig. 2.17: Simplified Polar Transmitter Block Diagram

Actual implementation is more complicated than the diagram would suggest. The output amplitude is not necessarily linear with respect to the PA supply voltage. Also, the phase shift of the PA can change as the supply varies; this is a variation of AM/PM (depending on AM of the output rather than the input). Thus, other linearization techniques such as predistortion or feedback are often used in conjunction with polar modulation.

While this approach allows the use of an efficient switch-mode PA that otherwise could not provide amplitude modulation, the burden of modulating the envelope is now pushed to the supply modulator, and matters of efficiency get passed on as well. With a linear series regulator to reduce the supply voltage, the combination of the supply regulator and the PA will draw supply current roughly proportional to the output amplitude, thus giving a power efficiency that varies with amplitude like a class B amplifier. Switch-mode supply regulation is often proposed, but the bandwidth requirements of the envelope signal make this challenging.

Also, the separation of phase modulation from amplitude modulation introduces issues of matching and synchronization. As the envelope and the envelope-eliminated phase signals are both nonlinear functions of I and Q, they also occupy wider bandwidths, making the design of the baseband circuitry more challenging. DC offsets in the amplitude path introduce leakage of the envelope-eliminated signal into the output, leaking spectral regrowth. Delay through the supply modulator also needs to be considered so that the restored envelope is still synchronized with the output phase modulation.

Sometimes, rather than synthesizing the envelope and phase signals directly, they are extracted from an already modulated signal. The phase contains no amplitude modulation, it is said to have had its envelope eliminated.

Supply modulation restores the envelope. Thus this approach is sometimes known as *envelope elimination & restoration* (EE&R).

Despite these challenges, several integrated polar-modulated transmitters have been published in recent years [17][18].

2.6.2 LINC: Linear Amplification using Non-Linear Components

An approach proposed by Cox[19] for sidestepping the difficulties in modulating the output envelope of a nonlinear amplifier revolves around discarding the assumption of only one signal path. The key concept behind this system is to look at the result of combining two signals of equal amplitudes: two equal-amplitude sinusoids will sum to a create a sinusoid of twice the amplitude when they are 0 degrees out of phase, or can cancel to zero amplitude if 180 degrees out of phase, with a continuum of amplitudes in between. Thus, a general non constant-envelope signal can in principle be generated by modulating the phase of two constituent constant-envelope signals, and taking their sum.

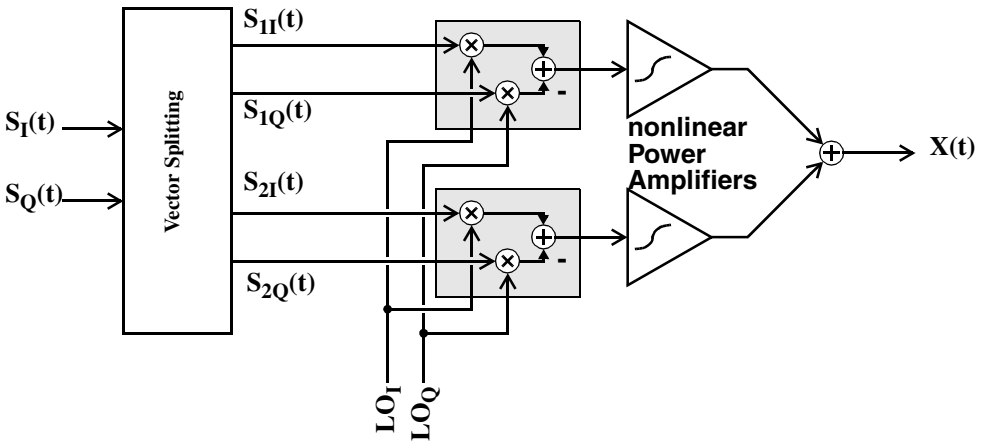


Fig. 2.18: LINC modulator

The decomposition of the nonconstant envelope signal into two constant envelope signals is relatively straightforward. A complex envelope of $\tilde{x}(t) \equiv E(t)e^{\phi(t)}$, can be expressed as the sum of two complex envelope components $\tilde{S_1}(t)$ and $\tilde{S_2}(t)$ as follows:

$$\tilde{S_1}(t) = E_{max}e^{(\phi(t)+\alpha(t))} \qquad \alpha(t) = \text{acos}\left[\frac{E(t)}{2E_{max}}\right] \qquad \text{(Eq 2-47)}$$
$$\tilde{S_2}(t) = E_{max}e^{(\phi(t)-\alpha(t))}$$

These signals are somewhat difficult to generate with analog techniques, but are a simple matter when signals are processed in the digital domain. A direct implementation has been reported by Hetzel et. al [20], achieving good linearity at the output.

There are problems with this scheme however. First, there is a singularity (more properly called a branch point) at $\tilde{x}(t) = 0$, where the phase

terms are undefined. Consider the simple case of $\tilde{x}(t)$ taking a trajectory directly from +1 to -1 in some time period. As it passes through the origin, its phase goes through a step transition of 180°, as will $\tilde{S_1}(t)$ and $\tilde{S_2}(t)$. In this particular case, the discontinuity could be remedied by exchanging the definitions for $\tilde{S_1}(t)$ and $\tilde{S_2}(t)$ when passing through the origin, however, it can be seen that passing near but not through the origin still causes sudden 180° shifts in $\tilde{S_1}(t)$ and $\tilde{S_2}(t)$ that can be problematic to generate.

Also, these signals, being highly nonlinear functions of the ideal complex envelope, are spread over a bandwidth much wider than the actual channel bandwidth. This increases complexity of the baseband circuitry, and the bandwidths required a no longer self-evident. Sundstrom [21] investigates this further.

Baseband issues aside, it would appear at first blush that mismatches between the two signal paths of gain and delay may be significant, and the effects of these impairments are analyzed by Casadevall[22]. However, there hides a much more serious problem with this architecture, in that there is no lossless way to take a linear sum of the two general signals. Hetzel et. al [20] use a terminated hybrid coupler to perform the summing function, and being a fully impedance matched network, the power amplifiers are presented with a constant, impedance matched load. However, with a constant load, the power

provided by the power amplifier, and thus, power consumed, is also constant, regardless of the final power delivered to the output (the balance of the power being sent to the dummy termination load on the hybrid's fourth port). Thus, using a hybrid coupler for the summation, the overall power consumption of the entire system will be independent of the output signal: such an LINC system becomes a complicated equivalent to a class A amplifier!

Subsequent papers [23] have looked into using feedback to accommodate some of the matching problems, however no true solution to the problem of power combination have been demonstrated.

2.6.3 Feedforward

Feedforward is a linearisation technique that starts with a more linear PA than used in LINC or EE&R. Rather than trying to achieve a perfect signal at the PA output, the PA is allowed to introduce some degree of distortion. This distortion is sensed, and amplified by a secondary error amplifier, and the outputs of the two amplifiers are combined to let the error amplifier cancel the

main amplifier's output distortion. Figure 2.19 shows a basic feedforward correction loop.

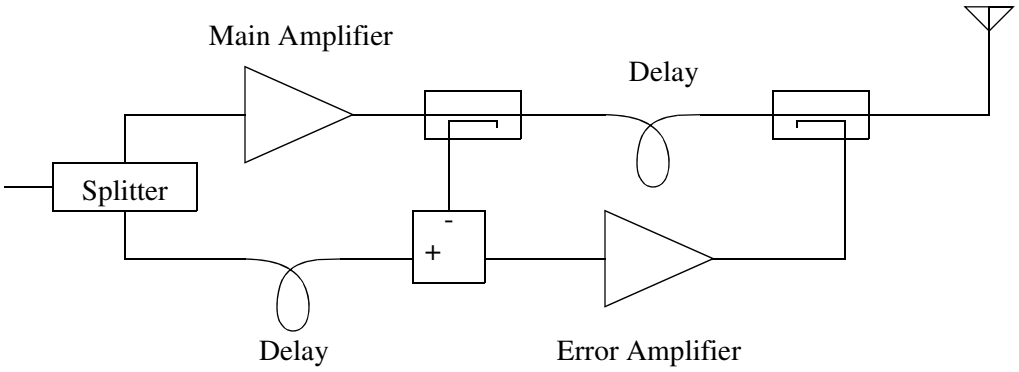


Fig. 2.19: Simplified Feedforward Block Diagram

Delay elements are included to correct for group delay through the main and error amplifiers. How well the linearisation suppresses distortion depends on the matching of the delays to the amplifiers, and matching of the main signal path to the error path. The signal combination at the final output is also somewhat problematic, introducing losses which impact power efficiency.

Rather than breaking the usual trade-off between linearity and power efficiency, feedforward functions more to extend the normal tradeoff, improving linearity beyond what is normally achievable with a single amplifier, at the expense of even worse power efficiency incurred from having to do power combining. It is commonly used in multicarrier applications such as CATV or cellular base stations, but is less useful for mobile applications with more modest linearity requirements that can be met by conventional means.

2.6.4 Predistortion

While feedforward allows correcting for distortion from the PA after it occurs, as long as the PA's AM/AM curve covers the necessary range of output amplitudes, in principle it is possible for the PA to generate the intended output directly. If the PA's AM/AM and AM/PM curves are known, then the inverse of the AM/AM curve and the AM/PM curve can be applied to the intended modulation to synthesize a *predistorted* signal which yields the intended modulation after being distorted by the PA.

The effectiveness with which this can be done depends on how well the PA's nonlinearity can be modelled. For relatively weak nonlinearities that have gain compression from third-order distortion, Yamauchi et. al. [24] describe a simple series-diode circuit which introduces gain expansion and AM/PM that can be tuned to cancel these effects of the PA. Many other circuit methods for creating a gain expansion are reviewed by Kenington [9] which devotes an entire chapter to predistortion.

For more severe PA nonlinearity, a more complicated distortion model is needed than can be realized by a simple circuit. One approach that is often suggested is to perform the predistortion in DSP. With the AM/AM and AM/PM maps in digital look-up tables, it is possible in principle to perform lookups of the inverse distortion, and generate appropriate baseband I(t) and Q(t) signals to feed into an otherwise conventional modulator and PA arrangement.

The predistorted signals will occupy somewhat wider bandwidths than the ideal signals, making the baseband design somewhat more difficult. The DSP required to implement predistortion adds complexity, although as digital performance progresses, this approach becomes more tangible.

A far more important issue however, is the need for an accurate model of the PA. In some circumstances, a static AM/AM and AM/PM table may be inadequate, as the PA's nonlinearity can vary with process, temperature, frequency, loading, or other factors. It is thus important to continually evaluate the AM/AM and AM/PM curves of the PA for the environment it is operating in, so that the curves used for predistortion are accurate.

To actually evaluate the PA's large-signal transfer function, requires observing the PA output. Figure 2.20 shows an example of an adaptive

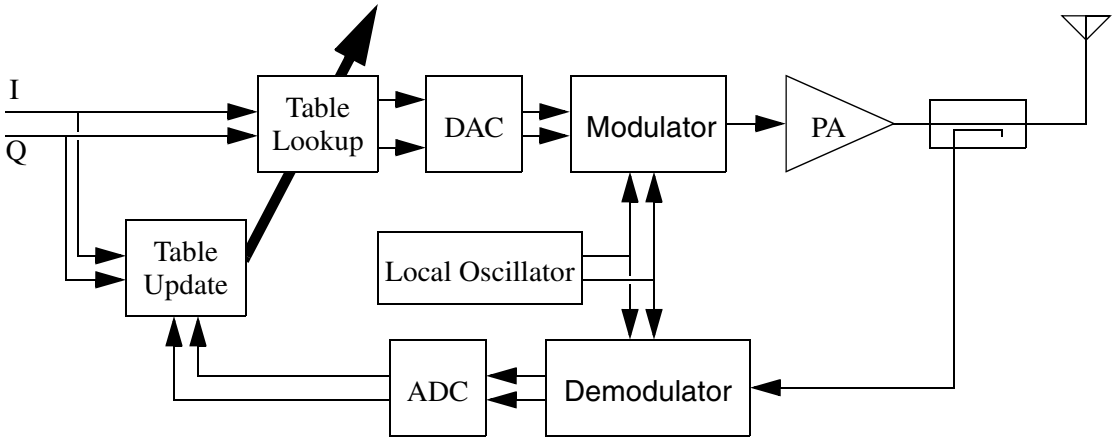


Fig. 2.20: Digital Adaptive Predistortion Loop

predistortion system. The forward path is a conventional transmitter with the

predistortion lookup done in front of the input. Extra hardware is needed for a receive path and adaptive table updates. The effectiveness of the linearization is also subject to the precision of the predistortion table lookup and the table adaption algorithm, subjects which will not be discussed here.

Complexity aside, one potential shortcoming of adaptive predistortion is latency of the adaptation. If the PA characteristics change suddenly - in particular, in the transition from standby to operating at power - the predistortion table may not match the PA, and distortion would be produced until the adaptation does its work.

Adaptation algorithms for predistortion are examined by Cavers [25]. Faulkner and Johansson [26] as well as Mansell and Bateman [27] present some experimental results for predistortion prototypes.

2.6.5 Cartesian Feedback

While adaptive predistortion takes some time to respond to changes of the PA nonlinearity, the actual distortion is available immediately in analog form at the output of the downconverter in Figure 2.20. Rather than converting the downconverted PA output to digital domain and making use of it there, a more immediate response can be had by working with the signal entirely in analog domain. By observing the output of the PA via the downconverter, the inputs to the upconverter can be adjusted in real-time to make the PA output

track the intended output modulation. This approach, when done with quadrature modulation, is known as *Cartesian Feedback*. A block diagram of a Cartesian Feedback transmitter is shown in Figure 2.21.

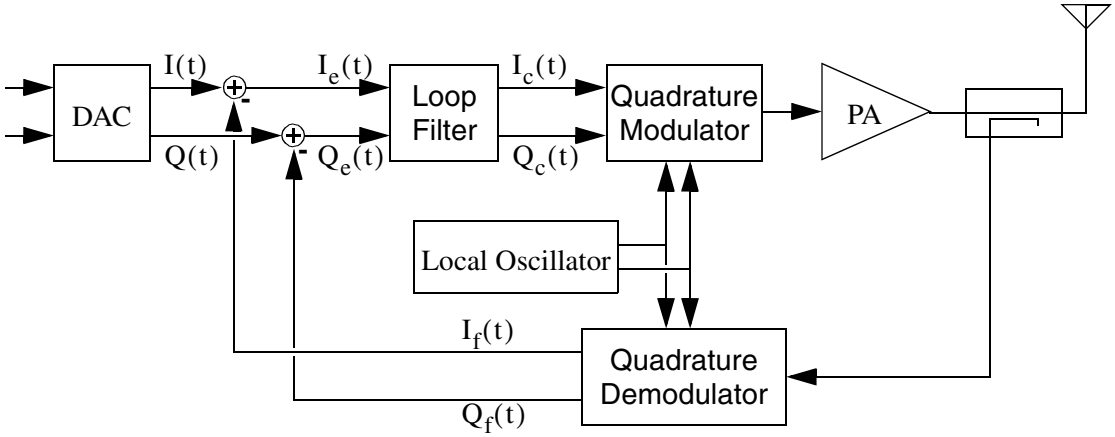


Fig. 2.21: Simplified Cartesian Feedback Loop Block Diagram

The output of the demodulator gives feedback signals $I_f(t)$ and $Q_f(t)$ representing the modulation of the PA output. These are compared against the intended $I(t)$ and $Q(t)$ signals, with the difference being error signals $I_e(t)$ and $Q_e(t)$. The loop filter takes these error signals to adjust the modulator inputs $I_c(t)$ and $Q_c(t)$ to change the PA's input, and hence output, towards what it should be.

For a reasonable PA input, the modulator inputs are reasonable in magnitude, and as long as the loop filter has a large gain, then its input signals - the error of the output modulation - will be small. As long as the loop filter

provides an adequately large gain, the linearity of the transmitter approaches that of the downconverter and is insensitive to what happens in the forward path.

The I(t) and Q(t) signals are straightforward to generate, being the same signals that would be needed for a conventional nonlinearized transmitter. Unlike polar approaches, there are no issues of synchronization - I and Q are symmetric with each other in this architecture and if implemented carefully, are inherently matched.

However, there is a catch: the feedback loop must be stable in order to work at all. Although some analyses exists in the literature [9][29][29] for stability of the cartesian feedback loop, the amplifier is assumed to be linear - hardly appropriate when trying to linearize a very nonlinear amplifier! In these analyses, distortion is treated as an independent additive input to the PA, but in reality, distortion is very much dependant on the signal fed through the amplifier. The effect of an added distortion, after going around the loop, affects the signal at the input of the PA, which can in turn cause more distortion, and so on: this is not taken into account in existing literature.

Despite this gap in the understanding of stability, cartesian feedback has found use in practice. It is common in TETRA (Trans-European Trunked RAdio) applications (with a relatively narrow channel bandwidth of around 25kHz), and has also been used in some integrated applications with relatively linear

amplifiers: a class AB amplifier in [30], and a bipolar amplifier of unspecified class in [31]. However, for use with PAs having more significant nonlinearity as might be found in an integrated CMOS transmitter, a better understanding of stability is called for. This is developed in the next chapter.

2.6.6 Hybrid Approaches

The linearization techniques of the preceding sections are not mutually exclusive. Various combinations have been proposed: Cartesian Feedback can be combined with EE&R [32], EE&R can be combined with adaptive predistortion [18], adaptive predistortion can be combined with Cartesian Feedback [33], and undoubtedly other combinations will be tried. These hybrids blend the benefits of the approaches they combine, but generally at the expense of combining complexity. While there may be applications where the benefits warrant the complexity, examples will not be pursued here.

Chapter 3

Cartesian Feedback Stability

Negative feedback is widely used in circuit design to reduce the effects of distortion in amplifiers as well as sensitivity to process, temperature and other parameters that may affect the behaviour of an amplifier. However, stability must be considered in any application of feedback or else unintended oscillations may occur.

Stability is well studied and understood for simple Single-Input/Single-Output (SISO) systems with a single feedback path. However, the Cartesian Feedback loop involves two interdependent feedback paths, with the coupling and feedback gains depending on the behaviour of the power amplifier. SISO techniques can be applied by breaking the feedback on one path, analyzing the feedback on the other, and then treating the resulting system as a simple block, around which feedback of the initially broken channel is re-applied [28][29].

Such an approach, however, does not provide a good intuition for evaluating the stability of a Cartesian Feedback loop; the behaviour of the amplifier is obscured behind too many steps of analysis. A better approach is to use Multiple-Input/Multiple-Output (MIMO) techniques, which are well established in the study of systems control, but not often presented in the context of circuit design.

This section will first review the reasoning behind the Nyquist stability criterion for SISO systems, pointing out assumptions that are usually made, and more importantly, what assumptions may not be necessary. The reader is assumed to have a basic knowledge of systems and signals, and the Nyquist criterion should be familiar, but is reviewed to support subsequent material where the MIMO equivalent is then presented, and applied to the Cartesian Feedback loop. Several applicable nonidealities are then considered with this technique.

3.1 SISO Feedback Stability

Consider the generalized feedback system of Fig. 3.1

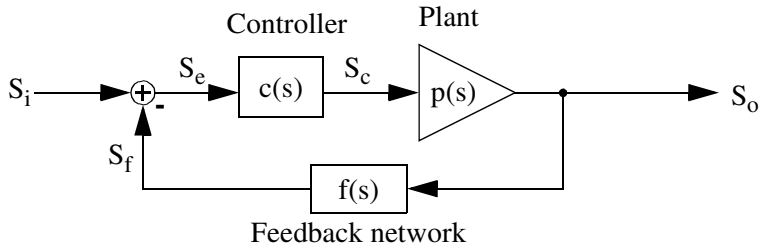


Fig. 3.1: Idealized feedback system

From the diagram, we have that

$$S_o(s) = p(s)c(s)S_e(s) \qquad \text{(Eq 3-1)}$$

$$S_f(s) = f(s)S_o(s) \qquad \text{(Eq 3-2)}$$

$$S_e(s) = S_i(s) - S_f(s) \qquad \text{(Eq 3-3)}$$

Substituting (Eq 3-2) into (Eq 3-3) and the result into (Eq 3-1) gives

$$S_o(s) = p(s)c(s)S_i(s) - p(s)c(s)f(s)S_o(s) \qquad \text{(Eq 3-4)}$$

Collecting like terms of (Eq 3-4) yields the closed loop transfer function:

$$A(s) = \frac{S_o(s)}{S_i(s)} = \frac{p(s)c(s)}{1 + p(s)c(s)f(s)} \qquad \text{(Eq 3-5)}$$

Assuming $p(s)$, $c(s)$ and $f(s)$ are rational transfer functions, we can write them as $p(s) = \frac{N_p(s)}{D_p(s)}, c(s) = \frac{N_c(s)}{D_c(s)}$ and $f(s) = \frac{N_f(s)}{D_f(s)}$, with each pair of

numerator/denominator polynomials irreducible (no common roots). Then, the loop transfer function can be written as:

$$A(s) = \frac{\frac{N_p(s)N_c(s)}{D_p(s)D_c(s)}}{1 + \frac{N_p(s)N_c(s)N_f(s)}{D_p(s)D_c(s)D_f(s)}} = \frac{N_p(s)N_c(s)D_f(s)}{D_p(s)D_c(s)D_f(s) + N_p(s)N_c(s)N_f(s)} \qquad \text{(Eq 3-6)}$$

Assume that there is no cancellation between poles and zeros of p(s), c(s) and f(s). Then the numerator and denominator of this representation of A(s) will also be irreducible.

For convenience, denote:

$$L(s) = p(s)c(s)f(s) \qquad \text{(Eq 3-7)}$$

$$\Phi_{OL}(s) = D_p(s)D_c(s)D_f(s) \qquad \text{(Eq 3-8)}$$

$$\Phi_{CL}(s) = D_p(s)D_c(s)D_f(s) + N_p(s)N_c(s)N_f(s) \qquad \text{(Eq 3-9)}$$

T(s) is referred to as the loop gain, or the open-loop transfer function (note that this is the transfer function from S_e to S_f, not to S_o).

$\Phi_{OL}(s)$ and $\Phi_{CL}(s)$ are characteristic polynomials of the open-loop and closed-loop transfer systems respectively. The roots of $\Phi_{OL}(s)$ are identical to poles of T(s), and roots of $\Phi_{CL}(s)$ are identical to poles of A(s).

Note that $1 + L(s) = \frac{\Phi_{CL}(s)}{\Phi_{OL}(s)}$ captures both characteristic polynomials with very little manipulation of L(s). The task of identifying whether the closed

feedback loop is stable can thus be reduced to looking for right-half-plane zeros of $1 + L(s)$.

Solving for zeros of $1 + L(s)$ would require manipulation of $L(s)$ to extract $\Phi_{CL}(s)$. However, the principle of the argument, as will be presented in the following section, allows the counting of poles and zeros in a region using only the value of the function along the boundary of the region, thus avoiding any need to extract $\Phi_{CL}(s)$.

3.1.1 Principle of the Argument

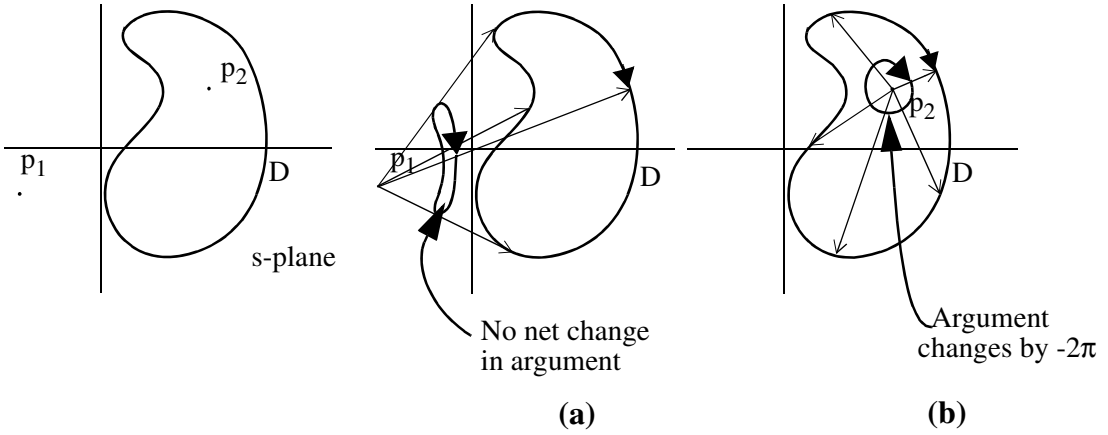


Fig. 3.2: Variation of $\angle(s - p)$ for s traversing D clockwise
(a) with p not enclosed by D, (b) p enclosed by D

Let D be a closed curve, and p be a point in the s plane, not directly on D. Consider the complex argument of $(s - p)$ for s traversing D clockwise. If p lies outside of D, then the argument of $(s - p)$ will vary as s changes, but will end

up at the same value it began at after a complete traversal of D. On the other hand, if D does enclose p, the argument will undergo a net change of -2π during this traversal. This change in the argument is equivalent to saying that the locus of $(s-p)$ encircles the origin once, clockwise.

This principle can be applied to a general rational transfer function, with zeros z_1, z_2... and poles p_1, p_2....

$$H(s) = \frac{(s-z_1)(s-z_2)\ldots}{(s-p_1)(s-p_2)\ldots} \qquad \text{(Eq 3-10)}$$

The argument of $H(s)$ can be expressed simply in terms of its factors:

$$\angle H(s) = \angle(s-z_1) + \angle(s-z_2) + \ldots - (\angle(s-p_1) + \angle(s-p_2) + \ldots) \qquad \text{(Eq 3-11)}$$

Assuming that none of the poles or zeros lie on D, it is apparent from this that as s traverses D clockwise, the net change of $\angle H(s)$ is the sum of the changes in the arguments for each of the (s-z) and the (s-p) binomials, each of which will be either 0 or -2π. The net change will be $2\pi(n_p - n_z)$ where n_p and n_z are the number of poles and zeros of $H(s)$ respectively which are enclosed by D. This is equivalent to saying that the locus of $H(s)$ encircles the origin $(n_p - n_z)$ times counter-clockwise.

3.1.2 The Nyquist Criterion

The previous section offers a means of counting the number of poles less the number of zeros enclosed by a general closed curve D. For evaluating stability, we are interested in poles and zeros of $1 + L(s)$ in the right half plane, thus it is useful to look at a curve enclosing the right half plane. For this, we can use a curve that follows the imaginary axis from $-i\infty$ to $i\infty$, and then follows a semicircle of infinite radius in the right half plane to close the curve. Denote this curve (known as the Nyquist contour) as D_+, and n_{p+} and n_{z+} as the number of poles and zeros respectively of $1 + L(s)$ enclosed by D_+. As has already been seen, these poles and zeros are the same as the open-loop and closed-loop poles of the system respectively.

Assuming that p(s), c(s) and f(s) have no poles on the imaginary axis, the following statements are equivalent:

1) The closed loop system is stable

$\Leftrightarrow$ $\Phi_{CL}(s)$ has no poles in the right half plane

2) $1 + L(s)$ has no zeros in the right half plane

$\Leftrightarrow$ $1 + L(s)$ has no zeros enclosed by D_+ (n_{z+}=0)

3) The locus of $1 + L(s)$ for s traversing D_+ clockwise will encircle the origin $n_{p+} - n_{z+} = n_{p+}$ times counter-clockwise. [By the Principle of the Argument]

4) The locus of $L(s)$ (known as the Nyquist plot of $L(s)$) will encircle the point -1 n_{p+} times counter-clockwise.

The equivalence of 4) to 1) offers a simple procedure for evaluating whether the closed feedback loop will be stable:

i) Plot the locus of $L(s)$ for s traversing D_+ clockwise (the Nyquist plot of $L(s)$)

ii) count the encirclements of -1

iii) compare with the number unstable open-loop poles n_{p+}.

This is of course, dependant on being able to determine the number of unstable open-loop poles. However, this is simpler than trying to directly count closed-loop poles, since with the loop opened, poles of each block can be evaluated individually without having to consider any interactions between blocks. Very often, all blocks are designed to be open-loop stable, in which case $n_{p+} = 0$ and *stability is indicated by the absence of any encirclements.*

One simplification to this procedure can be made by observing that $T(s)$ will typically take on the same value for s at $-j\infty$ as for $j\infty$ as well as for any s on the semi-circular portion of D_+ connecting the two. Thus, when plotting the

Nyquist plot, one only needs to consider s following the imaginary axis from $-j\infty$ to $j\infty$.

Although the procedure now appears to make no distinction between enclosing the right half plane clockwise and enclosing the left half plane counterclockwise, the two are actually equivalent. This can be seen by noting that $1 + L(s)$ has an equal number of zeros and poles, so $(n_p - n_z)$ for the two half planes must add to zero. Thus $(n_p - n_z)$ for the left half plane is $-(n_p - n_z)$ for the right half plane. This sign difference is accounted for by noting that the left half plane is encircled counter-clockwise instead of clockwise.

Some of the assumptions made so far will now be addressed. First, it was assumed that there is no cancellation between poles and zeros of p(s), c(s) and f(s). If cancellation does occur, the expressions given for the open and closed-loop characteristic polynomials will no longer be correct. (Eq 3-6) is no longer irreducible, and the system may have an unstable pole that is masked by a zero when looking at A(s).

In a real system, such exact cancellation is unlikely - any slight perturbation of parameters will move the pole and zero apart from each other, thus it is reasonable to assume they never coincide in the first place. If there is cancellation, $\Phi_{OL}(s)$, is no longer $D_p(s)D_c(s)D_f(s)$, with the cancelled poles/zeros

absent from $\Phi_{OL}(s)$. Cancellation can be accommodated by defining n_{p+} in terms of $D_p(s)D_c(s)D_f(s)$ rather than $\Phi_{OL}(s)$.

Next, it was assumed that the open-loop transfer function has no poles lying on the imaginary axis. This assumption was made so that D_+ would not pass through any critical points - a situation which has not been considered. If this were to happen, the argument of $L(s)$ would have a discontinuity when s passes through the pole, and the notion of counting encirclements would no longer be valid. This can be addressed by modifying the definition of D_+, indenting it with infinitesimal semi-circles as necessary to pass around these poles. D_+ is usually presented with these indentations protruding into the right half plane - a seemingly arbitrary decision of which way to pass - but indentations to the left can also be used: additional poles get counted into n_{p+} but the locus of $L(s)$ will also encircle -1 that many more times.

Lastly, it was assumed at the outset that all the individual block transfer functions are rational. Note that the procedure of plotting the Nyquist plot of $L(s)$ and counting encirclements is in no way dependent on $L(s)$ being rational. As long as the number of unstable open-loop poles can still be counted - trivial for a system which is open-loop stable - the procedure still be applied.

This can be argued by considering that blocks with a non-rational transfer function (say, a pure delay) can be modelled by rational functions of

arbitrarily high degree. The Nyquist criterion will tell whether these rational models are stable or not. This indeed reflects whether or not the non-rational system is itself stable: the criterion still holds even with nonrational transfer functions, as shown by Desoer[34].

3.2 Multivariate Nyquist Criterion

As has been seen, the SISO Nyquist Criterion revolves around the fact that $1 + L(s)$ has poles and zeros located at the open-loop and closed-loop poles of the system. The principle of the argument gives the number of poles less the number of zeros of $1 + L(s)$ in the right-half-plane - this being the difference between the number of unstable open-loop and closed-loop poles. Comparing this count with the number of open-loop poles in the right-half-plane thus tells whether or not the closed-loop system is stable.

It would be useful if the Nyquist Criterion could be applied to a multivariate system, however, one immediately faces a difficulty in trying to apply it literally to a multivariate system. The feedback loop having multiple inputs and multiple outputs means that $L(s)$ is no longer a scalar - it must be a matrix. However, the Nyquist plot and the notion of encirclements are intrinsically complex-scalar concepts - one can no longer plot the value of a matrix $L(s)$ much less see if it encircles -1.

Even worse, the loop transfer function $L(s)$ is no longer uniquely defined in the MIMO case. Different variations the derivation of (Eq 3-5) can also yield the transfer function in terms of $c(s)f(s)p(s)$ or $f(s)p(s)c(s)$ - all of which are identical for scalar transfer functions, but potentially quite different for matrices, possibly even being of different dimensions.

However, it can be shown that under certain circumstances, the determinant of $I + L(s)$ can serve the same role in a MIMO system that $1 + L(s)$ itself did in a SISO system. This determinant is the same regardless of which permutation of $L(s)$ is used, and has poles and zeros located where the poles of the open-loop and closed-loop system are. Justification for the use of $\det[I + L(s)]$ is presented in [35] but will not be covered here, however the following section will apply it to a simple Cartesian Feedback loop. The MIMO stability criterion will then be manipulated to a form that directly reflects the established SISO Nyquist criterion.

3.2.1 Cartesian Feedback

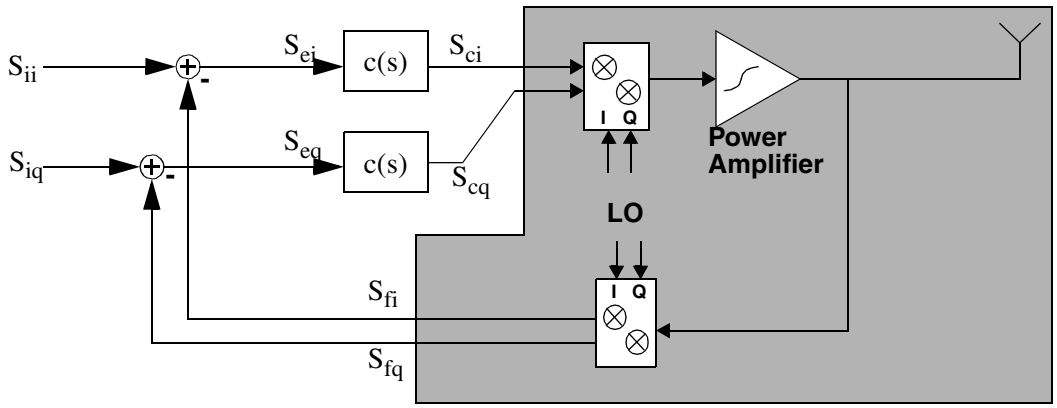


Fig. 3.3: Cartesian Feedback Loop

Consider the simplified block diagram of a Cartesian Feedback transmitter shown in Fig. 3.3. Assume that the loop filter (controller), upconversion mixers, power amplifier, downconversion mixers, and image filter, are all open-loop stable. Everything from the input of the upconversion mixers through to the output of the image filters can be considered as a two-input/two-output block, with a certain small-signal linearized model. The details of this model will be covered in more detail later, but for the time being, this can all be treated as a black box with an input-output matrix gain of A(s). The block diagram can then be simplified to that shown in Fig. 3.4

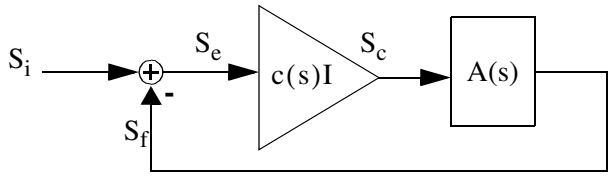


Fig. 3.4: Simplified Vector Feedback Model

Note that S_i, S_e, S_c and S_f are now vector signals, and the gains of A(s) and c(s)I are matrices. The loop transfer function from S_e to S_f is simply $L(s) = A(s) \cdot c(s)I = c(s)A(s)$.

Now, to apply the Nyquist criterion, we need to look for the locus of $\det[I + L(s)]$. But first, some manipulation. Denote $u = \frac{1}{c(s)}$. Then,

$$\det[I + L(s)] = \det\left[I + \frac{A(s)}{u}\right] = \frac{1}{u^2}\det[uI - (-A(s))] \qquad \text{(Eq 3-12)}$$

Note that this determinant is the recipe for finding the eigenvalues of $-A(s)$! If we denote the eigenvalues of A as $\lambda_1(s)$ and $\lambda_2(s)$, then this determinant is a polynomial in u with roots at $-\lambda_1(s)$ and $-\lambda_2(s)$.

$$\det[uI - (-A)] = (u - (-\lambda_1(s)))(u - (-\lambda_2(s))) = (u + \lambda_1(s))(u + \lambda_2(s)) \qquad \text{(Eq 3-13)}$$

Substituting this back in gives:

$$\det[I + L(s)] = \frac{1}{u^2}(u + \lambda_1(s))(u + \lambda_2(s)) = \left(1 + \frac{\lambda_1(s)}{u}\right)\left(1 + \frac{\lambda_2(s)}{u}\right) \qquad \text{(Eq 3-14)}$$

Now, substituting u back in, we have:

$$\det[I + L(s)] = (1 + c(s)\lambda_1(s))(1 + c(s)\lambda_2(s)) \qquad \text{(Eq 3-15)}$$

Now, returning to the principle of the argument, the system is stable if the argument of $(1 + c(s)\lambda_1(s))(1 + c(s)\lambda_2(s))$ does not change after one traversal

of D_+. A sufficient condition for this to occur is if the arguments of $(1 + c(s)\lambda_1(s))$ and $(1 + c(s)\lambda_2(s))$ individually do not change.

At this point, the Nyquist criteria says that if two SISO feedback systems, with open-loop transfer functions of $c(s)\lambda_1(s)$ and $c(s)\lambda_2(s)$ are both stable, then the MIMO system is stable. The amplifier can be thought of as having two different gains - $\lambda_1(s)$ and $\lambda_2(s)$. As long as the controller $c(s)$ is stable in a loop with each of these gains, then the Cartesian feedback loop is stable.

This perspective can be seen more directly by decomposing A(s). If A(s) has right-eigenvectors of $V_1(s)$ and $V_2(s)$ corresponding to its eigenvalues $\lambda_1(s)$ and $\lambda_2(s)$, then A(s) can be diagonalized as:

$$A(s) = V(s)\begin{bmatrix}\lambda_1(s) & 0\\ 0 & \lambda_2(s)\end{bmatrix}V(s)^{-1} \qquad \text{(Eq 3-16)}$$

where $V(s) = \begin{bmatrix}V_1(s) & V_2(s)\end{bmatrix}$. This decomposition can be put in place of the A(s) block of Fig. 3.3. The decomposition of A is effectively a coordinate transform for S_c and S_f, transforming from I and Q channels to what will be denoted here as channels 1 and 2. This coordinate transformation can also be

applied to S_i and S_e. Fig. 3.5. shows the system re-drawn with these transformations applied, and various blocks regrouped into new black boxes.

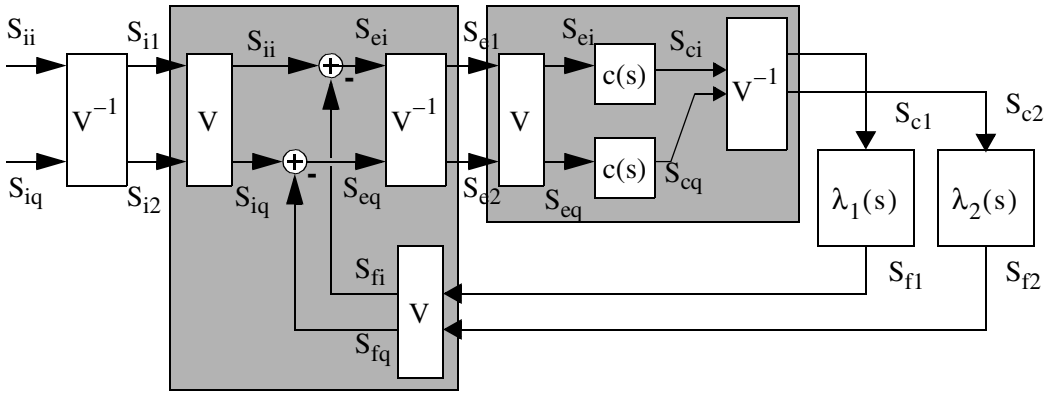


Fig. 3.5: Cartesian Feedback model with Coordinate Transforms

Note that the black boxes as drawn have coordinate transforms at their periphery, and identical, linear internal elements for the I and Q channels. The summation and c(s) are invariant under the coordinate transform, and thus the transformations are superfluous - the black boxes are identical to performing the

summation and $c(s)$ in the channel 1/channel 2 domain. Thus, the block diagram can be condensed significantly as shown in Fig. 3.6:

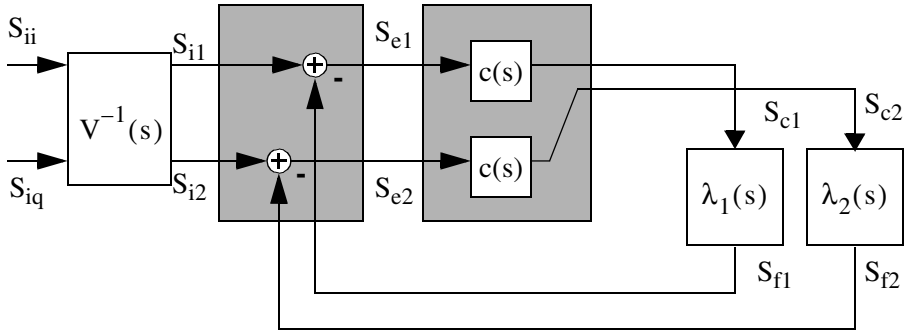


Fig. 3.6: Simplified Coordinate Transformed Feedback Loop

The system is now two independent SISO loops. Note that there may be situations where $\lambda_1(s)$ and $\lambda_2(s)$ are not necessarily conjugate symmetric (that is, $\lambda(-s) \neq \lambda^*(s)$) in which case the signals in this system may be complex rather than real. However, the SISO Nyquist criterion as has been presented, did not assume that signals are real, or that transfer functions are conjugate symmetric; each of these loops in this equivalent model, can be evaluated with the SISO Nyquist criterion.

Applying the criterion to these two loops literally, would require making two Nyquist plots - one for each of $\lambda_1(s)c(s)$ and $\lambda_2(s)c(s)$. However, if $\lambda_1(s)$ and $\lambda_2(s)$ are constant with respect to s, another modification can be made to the criterion. Note that the following are equivalent:

1) The closed loop system is stable

2) The locus of $1 + \lambda c(s)$ for s traversing D_+ does not encircle the origin (from before)

3) The locus of $\frac{1}{\lambda} + c(s)$ does not encircle the origin

4) The locus of $c(s)$ does not encircle the point $-\frac{1}{\lambda}$

Thus, only one locus - that of $c(s)$ - needs to be plotted, and as long as $-\frac{1}{\lambda_1}$ and $-\frac{1}{\lambda_2}$ lie outside of it, the Cartesian feedback loop will be stable.

For cases where $\lambda_1(s)$ and $\lambda_2(s)$ are frequency dependent, a sufficient (though not necessary) condition for stability is if the loci of $-(\lambda_1(s))^{-1}$ and $-(\lambda_2(s))^{-1}$ for s along D_+ such that $|\lambda c(s) \geq 1|$ (that is, for frequencies less than the loop's unity-gain bandwidth) do not encircle the origin and lie outside of the locus of $c(s)$.

Thus, the procedure for evaluating stability of the loop for a given operating point is:

i) Plot the locus of the loop filter transfer function $c(j\omega)$ for ω from $-\infty$ to ∞ (the Nyquist plot of $c(s)$)

ii) Determine eigenvalues of the upconverter/PA/downconverter and plot $-\lambda_1^{-1}$ and $-\lambda_2^{-1}$ or if frequency dependant, their loci for frequencies up to the loop's unity gain bandwidth

iii) Observe whether the Nyquist plot of $c(s)$ encircles the eigenvalue inverses (or loci)

If the loop filter and other blocks are open-loop stable, and there are no encirclements, then the feedback loop is stable.

3.3 Eigenvalue Examples

As was seen in the previous section, the combination of the upconversion mixers, power amplifier, and downconversion mixers in a Cartesian-feedback loop, can be thought of as having two possibly different transfer functions. These transfer functions set restrictions on the controller transfer function, thus it is important to develop some intuition for what they might actually be.

Ideally, these transfer functions would both be identical, frequency independent constant gains, with the gain simply being the combined gain of the power amplifier and both sets of mixers. However, nonlinearity in the power

amplifier, as well as memory effects in the RF path, will cause deviations from this ideal. Nonidealities within the mixers will also have some impact, however it is reasonable to assume that the magnitude of nonlinearities in the power amplifier - the very issue this architecture is intended to address - will dominate over these.

The following sections will look at various nonidealities, and the resulting effects on the system eigenvalues.

3.3.1 Mixer mismatch

With an ideal linear memoryless amplifier, the effect of mixer mismatch on the Jacobian is illustrated in Figure 2.6. For a simple gain mismatch, the two eigenvalues are just the I and Q channel gains, with eigenvectors pointing along the I and Q axis. These correspond to the major and minor axes of the ellipse in Fig. 2.6b. For a phase mismatch, the eigenvectors are no longer the I and Q axis, but are again the major and minor axes of the ellipse shown in Fig. 2.6c, with eigenvalues corresponding to the major and minor diameter of the ellipse. These effects are the same in form whether the mismatch is in the upconversion or the downconversion mixers.

Mixers for the I and Q channel can be matched very well with careful design, thus the eigenvalues will not be affected significantly by mismatch. The impact on stability can be minimal compared to other effects.

3.3.2 Memoryless AM/AM, AM/PM

Given the significance of amplifier nonlinearity, its effects on system eigenvalues is of obvious interest. Consider the block diagram in Fig. 3.7.

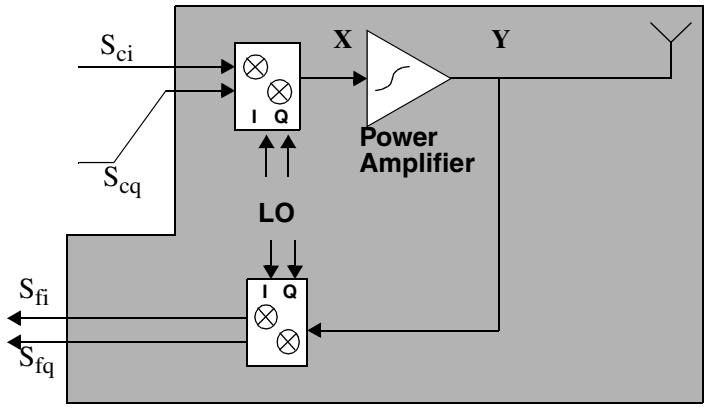


Fig. 3.7: Baseband to Baseband signal path

Assume that the mixers and the harmonic-rejection filters are ideal, with a mixer gain normalized to 1 (e.g. a 1V DC gets mixed to a 1V 0-p sinusoid, and vice-versa) and the Power Amplifier has a nonlinear large-signal envelope gain of $\overline{A}(u^2)$ where u is the magnitude of its input. Also assume that all harmonics of the Power Amplifier are filtered ideally before reaching the downconversion mixers.

Since we are assuming the system can be treated as memoryless, then we can characterize it by its DC input to output behaviour. Thus, we can assume that S_c is constant with respect to time, and look for the output given different values of this input.

Now, to trace the transfer function from S_c through to S_f. For convenience, denote $\tilde{S_c} = S_{ci} + jS_{cq}$ and note that $\left|\overline{S_c}\right|^2 = S_{ci}^2 + S_{cq}^2$.

First, we look at X, which is:

$$X(t) = Re[\tilde{S_c} e^{j\omega_c t}] \qquad \text{(Eq 3-17)}$$

The output of the power amplifier is simply this with the amplifier's gain applied:

$$Y(t) = Re[\overline{A}(\left|\tilde{S_c}\right|^2)\tilde{S_c} e^{j\omega_c t}] \qquad \text{(Eq 3-18)}$$

Now, coming back to baseband, the downconversion mixers simply extract the complex envelope of this signal:

$$S_{fi} = Re[\overline{A}(\left|\tilde{S_c}\right|^2)\tilde{S_c}] \qquad \text{(Eq 3-19)}$$

$$S_{fq} = Im[\overline{A}(\left|\tilde{S_c}\right|^2)\tilde{S_c}] \qquad \text{(Eq 3-20)}$$

Rather than characterize the amplifier's behaviour in terms of its input amplitude to large-signal-gain function $\overline{A}(\left|\tilde{S_c}\right|^2)$, it is sometimes convenient to instead look at the input envelope to output envelope function. The dependence of S_f on S_c is then captured by this directly. Denote $\tilde{F}(\tilde{S_c}) = \overline{A}(\left|\tilde{S_c}\right|^2)\tilde{S_c}$. S_{fi} and S_{fq} are thus simply the real and imaginary components of $\tilde{F}(\tilde{S_c})$. As was seen in Section 2.4.1, this is rotationally invariant, thus, one can assume that $\angle\tilde{S_c} = 0$

(equivalently, $S_{cq} = 0$) without loss of generality, since any operating point can be rotated an equivalent operating point that lies on the positive real axis.

The small signal matrix gain from S_c through to S_f is simply the Jacobian matrix of S_f with respect to S_c. Assuming an operating point with $S_{cq} = 0$, the elements of this matrix can be shown to be:

$$\frac{\partial S_f}{\partial S_c} = \begin{bmatrix} \frac{\partial \mathrm{Re}[\tilde{F}(\tilde{S_c})]}{\partial S_{ci}} & \frac{\partial \mathrm{Re}[\tilde{F}(\tilde{S_c})]}{\partial S_{cq}} \\ \frac{\partial \mathrm{Im}[\tilde{F}(\tilde{S_c})]}{\partial S_{ci}} & \frac{\partial \mathrm{Im}[\tilde{F}(\tilde{S_c})]}{\partial S_{cq}} \end{bmatrix} = \begin{bmatrix} \mathrm{Re}[2\bar{A}'(|\tilde{S_c}|^2)S_{ci} + \bar{A}(|\tilde{S_c}|^2)] & -\mathrm{Im}[\bar{A}(|\tilde{S_c}|^2)] \\ \mathrm{Im}[2\bar{A}'(|\tilde{S_c}|^2)S_{ci} + \bar{A}(|\tilde{S_c}|^2)] & \mathrm{Re}[\bar{A}(|\tilde{S_c}|^2)] \end{bmatrix} \qquad \text{(Eq 3-21)}$$

The elements of this matrix are easy to identify. The first column represents the incremental change at the output due to an incremental change of the input that is in the same direction as the input already present - a change in the input amplitude - and is just the derivative of $\tilde{F}(\tilde{S_c})$. The second column gives the output change due to a change perpendicular to the input - a change in phase, or a rotation - and is just a change perpendicular to the output - an identical rotation.

From here, extracting eigenvalues becomes straightforward. For notational convenience, note that:

$$\frac{\partial S_f}{\partial S_c} = \begin{bmatrix} \frac{\partial \mathrm{Re}[\tilde{F}(\tilde{S}_c)]}{\partial S_{ci}} & -\mathrm{Im}[\bar{A}(|\tilde{S}_c|^2)] \\ \frac{\partial \mathrm{Im}[\tilde{F}(\tilde{S}_c)]}{\partial S_{ci}} & \mathrm{Re}[\bar{A}(|\tilde{S}_c|^2)] \end{bmatrix} \quad \text{(Eq 3-22)}$$

From this, it can be shown that:

$$\lambda_1, \lambda_2 = \frac{\mathrm{Re}[\bar{A}(|\tilde{S}_c|^2)] + \mathrm{Re}\left[\frac{\partial \tilde{F}(\tilde{S}_c)}{\partial S_{ci}}\right] \pm \sqrt{\left(\mathrm{Re}[\bar{A}(|\tilde{S}_c|^2)] - \mathrm{Re}\left[\frac{\partial \tilde{F}(\tilde{S}_c)}{\partial S_{ci}}\right]\right)^2 - 4\mathrm{Im}[\bar{A}(|\tilde{S}_c|^2)]\mathrm{Im}\left[\frac{\partial \tilde{F}(\tilde{S}_c)}{\partial S_{ci}}\right]}}{2} \quad \text{(Eq 3-23)}$$

While this expression isn't very enlightening as it stands, various assumptions about $\bar{A}(|\tilde{S}_c|^2)$ allow some conclusions about the situations in which these assumptions arise.

3.3.2.1AM/AM distortion

First consider a power amplifier with AM/AM distortion, but no AM/PM distortion. For simplicity, assume that the amplifier introduces zero phase shift (this will be revisited later) from its input to output: that is, $\bar{A}(|\tilde{S}_c|^2)$ is real for any input. In this case, the two eigenvalues reduce to:

$$\lambda_1 = \bar{A}(|\tilde{S}_c|^2) = \frac{\tilde{F}(\tilde{S}_c)}{\tilde{S}_c} \quad \text{(Eq 3-24)}$$

$$\lambda_2 = \frac{\partial \tilde{F}(\tilde{S}_c)}{\partial S_{ci}} \quad \text{(Eq 3-25)}$$

These are simply the large-signal and the (in-phase) small-signal gains of the amplifier.

This presents a significant restriction on the choice of controller characteristics when using a class C or a switching PA. These amplifiers produce essentially no output when the input amplitude is small enough to not turn on their switching devices - the large-signal gain in this case is effectively zero. Increasing the input amplitude past the turn-on threshold then generates a non-zero output, and the large signal gain increases. The small-signal gain is initially larger than the large signal gain.

The controller must be able to accommodate such a large range of feedback gains - from zero up to the maximum large-signal gain encountered, plus some range of small-signal gains which may be even larger still from gain expansion. A single-pole controller is robust across such a range of gains, but if a higher-order loop filter is to be used, this range of gains may be a problem.

3.3.2.2AM/PM distortion

The presence of a weak phase shift in the amplifier's transfer function causes the second term in the discriminant of (Eq 3-23) to become smaller. This causes the two eigenvalues to shift closer to each other, but they remain real and stability isn't substantially affected relative to the AM/AM-only case.

For more severe phase shifts, several different things may happen. The first is that the discriminant of (Eq 3-23) may turn negative, causing the two eigenvalues to split into a complex-conjugate pair. As long as the real components of $\bar{A}(\tilde{S_c})$ and $\frac{\partial \bar{F}(\tilde{S_c})}{\partial S_{ci}}$ are both positive, this is not a problem for a single-pole controller.

However, with large phase shifts - in particular, those over 90 degrees - the real components of $\bar{A}(\tilde{S_c})$ and $\frac{\partial \bar{F}(\tilde{S_c})}{\partial S_{ci}}$ turn negative, at which point the loop will almost certainly turn unstable: the feedback gains have become *positive* feedback.

Thus, it is important to keep the phase shift through the amplifier under control. Ideally, the imaginary component of $\bar{A}(\tilde{S_c})$ would be kept zero - in this case, even a nonzero AM/PM introducing $\mathrm{Im}\left[\frac{\partial \bar{F}(\tilde{S_c})}{\partial S_{ci}}\right]$ of any magnitude would be tolerable as the second term in the discriminant of (Eq 3-23) would be kept at zero.

3.3.3 Frequency-dependant linear channel

The examples given so far have assumed a memoryless transmit path, however, the validity of this assumption warrants questioning. Extracting

eigenvalues for a transmit path possessing both memory and nonlinearity, is highly dependant on the particular amplifier being considered, and is not readily generalized. However, the analysis of a linear channel with memory is tractable and in itself can offer some insight into the significance of memory.

Again consider the block diagram in Fig. 3.7, with the same assumptions regarding the mixers as before. However, instead of the nonlinear power amplifier, assume a linear amplifier with a frequency domain response of $H(s)$.

Recall from Section 2.2.2, that this frequency-domain response is an eigenvalue function in and of itself. An eigenvector of e^{st} fed to the input of the amplifier at X, yields an output at Y of $H(s)e^{st}$ - the amplifier introduces a scalar gain of $H(s)$. X and Y remain real by virtue of the complex conjugate of this eigenvector (with its conjugate scalar gain) also being present.

To see the mapping from $H(s)$ to $\lambda_1(s)$ and $\lambda_2(s)$, it is simple enough to make an educated guess at what a baseband eigenvector might be, and then look for its eigenvalue. To this end, consider applying inputs of:

$$S_{ci} = e^{st} \qquad \text{(Eq 3-26)}$$

$$S_{cq} = -je^{st} \qquad \text{(Eq 3-27)}$$

Unfortunately, since these signals are complex rather than real, complex envelope notation is not practical. However, $X(t)$ is simple enough to write out explicitly:

$$\begin{aligned} X(t) &= S_{ci}(t)\cos(\omega_c t) - S_{cq}(t)\sin(\omega_c t) \\ &= e^{st}\cos(\omega_c t) + je^{st}\sin(\omega_c t) \\ &= e^{(s+j\omega_c)t} \end{aligned} \qquad \text{(Eq 3-28)}$$

Here we have stumbled onto an eigenvector for the amplifier! The associated eigenvalue is $H(s+j\omega_c)$, giving:

$$Y(t) = H(s+j\omega_c)e^{(s+j\omega_c)t} \qquad \text{(Eq 3-29)}$$

This output is no different from what would result if gains of $H(s+j\omega_c)$ were inserted at the baseband inputs, and the amplifier replaced with an ideal unity-gain amplifier. The system after these gains at this point is ideal and transparent, so the entire system effectively has a scalar gain of $H(s+j\omega_c)$.

It is easily verified that $S_{ci} = e^{st}$, $S_{cq} = je^{st}$ constitutes another eigenvector, with associated eigenvalue of $H(s-j\omega_c)$. Thus, the two eigenvalues for a particular frequency, are:

$$\lambda_1(s) = H(s+j\omega_c) \qquad \text{(Eq 3-30)}$$

$$\lambda_2(s) = H(s-j\omega_c) \qquad \text{(Eq 3-31)}$$

These are simply the RF path frequency transfer function, translated from ω_c and $-\omega_c$ to DC.

(Eq 3-30) and (Eq 3-31) show why it's often reasonable to treat the RF path in narrowband systems as being memoryless. For the range of frequencies spanned at baseband - often no more than several megahertz relative to a gigahertz carrier - the relative variation in $s \pm j\omega_c$ is small, and the variations in $H(s \pm j\omega_c)$ are insignificant relative to any frequency dependencies present at baseband, allowing these eigenvalues to be considered constant. This does assume that $H(s)$ is relatively insensitive to frequency, which may not be a valid assumption if highly selective filters (such as a frequency duplexer or other SAW filter) are present in the signal path. Thus, it is important to keep any such filters outside of a cartesian feedback loop, only using them open-loop at the output of the system.

The impact of channel memory on loop behaviour as seen at baseband, is now obvious: any gain and phase shift introduced by channel memory is seen as an identical gain and phase shift at baseband. Unlike in a real signal SISO system however, the gain and phase shift here are not necessarily symmetric - e^{st} may experience a different gain and phase shift from e^{-st} - but as has already been mentioned, the application of the Nyquist criteria does not depend on such symmetry.

The notion of phase shifts at RF mapping directly to baseband thus nests nicely with the concept of phase margin - any phase shift at RF takes away directly from the phase margin of the system as seen at baseband. This is recognized by Briffa and Faukner[29].

A simple example of this can be seen by considering a system with pure integrators for the controllers. $c(s) = \frac{1}{s}$, which has a Nyquist plot following the imaginary axis, and alone would have a phase margin of 90 degrees. With a no phase shift at RF, the I and Q channels operate independently, and both have the inherent 90 degree phase margin of the controller function. With a 90 degree phase shift at RF however, an input to the I-channel upconversion mixer comes out at the Q-channel downconversion mixer output, and vice-versa. The two channels are effectively chained in a series loop and become an undamped resonator - the phase margin has gone to zero.

Thus, as was seen in the discussion of AM/PM distortion, it is important to keep the phase shift through the RF path under control. This is discussed further in Section 3.4

3.3.4 Pure Delay

A pure delay in the RF path of τ is a transfer function of $H(s) = e^{s\tau}$. By (Eq 3-30) and (Eq 3-31), the baseband-referred eigenvalues for this are then:

$$\lambda_1(s) = H(s + j\omega_c) = e^{(s + j\omega_c)\tau} = e^{j\omega_c\tau}e^{s\tau} \quad \text{(Eq 3-32)}$$

$$\lambda_2(s) = H(s - j\omega_c) = e^{-j\omega_c\tau}e^{s\tau} \quad \text{(Eq 3-33)}$$

These are simply phase shifts of $\pm\omega_c\tau$ radians, plus the same delay of τ. The phase shift reflects the delay relative to the carrier frequency rather than the loop bandwidth, and may be significant and take away from the phase margin of the loop. This phase margin is essentially constant, however, and can be corrected for, as will be discussed in Section 3.4

The delay also impacts the phase margin, but as long as the delay is short relative to the unity-gain bandwidth of the loop, it can be ignored.

3.4 Phase alignment

As seen in Section 3.3.4, delays in the RF signal path can introduce a static phase shift. Other components in the RF signal path such as baluns or couplers, can introduce phase shifts as well: phase shifts may exist within the PA, mixers, or in the local oscillator distribution. The net effect is that the RF signal path will have some net phase shift which is difficult to predict a priori, which is a problem given that this phase shift directly impacts the system phase margin.

Although the phase shift is difficult to predict, its effects are easily understood. A static phase shift of δ is simply a transfer function of

$$\overline{A} = e^{j\delta} = \cos(\delta) + j\sin(\delta) \qquad \text{(Eq 3-34)}$$

Substituting this into (Eq 3-19) and (Eq 3-20) gives the relationship of:

$$\begin{bmatrix} S_{fi} \\ S_{fq} \end{bmatrix} = \begin{bmatrix} \cos(\delta) & -\sin(\delta) \\ \sin(\delta) & \cos(\delta) \end{bmatrix} \begin{bmatrix} S_{ci} \\ S_{cq} \end{bmatrix} \qquad \text{(Eq 3-35)}$$

This is a simple rotation.

3.4.1 Rotation Approaches

This rotation can be corrected either in the RF domain, or in baseband. In the baseband domain, correcting the rotation is a simple matter of recognizing that the rotation of (Eq 3-35) is easily inverted: applying the inverse in front of the modulator, is shown in Figure 3.8

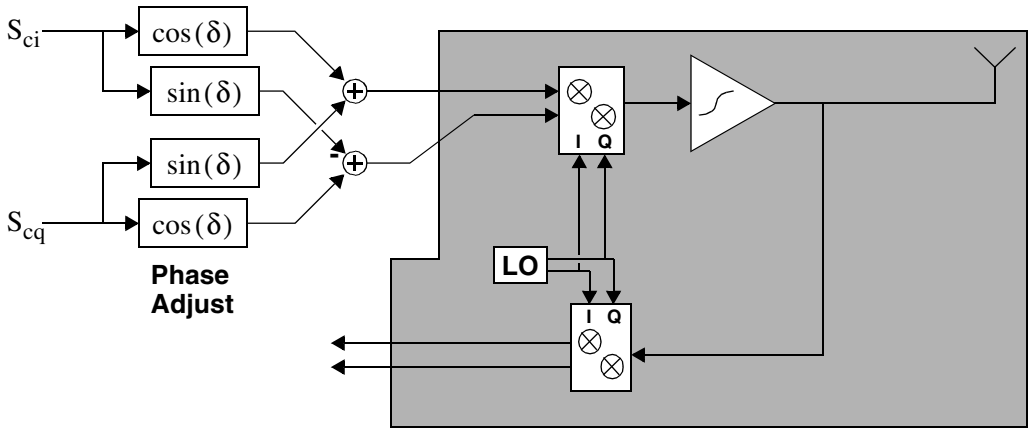


Fig. 3.8: Baseband Domain Phase Alignment

This approach depends on being able to perform reasonable baseband multiplications of S_{ci} and S_{cq} with $\cos(\delta)$ and $\sin(\delta)$. The correction could also be performed at the output of the downconverter, but working with the

upconverter signal is preferred as any low-frequency noise, multiplier linearity, or mismatch errors introduced in front of the upconverter gets suppressed by the loop feedback. Perraud et. al [30] use this approach to good effect.

In the RF domain, an RF phase shift inserted anywhere in the forward signal path is equivalent, and adding enough phase shift to make the total phase shift 2π radians (or any multiple of 2π) cancels the effect of static phase shift. A convenient place to insert this phase shift is in the local oscillator signal for the modulator: the signal fed through the phase shifter is constant in amplitude, thus avoiding any possible issues of AM/PM in the phase shifter. As with the baseband approach, shifting the upconverter LO results in noise introduced by the phase shifter close to the carrier frequency being suppressed by the feedback. Any minor I-Q phase mismatch introduced by the phase shifters would also be similarly suppressed.

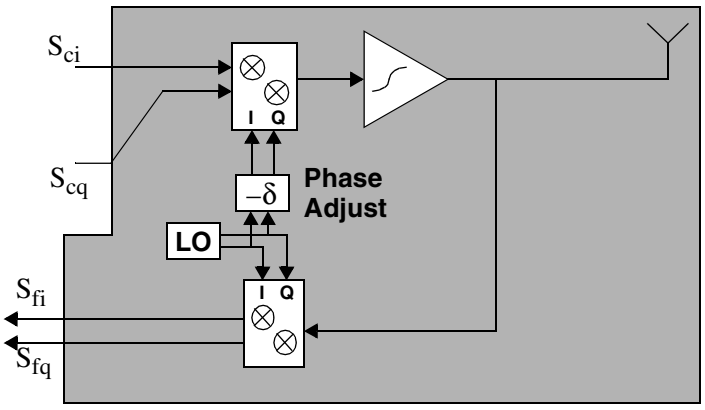


Fig. 3.9: RF Domain Phase Alignment

The need to compensate for these phase shifts was recognized very early on by Brown and Petrovic[36], and many variations have been proposed both for sensing the phase shift, and for introducing the correction.

Brown and Petrovic suggest several RF domain techniques for the phase shifting: tunable RC delays, PLL/counter approaches, and vector modulation. RC delays require several stages of delays to achieve a full 2π range of adjustability, but are subject to amplitude variation with phase adjustment, and are sensitive to parasitics. The PLL approach is practical only for very low carrier frequencies: each cycle of the carrier is subdivided into 2^n steps that are counted, and the phase shifted carrier is generated by comparing the counter value with the intended phase.

For high carrier frequencies, the vector modulation approach is the most practical, and is the approach taken by Brown and Petrovic, as well as in this thesis. The arbitrary phase shifted LO signal is itself simply a modulated signal with a static complex envelope of $e^{j\delta}$. The phase shifter can be implemented as a set of upconversion mixers, with I and Q baseband inputs of $\cos(\delta)$ and $\sin(\delta)$. This inherently has a full 2π range of output phase available.

The vector modulation approach to LO phase shifting is actually very closely related to the baseband vector rotation. The baseband rotation can be thought of as being a canonical implementation of:

$$x(t) = \mathrm{Re}[(\tilde{S_c}e^{-j\delta})e^{j\omega_c t}] \quad \text{(Eq 3-36)}$$

while the RF rotation would is a canonical implementation of:

$$x(t) = \mathrm{Re}[\tilde{S_c}(e^{-j\delta}e^{j\omega_c t})] \quad \text{(Eq 3-37)}$$

3.4.2 Phase Error Detection

Correcting for phase error, whether in the baseband or RF domain, requires being able to identify the phase error in the first place. The effective RF phase shift can be found by observing S_f and S_c and noting the angle between these vectors.

Brown and Petrovic [36] suggest an approach for detection. The transmitter is operated open-loop for calibration, and an SSB signal is used for S_f, with cosine and sine of a test tone on the I and Q channels. The resulting S_c is also an SSB signal, subject to the effective RF phase shift. A phase detector (such as used in a PLL) can be used to compare the baseband tones of S_f and S_c, and used to trim the correction angle. This approach is slow: the baseband tones are necessarily of low frequency, and the loop bandwidth of the trimming must be even slower, taking several seconds to complete acquisition. This approach also requires pauses in transmission: the loop cannot transmit arbitrary modulation while it is busy with calibration.

A better approach comes from noting that the sign of the angle - that is, whether the total phase shift, including correction, is a phase lead or a phase lag - can be found by observing the sign of the vector cross product between S_f and S_c. If S_f is held to be purely in the I direction, then this cross product reduces to observing S_{cq}. This can be accomplished operating the transmitter open-loop, feeding the upconversion mixers a fixed input instead of the loop filter output. Knowing the sign of the net phase error, a successive-approximation approach can be used to zero in on the correct angle. This approach is used in [30] during signal ramp-up before switching to closed-loop operation for a time slot. The correct phase is acquired in under a microsecond in this design.

The two preceding approaches depend on operating the transmitter open-loop. While this works for time-duplexed applications, continuous-modulation systems do not allow for these approaches once data transmission has started. To accommodate possible changes in the phase error during transmission, the vector cross product of S_f and S_c would have to be implemented for on-line detection. This is approach is used by Dawson [37] together with chopper stabilization techniques to cancel the effects of DC offsets.

3.4.3 Static vs. Dynamic Correction

Whether on-line detection is necessary depends on what phase error is being corrected for. Phase error consists of static phase error from small RF path delays, but there can also be a dynamic signal-dependant component from amplifier AM/PM. The offline detection described cannot correct for AM/PM, but as long as amplifier AM/PM is not so severe to cause instability, the cartesian feedback can correct for it. Thus, the main concern to consider online detection would be to accommodate drift.

Brown and Petrovic report having tested a transmitter for several hours without observing any significant drift. This suggests that under similarly controlled conditions, a one-time manual phase adjustment may be adequate. Thus, this thesis does not pursue implementation of phase error correction, and for use in a TDMA environment, the ramp-up time calibration used in [30] is deemed adequate.

While static correction of phase error is adequate for weak AM/PM, a fast online phase correction could allow a cartesian feedback transmitter to accommodate more severe AM/PM. Performing a fast phase correction amounts to building the phase loop of a polar-feedback modulator. Adaptive predistortion could be applied to the AM/PM correction as well. These approaches could be worthy of future work.

3.5 Local and Global Stability

The stability analysis thus far has only been of small-signal stability of the feedback loop - that is, with a constant input S_i, for small values of S_e, the system appears linear, and the feedback will cause S_e to converge towards zero as intended. However, in a general nonlinear system, small-signal stability only ensures this convergence for S_e within some neighbourhood of the origin. For a general nonlinear system, it is possible that for sufficiently large S_e, the feedback behaviour will be adequately different from that around the intended operating point, that the system never converges as intended, instead orbiting around some limit cycle.

It is important to verify that either these limit cycles do not exist, or if they do, the error S_e never becomes so large in operation to fall into these limit cycles.

Nonexistence of such limit cycles in a cartesian feedback loop can be easily verified under certain circumstances. If the power amplifier is memoryless and single-pole controllers are used, then it is possible to empirically plot a phase portrait of S_e (trajectories of S_e across all possible starting values) for any given input S_i. Any limit cycle for a given input can be readily seen on the phase portrait as a loop of some sort which trajectories for

nearby initial conditions converge towards. Plotting phase portraits for a set of S_i spanning the entire range of output amplitudes to be used, one can see if any limit cycles exist for constant S_i.

However, even in the absence of any such limit cycles, the system being small-signal stable for all constant S_i still does not in general ensure stability when S_i varies with time[1]. But as long as S_i changes slowly relative to the asymptotic small-signal settling time of the system across all S_i, then it is not unreasonable to expect everything to behave as intended. Although there are more sophisticated analytical methods for analyzing stability, in a cartesian feedback system any potential for instability in practice should surface in a behavioural simulation. Since these simulations ultimately need to be run in verifying system performance, it is reasonable to forego further analytical stability analysis once small-signal stability is established, and rely on these simulations to verify dynamic stability.

1. With certain additional assumptions, the Cartesian feedback system can be transformed into the form of the *Lur'e problem.* M.A. Aizerman conjectured that a Lur'e system with time-varying nonlinear feedback whose large-signal gain is bounded by k_1 and k_2 will be stable if the system is stable with linear feedback for all gains between k_1 and k_2. Aizerman's conjecture is known to be false [38][39].

Chapter 4

Prototype System Design and Simulations

To demonstrate the feasibility of applying linearization to allow the use of a nonlinear integrated CMOS PA, a prototype Cartesian-Feedback transmitter was designed and fabricated. The prototype was designed to meet GSM EDGE specifications, operating in the DCS1800 band. EDGE is an extension of the GSM digital cellular telecommunications standard, which by using a nonconstant-envelope, 8-PSK modulation, carries three times the data rate of the constant-envelope GMSK modulation originally specified.

This chapter describes the system-level design of the prototype. Some relevant specifications from the GSM standard are interpreted into linearity, matching, and noise requirements for various transmitter blocks. Nonlinearity of the PA is examined to find the loop gain required for linearization. The tradeoff between loop gain and bandwidth is considered, and a loop filter design is proposed, with stability verified following the analysis given in Chapter 3. A

transient simulation of the closed-loop transmitter is run to verify successful linearization.

GSM standards specify output power requirements for several different classes of handsets. GSM class 1 handsets must produce up to 30dBm (1W) of output power when transmitting a GMSK signal, and the prototype PA was designed to produce this power level. EDGE modulation has a PAR of 3.2dB, so a 500mW modulated signal can be produced within this maximum. GSM has another set of handset power class designations for 8-PSK modulation, with power class E2 handsets delivering up to 26dBm (400mW) of output power. The system analysis of this chapter assumes a power of 500mW being transmitted however.

4.1 Downconverter linearity requirements

With a large loop gain, the closed-loop linearity of the transmitter depends on the downconverter, and thus the linearity of the downconverter is central to whether the closed-loop transmitter can meet spectral mask and EVM requirements. As was seen in Section 2.4, memoryless nonlinearity can be characterized in terms of a power-series from a block's input to output. In closed-loop operation, the downconverter has an ideal signal at its output, and the sensed PA output at its input, thus the mapping of interest is actually a power series representing the downconverter input as a function of its output. This is

the inverse of the usual power series, but it can be shown that for weak nonlinearities, with appropriate normalization, the coefficients of the inverse series are equivalent to the forward series.

This power series representing the downconverter can be written as:

$$\tilde{x}(t) = G_1\tilde{y}(t) + G_3\tilde{y}(t)|\tilde{y}(t)|^2 + G_5\tilde{y}(t)|\tilde{y}(t)|^4 + \ldots \quad \text{(Eq 4-1)}$$

where $\tilde{x}(t)$ represents the sensed PA output, and $\tilde{y}(t)$ is the downconverter output which is forced by the feedback to track an ideal undistorted signal. $\tilde{y}(t)|\tilde{y}(t)|^2$ is a third-order product, weighted by coefficient G_3, and similarly, higher-order distortion products of the linear term are also present.

In practice, downconverter linearity is usually measured using a two-tone test to find intermod-intercept points, and the magnitudes of the power series coefficients is captured in these intercept points. This section examines GSM specifications to estimate limits on these coefficients for the downconverter, expressed in terms of intermod-intercept points.

4.1.1 Spectral Mask

Figure 4.1 shows the transmit mask and spectrum of a normalized, ideal undistorted EDGE modulated signal $\tilde{y}(t)$, and the spectrum of its normalized

third-order product $\tilde{y}(t)|\tilde{y}(t)|^2$. The fifth and seventh order products, $\tilde{y}(t)|\tilde{y}(t)|^4$, and $\tilde{y}(t)|\tilde{y}(t)|^6$ are shown as well. These spectra are given with a 30kHz video-filter applied as specified by the GSM standard.

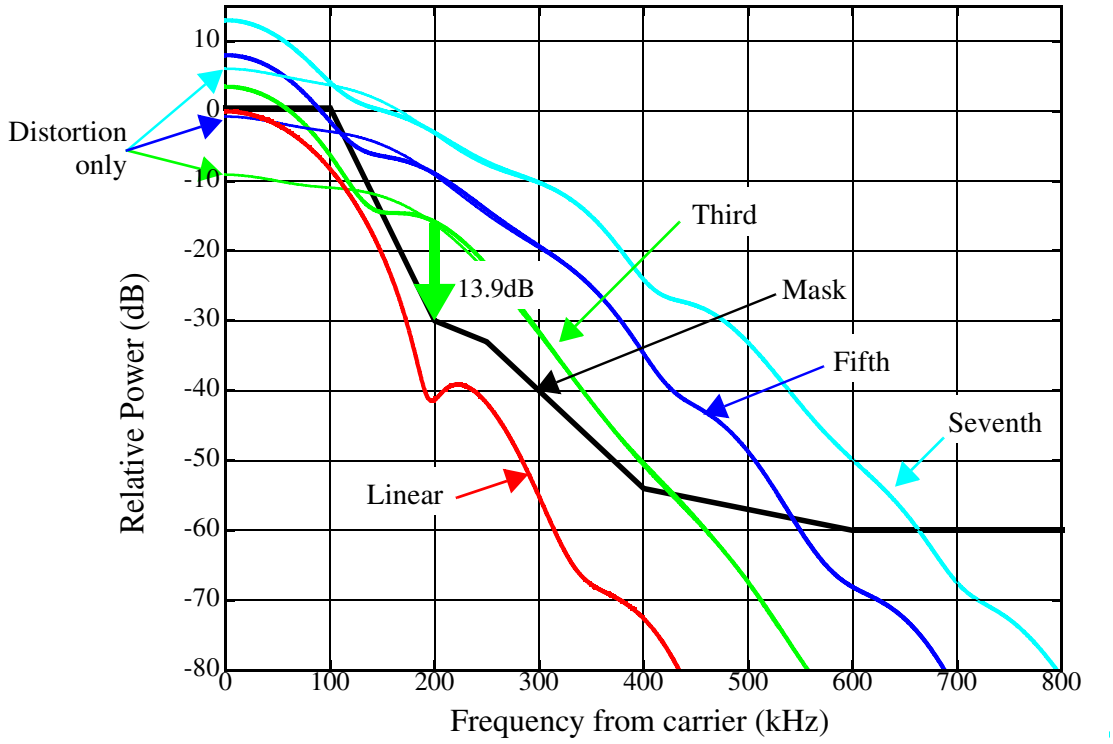


Fig. 4.1: Spectrum of GSM EDGE modulated signal $\tilde{y}$ and odd-order products

The linear signal $\tilde{y}(t)$ is normalized to have an RMS magnitude of 1. The spectrum of $\tilde{y}(t)|\tilde{y}(t)|^2$ is what would be produced if the modulated signal were presented to a system at the same power as its third-order intercept point IP_3 - this relationship is verified in Appendix A. The third-order product includes a component that is correlated with the linear component, and from the perspective of spectral mask and EVM measurements, appears as though it had

been produced by a linear gain: indeed, this component is larger than what the linear gain itself produces, being 1.9dB higher. Subtracting this component from the third-order product leaves the unwanted distortion. This distortion is shown by the thinner curves in the graph visible at frequency offsets under 200kHz.

To clear the spectral mask, this distortion must be suppressed by 13.9dB. A power back-off of 7dB would be enough to just barely clear the mask at 200kHz, but the component of the third-order product that is correlated with the linear product is still large enough to need consideration. This component was 1.9dB larger than the linear signal, and would be reduced to -12dB after this backoff. If the third-order coefficient gives a gain compression, this component would subtract from the linear component, reducing it by $-20\log\left(1-10^{\frac{-12}{20}}\right) = 2.5\text{dB}$ This means the transmit mask itself is actually this amount lower than originally expected.

Adding 1dB more power backoff would lower the third-order distortion by 2dB, and the reduced gain compression would raise the mask by 0.5dB. The total 8dB of backoff represents a minimum requirement for meeting the spectral mask - the downconversion path's input-referred IP_3 must be at least 8dB larger than the signal power.

Similarly, the fifth and seventh order products are 21.8dB and 31.9dB over the spectral mask around 350kHz respectively. These products have components correlated with the linear term that are even larger than the third-order product, but relative to these margins, are comparable or smaller than for the third-order term. IP_5 and IP_7 (defined for two-tone test products at $3\omega_1 - 2\omega_2$ and $4\omega_1 - 3\omega_2$ respectively rather than at $2\omega_1 - \omega_2$) of 6dB over signal power are enough to clear the mask.

Note that these minimum values for third, fifth and seventh order linearity assume that each distortion exists alone: if all orders are present at magnitudes near these values, their contribution to the spectrum will add and exceed the mask. However, it is expected that as is typical for receiver circuits, the third-order distortion of the downconverter will dominate, while higher-order distortion can be neglected.

The third-order product first violates the spectral mask at 200kHz, while the fifth and seventh order products, having broader spectrum lobes, first meet the mask around 350kHz. The 400kHz and 600kHz corners of the spectral mask are not an issue for these kernels as the mask would be violated at lower frequencies first. Any distortion products that appear to approach these higher frequency corners first are thus the products of higher-order distortion, likely relating to PA clipping.

4.1.2 EVM

GSM standards specify an RMS EVM of 9.0%, or -21dB. With the third-order distortion suppressed to marginally clear the spectral mask, the third-order distortion product has a spectral density of -9.1-14-2.5=-25dB below the desired signal at the carrier frequency. This margin decreases with increasing offset from carrier, but the importance to EVM also decreases with offset: the EVM calculation includes a 90kHz measurement filter which excludes both the signal and distortion products at higher offsets.

There being greater than 21dB of separation between the desired signal and distortion suggests that the EVM requirement is met by any signal whose third-order distortion clears the spectral mask; if a system clears the spectral mask with a reasonable margin, then linearity should not be an issue in EVM performance.

The fifth and seventh order products similarly meet the EVM requirement as long as the spectral mask is not violated.

4.2 Downconverter Matching Requirements

The 9% EVM requirement is marginally met if α of (Eq 2-21) is 0.09. This represents I and Q channel gains of 1.09 and 0.91, or a $20\log\left(\frac{1.09}{0.91}\right) = 1.6\text{dB}$ mismatch between channels.

Similarly, a quadrature mismatch of β in (Eq 2-22) introduces an error of $\tan(\beta)$. To marginally meet 9% EVM, this implies $\beta = \text{atan}(0.09) = 5.14°$. Any error in quadrature of 2β must therefore be less than 10.3°.

4.3 PA model

A CMOS three-stage power amplifier was designed and is described in more detail in Section 5.1. The PA has a class C output stage with a class AB

helper. AM/AM and AM/PM curves were extracted from Spectre simulations and are shown in Figure 4.2:

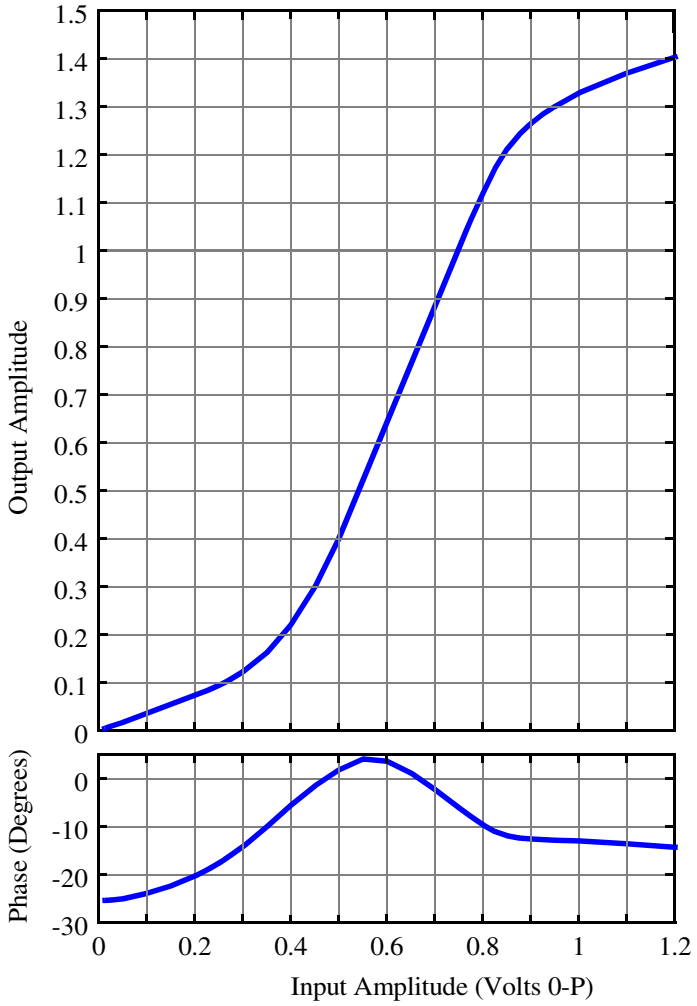


Fig. 4.2: Power Amplifier AM/AM and AM/PM curves

The input amplitude is given in Volts, 0-peak single-sided, while the output amplitude is normalized relative to a 500mW output power, the mean power of an EDGE signal to be transmitted. Amplitudes will be treated as dimensionless from this point on, but are normalized to these values.

Gain expansion of the class C stage turning on is apparent for inputs from 0.4 to 1.2V, and then output saturation can be seen at higher output amplitudes. For small inputs, the class C devices remain turned off, and the gain seen comes from the class AB helper stage. The 25° phase lag seen for very small signals is from the class AB stage working against the capacitances of the turned-off class C devices. Without the class AB helper, the phase shift here was much greater, on the order of -60°.

To see this PA's suitability for cartesian feedback, eigenvalues of the input/output jacobian are found per (Eq 3-23), and are plotted as a function of input amplitude in Figure 4.3. A constant 12° phase lead was added to minimize

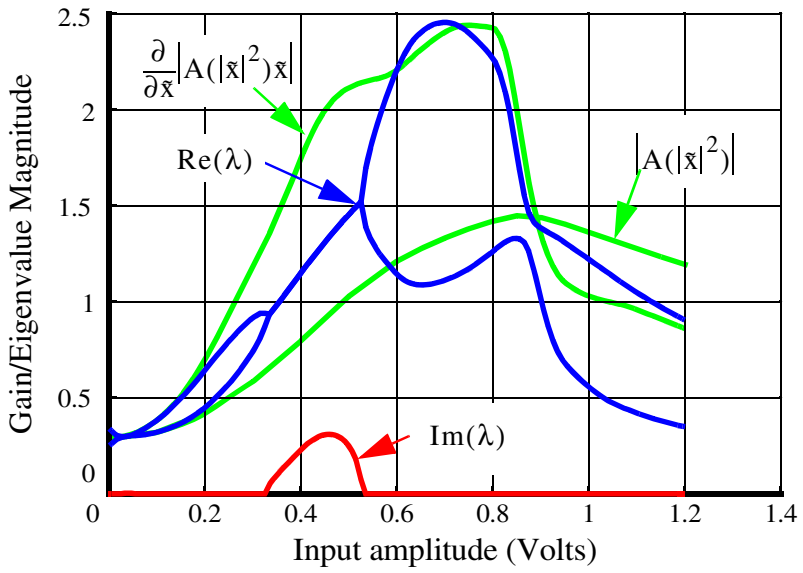


Fig. 4.3: PA transfer function eigenvalues

the ranges in which the eigenvectors go complex. This phase lead can be added in a system per Section 3.4.

These eigenvalues are difficult to measure in a typical system as it is usually difficult to measure phase shift through the PA. However the large signal gain $A(|\bar{x}|^2)$ is more accessible as signal power levels are more easily measured. If AM/PM is ignored, an incremental small-signal gain $\frac{\partial}{\partial \bar{x}}\left|A(|\bar{x}|^2)\bar{x}\right|$ can also be inferred from the large signal gain. These gains are also shown for comparison. It can be seen that the inferred small-signal gain varies over a range that is quite comparable to that of the eigenvalues. Thus, the AM/AM measurement alone can sometimes give a reasonable idea of what gains the loop must be designed for.

It is seen that for small input amplitudes where the gain comes from the class AB helper stage, the two eigenvalues are real. This would imply that the AM/AM distortion of the PA dominates over the effect of phase shift in this range. Then for input amplitudes from 0.33 to 0.53, the eigenvalues are seen to turn complex. The class C amplifier is turning on in this range, and although there is significant AM/AM as this happens, the phase lag decreases quickly, turning into a phase lead, and this AM/PM effect dominates stability. For further higher amplitudes, the eigenvalues become a a real pair again, with their magnitude falling off as gain compression from output saturation takes effect.

The magnitudes of the eigenvalues (more specifically, their real components) relate to how closely the system tracks its input, and how errors

settle out. Figure 4.4 shows a visualization of how the residual error settling behaves.

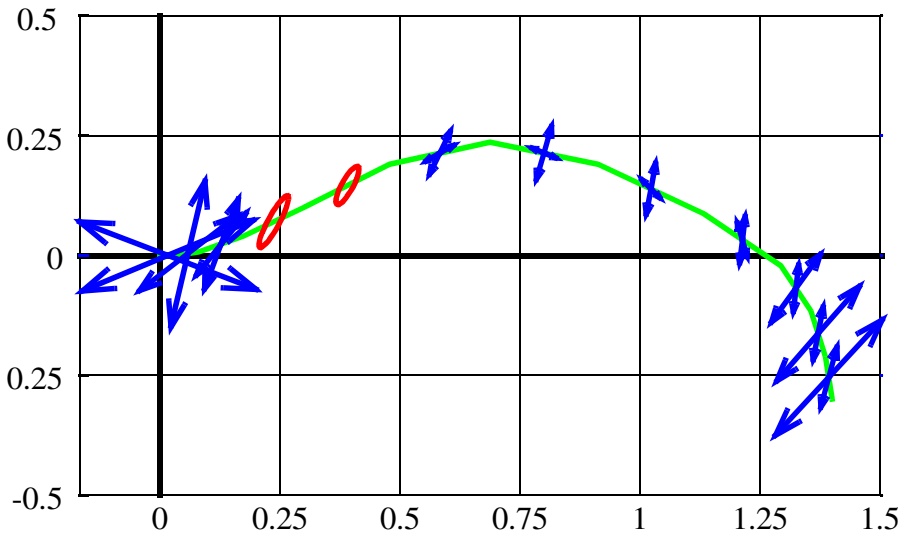


Fig. 4.4: Settling Eigenvector/Eigenvalue plot

The spine of this plot represents the locus in the output IQ plane as the input goes from zero to full amplitude along the I axis. This represents one of the radial rays in Figure 2.8b. Along each point in the plane, any error of the output can be decomposed into two eigenvectors, each of which settles according to the respective eigenvalue.

With a single-pole loop filter, settling towards equilibrium is exponential with real eigenvalues, while with complex eigenvalues, settling is oscillatory, following decaying elliptical orbits. Where the eigenvalues are real, arrows are given showing the eigenvector directions, with lengths proportional to λ^{-1} - larger eigenvalues mean faster settling, or a smaller error, thus short

arrows represent small tracking errors. For complex eigenvalues, the ellipse of the settling orbit is shown, with the major axis scaled according to $\mathrm{Re}[\lambda^{-1}]$ for consistency with the real eigenvalue arrows.

The error is seen to remain relatively small for moderate output amplitudes, but is large for both very small and very large amplitudes, where large-signal and small-signal gains are small respectively.

For use with the modified Nyquist criterion as given in Section 3.2, Figure 4.5 shows $-\lambda_1^{-1}$ and $-\lambda_2^{-1}$ plotted across all amplitudes. The eigenvalues

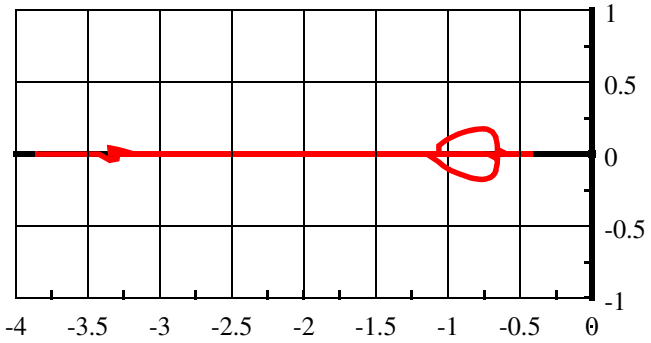


Fig. 4.5: Locus of Inverse PA Eigenvalues

mostly remain along the real axis, aside from small loops around -1 and -3.3. The loop around -1 is the most likely to affect closed-loop phase margin, and comes from AM/PM of the class-C PA action. The loop around -3.3 occurs for small signal amplitudes and is from phase shift when the amplifier operates in class AB.

4.4 Upconverter input spectrum

The inverse of the AM/AM and AM/PM curves in Figure 4.2 was applied to an ideal EDGE modulated signal to find the predistorted signal that must be fed to the PA to produce the intended output signal. The ideal output and predistorted signal are shown in Figure 4.6.

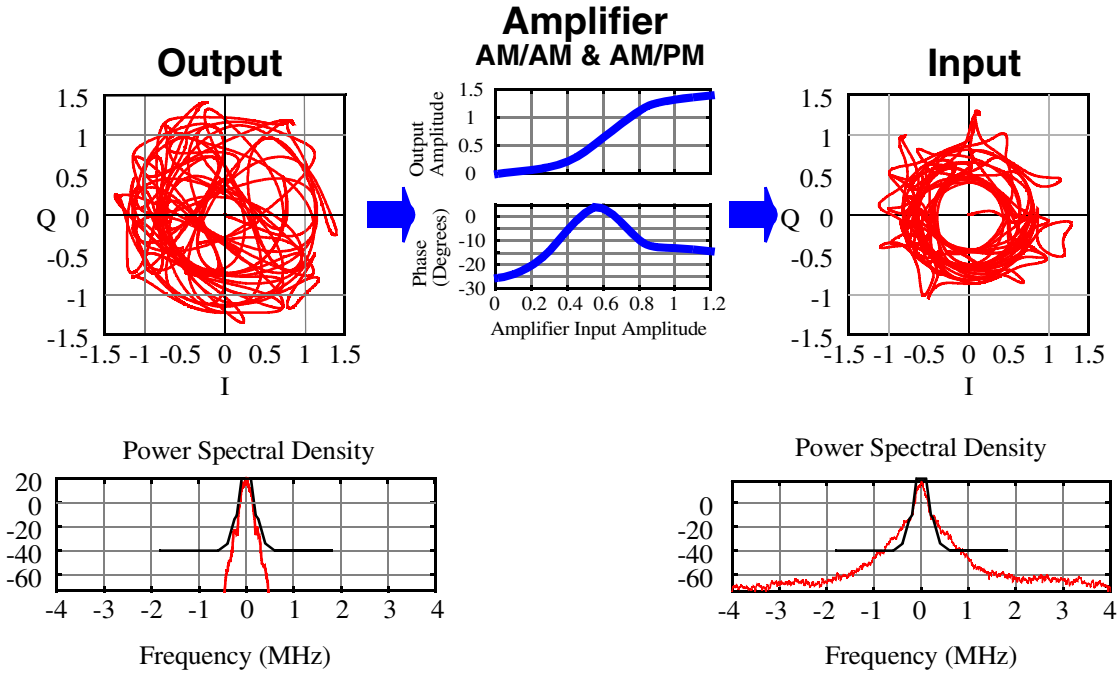


Fig. 4.6: Ideally Predistorted EDGE modulation

The gain expansion of the class C PA is corrected by a gain compression that can be seen in the predistortion: the relatively uniformly distributed range of amplitudes in the ideal output signal is visibly limited to a narrower range of amplitudes in the predistorted input. Conversely, gain compression of the PA for large amplitudes requires a significant gain expansion in the predistortion - this

is seen as the dangling loops seen around the edges of the predistorted modulation in the IQ plane.

The spectrum of the predistorted signal shows the same sort of spectral regrowth as expected from running a nonlinear amplifier open-loop. The predistorted signal's spectrum exceeds the normalized transmit mask by 20dB at 600kHz: this gives a minimum loop gain required from the loop filter.

4.5 Loop filter design

From the spectrum of the ideally predistorted signal, the loop filter must provide a gain of at least 20dB at 600kHz, as referred from the PA output (in $\sqrt{\text{Watt}}$) through downconverter, loop filter, and upconverter, back to the PA input (in Volts). However, as seen in Figure 4.3, the effective gain of the PA can be as large as 2.5 (in corresponding units). With a single-pole loop filter, this would result in a unity-gain bandwidth of $2.5 \cdot 10 \cdot 600\text{kHz} = 18\text{MHz}$ at baseband when the PA is in operating in the class-C gain expansion region.

This 18MHz at baseband means that at RF, 18MHz above and 18MHz below the carrier frequency is within the loop bandwidth, and over this 36MHz range of frequencies it is questionable how well the assumption of a memoryless channel holds. If the transfer function of the RF channel changes substantially in this bandwidth (as may occur if operating near the band-edge for a SAW

diplexing filter), then frequency-independent eigenvalues extracted for the PA are not adequate to represent the feedback, and ensuring stability is more difficult. If possible, it is preferable to reduce the loop bandwidth to avoid unexpected effects from channel memory.

The gap between this 600kHz corner and the unity-gain bandwidth depends on the effective number of poles in this frequency range. With a single pole, the rolloff is 20dB/decade. With two poles, the rolloff could be 40dB/ decade giving a worst-case unity-gain bandwidth of 3MHz (baseband).

The problem with using two poles however, is the extra phase shift: each pole contributes another 90° of phase shift together with its 20dB/decade of rolloff. With two poles, the net phase shift of 180°, if present around the unity-gain frequency, leaves essentially no phase margin. The range in which two poles are acting, if kept away from the unity gain frequency, might not be a problem. However, with the PA gain varying by a decade - from 0.25 up to 2.5, or -12dB to 8dB - where the unity-gain frequency ends up can vary, so the two-pole rolloff needs to avoid wherever the unity-gain frequency may end up.

It is possible to compromise between one and two poles. As the frequency passes through a pole frequency, the phase lag does not change instantly, but rather goes gradually in an arctan curve. If a pole is followed by a zero close enough in frequency, the phase lead from the zero cancels some of the

pole's phase lag, keeping the total lag from becoming too large, but at the same time also limits further gain rolloff.

A single pole/zero pair is known as lag compensation, and Boloorian and McGeehan[40] present using a widely-spaced pole/zero pair to improve low-frequency gain without impacting behaviour around the unity-gain bandwidth.

The technique need not be limited to a single pole/zero pair though: successive pole/zero pairs can continue the extra gain rolloff, and as long as the total phase shift is not too large, this can be continued through the unity-gain frequency. (Eq 4-2) gives an example loop transfer function, with a low-frequency dominant pole, and three half-decade pole/zero pairs spaced over three decades.

$$L(s) = \frac{2\pi 10^6 (s + 2\pi 10^5)(s + 2\pi 10^6)(s + 2\pi 10^7)}{(s + 2\pi 100)(s + 2\pi 10^{4.5})(s + 2\pi 10^{5.5})(s + 2\pi 10^{6.5})} \qquad \text{(Eq 4-2)}$$

Figure 4.7 shows the Bode plot for this transfer function.

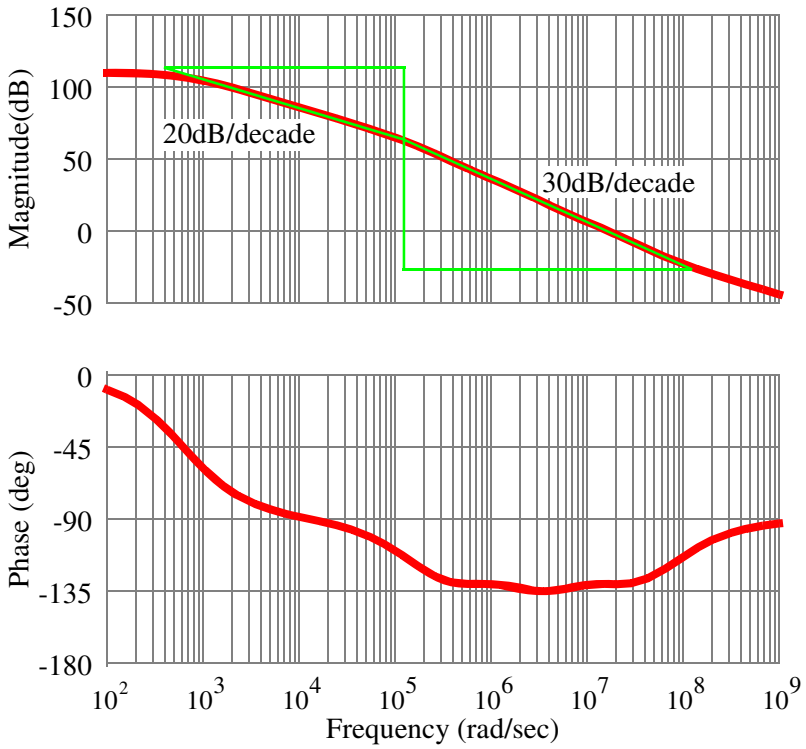


Fig. 4.7: Bode Plot of "1-1/2 pole" Loop Filter

It can be seen that the gain plot follows 30dB/decade, with only 135° total phase shift in the range from tens of kilohertz to tens of megahertz, leaving about 45° phase margin (not counting eigenvalue phase from the PA). The bode plot has a gain rolloff and phase shift of effectively 1-1/2 poles, giving better rolloff than one pole, but still retaining more phase margin than two poles would.

With the required 20dB gain at 600kHz and an amplifier gain of 2.5 (8dB), the worst-case loop bandwidth is 6MHz: much better than the 18MHz for a single-pole loop filter. To verify stability of using this loop filter together with

the designed PA, Figure 4.8 shows the Nyquist plot of this transfer function together with the λ^{-1} locus from Figure 4.5. The Nyquist plot is seen to clear the

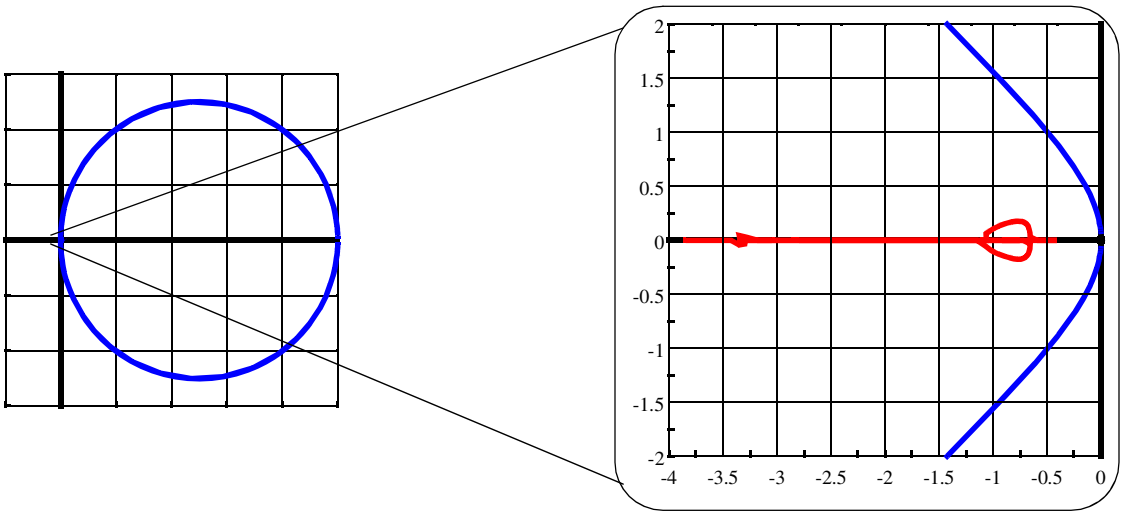


Fig. 4.8: Nyquist Plot with Inverse Eigenvalue Locus

λ^{-1} locus: this feedback system should be stable.

The choice of using an extra "1/2 pole" as opposed to, say, 0.4 or 0.6 poles or some other value was not explored, but the differences amount to adjusting the trade-off between phase margin and limiting the loop bandwidth. Difficulty with the loop phase correction adjustments, or a more accurate picture of RF channel memory may suggest adjusting this one way or the other, but the results found for 0.5 appear to be reasonable.

While the three-decade frequency range in which the extra "1/2 pole" acts is determined by the spectrum requirements of the predistorted signal, the choice of using three poles and three zeros with uniform spacing was arbitrary.

The spacing of the poles and zeros need not have been uniform: there is some small variation of the phase lag in the frequency range of interest, and adjusting the pole/zero locations may gain a few extra degrees of phase margin without significantly affecting the gain rolloff. This could be achieved by using a variation of the Remez algorithm or other numerical methods to optimize pole/ zero locations, but this was not pursued.

The number of poles and zeros used was also arbitrary: six poles and six zeros at quarter-decade spacing could well have been used, or for that matter, any other arbitrary number of poles and zeros, suitably spaced. Fewer pole/zero pairs with larger spacing would result in more variation of the phase and likely reduce phase margin, but the downside to using more poles and zeros spaced more densely is the hardware cost of implementing extra pole/zero pairs (this is examined in Appendix B). The three pole/three zero implementation seems to be adequate though.

The "1/2 pole" component of the transfer function approximates an amplitude response satisfying $|H(j\omega)|^2 \propto \frac{1}{f}$. As $\frac{1}{f}$ noise is sometimes referred to as "pink noise", such a "1/2 pole" response can be called a 'pinking filter' as such filters are sometimes used to synthesize pink noise from white noise.

4.6 Closed-loop simulation

A transient simulation of the closed-loop system, including the PA model from Section 4.3 and loop filter design of Section 4.5 was performed in Simulink/MATLAB. The simulation was behavioural, not simulating RF waveforms, but only baseband representations: everything from the inputs of the upconverter to the output of the downconverter was modelled according to lookup tables for the PA's AM/AM and AM/PM curves. Circuit noise is not modelled in the transient simulation.

IQ plots of the modulation, and the spectrum at both the PA input and output are shown in Figure 4.9.

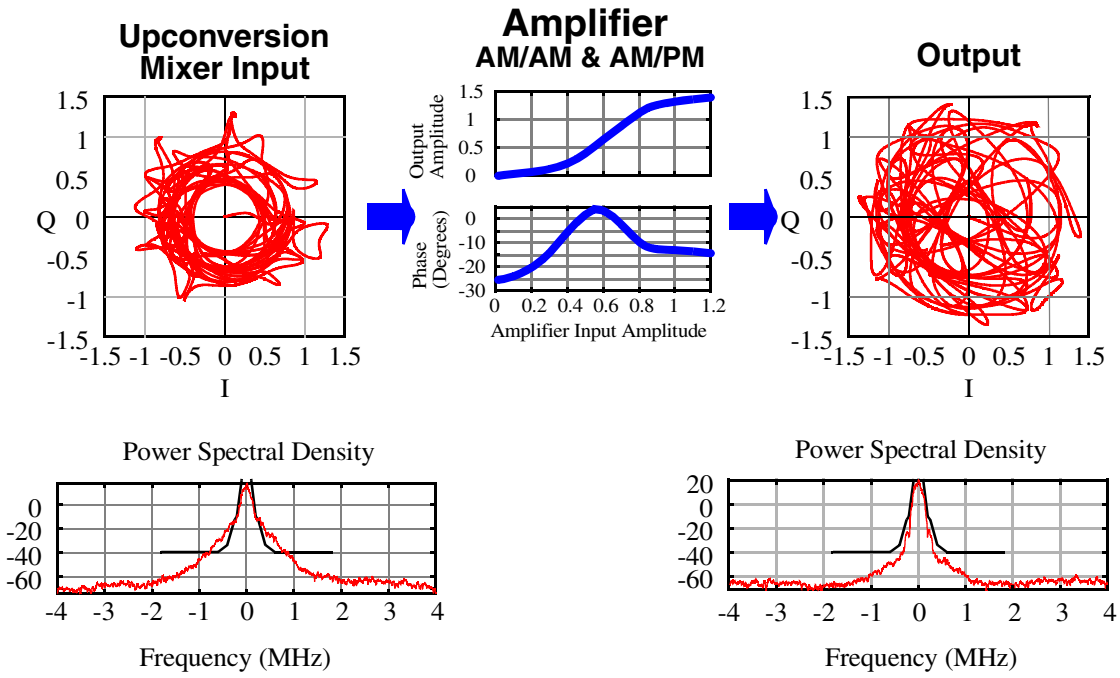


Fig. 4.9: IQ Modulation and Spectrum from closed-loop simulation

The waveforms and spectra are largely indistinguishable from the ideal predistortion shown in Figure 4.6, although the PA output spectrum does show a noise floor that was not visible in the ideal signal.

The magnitude of the input to the loop filters remained well under 0.1% of the modulated signal, representing an EVM two orders of magnitude better than the 9% specified. The waveforms observed for this are of limited value though: they do not exhibit the same continuity of the modulated and predistorted waveforms, but shoot around sharply from timestep to timestep in the simulation. This represents the numeric noise and accuracy limitations of the simulator rather than behaviour of the feedback system itself, but the residual error from the linearization is presumed to fall under this noise floor.

4.7 Noise

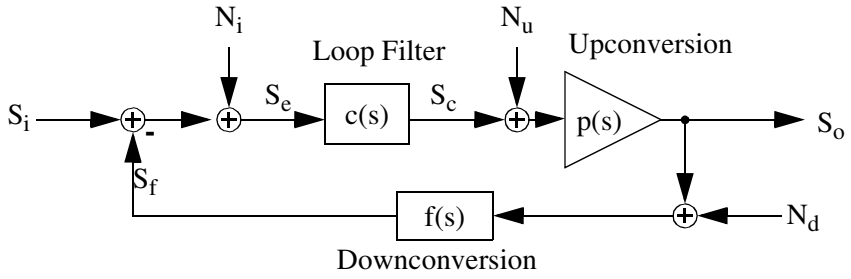


Fig. 4.10: Feedback system with noise

Figure 4.10 shows the linear system block diagram of Figure 3.1 with noise sources added to represent input-referred noise of the loop filter N_i, downconverter N_d, and upconverter N_u. The output signal can be shown to be:

$$S_o = \frac{p(s)c(s)S_i}{1 + p(s)c(s)f(s)} + \frac{p(s)c(s)N_i}{1 + p(s)c(s)f(s)} + \frac{p(s)c(s)f(s)N_d}{1 + p(s)c(s)f(s)} + \frac{p(s)N_u}{1 + p(s)c(s)f(s)} \quad \text{(Eq 4-3)}$$

The first term is the desired modulated signal, while the other three are noise added by the system.

The spectral mask given in Figure 4.1 gives the basis for noise performance requirements near the carrier. The noise spectral density from 600kHz to 1.8MHz must be 60dB lower than the spectral density of the modulated signal at the carrier frequency. This range of frequencies, if within the cartesian feedback loop bandwidth, refers directly to the noise performance requirement for the downconversion path and loop filter.

The spectral densities given in spectral mask specifications are for power measured in 30kHz bandwidths. In computing the spectrum given in the figure, the RMS power in 30kHz centered at the carrier is found to be -6.1dB relative to the total RMS power of the modulated signal. This means that the downconverter noise floor can at worst be -66dBc/30kHz or better from 600kHz to 1.8MHz.

GSM specifications for the spectral density also describe averaging across at least 200 sweeps to obtain the spectrum measurement. The -6.1dBc/ 30kHz figure was for an RMS (power) average, but it is also common for spectrum analyzers to perform a log-power (video) average, and the standard does not specify what form of average is to be taken. Measuring a modulated signal with a spectrum analyzer, the difference between an 8MHz bandwidth (capturing the entire channel) and a 30kHz bandwidth measurement gives a difference of -7.2dB instead. Using this measure for a basis, the downconverter noise floor can at worst be -67dBc/30kHz, or -112dBc/Hz.

The spectral mask for offsets from 1.8MHz through 6MHz varies with signal power, being 4dB lower for powers of 24dBm or lower. For higher output powers, the mask does not follow the signal power and remains fixed in terms of absolute power density. From 6MHz to the edge of the transmit band, the mask is another 8dB lower, being either -124dBc/Hz or -100dBm/Hz, whichever is lower.

Upconverter noise within the loop bandwidth is suppressed by the feedback, but outside the loop bandwidth, contributes directly to the output. The mask for >6MHz offset thus applies to the noise floor of the upconverter/PA.

The preceding estimates apply inside, and outside of the loop bandwidth respectively, however the transition region around the loop bandwidth requires caution. Both upconverter and downconverter noise contribute to the output,

with neither the $1 + p(s)c(s)f(s)$ denominators in (Eq 4-3) going to infinity, nor numerators going to zero. The noise gains are enhanced by a factor of $\frac{1}{1 + p(s)c(s)f(s)}$. For a phase margin of $45°$, this is $-20\log\left(\left|1 - e^{j\frac{\pi}{4}}\right|\right) = 2.32\text{dB}$ worse.

Chapter 5

Transmitter Prototype

To demonstrate the linearization of an integrated CMOS PA used with non-constant envelope modulation, a prototype transmitter was designed and fabricated. A PA was designed to produce up to 1W of output power operating in the DCS1800 band (from 1.710GHz to 1.785GHz). The transmitter linearizes this PA and was designed to meet EDGE linearity requirements for an 8-PSK modulated signal with this peak power (500mW average power). This chapter describes the design of the prototype chip.

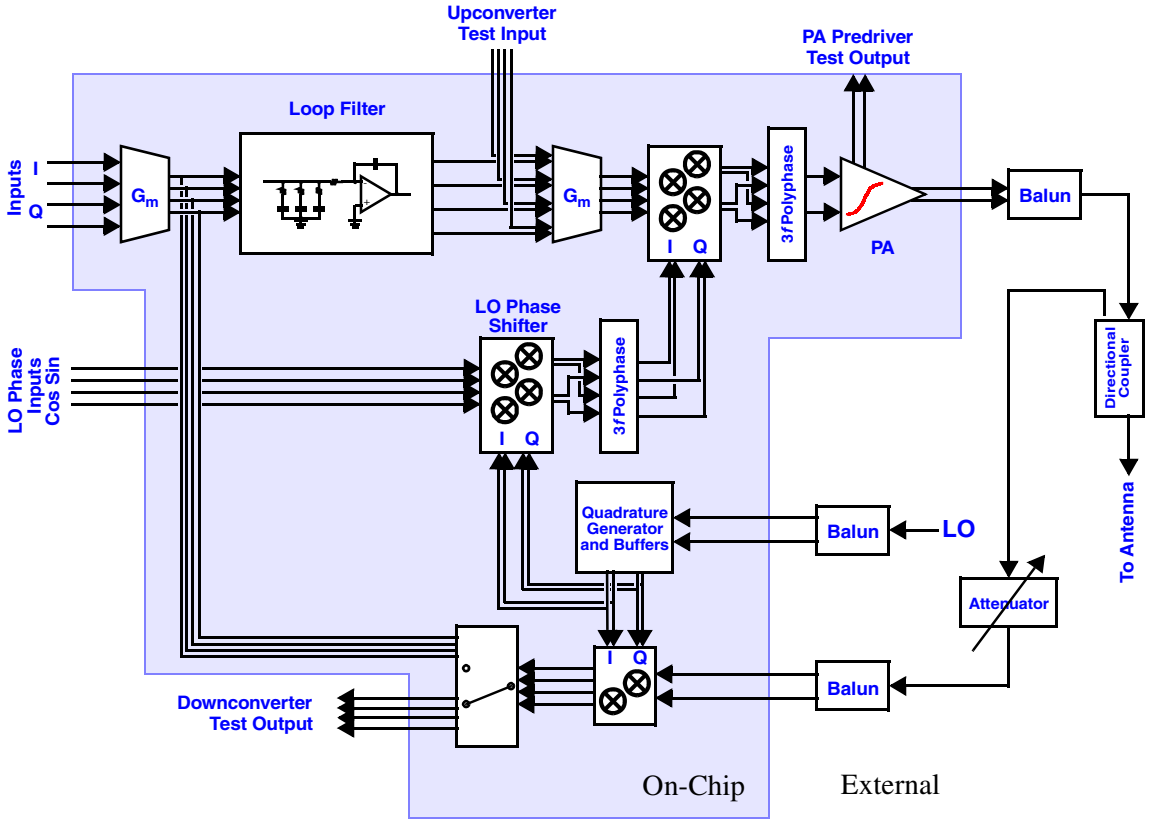


Fig. 5.1: Transmitter Block Diagram

A block diagram of the transmitter is shown in Figure 5.1, which is essentially a detail view of Figure 2.21 from the DAC output onwards. The prototype IC integrates a PA together with all the active circuitry needed to linearize it using Cartesian Feedback. The chip contains a PA, upconversion and downconversion mixers, baseband loop filters, and LO phase shifter circuitry. All on-chip signal paths are differential to reject coupling. Baseband signals are generally distributed in current mode into virtual grounds, facilitating the

summing of signals, and avoiding issues of voltage swing and voltage dependent distortion.

The output of the PA is taken off-chip, and the input for the downconversion mixers, rather than being tapped from the PA on-chip, is brought back in from off-chip allowing for flexibility in testing. The loop filter can be switched off, and an additional input is provided to the upconverter, to allow for open-loop testing of just the modulator and PA. Similarly, the output of the downconverter can be directed off-chip to allow testing of the downconverter on its own. Adjustable integrating capacitors in the loop filters and a programmable off-chip attenuator allow for errors in signal path gains to be corrected for.

These adjustable gains allow for flexibility of signal levels through the loop, but nominal levels were chosen for key points in the signal path. These levels are somewhat arbitrary and not aggressively optimized, but were found to be reasonable to design for. The PA output was designed to be 500mW average, while the downconverter was designed for a 0.3mW input signal, and produce a 1mA (0-peak) output current. The upconverter takes a 0-peak input signal of 250mV to produce a 560mV amplitude signal for the PA. Baseband I-Q loop-input signals are 500mV. These figures are single-ended, RMS figures for the modulated signal, with peaks from modulation being nominally $\sqrt{2}$ times larger.

The transmitter was implemented in a 0.18µm triple-well CMOS technology provided by STMicroelectronics with a MIM capacitor option. Transistors in this technology are rated for 1.8V operation, with 3.3V-tolerant thick-oxide transistors available.

All blocks operate on 1.8V supplies except for the PA which uses 2.5V supplies. The following sections describe the major circuit blocks in more detail. Some approaches that were not used in the prototype but were tried or considered during design are also mentioned for perspective.

5.1 Power Amplifier

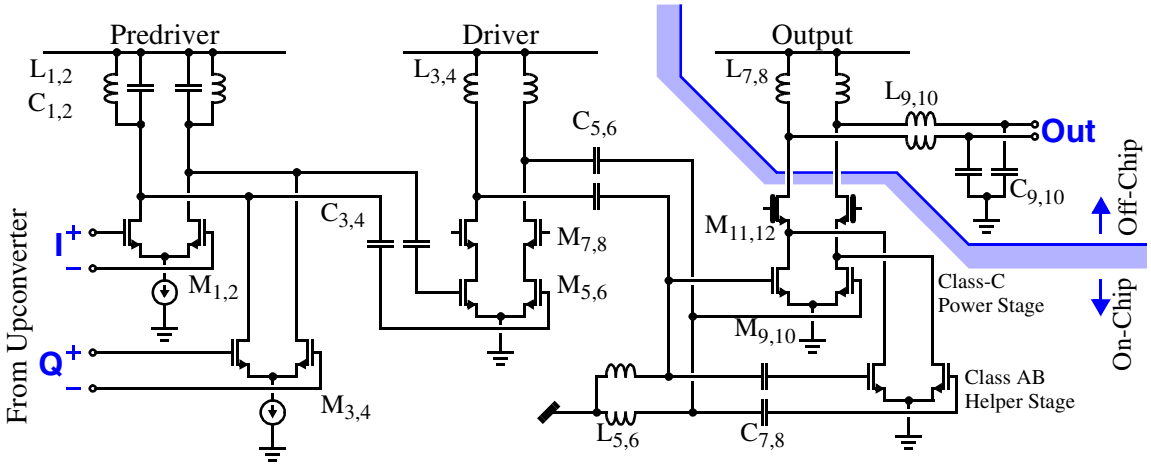


Fig. 5.2: Class C/AB Power Amplifier

	$M_{1\text{-}4}$	$L_{1\text{-}4}$	$C_{1,2}$	$C_{3,4}$	$M_{5,6}$	$M_{7,8}$	$C_{5,6}$	$L_{5,6}$	$M_{9,10}$	$M_{11,12}$	$C_{7,8}$	$M_{13,14}$	$L_{7,8}$	$L_{9,10}$	$C_{9,10}$
Size	80µm/ 0.18µm	6nH	0.25pF	10pF	500µm/ 0.18µm	250µm/ 0.18µm	10pF	0.6nH	7680µm/ 0.18µm	15360µm/ 0.35µm	3pF	480µm/ 0.18µm	0.6nH	1.5nH	4.8pF

Table 5.1: PA Device Sizes

A schematic of the power amplifier used in the prototype is shown in Figure 5.2 with device sizes given in Table 5.1. The design is almost identical in topology to the final three stages of a class C PA designed in 0.35µm CMOS by Narayanaswami[41]. This section will briefly describe the PA, highlighting differences from the previous design.

The PA consists of three stages, ending with a class C output stage. The input of this stage is a large gate capacitance, which is driven by a class AB driver stage. The input of the driver is a smaller capacitance, which is driven by a class A predriver that presents an even smaller capacitance, as well as requiring less voltage swing from the upconverter.

Inductor pullups are used for all three stages: this avoids the IR drops that resistive loads would experience. The inductors also resonate with capacitances at their respective output loads, forming tuned loads that reduce the signal current needed from the active devices compared to if the inductors were not present. The inductors for the predriver and driver are 6nH on-chip spiral inductors, with a Q of 4.8 as estimated by IE3D simulations. The output stage has bondwire inductors for its pullup and output matching network, with an estimated Q of 20. All three stages operate from supplies of 2.5V.

The use of tuned loads means that each stage's output voltage swings above and below the respective supply voltage. The voltage swings above

supply voltage call for attention, as gate oxides can be damaged if stressed with too much voltage being applied.

The predriver takes an input that is biased at 1.8v (from inductor loading of the upconverter outputs), and has an output that swings about 2.5v, and oxide stress is not a concern with the voltages seen here. The predriver takes two inputs in quadrature and simply sums them: this is done to provide roughly symmetric loading for four input phases from the upconverter: the need for more than a single differential input is explained in Section 5.2.2. Each input pair is biased with a tail current of 3mA.

The driver on the other hand, is configured with $M_{5,6}$ as common-source devices biased in class AB, with a quiescent current of about 8mA each set by an off-chip reference current. These devices are biased with a gate voltage around 0.55V, and with an output voltage swinging above 2.5V, would overstress their gate oxides. Cascode transistors $M_{7,8}$ shield the common-source devices from the output voltage swings. These cascode devices are biased with a 2.0V gate voltage, brought in from an external pad, and can pull their source voltages up to within V_t of this, giving the common-source devices a maximum drain voltage of about 1V. Capacitances at this node could be tuned with inductors, as was done in [41], but this was not attempted here.

The output stage is similar in topology, but common-source transistors $M_{9,10}$ are biased for class-C operation. Cascode transistors $M_{11,12}$ are thick-oxide devices to accommodate the larger voltage swings seen here. The transistors in the output stage are very large, and on-chip spiral inductors $L_{5,6}$ (with Q of about 3.8) are used to tune out their gate capacitance, allowing DC blocking capacitors $C_{5,6}$ to be reasonably sized. The DC bias voltage for the class-C transistors is brought in from off-chip and is below threshold at around 0.2V. The supply and cascode voltages are both 2.5V. The output pullup inductors $L_{7,8}$ and matching network of $L_{9,10}$ and $C_{9,10}$ are sized almost identically to [41]. This pullup and matching network is discussed further in Section 5.1.2

The class C amplifier does pose a potential problem for cartesian feedback, however. In normal large-signal operation, the signal path is an inverting gain through the active device. However, when the input is small enough that the transistor does not turn on, the signal path becomes the capacitive feed-forward path through the transistors' gate-to-drain miller capacitance. This is a non-inverting signal path, and the phase shift through the amplifier is very different from when turned on. This AM/PM was found to be about a 60° shift going from feedforward to inverting operation, and is significant enough to compromise stability of the feedback.

To reduce this AM/PM effect, a class AB helper was added to the output stage. This consists of transistors $M_{13,14}$ and DC blocking capacitors $C_{7,8}$ that allow these devices to be biased independently from the class C devices. These transistors draw a quiescent current of about 8mA each, about the same as the driver stage. For very small inputs, these devices operate in class A and provide an inverting gain when the class C transistors stay turned off.

Some AM/PM still occurs, however: the gain across the gate-to-drain capacitances still varies significantly with the class AB amplifier operating alone, as compared with the much larger class C devices working. This varying gain means the miller effect acting on the gate-to-drain capacitors presents a varying apparent capacitance for $L_{5,6}$ to resonate against. The amount of AM/PM left is reasonable, though, and from the simulations presented in Chapter 4, appear acceptable.

To further reduce the effect of this varying miller effect on the gate-to-drain capacitance, capacitor neutralization was considered, adding capacitors

between these gates and the opposite-phase drains as shown in Figure 5.3. The

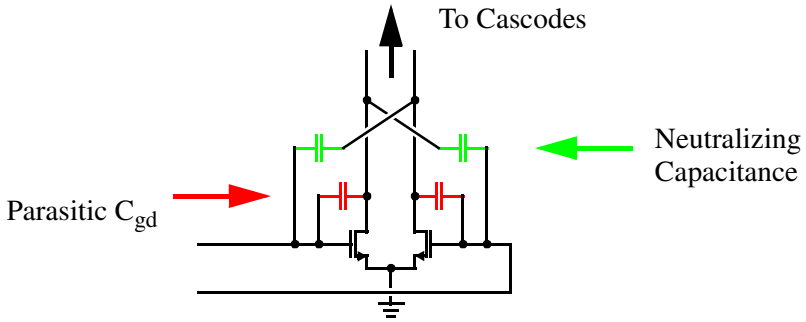


Fig. 5.3: Capacitor Neutralization

extra capacitors serve as a local positive feedback from the transistors drains back to the gates, and it was feared that with the inductor tuning at the gates, this arrangement could become unstable, so this prototype relies only on the class AB helper to mitigate AM/PM. The neutralization technique was subsequently used with success in [42] though.

AM/AM and AM/PM curves for the PA were shown in Section 4.3. Note that Figure 4.2 shows values for operation using only one of the predriver inputs - in operation with both inputs active, actual input voltages are only a factor of $1/\sqrt{2}$ of what the horizontal scale indicates.

Figure 5.4 shows a plot of the simulated output-stage drain efficiency as a function of output amplitude (in $\sqrt{\text{Watts}}$, together with ideal linear-conductor class A and B efficiency curves, for comparison.

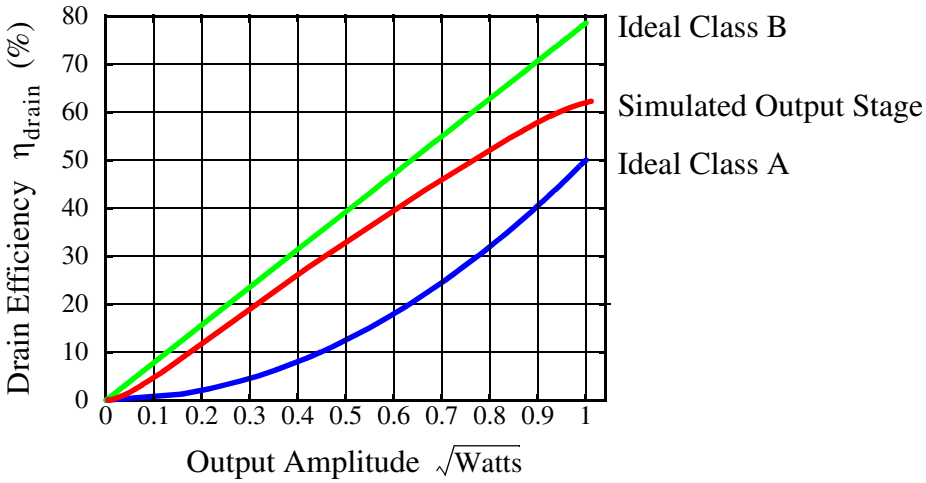


Fig. 5.4: PA Output Stage Drain Efficiency

Overall efficiency was also simulated, and found to be 55% peak, for a 1W signal, and 38% for a 500mW 8-PSK modulated signal.

5.1.1 PA Layout

Given the large currents seen in the output stage, and the very large device sizes required to deliver them, the effects of parasitic resistances and capacitances are significant and deserve special attention. Layout plays a major role in how these parasitics turn out, and some of the considerations that went into the output stage layout will be discussed here.

The output stage was placed in the corner of the prototype chip in an effort to maximize the number of bonding pads available. Extra bonding pads

for grounds and supplies allows more bondwires for these nodes to keep parasitic bondwire inductances low.

The corner placement makes device matching a potential issue: with a diagonal symmetry between halves of the circuit, the reflection of horizontal devices in one half are vertical devices in the other, and the different orientations mean the devices could potentially be poorly matched. As the large transistors of the output stage are implemented as numerous short fingers to keep gate resistance down, the matching issue is addressed by dividing each transistor into an equal number of horizontal and vertically oriented fingers in a basketweave arrangement.

The combination of individual fingers is done in roughly square blocks of three (for the cascode) or four (for the common-source devices) pairs. Each pair of fingers is surrounded with substrate contacts to try and collect any substrate noise injection. The diffusion shared by each finger pair is used for the drain of the common-source devices, and the source of the cascode transistors to minimize diffusion capacitance on this cascode node. The unshared diffusion outside each pair goes to ground for the common-source devices - where capacitance to ground is a non-issue - or to the output node for the cascodes, where the drain capacitance is tuned against the inductor load anyway.

The finger pairs within each block are staggered to optimize parasitic resistances, putting extra width where the most current is needed, as

conceptually illustrated in Figure 5.5. The slope of the staggering used was

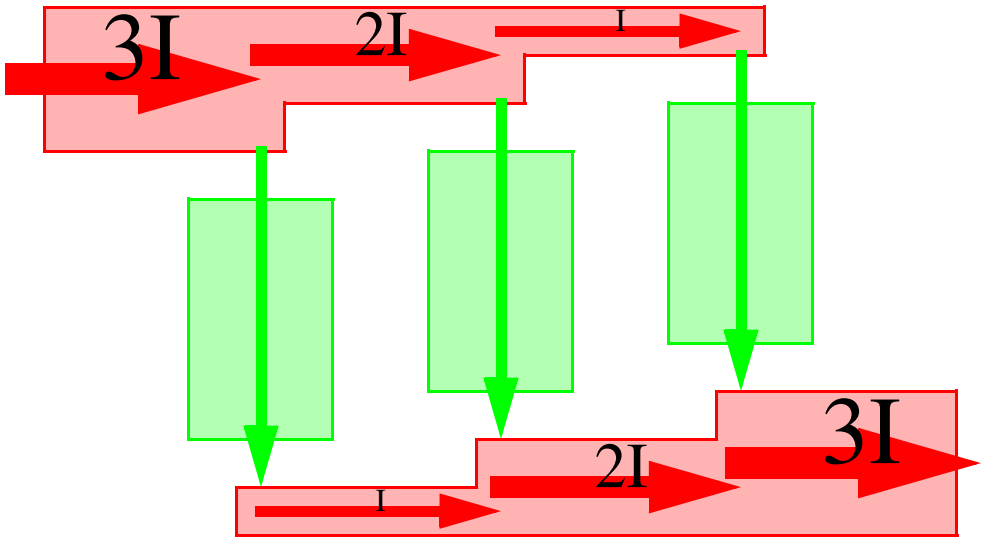


Fig. 5.5: Device Staggering

generally chosen to keep current densities roughly constant through the large collection lines.

This staggering technique was used on several levels. The fingers within each block are staggered. Pairs of cascode transistors are staggered in their blocks as well, and groups of eight of these pairs are combined using this staggering too.

Parasitics of the cascode node were minimized by keeping the common-source transistors close to the cascode transistors. Blocks of the common-source transistors are distributed amongst cascode transistor blocks, minimizing the distance that currents have to travel between these devices.

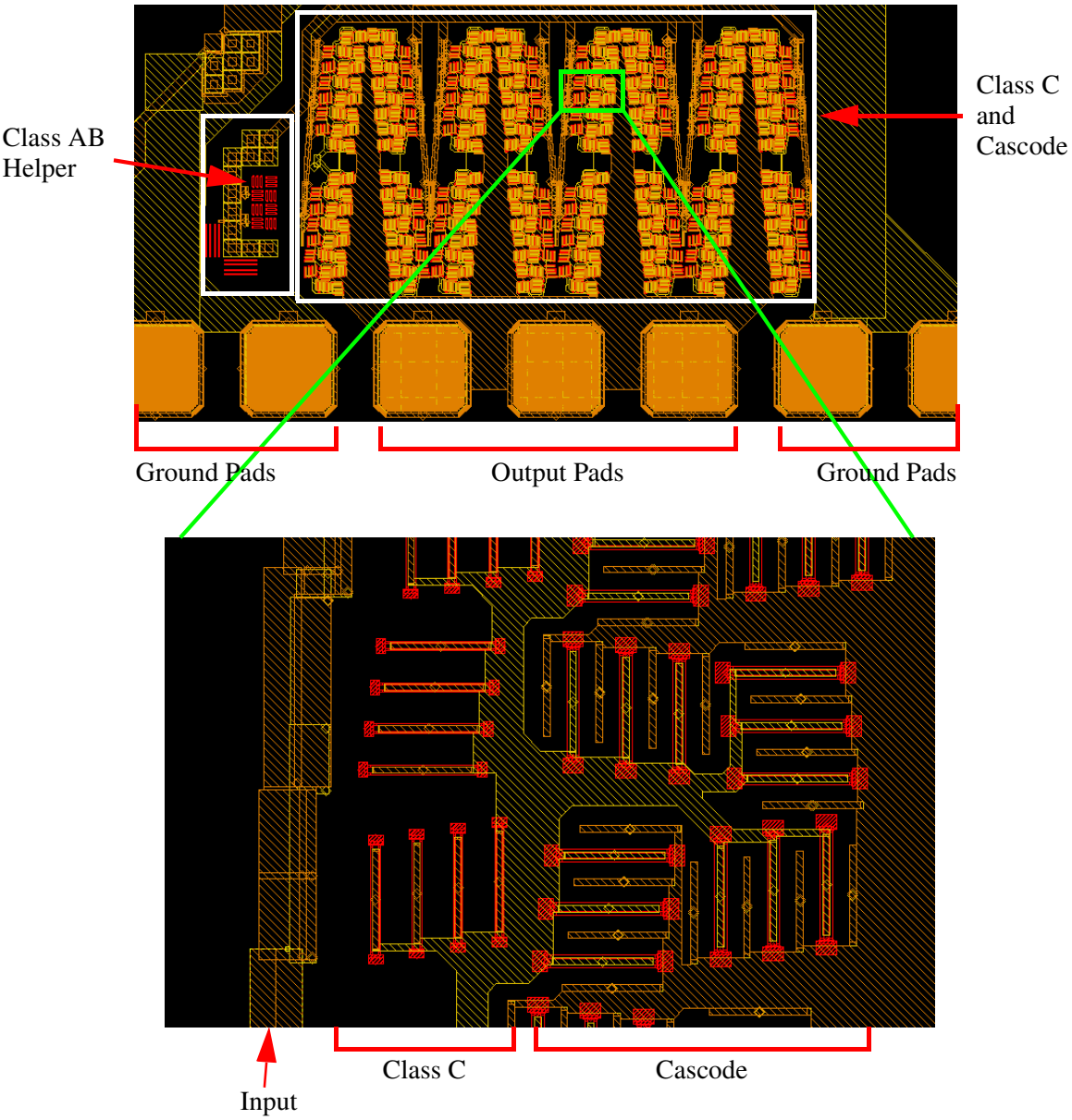


Fig. 5.6: Output Stage Layout

The layout of the devices for half of the output stage is shown in Figure 5.6, with an enlarged view of several blocks showing finer details of the staggering. The gap between the devices in the top and bottom sets of devices allows a path for ground currents to travel horizontally to the complementary

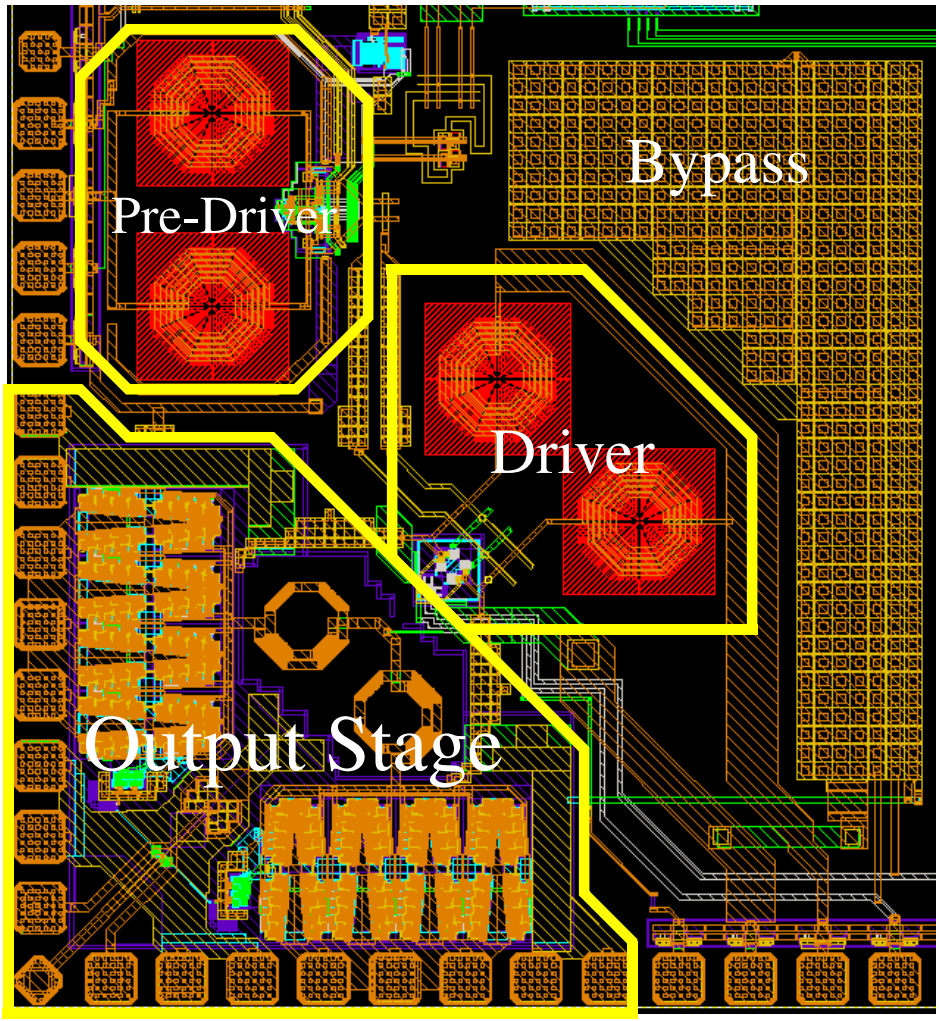


Fig. 5.7: Complete PA Layout

other half of the stage. The signal input is carried in the metal line shown on the left side of the enlarged view, with connections to transistor gates being in lower metal layers not shown.

The complete PA is shown in Figure 5.7. The layout of the driver and predriver are less noteworthy than the output stage, consisting of much smaller devices that occupy little area relative to their spiral inductor loads.

The process used includes a triple-well option, and this was used to provide some degree of substrate isolation. Each stage of the PA is contained in its own P-well. The buried N-layer and surrounding N well that isolate these P-wells are connected to the respective stages' supply voltages through large resistors. These resistors set the bias voltage for the N-well, reverse biasing their junctions to the substrate and the contained p-wells, but shield the supply from any substrate noise that may get coupled into the buried layer. This same technique was used throughout the rest of the chip as well, with each of the other major circuit blocks being isolated in its own p-well.

5.1.2 Output matching

The transition from the on-chip circuitry of the PA to off-chip signals on the board makes use of bondwire inductors, and an ideal schematic of the interface is shown again in Figure 5.8 for reference. $L_{9,10}$ and $C_{9,10}$ form a standard L-match network providing the impedance transformation between an on-board impedance of 50Ω to impedance of about 6Ω seen at the bond pads. This 6Ω impedance needed at the bond pads is determined by the need to deliver up to 500mW using a voltage swing on the order of V_{dd} or 2.5V. Supply

bondwires $L_{7,8}$ provide supply current and tune out output-node capacitance from the cascode device.

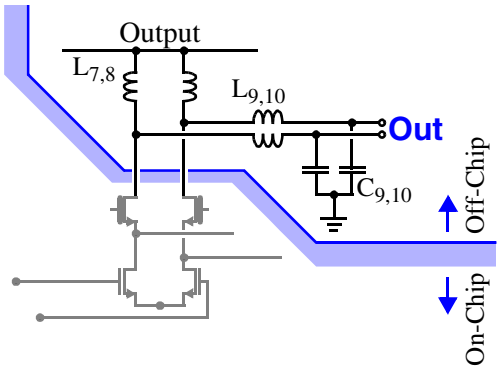


Fig. 5.8: Ideal PA output network

The inductances used in this design are comparable to those used in [41], so the intended values seem to be reasonably achievable, but the values of these bondwire inductances are not well controlled. Although there is a rule of thumb that each millimetre of bondwire length accounts for about a nanohenry of inductance, as a practical matter, achieving the correct values is a trial-and-error process involving bonding and re-bonding these bondwires to find the necessary lengths unless inductance errors can be absorbed into other adjustable parameters.

To accommodate bondwire variation, the output network as shown in Figure 5.9 is used in going off-chip.

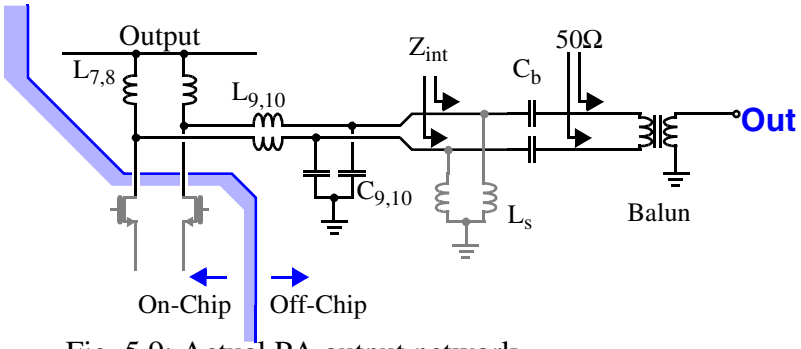


Fig. 5.9: Actual PA output network

Off-chip components L_s and C_b form an L-match network that converts the 50Ω impedance of the balun to a higher intermediate impedance Z_{int}. $L_{9,10}$ and $C_{9,10}$ match to this impedance instead. $L_{9,10}$ is anticipated to be larger than the ideal value after taking into account extra inductance from on-board routing to reach capacitors $C_{9,10}$, thus this impedance naturally matches a higher board impedance anyway. The inductor L_s is absorbed into $C_{9,10}$, reducing the capacitor value needed. C_b also serves as a needed DC block between the supply voltage present at the PA output nodes and the DC ground presented by the balun.

The values of $C_{9,10}$ and C_b provide two degrees of freedom that can accommodate variations in $L_{7,8}$ and $L_{9,10}$. These capacitors are implemented on the board as parallel combinations of fixed and trimmable capacitors.

5.1.3 Test Output

Not shown in the schematic, a scaled replica of the predriver was included, sharing inputs with the PA predriver, but with outputs brought to bonding pads instead of an on-chip load. This output was included to allow observing the upconverter output without involving later stages of the PA.

5.2 Upconversion Mixers

The PA is driven by a set of on-chip direct-conversion mixers. These mixers, their output polyphase filter, and LO phase shifting mixers are described in detail in [43] and are summarized here.

5.2.1 Upconverter Core

A simplified circuit diagram for the upconverter is shown in Figure 5.10, and is essentially a modified version of Figure 2.5. The circuit generates quadrature outputs I_{out} and Q_{out} satisfying $I_{out} + jQ_{out} = (LO_I + jLO_Q)(I_{in} + jQ_{in})$. Aside from generating a quadrature output instead of a single differential pair, the circuit is functionally identical to what was described in Section 2.3.1. The

additional output phases facilitate filtering of the output signals as described in Section 5.2.2.

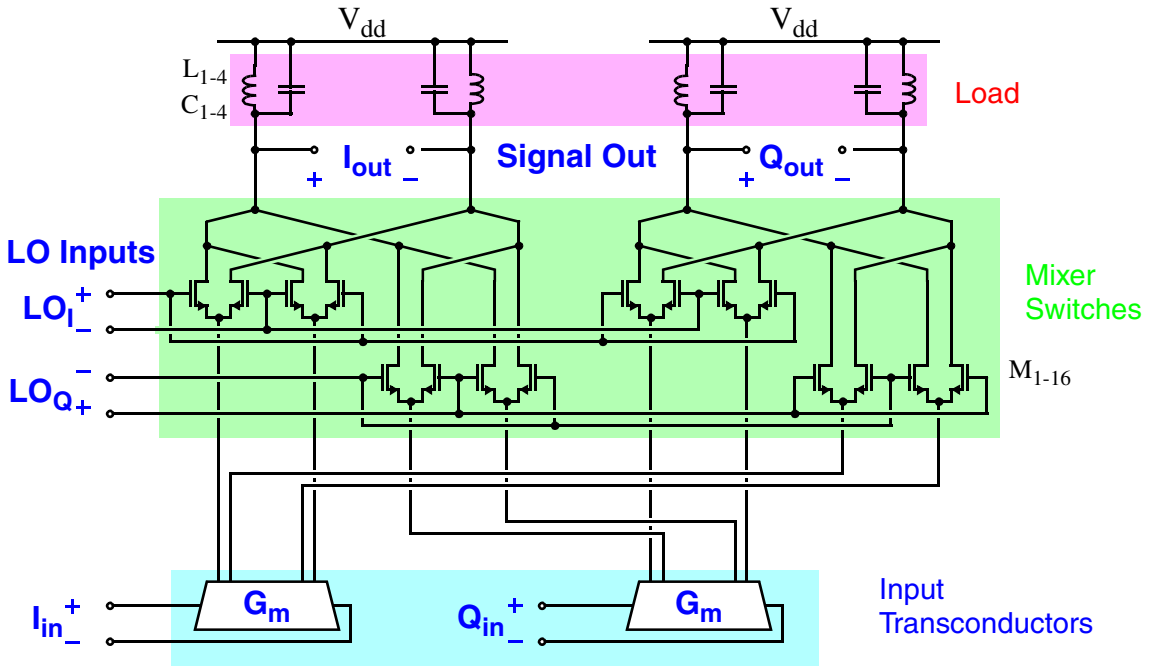


Fig. 5.10: Simplified CMOS Direct-conversion modulator

	M_{1-16}	L_{1-4}	$C_{1,4}$
Size	70µm/ 0.18µm	6nH	0.32pF

Table 5.2: Upconverter Device Sizes

The LO input signal is taken with a common-mode voltage of 900mV instead of V_{dd}. This low input bias voltage helps keep the switching transistors, whose drain voltages swing around V_{dd}, in saturation during switching transitions. This helps linearity by keeping the transistors insensitive to the magnitude of the output voltage swings.

The differential-pair input transconductors in Figure 2.5 depend on the linearity of the input transistors, and stack two transistors between the mixer switches and ground, limiting headroom. As the transconductors only deal with baseband signals, more complicated circuitry is easily substituted. An active transconductor circuit was used instead of simple differential pairs, offering linearity that depends on passive resistors rather than active devices, and inserts only one device underneath the mixer switches, A diagram of the transconductor is shown in Figure 5.11.

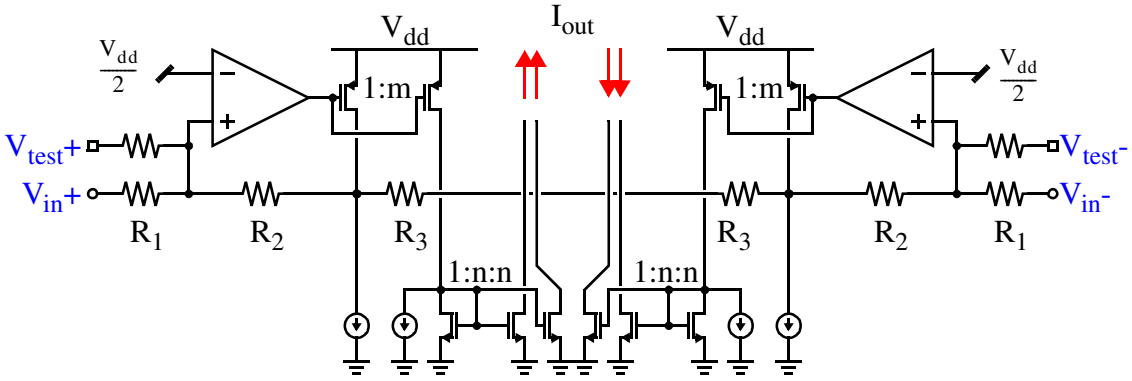


Fig. 5.11: Upconverter Transconductor

	R_1	R_2	R_3	m	n
Size	8kΩ	5kΩ	1kΩ	1	8

Table 5.3: Transconductor Component Sizes

The conversion from voltage to current is performed by the input resistors R_1 between the input voltages and virtual grounds presented by the op-amps. Two resistors are used to provide two sets of inputs: one from the loop filter for closed-loop operation, and another off-chip input for testing purposes. The virtual ground isolates each input from the other and each input is invisible

to the circuit if the respective driving source is left open-circuit, whether by turning off the loop-filter, or leaving the test inputs disconnected.

Resistor R_2 converts the net input current back into a voltage that creates an additional signal current from R_3. This signal current comes from the PMOS transistor at the op-amp output, and an identically driven PMOS transistor feeds a 1:n current mirror that provides the transconductor's output current that goes to the mixer switches. The overall transconductance of the circuit is:

$$G_m = \frac{mn}{R_1}\frac{R_2 + R_3}{R_3} \qquad \text{(Eq 5-1)}$$

For a baseband input signal swing of 350mV 0-p, the transconductor delivers an output current of 2.1mA 0-p for the mixer switches on a bias current of 2.275mA on each output. Each op-amp consumes 0.55mA, and another 0.82mA is consumed by current mirroring associated with each op-amp.

No deliberate common-mode rejection circuitry is used on the transconductor as the inductor loads fix the mixer output to V_{dd} regardless of common-mode current. Some common-mode rejection occurs as a side-effect of R_3 being essentially an open-circuit in common mode; this makes the common-mode transconductance less than the differential transconductance by a factor of $\frac{R_2 + R_3}{R_3}$.

5.2.2 Harmonic Reduction for Commutated Waveforms

One issue with using current-commutated mixers is that commutation amounts to multiplying with a squarewave which contains significant harmonic content. Characterization of the PA is generally done assuming a sinusoidal input for the PA, however the shape of the waveforms coming out of current-commutated mixers is not a pure sinewave. Figure 5.12 shows idealized current waveforms for three different phases of output coming from a current-commutating upconverter, together with the ideal sinewaves they are meant to represent.

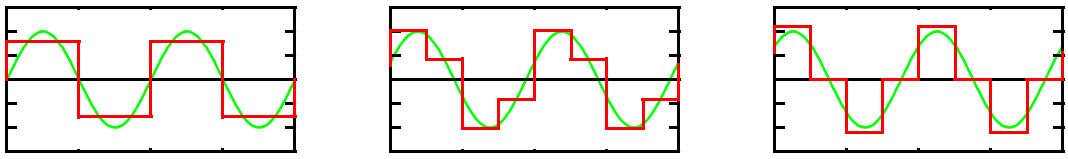


Fig. 5.12: Current-commutated waveforms for $\sin(\omega t + \theta)$

The staircased shape of these waveforms reflects the harmonics of the carrier that are created by commutation. Note that the shape of these waveforms varies with phase. If the PA is sensitive to these differences, then AM/AM and AM/PM curves are inadequate to characterize it as the overall transmitter can have PM/AM and PM/PM distortion as well. The mapping from the upconverter output's phasor to the PA output phasor is no longer rotationally symmetric, and characterizing PA distortion becomes more complicated.

The choice of upconverter architecture affects this problem. In a dual-conversion modulator, these harmonics are of the upconverter LO frequency rather than the carrier frequency. The staircasing of the LO signal in effect slides constantly across the carrier sinewave and distortion around the carrier gets averaged out. AM/AM and AM/PM curves apply to the averaged behaviour, and what distortion products of the PA remain are modulated by the intermediate frequency LO, ending up away from the carrier.

A direct-conversion upconverter as used in the prototype does not benefit from such averaging and suffers from carrier harmonics, but filtering the harmonics helps. Typical transceiver designs using external power amplifiers benefit from filtering that occurs in going off-chip, as well as filtering built into the input of the PA, but with an integrated PA, some attention is required. Some filtering happens by virtue of the LC tuned loads of the upconverter, but the third harmonic can still be a concern: it has the largest amplitude of the harmonics, being 1/3 of the fundamental amplitude, and being at the lowest frequency, is the least attenuated by the LC tuning.

5.2.2.1 3*f* Post-Modulator Polyphase Filter

To attenuate the third harmonic content, the prototype uses a sequence-asymmetric polyphase filter at the upconverter output. Sequence-asymmetric polyphase filters were originally proposed in the 70's by Gingell [44] for generating SSB signals. The topology seems to have been largely forgotten until

resurfacing in the mid-90's as a way to filter images in integrated low-IF receivers[45][46] as well as for generating quadrature LO signals [47]. Passive polyphase filters are now a well known approach for these applications, and a very good description of their function is given in [48].

Polyphase filters have also been used for suppressing harmonics of quadrature signals [49], and a single stage of this approach is used in the prototype. A schematic of the post-modulator filter is shown in Figure 5.13.

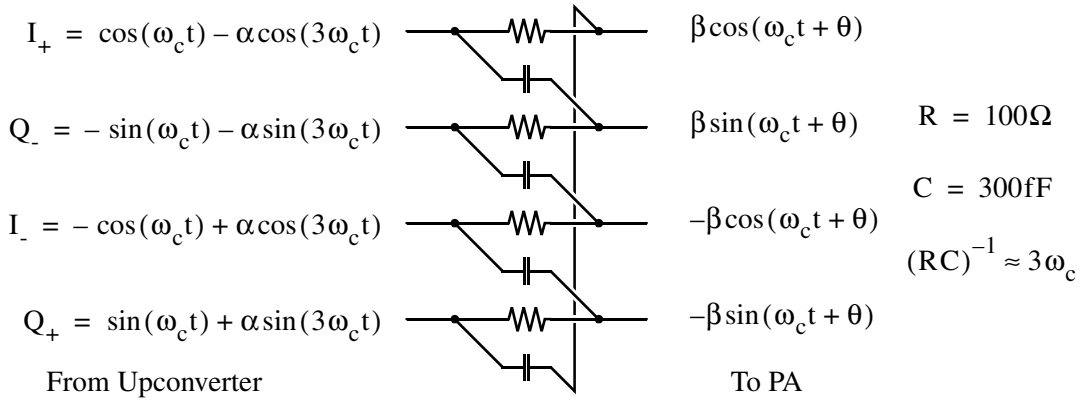


Fig. 5.13: Asymmetric-sequence Polyphase Filter

Polyphase circuits can be thought of as an extension of differential circuit techniques, where a phase shift of 180° can be implemented by simply swapping signals. Although in principle any number of phases could be used, it is common to use four phases in quadrature; by adding quadrature signals, phase shifts of 90° are similarly accomplished by tapping adjacent signals. A capacitor has an impedance that is 90° away from that of a resistor, so by combining signal current through resistors with signal current through capacitors with

phase re-ordering, a 180° phase difference exists. In the prototype's post-modulator filter, component values are chosen so that the magnitude of third-harmonic currents through resistors and capacitors match, thus nulling the unwanted third-harmonic at the output.

As four signal phases are needed, this doubles the hardware required of the upconverter to provide the additional quadrature outputs as compared to a single differential signal pair, but at under 10mA of extra current draw, this was deemed an acceptable cost. The PA predriver also requires an extra input to accommodate, but this is also a minor cost, the two individual inputs each being somewhat smaller than what a single-input predriver would need to be.

5.2.2.2 Higher-Order Oversampling (not used)

Another approach was considered for harmonic suppression. This approach was first introduced by Davies[50] who considers the use of many phases of squarewaves to synthesize low-frequency sinewaves, but the basic approach can be applied to higher frequency sinewaves as well.

Combining quadrature squarewaves to synthesize a sinewave is like sampling the sinewave four times each period, holding each sampled value until the next sample. The sinewave being sampled exists at frequencies of $\pm f_c$, and sampling at $4f_c$ creates aliases at frequencies of $n \cdot 4f_c \pm f_c$, with the lowest of these being the $3f_c$ harmonic. Sampling more points through each cycle would

increase the lowest alias frequency. Figure 5.14 shows waveforms for sinewaves synthesized by combining four squarewaves spaced 45° apart, effectively sampling at $8f_c$. The extra phases result in a waveform that better approximates

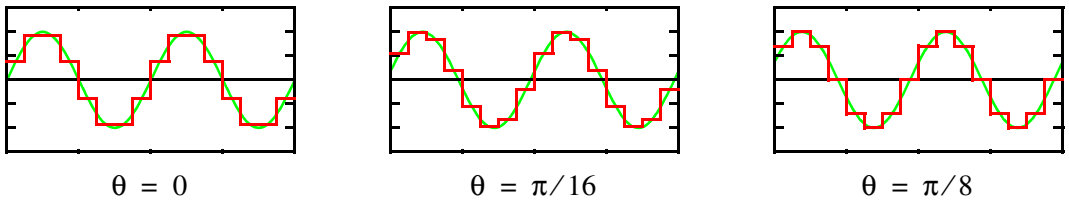


Fig. 5.14: 8x commutated waveforms for $\sin(\omega t + \theta)$

the sinewave expected by the PA, having no third or fifth harmonic content.

The drawback is the extra hardware required: to go to twice as many phases, twice as many mixers are needed (although the output signal current required of each individual mixer can also be somewhat smaller). More importantly, additional phases of the LO signal are needed, and the extra hardware to generate them was deemed too expensive to be practical for this prototype. The approach was subsequently used to good effect in [51] for a double-conversion modulator where the first LO is at a low enough frequency to not need LC tuning and the additional phases are available for free from frequency division: this first conversion is more in keeping with the low-frequency applications for which the approach was first proposed.

5.2.3 LO Phase Shifter

As was discussed in Section 3.4, the phase of the upconverter LO needs to be adjustable relative to the downconverter LO. The prototype introduces this adjustment by taking the downconverter's LO signal, and synthesizing the phase-shifted LO signal from that, based on DC sine/cosine inputs to select the phase angle. As the phase-adjusted LO signal is itself essentially a modulated signal with a constant complex-envelope, with slight modification, the same

modulator circuit as for the upconverter core is used for the LO phase shifter. The phase shifter is shown in Figure 5.15.

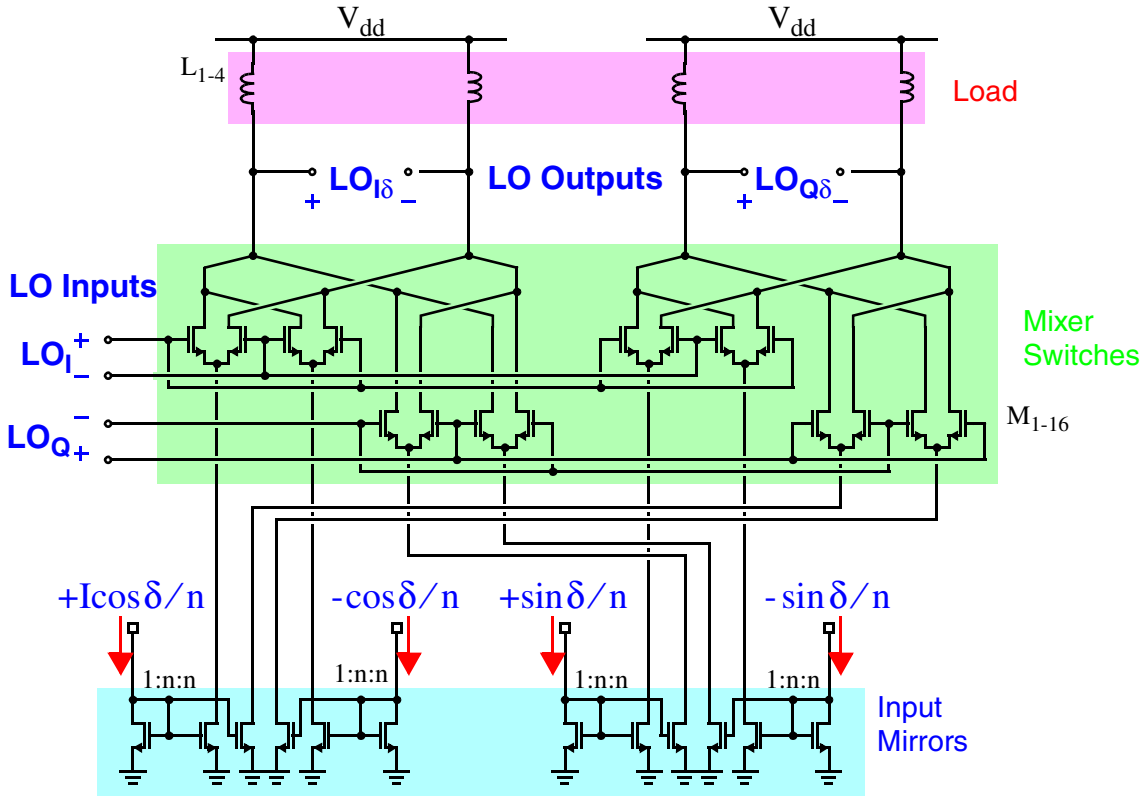


Fig. 5.15: LO Phase Shifter

	M_{1-16}	L_{1-4}	n
Size	136µm/ 0.18µm	6nH	10

Table 5.4: Phase Shifter Device Sizes

The transistors are sized larger than for the upconverter core to allow for larger DC currents, and capacitors were removed from the tuned load in light of extra drain capacitance from these devices. Also, as linearity from the baseband inputs to the output envelope is not critical for these mixers, the

switch transistors have their LO inputs biased about V_{dd} rather than $\frac{V_{dd}}{2}$, saving the need for a DC level shift.

The baseband inputs are current-mode and come from off-chip. Input currents can be brought in differentially on top of common-mode bias currents, but can also be fed single-ended, with only one cosine and one sine input being given a current, while the other two inputs are grounded to turn off their respective branches.

The phase-shift mixers were designed for tail currents with a DC vector magnitude of 5mA, e.g. $\sqrt{I_{cos}^2 + I_{sin}^2} = 5\text{mA}$. The actual current consumed varies, but for single-ended inputs, is on the order of $5\text{mA} \cdot \cos\left(\frac{\pi}{4}\right) \cdot 4 \approx 14\text{mA}$ at most.

The current inputs also have 12pF bypass capacitors and 100Ω series resistors between input pads and bypass, not shown in the schematic. The bypass capacitors were sized to fill available area in the layout, and the resistors were added to damp any parasitic resonances that could otherwise occur between these capacitors and bondwire inductances.

The output of the phase-shifters is DC level-shifted as shown in Figure 5.16 and goes to a 3*f* polyphase filter identical to that used for the upconverter

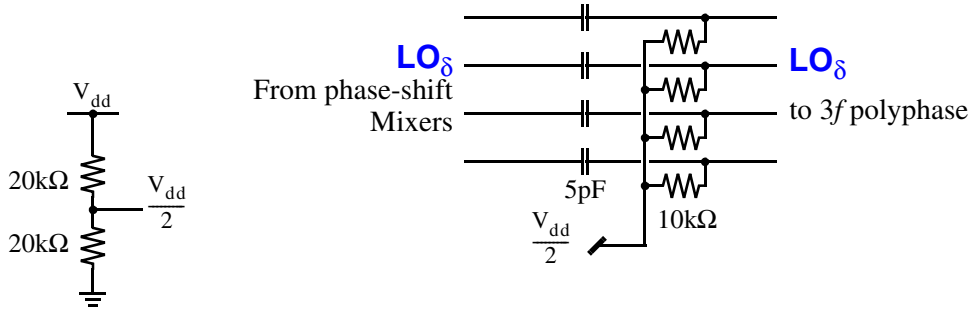


Fig. 5.16: Phase-Shifted LO DC-Level Shift

core before driving the upconverter switch transistors. The same $\frac{V_{dd}}{2}$ voltage is used as the input reference voltage for the transconductors.

5.3 Quadrature LO Generation

The upconverter and downconverter mixers require a quadrature LO signal, but external signal sources generally start as single-ended signals. A passive off-chip balun converts the signal-ended LO signal to differential, and quadrature signals are generated on-chip from that. The circuitry to generate LO signals for the mixers is shown in Figure 5.17 and is based closely on [52] where

various approaches for generating quadrature are studied, and the design of each block is described in depth.

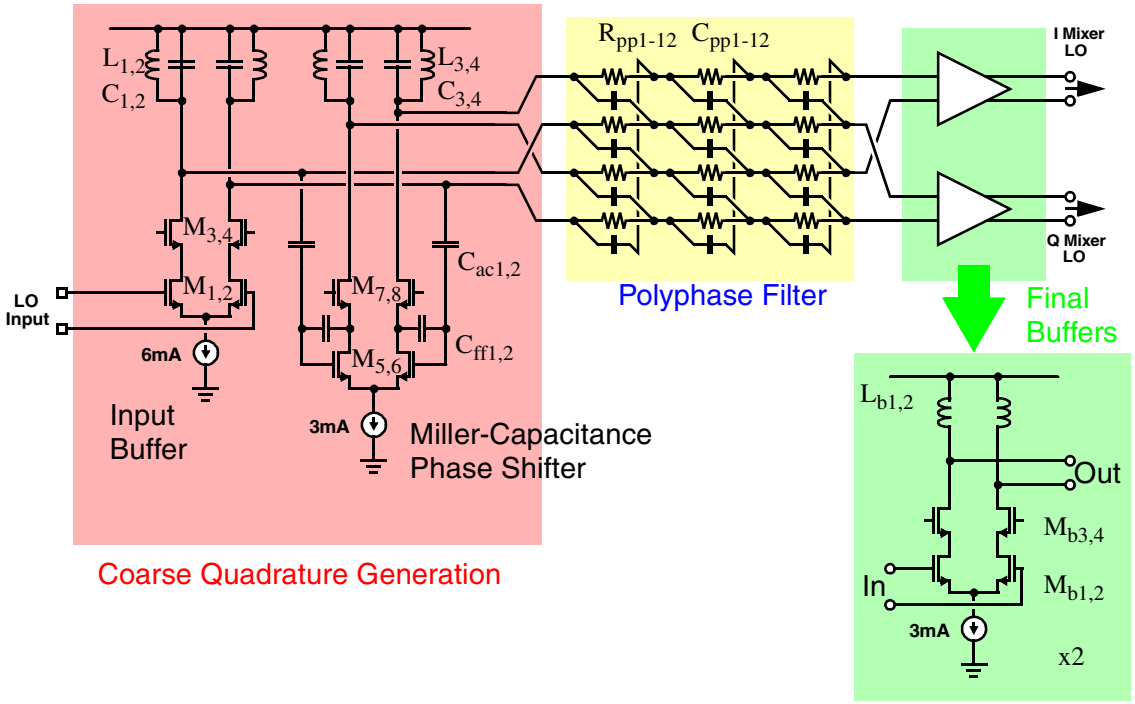


Fig. 5.17: Quadrature LO Generation Circuitry

	$M_{1,2}$	$M_{3\text{-}8}$	$L_{1,2}$	$L_{3,4}$	$C_{ac1,2}$	$C_{ff1,2}$	$R_{p1\text{-}12}$	$C_{p1\text{-}12}$	$M_{b1,2}$	$M_{b3,4}$	$L_{b1,2}$
Size	120µm/ 0.18µm	100µm/ 0.18µm	6nH	10nH	800fF	2.75pF	606	150fF	150µm/ 0.18µm	100µm/ 0.18µm	8nH

Table 5.5: Quadrature LO Generator Device Sizes

$M_{1\text{-}4}$ and associated load form an LO signal driver, with input gates connected directly to bonding pads. Input termination is done with off-chip resistors, and DC blocking of the gate bias is also off-chip. Transistors $M_{5\text{-}8}$ and their associated load, with Miller feedback capacitances $C_{ff1,2}$ generate a quadrature version of the LO signal[53]. The accuracy of the resulting 90° shift depends on the matching of the device transconductance to the Miller capacitor

impedance at the LO frequency, and with their different loading (one phase being loaded by the quadrature buffer and the other not), the gains of the in-phase and quadrature LO signals are not inherently well matched, so the LO signal from these drivers is considered only a coarse quadrature.

This coarse quadrature signal is then fed through a 3-stage asymmetric polyphase filter, each stage being identical in topology to that described in Section 5.2.2.1, except the component values are chosen to null the unwanted fundamental-frequency image rather than the third harmonic. All three stages were mistakenly implemented with the same notch frequency - better tolerance to component values could have been achieved had the notch frequencies been staggered.

The polyphase filter improves the matching and phase relationships of the signals at the expense of signal amplitude, thus a set of buffers is added at the polyphase filter output to restore signal swing and drive the mixers.

5.4 Downconversion Mixers

The function of the downconversion mixer is complementary to the upconversion mixer, taking the modulated RF signal and demodulating it back into its complex envelope. Recall from (Eq 2-8) that:

$$x(t) = \frac{\tilde{x}}{2}e^{j\omega_c t} + \frac{\tilde{x}^*}{2}e^{-j\omega_c t} \quad \text{(Eq 5-1)}$$

and consider that multiplying by $2e^{-j\omega_c t}$ yields:

$$x(t) \cdot 2e^{-j\omega_c t} = \tilde{x} + \tilde{x}^* e^{-2j\omega_c t} \quad \text{(Eq 5-2)}$$

The $\tilde{x}^* e^{-2j\omega_c t}$ term is a modulated signal at a high frequency of $2\omega_c$, and is easily removed with low-pass filtering. Thus, the complex envelope is readily recovered by multiplying the RF signal by $e^{-j\omega_c t}$, the real and imaginary parts of which are just I and Q phases of the local-oscillator signal. This is the same basic operation as the upconversion mixer, so the current-commutated mixer topology shown in Figure 2.5 is a reasonable starting point for designing a downconverter. One channel of a quadrature downconverter is shown in Figure 5.18

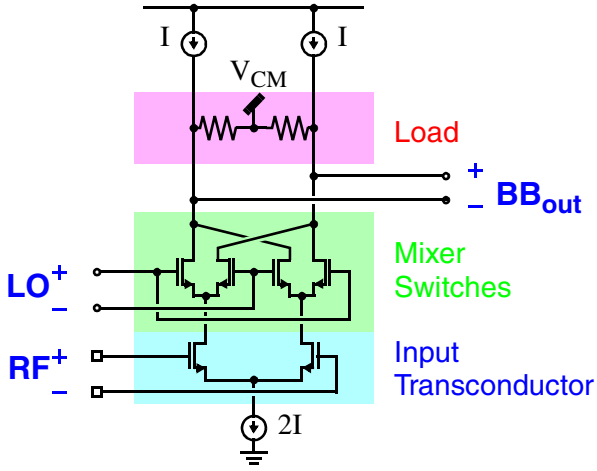


Fig. 5.18: CMOS Downconverter

The tuned LC load of the upconverter is replaced with a resistor load as the outputs are baseband rather than at RF, but otherwise the topology is the same. This topology was used as the initial design for the prototype's downconverter, however it was found that the mixer switches contributed significant 1/*f* noise at the mixer output, with the 1/*f* noise corner of the initial design being around 1.5MHz. As the function of Cartesian Feedback makes the closed-loop transmitter operation depend on the downconverter rather than the forward path, this downconverter noise would be converted into close-in noise at the PA output.

The 1/*f* noise of the switch transistors depends on the transistor gate area and on the DC current. Increasing the size of the switches to reduce their noise is not practical as by the time any substantial noise improvement is achieved, the gate capacitances that the LO must drive become unreasonably

large. Thus, the circuit was instead re-arranged to eliminate the DC current through the mixer switches, as shown in Figure 5.19

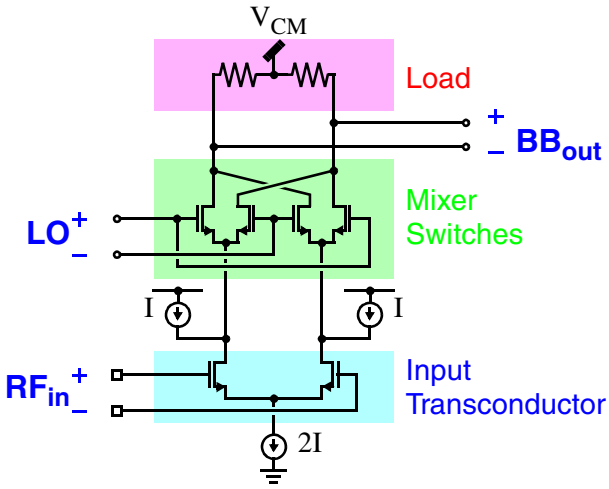


Fig. 5.19: Passive CMOS Downconverter

This topology is in effect a standard passive mixer, with the input transconductor and current bleeds acting as an RF buffer in front of the mixer, and a resistor loading the mixer output. The flicker noise of the mixer is greatly improved, however linearity suffers. In the initial design, the switch transistors would switch between cutoff and saturation mode operation, both of which are insensitive to drain voltage. With the DC bias current removed, the switch transistors would operate in triode mode at times, and the current they pass is then a function of the output voltage. The output voltage skews the time in the LO cycle that the RF signal current is transferred from one switch transistor to its partner; this modulation of the switching by the output signal introduces unwanted distortion.

As the closed-loop operation of the transmitter depends on linearity of this mixer, this distortion is best avoided. The baseband output voltage from the mixer switches is central to this distortion, but this is not a necessary or even useful output! The mixer output needs to be subtracted from the transmitter's baseband input signal, and such a subtraction is inherently current mode, so any output voltage of the mixer is only an intermediate step from the mixer output current before being converted back into a current for the subtraction. The mixer was further modified to suppress this voltage as shown in Figure 5.20

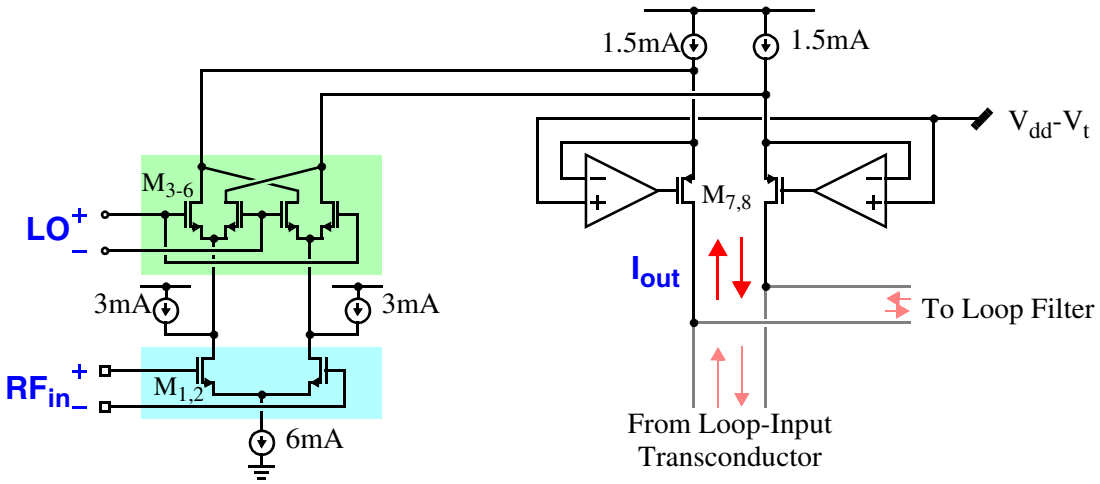


Fig. 5.20: Final CMOS Downconverter Design

	$M_{1,2}$	$M_{3\text{-}6}$	$M_{7,8}$
Size	150µm/ 0.2µm	20µm/ 0.18µm	200µm/ 0.18µm

Table 5.6: Downconverter Device Sizes

Cascode transistors $M_{7,8}$ present a low g_m^{-1} impedance to the mixer core, and associated op-amps further improve the quality of the virtual ground. The op-amps are given a reference voltage V_t below the LO common-mode voltage

of V_{dd}; this places the switch transistors on the edge of turn-on during the LO zero crossings. Signal current from the mixer switches passes straight through the cascode transistors and is delivered to the summing node where it is combined with the loop-input current and sent to the loop filter.

The op-amps used for this active cascode structure consume 1.78mA each, not counting the 1.5mA going through the cascode device itself which then goes through the loop-input transconductor.

The complete mixer consists of two instances of the circuitry shown in Figure 5.20, one for I and one for Q, with the RF inputs for the two instances connected together and biased on-chip and brought to bonding pads. As with the LO signal input, DC blocking of the bias voltage, and termination is are done off-chip with two 25Ω resistors.

Noise performance of the downconverter was simulated with Spectre-RF, and input-referred noise is shown in Figure 5.21. The 1/*f* corner is around

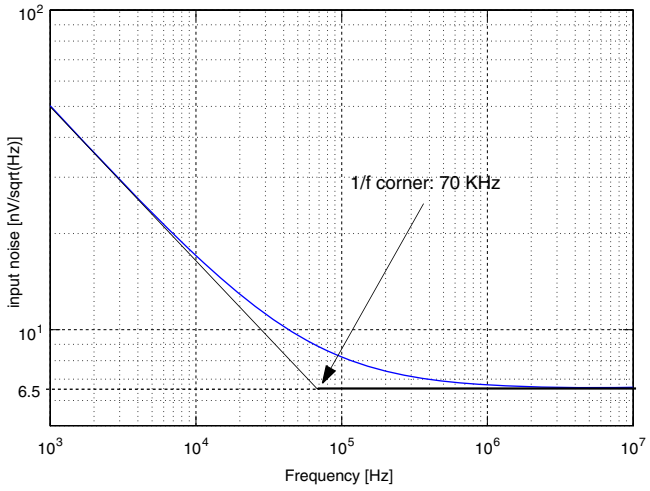


Fig. 5.21: Downconverter Noise Performance

70kHz, and is greatly improved from the initial design. Noise above the 1/*f* corner is dominated by the input transconductor devices $M_{1,2}$.

Two-tone tests were also simulated as shown in Figure 5.22. Equal-

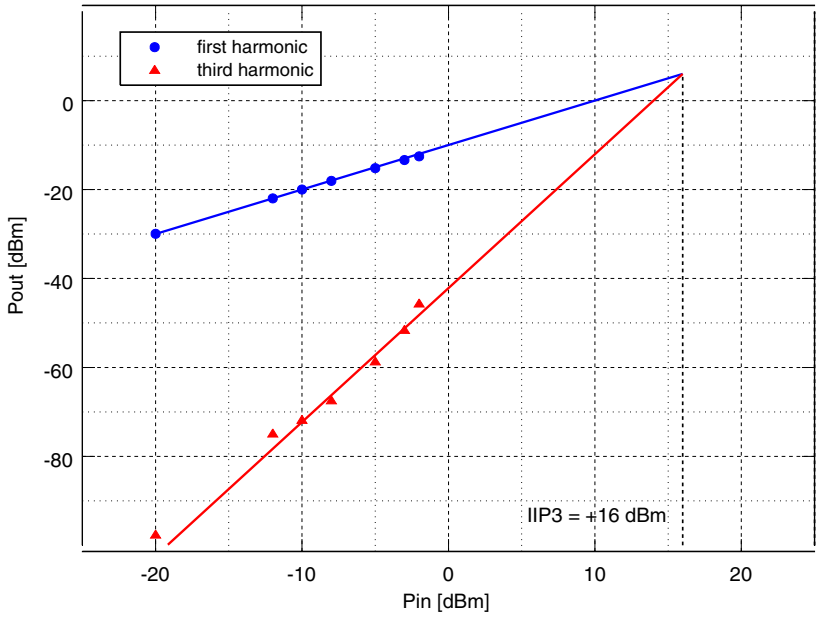


Fig. 5.22: Downconverter two-tone test simulation results

power input tones at 1.748 and 1.749 GHz were applied, with an LO input at 1.75GHz, producing linear downconversion products at 1MHz and 2MHz, and third-order products at DC and 3MHz, and current output after the cascodes is put on simple resistive loads. The input-referred IP3 intercept is extrapolated to be 16dBm. Again, the dominant source of distortion is the input transconductor.

A direct descendant of the downconverter in this prototype, integrating a more carefully designed current-recycling LNA, passive mixer core, and virtual-ground transimpedance amplifier is found [54].

Subsequent to the design of this downconverter, earlier examples were found of this technique of eliminating DC current from the switch devices and

loading the switches with a virtual ground, apparently starting with [55]. An investigation of 1/*f* noise in this topology is presented in [56].

5.4.1 V_{dd}-V_t Reference Voltage Generation

The V_{dd}-V_t reference voltage for the active-cascodes is generated by the circuit shown in Figure 5.23

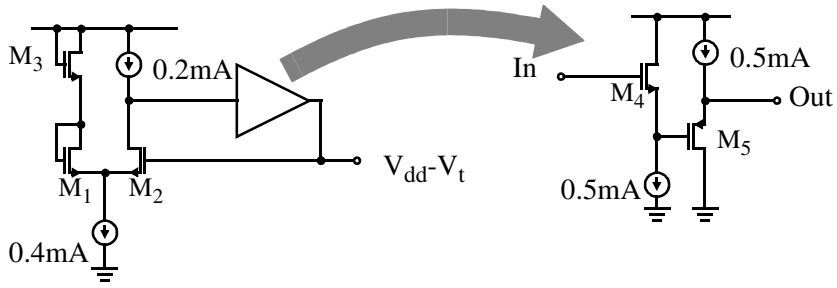


Fig. 5.23: Downconverter Voltage Reference

	$M_{1,3}$	M_2	M_4	M_5
Size	20µm/ 0.3µm	5µm/ 0.3µm	40µm/ 0.2µm	40µm/ 0.4µm

Table 5.7: Downconverter Reference Device Sizes

Assuming square-law operation, diode-connected transistor M_3 drops V_t+V_{dsat} from V_{dd}. Transistors M_1 and M_2 form a differential pair that compares this voltage with the reference output. M_2 is sized with one quarter the W/L of M_1, and thus needs twice the V_{gs}-V_t of M_1; this gives the differential pair a built-in input offset voltage of V_{dsat-1}, which sets the M_2 input at V_{dd}-V_t. Transistors M_4 performs a level-shift to drive source-follower M_5 that gives a low output impedance.

The 0.2mA current source at the drain of M_2 is the output of a current mirror that consumes another 0.2mA on its input, and counting this current, the reference consumes a total of 1.6mA. This reference was overdesigned, however the circuit was not re-visited after initial rough design. The M_4,M_5 followers are overkill given that the circuit only needs to drive the input gates for the active-cascode op-amps, and the currents through the other devices could have been scaled down with no ill effect. Eliminating the source-followers, M_2 is just another diode-connected transistor along with M_1 and M_3.

5.4.2 Downconverter Test Outputs

The active-cascodes shown in Figure 5.20 are simplified from the actual circuitry. With the output current being delivered to the summing node, the downconverter output is not readily accessible for testing. Although the summing node itself could have been brought off chip, a switchable output for

the mixer was instead implemented, one channel of which is shown in Figure 5.24.

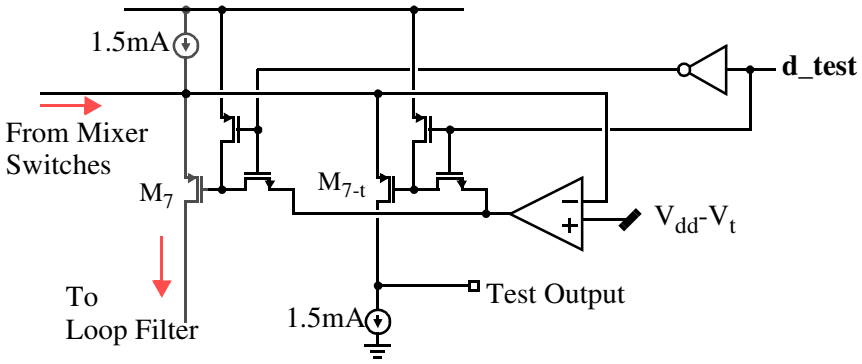


Fig. 5.24: Downconverter Test Output Switch

M_7 corresponds to the same device from Figure 5.20, and is twinned with a test replica $M_{7\text{-}t}$. The feedback op-amp output is fed to only one of the two devices, as selected by the digital control voltage, **d_test**, and the gate of the unused cascode device is pulled to V_{dd}. The test output is sent to a bonding pad where it is taken off-chip. A 1.5mA dummy current sink substitutes for the input transconductor on this test channel.

This switching consumes no static power, and the effect of the series resistance between the op-amp and cascode gate is negligible.

5.5 Loop Filter

The loop filter implements the Controller transfer function C(s) from the downconverter output to the upconverter input. C(s) creates a loop transfer-function L(s) satisfying (Eq 4-2). Within the loop bandwidth, the downconverter and upconverter are assumed memoryless, and are thus simply constant factors within the loop transfer function. The upconverter was designed to take a 0-p voltage of 250mV for nominal signal levels, while the downconverter was designed to produce 1mA in response: this is effectively a transconductance of $(250\Omega)^{-1}$ in the loop, thus the loop filter has a transfer function of:

$$C(s) = 250\Omega \cdot L(s) = \frac{250\Omega \cdot 2\pi 10^6 (s + 2\pi 10^5)(s + 2\pi 10^6)(s + 2\pi 10^7)}{(s + 2\pi 100)(s + 2\pi 10^{4.5})(s + 2\pi 10^{5.5})(s + 2\pi 10^{6.5})} \qquad \text{(Eq 5-1)}$$

which can be broken into two portions, the dominant pole (integrator) and lag-compensation network:

$$C(s) = C_{int}(s) L_{lag}(s) \qquad \text{(Eq 5-2)}$$

where:

$$C_{int}(s) = \frac{250\Omega \cdot 2\pi 10^6 10^{1.5}}{(s + 2\pi 100)} \qquad \text{(Eq 5-3)}$$

and:

$$L_{lag}(s) = \frac{(s + 2\pi 10^5)(s + 2\pi 10^6)(s + 2\pi 10^7)}{10^{1.5}(s + 2\pi 10^{4.5})(s + 2\pi 10^{5.5})(s + 2\pi 10^{6.5})} \qquad \text{(Eq 5-4)}$$

These two portions of the transfer function are implemented in separate sections of the loop filter. Both I and Q paths of the loop filter are identical, one path of which is shown in Figure 5.25.

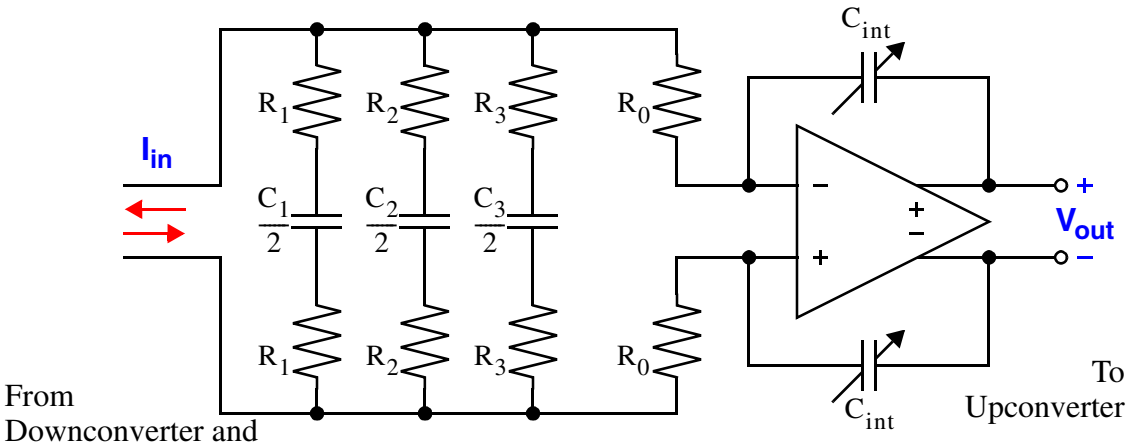


Fig. 5.25: Loop Filter

	R_1	C_1	R_2	C_2	R_3	C_3	R_0	C_{int}
Size	171Ω	93pF	716Ω	222pF	2490Ω	638pF	4kΩ	selectable 75fF to 19.125pF

Table 5.8: Loop Filter Device Sizes

The lag compensation is implemented as a passive current divider consisting of resistors $R_{1\text{-}3}$ and capacitors $C_{1\text{-}3}$, which bypass high-frequency current that would otherwise go to R_0. Variants of this network are well known for audio applications for pink-noise generation [57][58] and have also been put forth for loudspeaker impedance compensation [59].

Derivation of the component values for the lag compensator is given in Appendix B, with the value of R_0 being somewhat arbitrary. Component values

were originally selected for use as a voltage divider that would share the upconverter's input resistor (R_1 in Figure 5.11), but sizes were kept after the topology was rearranged to the final design. Using larger resistor values would allow smaller capacitances at the expense of larger voltage swing at the filter input.

Current from the lag compensator goes to a Miller integrator consisting of an op-amp and integrating capacitors C_{int}, that implement the dominant pole. The actual pole frequency in (Eq 5-3) is relatively unimportant as the dominant pole in (Eq 4-2) was only meant to model a practical implementation of an ideal integrator. To estimate the integrating capacitance needed, note that:

$$C_{int}(s) \approx \frac{250\Omega \cdot 2\pi 10^6 10^{1.5}}{s} \approx \frac{1}{s \cdot 20.1\text{pF}} \qquad \text{(Eq 5-1)}$$

The actual integrating capacitors are implemented as switchable binary-weighted capacitor arrays allowing freedom to adjust the loop bandwidth in testing. The arrays each have a total of 255 75fF MIM capacitors for a maximum integrating capacitor of 19.125pF, which is slightly shy of the calculated value. This makes the minimum integrator gain slightly in excess of what would implement (Eq 4-2), but it was presumed that insufficient loop gain is more likely to be a problem than excess gain, and extra attenuation is readily added in the off-chip portion of the loop if needed.

The op-amp for each channel consumes 1.4mA, plus another 200µA for output common-mode feedback. The biasing circuitry for the op-amp allows for the op-amp output stage to be completely turned off, leaving the output open-circuit. This together with switching out all the integrating capacitors leaves the loop filter invisible to the upconverter when testing the upconverter by its external inputs.

5.6 Loop-Input Transconductor

The baseband input signal for closed-loop operation is injected into the feedback loop where the downconverter output is fed to the loop filter. The transconductor used for this input is based on the transconductor from the upconverter, and is shown in Figure 5.26.

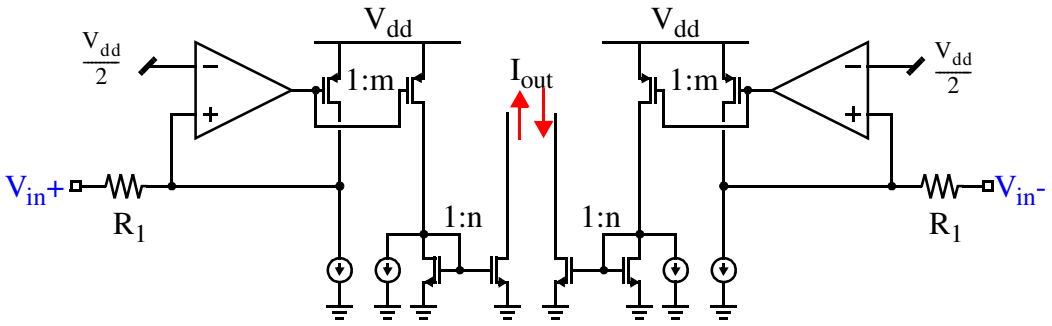


Fig. 5.26: Loop-Input Transconductor

	R_1	m	n
Size	2kΩ	1	4

Table 5.9: Transconductor Component Sizes

This transconductor functions identically to the upconverter, with some minor scaling differences as the operating requirements are somewhat different. The linearity of this transconductor is more important than in the upconverter, as this block is outside of the feedback loop. Eliminating R_2 and R_3 of the upconverter transconductor helps linearity, as the PMOS transistor at the output of the op-amp then sees a larger load impedance, and the feedback factor back to the op-amp input is increased to unity; both effects increase loop gain for this servo loop, at the expense of losing the $\frac{R_2 + R_3}{R_3}$ factor in transconductance. However, less transconductance is needed, as the output current required is much weaker than in the upconverter; the output current here only needs to match the output current produced by the downconverter rather than drive the upconverter's output load. As the input is driven by an external signal source (with 50Ω source impedance) rather than a weaker on-chip amplifier (of the loop filter), a smaller input resistance R_1 can be used making up some of the lost transconductance.

The op-amps used in this transconductor are identical to the upconverter, as are the PMOS transistors on their outputs. More current is sent to the final NMOS output mirror instead of being shunted into the DC current sink though; this is to improve the linearity of the NMOS current mirror. Active power consumption (excluding final output current) is identical to the

upconverter transconductor, with 0.55mA consumed in each op-amp, plus 0.82mA per op-amp for their PMOS output and replica devices. The final output signal is on top of a DC output current of 1.5mA.

5.6.1 Common-Mode Feedback

The output of this transconductor goes into the same summing node as the downconverter output, and this node is loaded by the loop-filter input which is a DC open-circuit. The differential voltage here is stable by virtue of the Cartesian-Feedback loop operation, but there is no inherent mechanism to set the common-mode voltage, thus a common-mode feedback (CMFB) loop is needed. The CMFB circuit is shown in Figure 5.27.

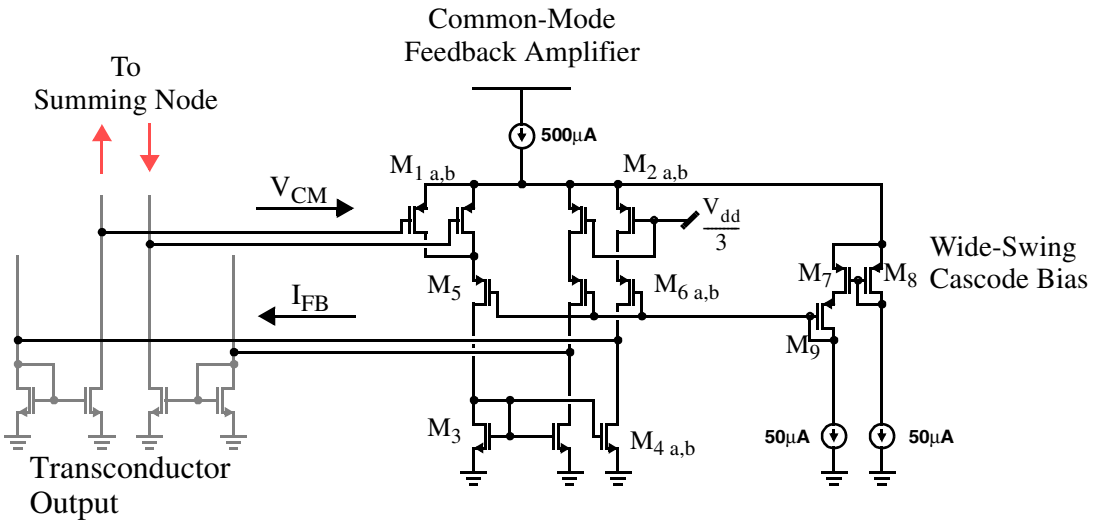


Fig. 5.27: Input Transconductor Common-Mode Feedback Circuit

	$M_{1a,b}$ $M_{2a,b}$	M_3	$M_{4a,b}$	M_5	$M_{6a,b}$	M_7	M_8	M_9
Size	23.2µm/ 0.18µm	40µm/ 1µm	20µm/ 1µm	38.4µm/ 0.18µm	19.2µm/ 0.18µm	1.16µm/ 0.18µm	0.58µm/ 0.18µm	9.6µm/ 0.18µm

Table 5.10: CMFB Device Sizes

This circuit is essentially a modified single-ended op-amp. M_1 and M_2 form a differential 'pair' that compare the output common-mode voltage with the reference. M_3 and M_4 form a current mirror that takes the drain current of M_1 and adds it to that of M_2, and the sum is fed back to the transconductor's output current mirror to adjust the common-mode voltage. Cascode transistors M_5 and M_6 hold the drains of M_1 and M_2 at about the same voltage to improve input offset.

The common-mode output voltage is sensed by splitting M_1 in two and sensing each side of the summing node voltage independently, rather than the more typical method of using a pair of resistors across the output. Resistor loading would degrade the differential-mode output impedance and divert signal current that should go to the loop filter. Splitting M_1 in this manner is not robust in the presence of large differential voltages, but the differential voltage remains small by operation of the cartesian-feedback loop, thus the two constituent transistors remain biased the same.

M_2, M_4 and M_6 are similarly split to provide two output currents instead of one.

M_9 sets the gate voltage of cascode transistors M_5 and M_6 so that M_1 and M_2 have about the same V_{ds} as M_7. M_7 is sized twice as wide as M_8 which carries the same drain current with the same V_{gs}, thus M_7 operates in triode

region with a V_{ds} that reflects the V_{dsat} of M_8. M_8 is narrower than M_1 and M_2 and operates with a larger V_{dsat}, thus keeping the input devices out of saturation.

5.7 Transmitter Test Chip

The prototype was fabricated in a 0.18µm CMOS process by STMicroelectronics with the MIM capacitor option. The process is a triple-well process, and the major circuit blocks are each isolated in their own well. A die micrograph of the transmitter prototype chip is shown in Figure 5.28. Unlabelled structures in the figure are bypass capacitors.

The die is $3.5 \times 6\text{mm}^2$ including the pad ring. The LO input circuitry was placed in the opposite corner from the PA to minimize coupling that could cause distortion. Special RF bondpads with minimal ESD protection diodes were used for the LO and the downconverter input to minimize parasitic capacitance. Pads are spaced at 152.4µm pitch to match the minimum pitch of traces on the test board.

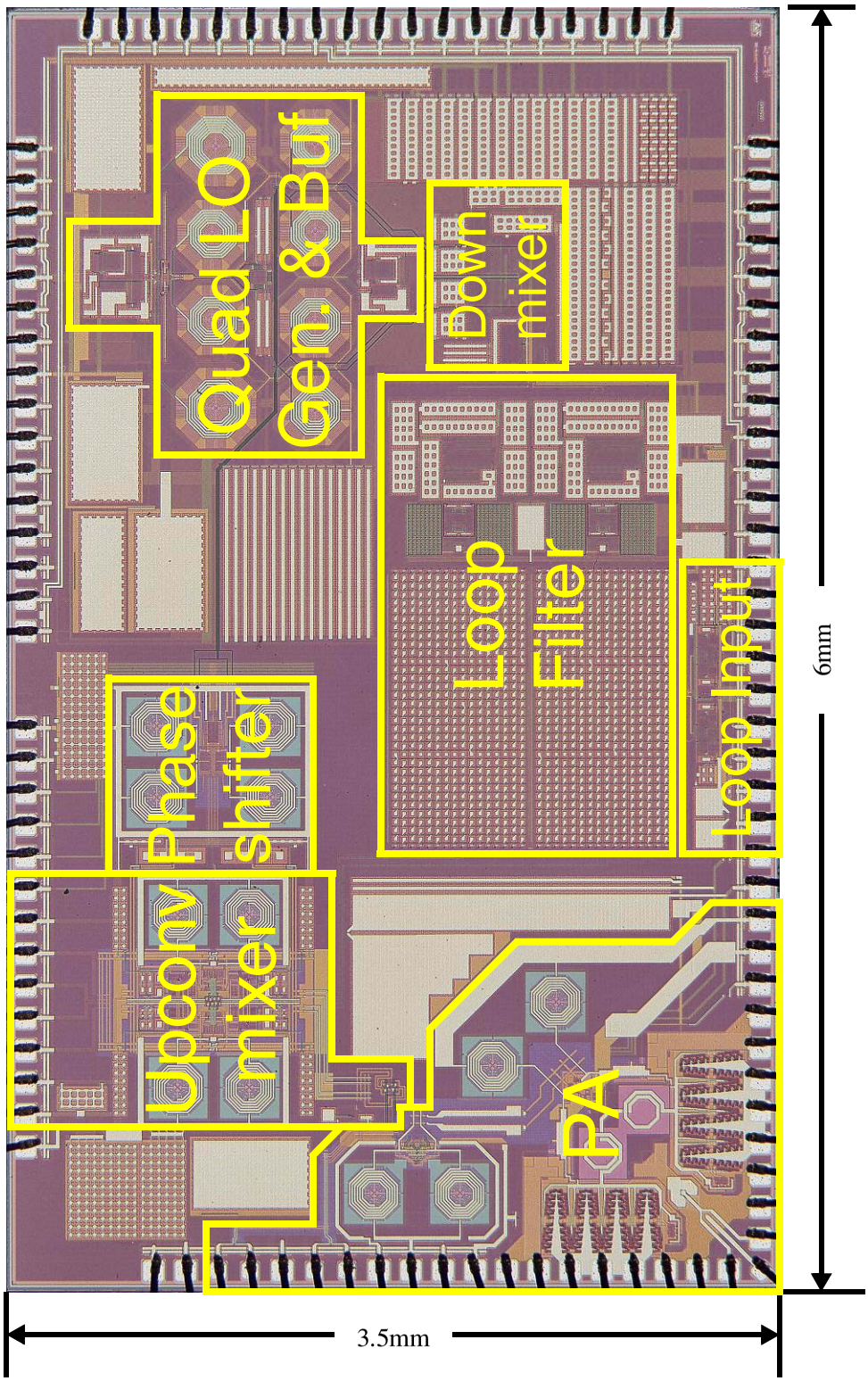


Fig. 5.28: Transmitter Test-Chip Micrograph

Chapter 6

Measurement Results

This chapter summarizes measurements that were performed on the experimental prototype. The goal of the measurements was to demonstrate linearized closed-loop operation of the PA, and as such, open-loop measurements of the individual circuit blocks was largely diagnostic, looking for 'signs of life' and not necessarily to fully characterize each block.

Although the transmitter was designed to operate at 1.745GHz, it was found in testing that the PA produced the strongest output at 1.55GHz. Some of the earlier measurements that had been performed at the design frequency were re-done at this frequency. Most of measurements described here are given at this operating frequency.

Due to equipment sharing contentions in the lab, many of the measurements were performed using a spectrum analyzer and analog oscilloscope that did not have a means to export data, thus spectrum plots and baseband waveforms were not recorded for these measurements.

6.1 Test Board

A test board was designed and fabricated with a standard FR4 material. Chip-On-Board assembly was used to mount the prototype chip, with the unpackaged die being directly attached to the circuit board. Traces on the board are gold-plated for bondability and 1.25mil gold bondwires connect the on-chip bonding pads to landing areas on the board.

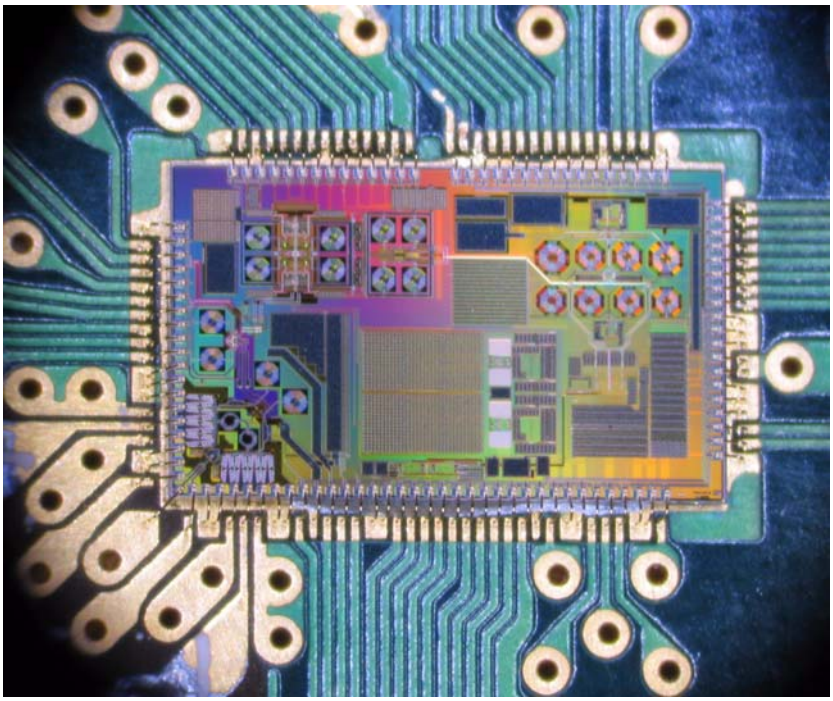


Fig. 6.1: Prototype chip on test board

The board was a 4-layer board, with most signal routing being done in the top layer, and some bias lines being diverted to the backside where necessary. The inner plane layer closest to the top is used as a ground plane,

while supply voltages for the chip are distributed in the bottom inner plane layer.

The test board is powered by a triple-output bench supply, with one positive supply providing power for the PA, and the other outputs providing +5V and -5V to power the prototype chip and supporting circuits on the board. The remainder of this section describes the supporting circuitry on the board.

6.1.1 RF Loop

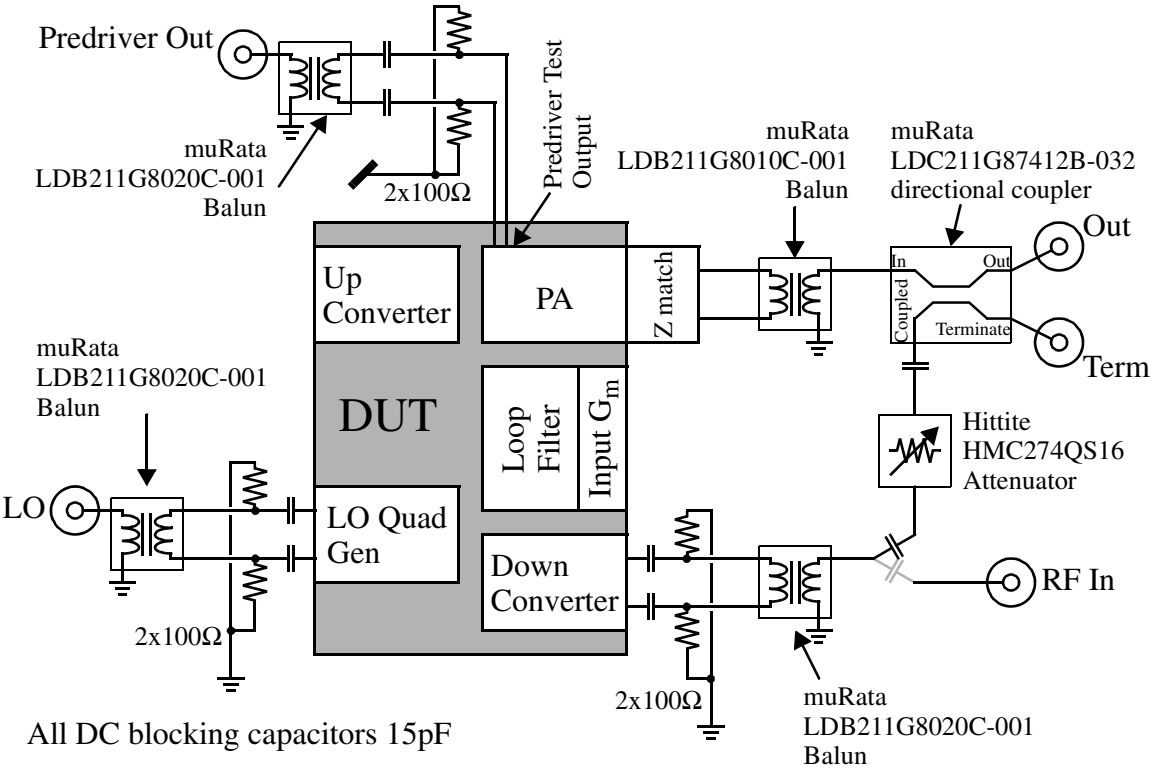


Fig. 6.2: Off-chip RF components

RF signals are carried on/off chip differentially, and are converted from/ to 50Ω single-ended signals with commercial baluns from muRata. An LDB211G8010C-001 balun (1.8GHz, 50Ω per side on the differential port) is used at the PA output, while the LO and downconverter RF signal inputs use LDB211G8020C-001 baluns (1.8GHz, 100Ω per side) instead. Although the chip was originally designed and simulated with 50Ω per side on these inputs, this termination is done off-chip, and was changed to 100Ω during board design for better voltage gain. The PA predriver test output also uses this balun.

The output of the PA passes through a muRata LDC211G7412B-032 directional coupler before going to an SMA connector. This coupler is designed for use at 1.74GHz, with a coupling factor of 12.8dB. The coupler has two other ports for the coupled line: one is the coupled output which goes back to the downconverter, and the other port is sent to another SMA connector for termination. At 1.55GHz, the coupler was measured as having an insertion loss of 0.5dB on its main line, and a coupling factor of 14.6dB to the coupled output.

Between the directional coupler and the downconverter input balun, the loop is closed with a Hittite HMC274QS16 programmable attenuator which provides from 0-31dB of attenuation (plus insertion loss of about 2dB), selectable in 1dB steps. The attenuator is biased by its RF terminals and requires DC blocking capacitors on both its input and output.

An extra solder pad leading to an SMA connector is placed beside the blocking capacitor between the attenuator and the downconverter input balun. By re-soldering the bypass capacitor between this pad and the pad for the downconverter input, the RF loop can be broken, and a signal fed directly to the downconverter.

6.1.2 Downconverter Test Output

The downconverter test output is a differential current signal. To convert this to a single-ended voltage for measurement, two Maxim MAX4145 Differential Line Receiver chips were used, one each for the I and Q test outputs. The circuit for one path is shown in Figure 6.3

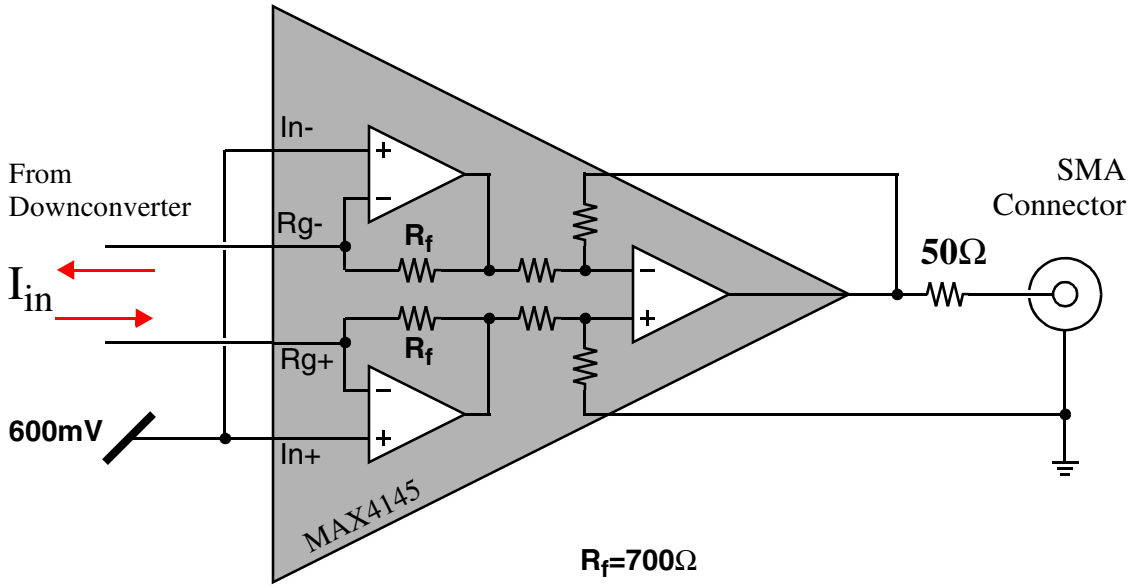


Fig. 6.3: Downconverter test output trans-resistance amplifier

The MAX4145 is intended for use as a voltage amplifier, taking a differential voltage input on In+ and In-, and produces a single-ended output. Gain resistor pins Rg+ and Rg- allow setting higher gains by adding an external resistor, analogous to R_3 in the transconductor shown in Figure 5.12. Connecting the voltage inputs to a reference voltage (taken from a resistor divider on a regulated supply voltage), the gain resistor pins become virtual grounds biased at the reference voltage and can be used as current-mode inputs. The receiver chip can thus function as a trans-resistance, the output of which can be observed on an oscilloscope.

6.1.3 Other supporting circuitry

Voltage supplies for all blocks of the prototype except the PA output stage are regulated on-board using National Semiconductor LM317 3-terminal voltage regulators. Each supply is bypassed with a 10μF tantalum capacitor near the regulator, and a series of progressively smaller capacitors down to 3pF closer to the chip. A series jumper and choke inductor are inserted between each regulator and the bypass capacitors. The jumpers allow insertion of a multimeter to measure current provided by the regulator.

Reference voltages for the chip are generated on the board with resistor dividers across these regulated voltages, and are buffered with National Semiconductor LM8272 Op-Amps configured as voltage followers. An LM317

regulates a dedicated positive supply for these op-amps, and the negative supply needed by the op-amps is regulated with a National Semiconductor LM337.

Current references for the chip are implemented using National Semiconductor LM334 current sources. These sources are actually voltage regulators of a fashion, series-regulating a fixed 64mV voltage across a SET resistor. The resistor current, plus the regulator's bias current (about 6-7% of the set resistor current) sum to a total output current of $I_{set} \approx \frac{68mV}{R_{set}}$. A typical current source circuit is shown in Figure 6.4.

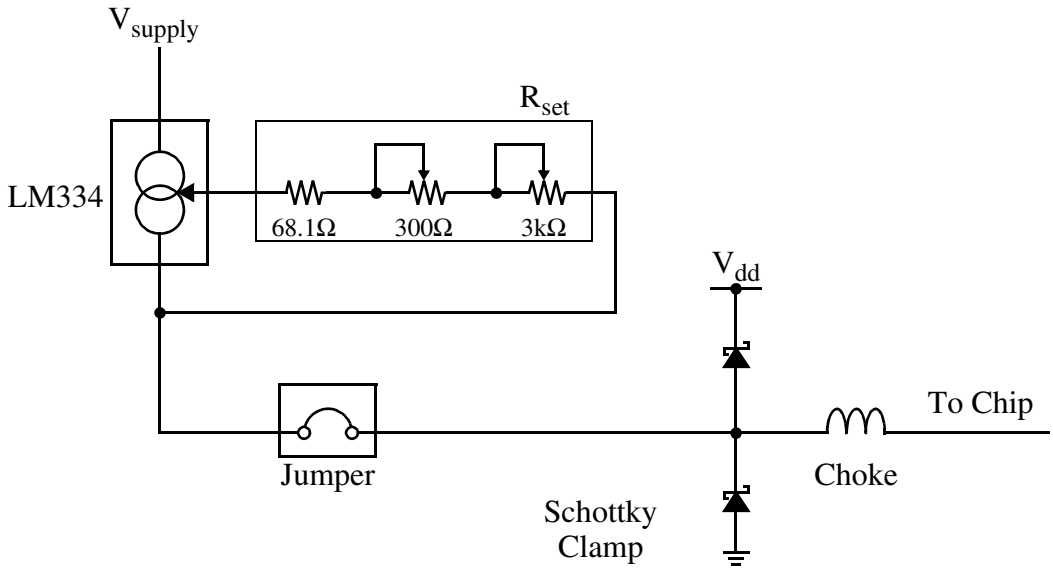


Fig. 6.4: Typical test-board current reference

The series combination of one fixed and two variable resistors forms the set resistor. The fixed resistor limits the minimum resistance, hence maximum current from the current source, while the two variable resistors allow for

coarse/fine adjustment. Resistor values shown are for the LO phase shifter sin/cos current inputs - other current sources use different resistor values for other current levels. The jumper allows measuring or disconnecting the delivered current. Schottky diodes to V_{dd} and ground are used for extra protection from ESD events that could happen at the jumper. A series choke between the diodes and the chip offers further protection, as well blocking any high-frequency noise.

6.1.4 Bugs

Several minor errors were made in the design of the test board. The most serious of these was that the wrong pinout had been used for the baluns when laying out the board. The baluns were re-mounted on the board upside-down and rotated from the footprints to put the signal inputs in the proper order, and jumpers soldered in place to complete the ground connections. Inductance of these ground jumpers likely affects balun operation, but revised boards were not fabricated to properly correct for this.

6.2 Downconverter Test

The downconverter was tested with the DC blocking capacitor at its balun input routed to an SMA connector, and signal generators are used to drive both the LO and RF signal input. The downconverter was set to send its output

to the off-chip test outputs instead of the loop filter. These outputs were observed on an oscilloscope for I-Q testing. For two-tone and spectral mask tests, a spectrum analyzer was connected to one of the baseband outputs.

A block diagram of the test setup is shown in Figure 6.5

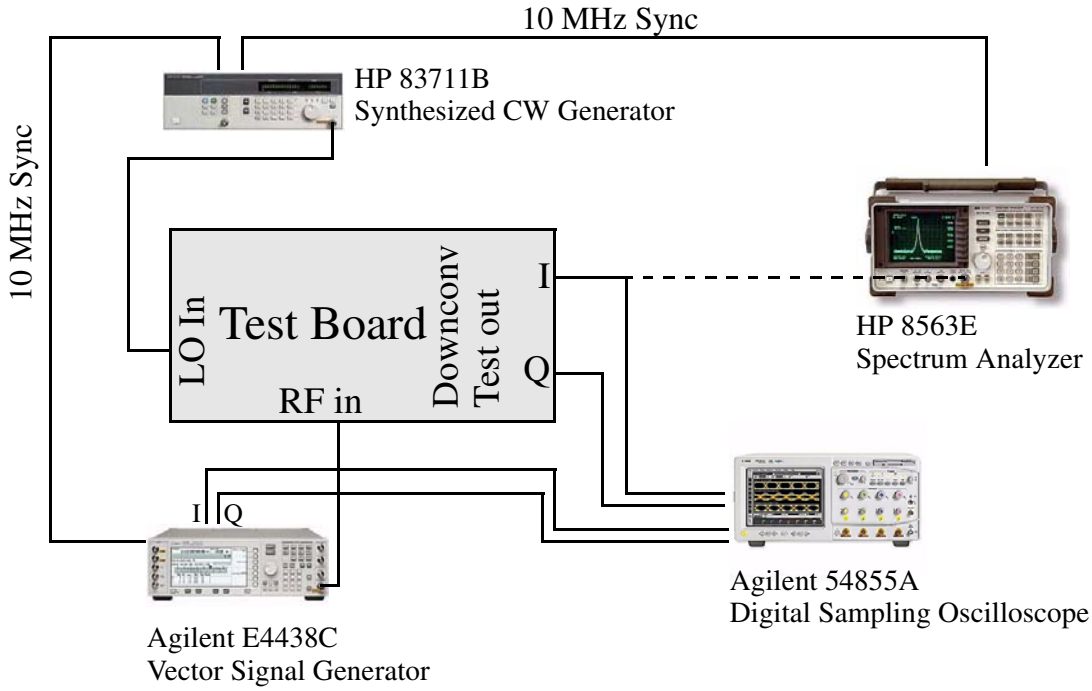


Fig. 6.5: Downconverter Test Setup

The Agilent E4438C Vector Signal Generator (VSG) has an Arbitrary Waveform Generator (AWG) module that can produce two differential baseband output voltages in addition to its RF output which can optionally be modulated by this AWG output. The AWG was programmed with waveforms for EDGE modulation, and also has built-in sinewave generation routines. Although the EDGE waveforms were available from the original programming process, they

were also captured with the oscilloscope for more convenient comparison against the measured downconverter output.

6.2.1 I-Q Demodulation Test

The downconverter was first tested with pure sinusoidal inputs for the LO and RF input signals at slightly different frequencies. The gain of the downconverter showed sensitivity to the LO signal level for small LO inputs, but this effect tapered off as the LO signal was increased to 5dBm and no differences were observed for LO signals larger than that. Presumably the input at this point is enough to fully current-commutate the LO input buffers, and additional input power causes no further changes. All measurements from this point on were performed with a 5dBm LO signal.

Gain mismatch and quadrature error were immediately apparent - the downconverted baseband sinewaves differed in amplitude by a factor of about 1.14x (1.15dB), and the phase difference between I and Q was about 110° (an error of 20° from quadrature). These errors appeared to hold steady over a wide range of power levels, and did not vary with the offset of the RF to LO frequency from several hertz up to several megahertz, and the same errors were seen on all three of the boards tested. There appears to be a systematic gain/ quadrature problem inherent in the implementation of the prototype's LO quadrature generation and downconverter, but this was not investigated further.

The downconverter was then tested with a modulated signal. A 1.55GHz LO signal was used, and a -5dBm EDGE modulated carrier was applied to the downconverter input. Figure 6.6 shows I-Q plots of the input (taken directly from the AWG outputs) and demodulator output signals.

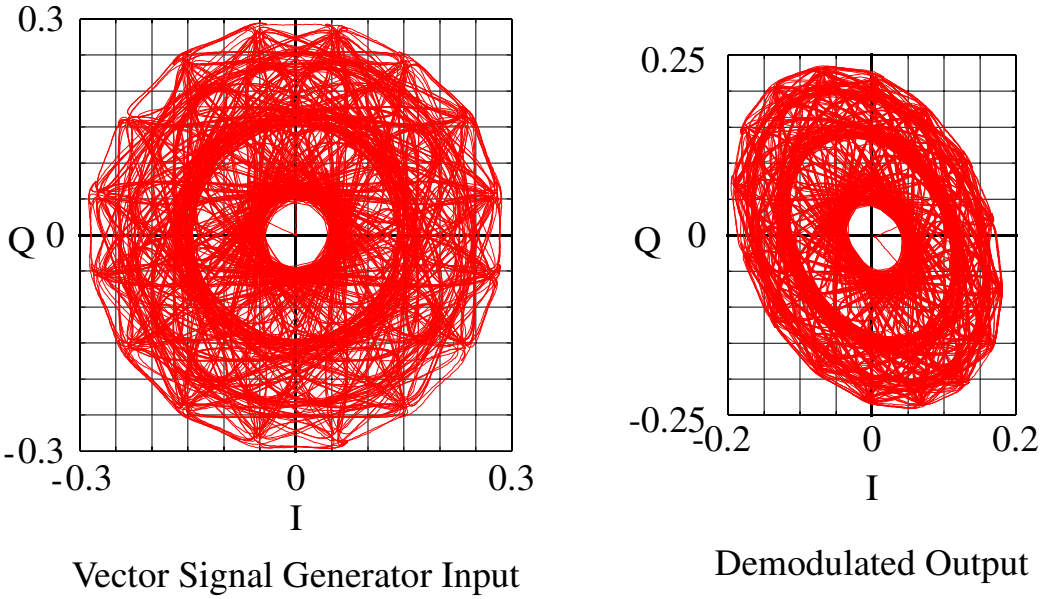


Fig. 6.6: Downconverter IQ demodulation test

The gain mismatch and quadrature error are apparent in the elliptical shape of the demodulated signal, but the output looks reasonable otherwise. No gross nonlinearity is observed, although the gain of the I channel appears to be larger for negative I than for positive I - this would imply that some even-order distortion is present. This even-order distortion indicates that some asymmetry is present in what is supposed to be symmetric differential circuitry. The source of this asymmetry was not pursued, although the input balun is suspected to play a role.

6.2.2 Two-Tone Test

To characterize the linearity of the downconverter, a standard two-tone test was performed. The VSG was set to produce two tones at $1.551\text{GHz} \pm 50\text{kHz}$. With the 1.55GHz LO signal, the linear downconversion products are at 950kHz and 1050kHz, while third-order products appear at 850kHz and 1150kHz. One of the baseband output channels was fed to the spectrum analyzer, and the observed magnitude of the 1050kHz and 1150kHz products are plotted as a function of the total input power in Figure 6.7.

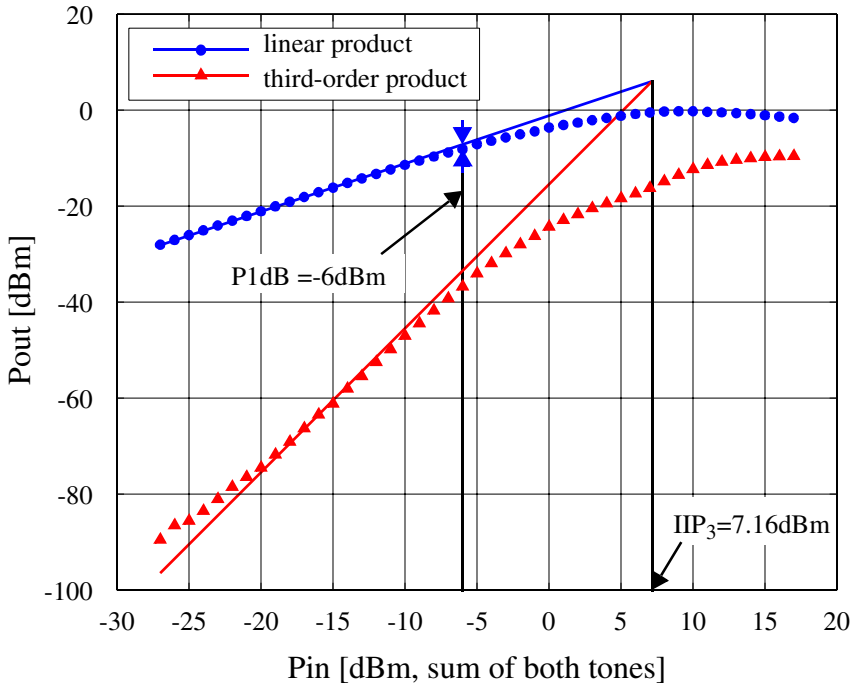


Fig. 6.7: Downconverter two-tone test

The input-referred IP_3 intercept point was found to be 4dBm per-tone (7dB total power for both tones) and the 1dB compression point is -9dBm per tone.

6.2.3 Spectral Mask

Although the 1dB compression and IP3 intercept points are common metrics for comparison, linearity requirements for GSM are captured in the spectral mask measurement rather than these two-tone test metrics. While the spectral mask is meant to be applied to the output of the PA, closed-loop operation of the transmitter relies on the linearity of the downconverter, and it is reasonable to apply the spectral mask here instead.

The spectrum analyzer used is only calibrated to measure down to tens of kilohertz not to DC, but the spectral mask is defined in terms of the spectral density at the channel centre, thus the spectral mask cannot be applied directly to the observed spectrum when downconverting to DC: the origin for the mask is unobservable. The spectral mask can be applied, however, to a low-IF downconversion.

An EDGE modulated 1.5506GHz signal was fed to the downconverter with a 1.55GHz LO, producing a low-IF received signal at 600kHz. The spectrum of this signal clears the spectral mask for positive frequency offsets (analyzer noise overwhelms the mask around DC or -600kHz offset) for input

powers up to –5dBm. At this power level, regrowth just clears the spectral mask at +400kHz offset.

Changing the signal's carrier frequency to match the LO, the downconverted spectrum shifts to DC and folds over from negative frequency into positive. While the spectral density at DC is obscured by the spectrum analyzer's noise, it can be estimated as being 3dB higher (from folding) than the channel-centre density of the low-IF signal: this estimate is used as the origin for the spectral mask test in lieu of a direct measurement of the density at DC. The spectrum analyzer noise falls below signal power in the tens of kilohertz, and for the mask from 200kHz offset and higher, the received signal is observed to clear the mask for signal powers up to -5dBm just as with the low-IF measurement.

Reducing the input signal level, the mask was cleared for signal powers down to -18dBm. At this point, the mask for >=600kHz offsets runs into the noise floor of the measurement. An Agilent E4440A spectrum analyzer was available briefly to repeat this measurement with, and using it, the spectral mask was cleared for signals down to -29dBm of power. It was later realized that the default input attenuator settings on the different spectrum analyzers were different, with the HP 8563E analyzer originally used having a 10dB attenuation on its input by default, raising its noise floor by the same amount. Thus, it is believed that the noise floor seen in these measurements is from the spectrum

analyzer and not the downconverter itself, and that the downconverter would still meet spectral mask requirements with even smaller signals at the downconverter input.

6.3 Upconverter/PA Test

To test the upconverter, the prototype's loop filter was deactivated and the upconverter test input was connected to the baseband AWG outputs of the Vector Signal Generator. A spectrum analyzer was used to observe the PA predriver test output, bypassing later PA stages. For PA testing, the PA output was measured at the directional coupler output terminal, and a 50Ω terminating resistor was connected to the directional coupler's TERM terminal. The DC blocking capacitor at the downconverter input was connected to the attenuator, closing the RF loop.

The LO signal was provided by a signal generator as with the downconverter tests, while the Vector Signal Generator's RF output was unused. The downconverter was left online for the upconverter tests, and its outputs

observed on an oscilloscope together with the AWG outputs. A block diagram of the test setup is shown in Figure 6.8

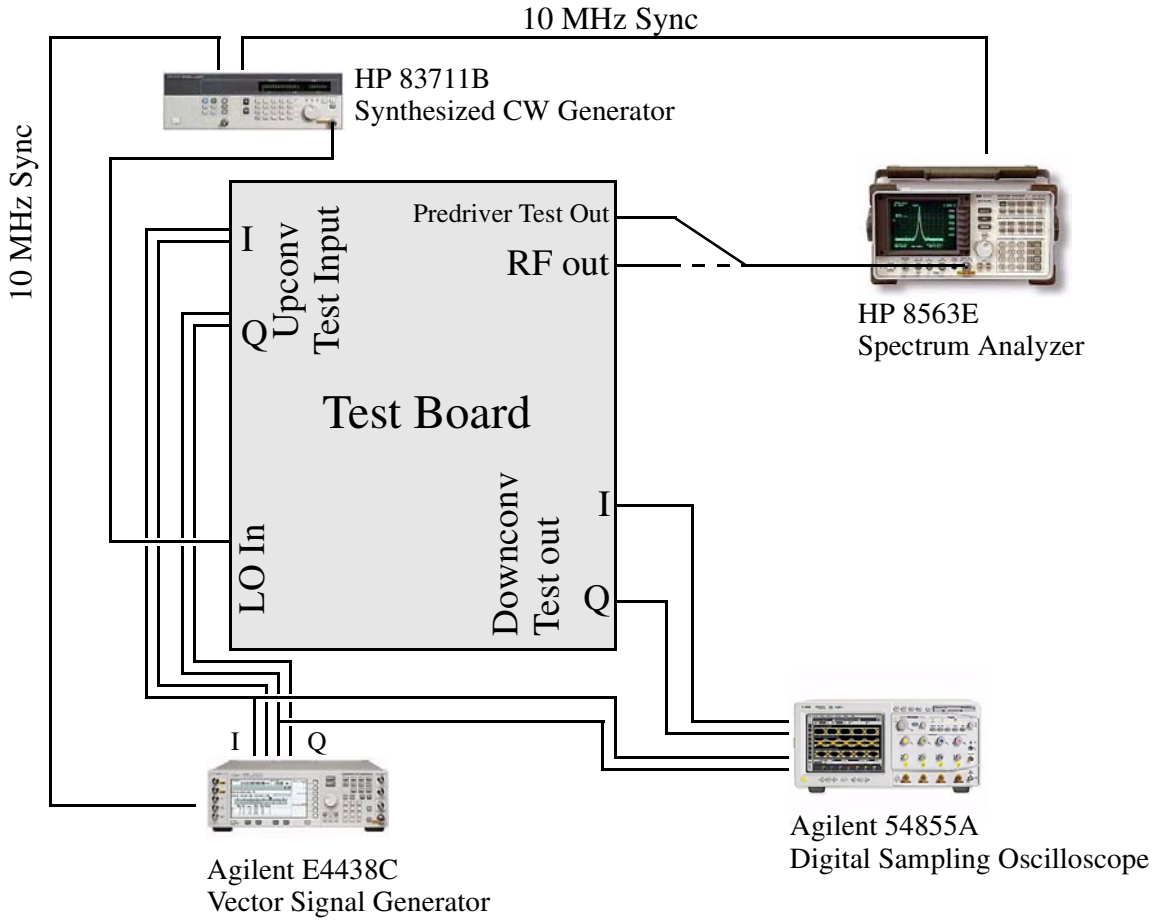


Fig. 6.8: Upconverter Test Setup

6.3.1 Upconverter SSB test

The AWG was configured to output 1MHz quadrature sinewaves for the upconverter I and Q test inputs. A 1.55GHz LO signal used, and the PA predriver test output was observed with a spectrum analyzer. The negative cos

input for the upconverter phase shifter was set to 0.5mA, and the other three inputs were left open-circuit.

DC offsets on the AWG outputs were adjusted to minimize the LO leakage of the upconverter, and the amplitude of the I channel relative to the Q was adjusted to minimize the image ratio. With a baseband amplitude of 300mV 0-peak (per-input single-sided), a -34dBm 1.549GHz output was observed from the predriver test output, and the 1.551GHz image was found to be 30dB smaller. LO leakage was smaller than the image, and second-order products at 1.552GHz and 1.548GHz were observed as being even smaller still (actual levels were not recorded). Third-order products were also observed, being between the LO leakage and the second-order product in magnitudes.

Reducing the input signal level, the image ratio remained relatively constant, while the third order products fell by three times as much as the linear tones: this is as expected. Increasing the input signal amplitude from this 300mV, the output tone increased less than the input signal (increasing only 3dB more for a 10dB larger baseband signal). More tones from higher-order distortion would rise out of the noise floor, and the image ratio became somewhat worse (decreasing to 24dB); this is presumably from the upconverter clipping and disrupting the balance of I to Q amplitudes. Changing the frequency of the baseband signals from tens of kilohertz up to a few megahertz did not appreciably change any of these effects.

In linear operation, with gain mismatches having been empirically adjusted out, the 30dB image ratio is presumed to be entirely from quadrature phase mismatch of the upconverter, with the phase error being estimated as:

$$\Delta\theta = 2\,\mathrm{atan}\left(10^{\left(\frac{-30\mathrm{dB}}{20}\right)}\right) \approx 3.6^\circ \qquad \text{(Eq 6-1)}$$

This is significantly better than the estimated quadrature phase error of the downconverter.

6.3.2 PA Output Power

The spectrum analyzer was moved to the RF output terminal of the directional coupler to observe the PA output power. Baseband sinewaves of the SSB test were still fed to the upconverter input, but with signal levels increased to about 1V 0-p: the upconverter is presumed to be saturated with this input and providing the largest signal it can to the PA. The PA was powered from a supply voltage of 2.3V, which is slightly reduced from the design value out of paranoia fear of overstressing the prototype.

Various values of shunt and series capacitances were tried, and the PA output power observed across frequencies from about 1.4GHz to 1.8GHz. The best power output was achieved with 3.0pF series capacitors and 4.3pF shunt capacitors in the matching network, producing just over 21dBm of output power at 1.55GHz. This is the power as measured after the directional coupler, after

balun and directional coupler insertion losses: the actual power at the PA output is slightly more.

Sweeping the baseband input amplitude (in dB scale), the output power and supply power for the PA output stage were recorded. The output power and overall large-signal gain (from baseband input to PA output) as a function of input power are shown in Figure 6.9.

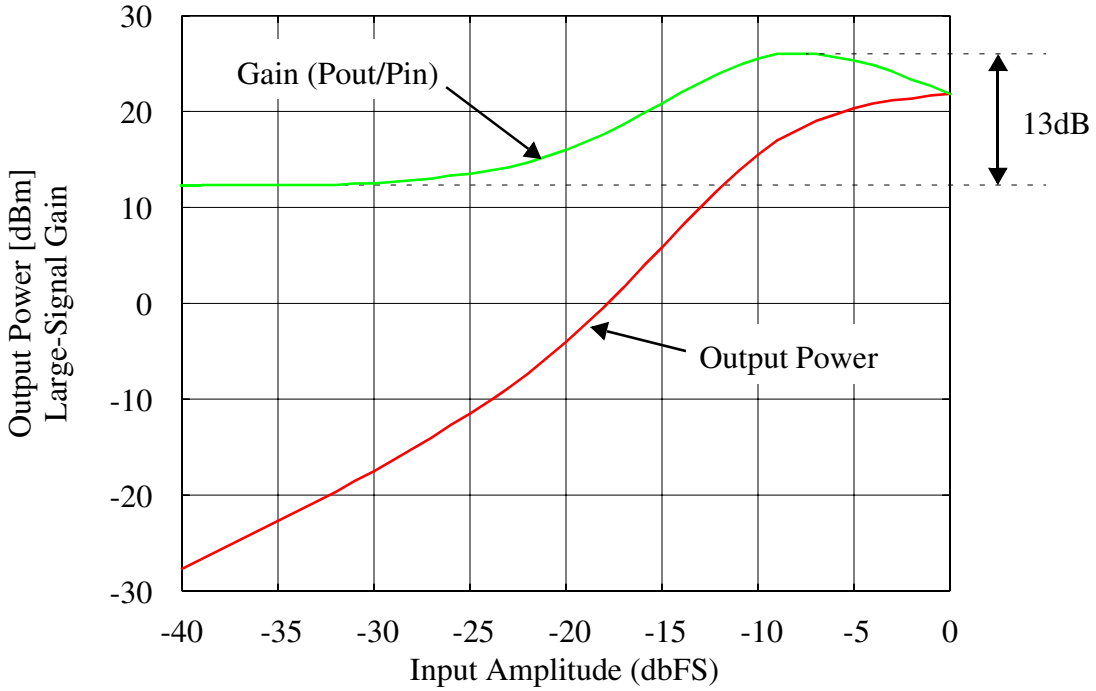


Fig. 6.9: Upconverter/PA AM/AM Transfer Function

The large-signal gain is seen to vary by 13dB, and the feedback needs to be robust across at least this range.

The PA drain efficiency as a function of the output amplitude (linear scale normalized to maximum power) is shown in Figure 6.10. The peak drain efficiency observed is just over 20%.

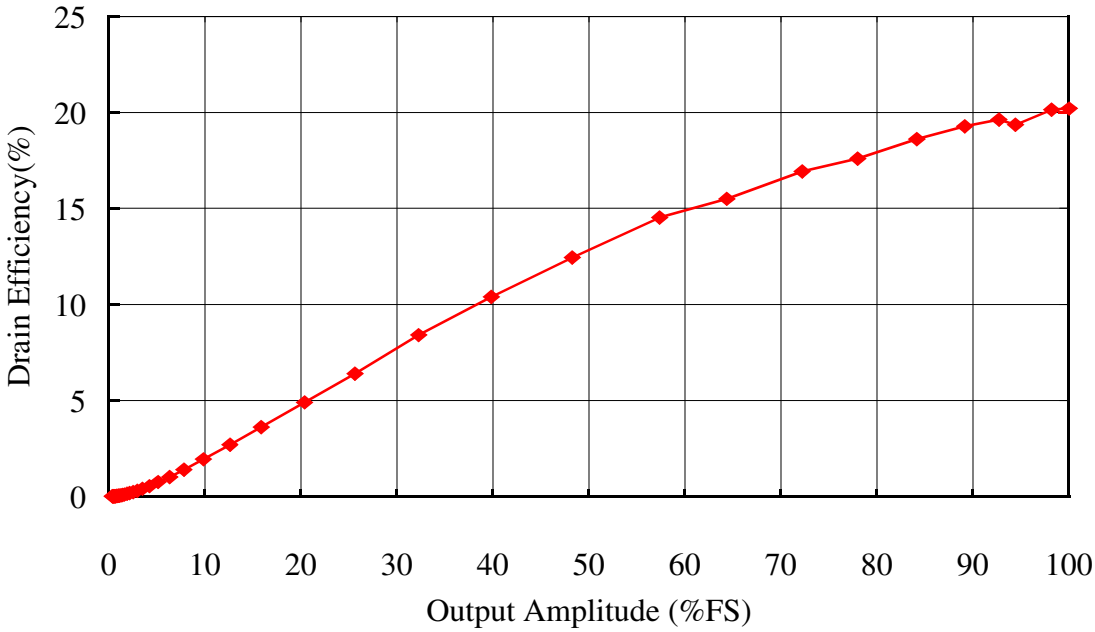


Fig. 6.10: PA Drain Efficiency

6.3.3 Spectral Mask

Although the PA was not intended to be used open-loop with a modulated signal, the VSG was configured to drive the upconverter and PA with ideal EDGE waveforms as a point for comparison. The amplitude of these baseband inputs was swept and the open-loop output spectrum of the PA was observed.

The spectral mask was not met at any power level. For large power levels, the spectrum grossly violates the spectral mask from below 200kHz up to

600kHz or beyond. The spectrum for an 18dBm output signal is shown later in Figure 6.17. Reducing the power level, the 200kHz corner is cleared when the output power is -10.6dBm, but the 400kHz corner is still not met. This is shown in Figure 6.11. Further reducing the power level, the bottom of the mask remains lost under the observed noise floor.

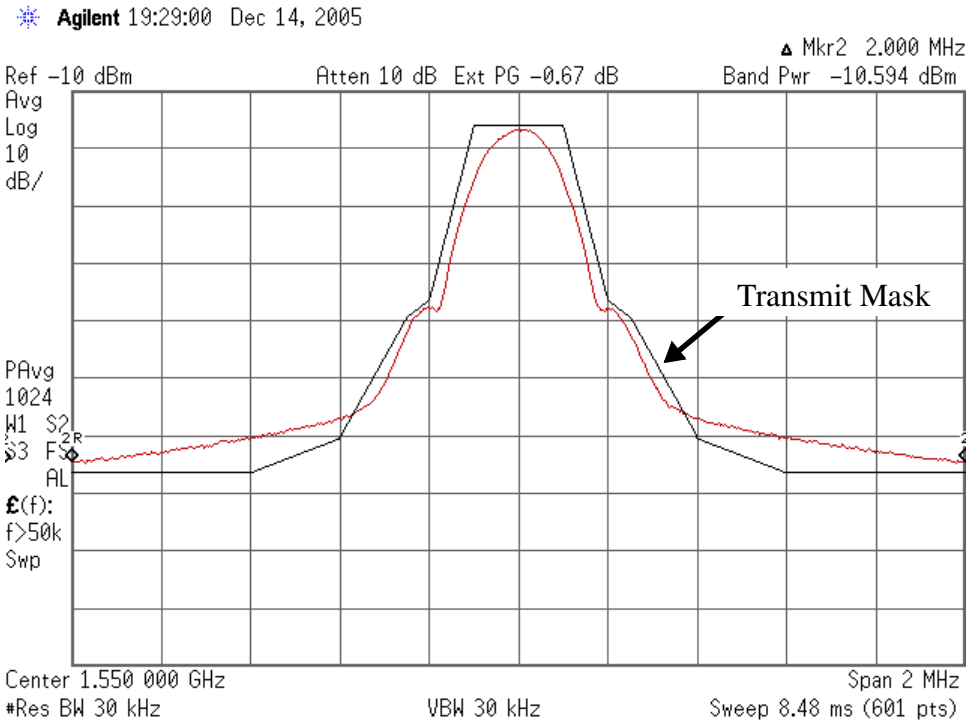


Fig. 6.11: Open-Loop Output Spectrum for -10.6dBm Modulated Signal

6.4 PA to Downconverter Feedback

With operation of both the downconverter and the upconverter individually verified, attention was turned to their combined operation. The upconverter, PA and downconverter were tested together with the downconverter

sensing the PA output as would be done in closed-loop operation. The phase relationship between the upconverter and downconverter was then adjusted with this feedback sensing in place.

6.4.1 IQ Modulation/Demodulation

The upconverter was given baseband EDGE signals, with the I to Q gain ratio adjusted for best image ratio (from the SSB test), and the downconverter left online to observe the PA output. This in essence replicates the test of Section 6.2.1, using the prototype's forward signal path instead of the vector signal generator's internal modulator. The input was scaled to produce a PA output signal of 10dBm (at the directional coupler output), and the downconverter input is estimated to be about -7dBm (after 14.6dB of coupling loss from the directional coupler, and another 2dB of insertion loss from the on-board attenuator).

I-Q plots for the upconverter input and downconverter output are shown in Figure 6.12.

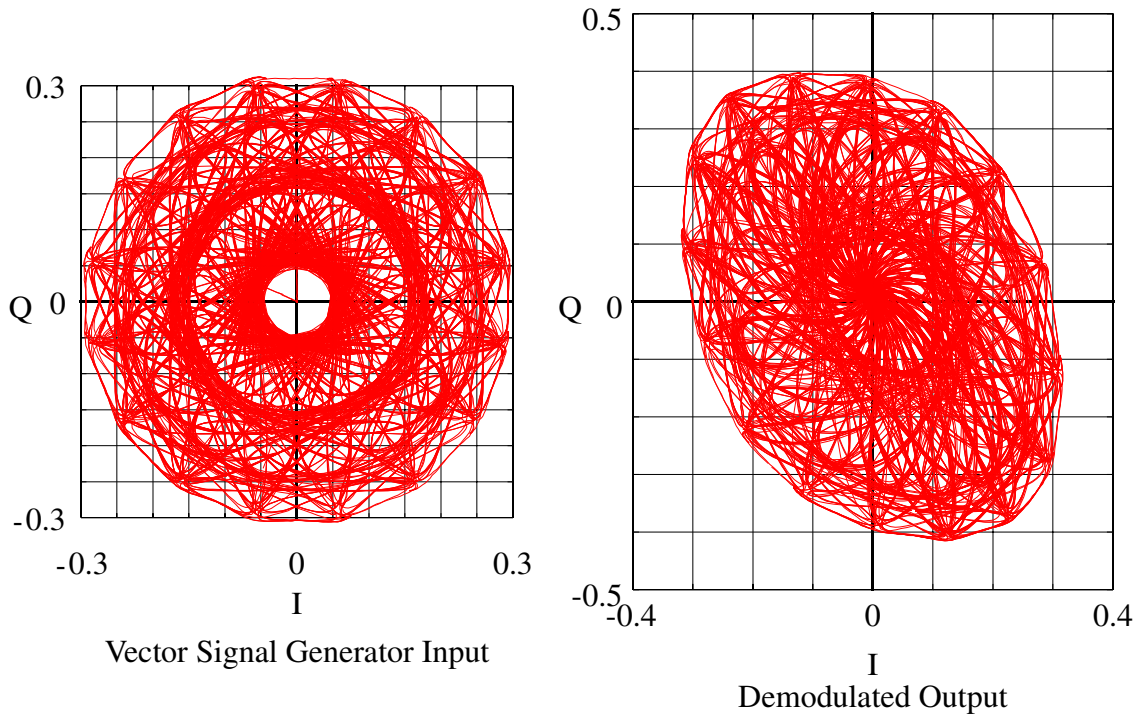


Fig. 6.12: Upconverter IQ modulation test

The same I/Q gain mismatch and quadrature error observed in the downconverter test is seen in the elliptical shape of the demodulated signal. More distortion is now visible though: this is most apparent in how the 'hole' at the centre of the input modulation has collapsed at the output - this is consistent with the loss in gain seen at small signal levels when the class C stage does not turn on effectively, and the PA has only the minimal gain of the class AB helper. The outer boundary of the I-Q plot is now larger even though the average power is 2dB less: with the reduced gain when the amplitude is small, the large-amplitude signals are larger to keep the average comparable.

From this test, it is seen that board-level portion of the feedback loop is functional, with baseband inputs to the upconverter being observable at baseband outputs of the downconverter.

6.4.2 LO Alignment

Preparing for closed-loop operation, the alignment of the upconverter LO phase to the downconverter LO was adjusted. The AWG was configured to produce a sawtooth waveform between 0 and 1V DC on its I output and no signal on its Q output (a ray on the IQ plane), and the downconverter output observed.

The downconverted I-Q locus was not captured, but resembled the simulated curve shown in Figure 4.4 in form. The phase shifter cosine and sine reference currents were adjusted to put the curve in roughly the intended direction, aligned with the I axis. This adjustment was only intended as a coarse adjustment, with the expectation that finer trimming would be made once the chip is operated closed-loop.

6.5 Closed-Loop Operation

With both the upconversion and downconversion paths and off-chip feedback between them being verified as functional, the prototype was ready for closed-loop testing. Linearly predistorted EDGE waveforms were generated to

accommodate the linear gain and quadrature phase errors observed in testing the downconverter: a linear least-squares fit of the downconverted I and Q signals to ideal input signals was performed, and the coefficients from the fit were used to generate an linearly distorted ideal signal to represent the downconverter output. These waveforms were loaded into the AWG.

The loop filter was enabled, and the downconverter configured to send its output there instead of to the test outputs, and the AWG was connected to the prototype's loop inputs instead of upconverter test inputs. It was realized that the VSG can produce an unmodulated RF output even when being used for its AWG baseband outputs, so the VSG was used to generate the LO, freeing up the

signal generator that had been used to generate it before. A diagram of the test setup is shown in Figure 6.13.

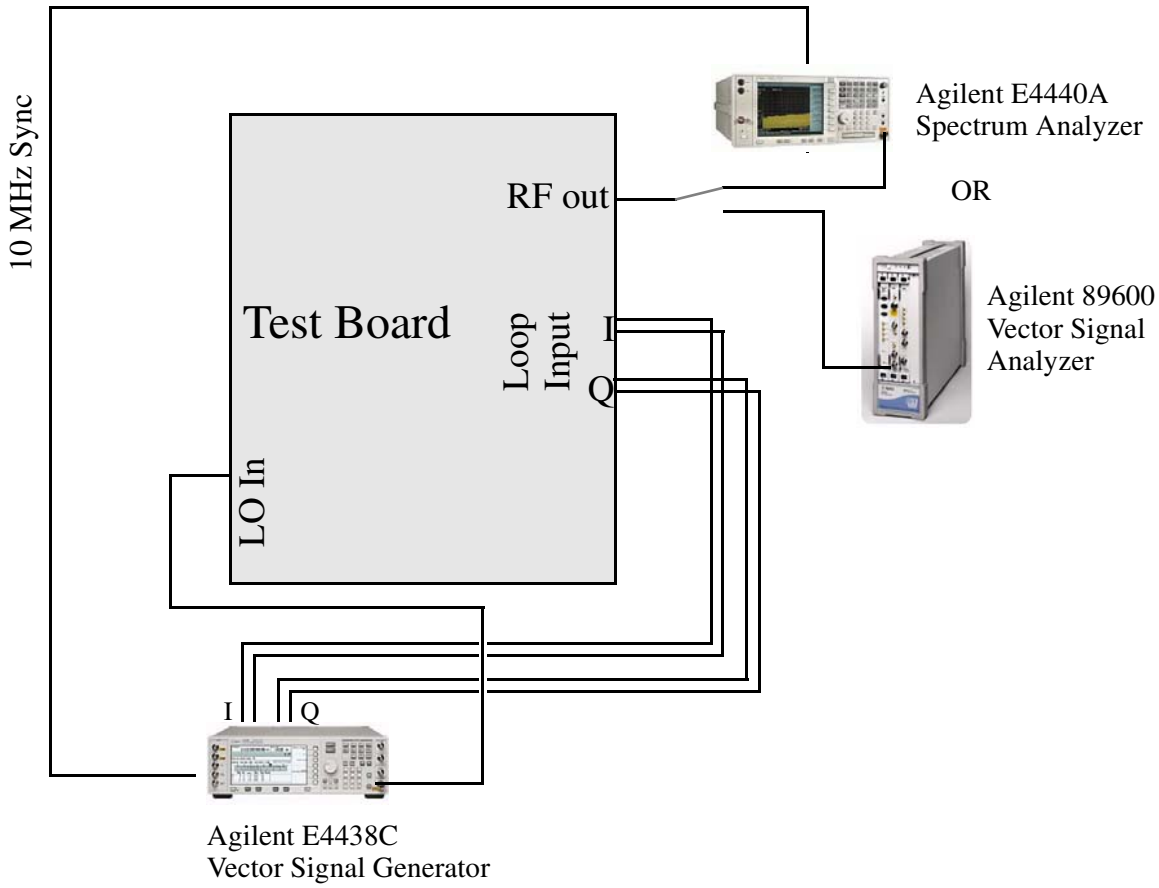


Fig. 6.13: Closed-Loop Test Setup

6.5.1 Spectral Mask

With the prototype configured to operate closed-loop, the output spectrum was observed on the spectrum analyzer. With very little trimming, the output spectrum cleared the transmit mask for output powers up to 18dBm. The maximum output power achieved in open-loop testing was just over 21dBm, and

EDGE modulation has a peak-average ratio of 3dBm, so this 18dBm represents operating at the maximum linear power that could be hoped for!

The initial coarse adjustment of the LO phase did not need further adjustment. The attenuator in the feedback loop and the baseband signal levels were adjusted to control the output signal level, and the loop-filter integrating capacitor size was adjusted to control the loop bandwidth.

6.5.1.1 Input Scaling

In principle, the input signal level is unconstrained and can be scaled arbitrarily. Any change in the input signal level, if done together with a matching change in the feedback attenuation, will make no change in the output signal level. In practice, the output spectrum does vary somewhat with this scaling. Figure 6.14 shows the output spectrum for closed-loop operation with

three different input levels, with feedback attenuator adjusted accordingly, and loop filter integrating capacitor size kept constant.

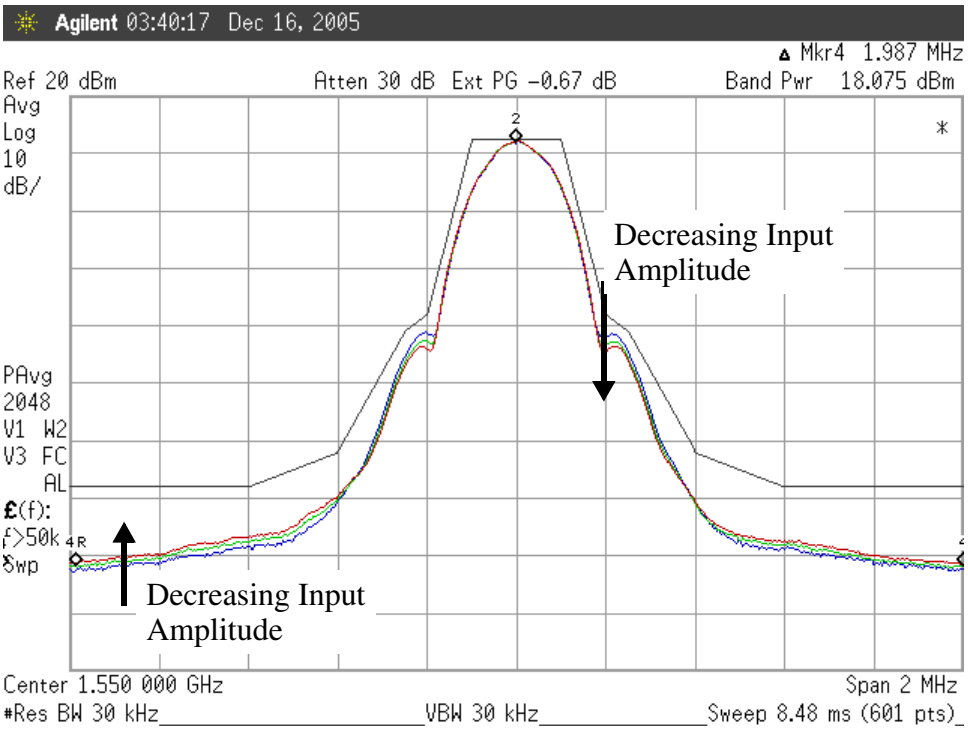


Fig. 6.14: Effect of Input Scaling on Closed-Loop PA Output Spectrum

As the input level is reduced, the shoulders of the spectrum at 200kHz offset falls. This reflects a reduction in third-order distortion from either the input transconductor or the downconversion mixer (or both). The spectrum beyond offsets of 400kHz increases as the input levels are reduced - this is believed to be primarily from the reduced loop gain as the feedback attenuation is increased. The input-referred noise of the downconverter and input transconductor both become larger relative to the signal as the input amplitude

decreases, but from the downconverter measurements, this noise floor is not believed to be significant.

The largest of the input amplitudes used for Figure 6.14 was used in subsequent testing described as third-order distortion performance for this setting is already adequate for meeting the spectral mask.

6.5.1.2 Loop Gain Adjustment

For a given output power and baseband input signal level, a degree of freedom remains in the integrating capacitor. Adjusting this capacitance on its own affects the loop gain. Figure 6.15 shows the effect of changes in the loop gain on the close-in output spectrum.

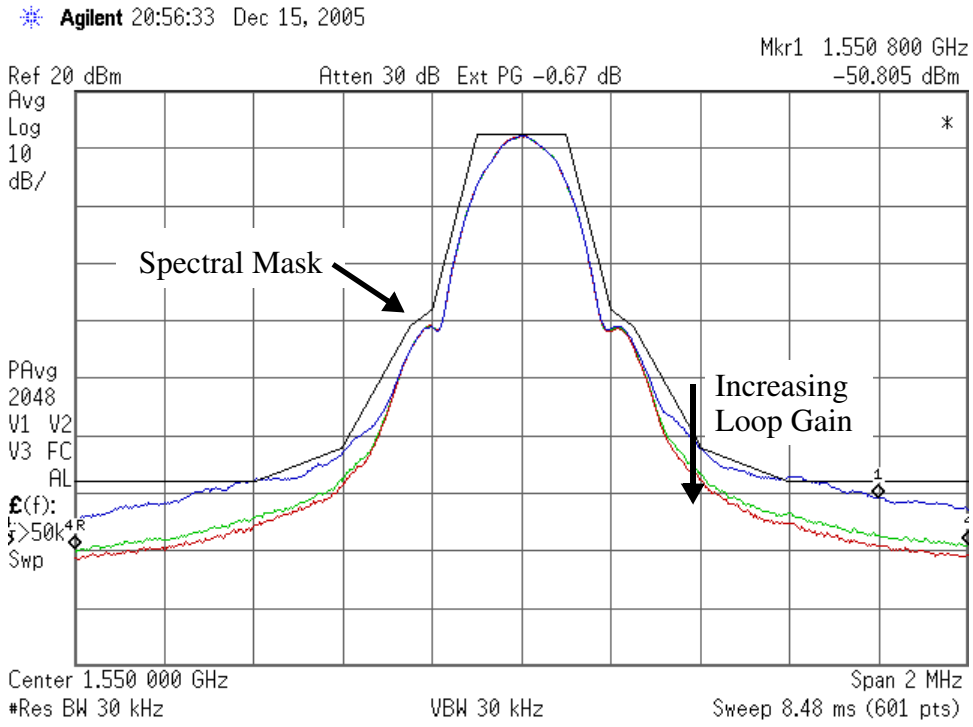


Fig. 6.15: Effect of Loop Gain on Closed-Loop PA Output Spectrum

It is apparent that as loop gain increases, the spectrum improves, with distortion products of the PA being suppressed more. No change is seen at or below the 200kHz shoulder - here the feedback gain is large enough that closed-loop operation has reached its asymptote of tracking the downconverter's performance even with the smallest loop gain tried.

Although increasing the loop gain is seen to improve performance near the carrier, the drawback to excessive loop gain is seen at frequencies farther away. Figure 6.16 shows the output spectrum over a wider bandwidth for the same integrating capacitor settings as Figure 6.15.

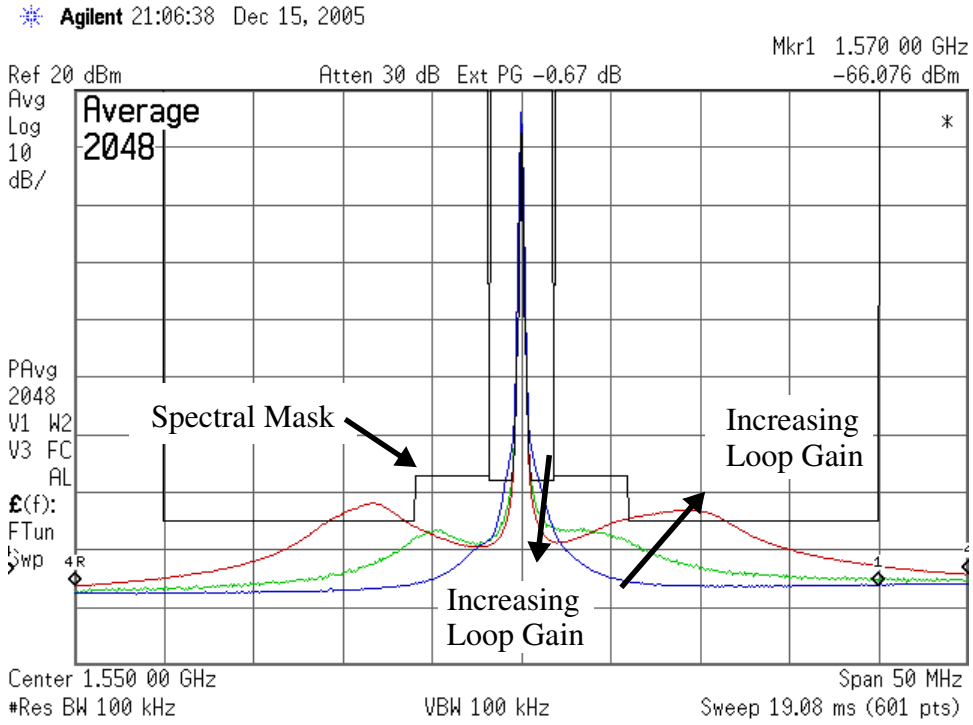


Fig. 6.16: Effect of Loop Gain on Closed-Loop PA Output Spectrum

While the close-in spectrum improves with increasing loop gain, the spectrum farther out worsens. As the loop gain is increased, the loop bandwidth also increases, and extra phase lag from parasitic poles comes into play, decreasing phase margin. This peaking in the spectrum is believed to be noise peaking from reduced phase margin. The unity-gain bandwidth of the loop can be estimated from the frequency of these peaks.

The middle of the three loop bandwidths shown here was used for subsequent measurements.

6.5.1.3 Final Spectrum

Figure 6.17 shows the measured close-in output spectrum for closed-loop operation. The spectral mask, output spectrum for open-loop operation at 18dBm output power, and a reference spectrum of an ideal EDGE signal

generated by the VSG are also shown for comparison. Distortion at 400kHz is suppressed by 21dB by the feedback.

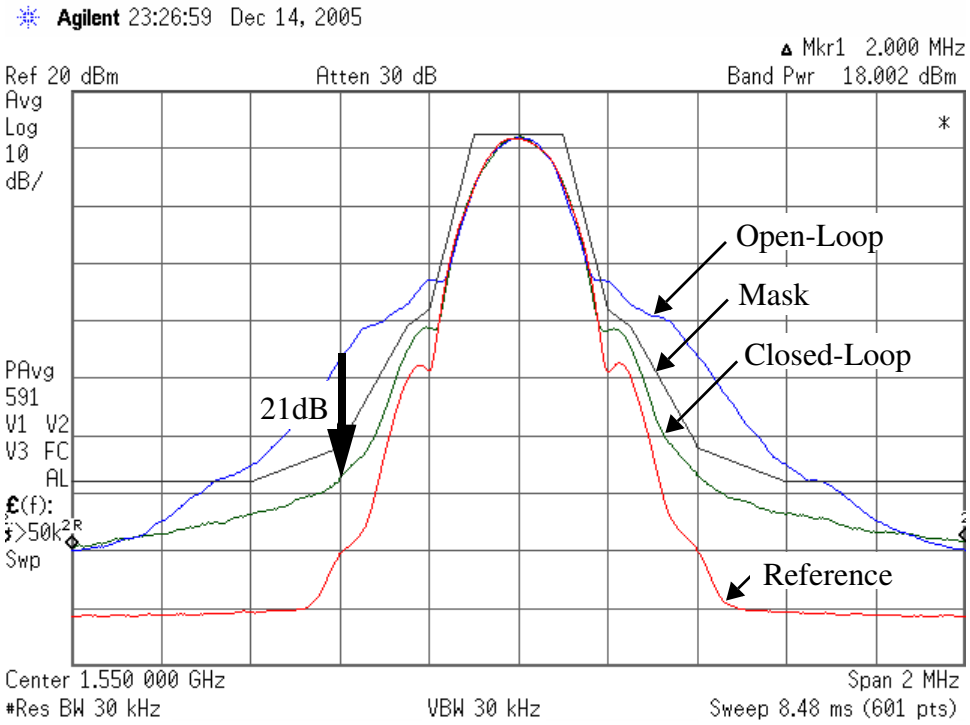


Fig. 6.17: Closed-Loop PA Output Spectrum

6.5.2 Error-Vector Magnitude

The Error-Vector Magnitude (EVM) performance of the transmitter was measured with an Agilent 89600 Vector-Signal Analyzer setup. The results for closed-loop operation with an 18dBm output is shown in Figure 6.18.

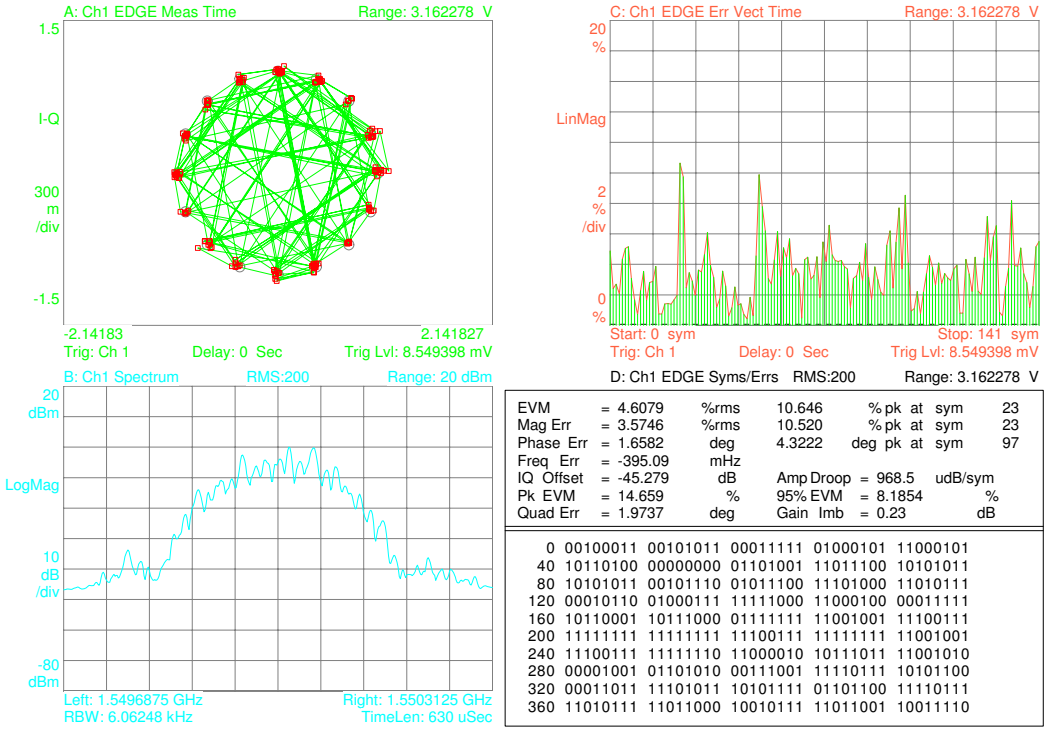


Fig. 6.18: Error-Vector Magnitude Measurement

The key numbers in this measurement are the RMS and Peak EVM, which GSM standards specify must be below of 9% and 30% respectively. These were measured to be 4.6% and 14.6% peak respectively, easily meeting requirements.

6.5.3 Power Consumption

Power consumed from each supply was measured for closed-loop operation with an 18dBm output, and is summarized in Table 6.1.

Block	Power (mW)
Loop Filter	28.8
Downconverter and LO Quadrature Generation	117
Upconverter and LO phase shifter	79.2
PA Driver	69.3
PA Output	341

Table 6.1: Measured Power Consumption

The drain efficiency is $\frac{10^{\frac{18\text{dBm}}{10}}}{341} = \frac{63}{341} = 18.4\%$.

Of the total powered consumed, supplies for the loop filter and downconverter account for under 23% of the total. This roughly represents the power consumed for linearization, the remaining components being the forward path of a nonlinearized transmitter. This breakdown is not exact though as the downconverter supply power includes LO quadrature generation which would be needed even without linearization, while upconverter power includes the LO phase shifter which is not needed without the linearization.

6.5.4 Harmonic Content

Signal powers for harmonics of the carrier were measured and are summarized in Table 6.2.

Harmonic	Power (dBm)
1st (1.55GHz)	18
2nd (3.1GHz)	-24
3rd (4.65GHz)	-38
4th (6.2GHz)	-41
5th (7.75GHz)	-51

Table 6.2: Harmonic Power

6.6 Summary

Measurements of the prototype have been discussed. The performance of individual components differed significantly from what was originally simulated, with the PA producing significantly less power than designed, and downconverter linearity being somewhat worse than designed. Closed-loop linearization was still demonstrated, however. The effect of scaling input signal level and loop bandwidth on closed-loop operation are examined. Nonlinearity of the PA was bad enough to not meet spectrum mask specifications for GSM in open-loop operation, but with feedback enabled, spectral mask and EVM specifications are met with the PA operating at its clipping limit.

Chapter 7
Conclusions

7.1 Research Summary

It is clear that growth in demand for wireless voice and data communications has driven recent efforts to develop highly-integrated radio transceivers, and to minimize costs, much effort has gone into implementing these transceivers in standard CMOS technology. Although CMOS radios now exist on the market, for high data-rate applications the power amplifier typically remains a separate component due to generally poor performance of CMOS PAs.

Performance of the PA is a compromise between power efficiency and linearity, and this thesis examined a number of architectures that can relax linearity requirements for the PA. Cartesian Feedback is identified as a viable approach and this work sought to demonstrate its application to linearizing an integrated CMOS PA.

The contribution of this work is two-fold: theoretical and practical. On the theory side, the existing literature for cartesian feedback has been somewhat

lacking in its analysis, generally assuming a linear PA when considering feedback stability, but this is not a good assumption when trying to enable the use of a very nonlinear PA. Stability of the cartesian feedback loop is examined using MIMO techniques, and it is shown that the effect of the PA on feedback can be captured in eigenvalues representing the transfer function from baseband upconverter inputs to baseband downconverter outputs. For the linear PA assumed in existing analyses, this eigenvalue is simply the linear amplifier gain assumed of the PA, but for a nonlinear memoryless PA, the eigenvalues are extracted from its incremental envelope transfer function. The effect of some linear channel effects are also considered in terms of feedback eigenvalues.

With a better understanding of the effect of the PA on feedback, the design of the loop transfer function is also considered. While single-pole loop transfer functions are common, higher-order transfer functions are not typically seen, owing in part to the challenges of stabilizing them with nonlinearity present. A compromise between first and second-order loop transfer functions is proposed, offering faster gain rolloff with frequency than a first-order while maintaining better phase margin than a second-order loop for robustness against and channel memory effects and PA nonlinearity.

On the practical side, a prototype transmitter including an integrated class-C PA, was designed and fabricated in a standard 0.18μm CMOS process. While none of the blocks of the transmitter advanced the state of the art for their respective functions, a number of interesting circuit techniques were found and used, including a passive downconverter topology offering good 1/f noise and linearity, a simple passive pinking filter design for providing a ‘half pole’ in the

chosen loop transfer function, and the use of sequence-asymmetric polyphase filters for harmonic suppression. AM/PM distortion from the class-C PA is also considered, and a class-AB helper stage used to keep it from impacting stability.

The prototype was tested using EDGE modulated signals, and in open-loop operation the PA as found to violate GSM spectral mask specifications at all power levels. In closed-loop operation, GSM spectral mask and EVM specifications were met while producing an 18dBm output. This power level is a fundamental clipping limit of the PA that would also affect any other linearization method used on the same PA. Cartesian Feedback is thus demonstrated to take a PA that without linearization is unusable at any power level, and linearize it to meet specifications for output powers right up to the PA's theoretical limit.

7.2 Future Work

Although linearization was successfully demonstrated, the prototype was still limited by poor performance of its PA. Even if the PA had delivered the output power and power efficiency of the original circuit simulations, the maximum efficiency comes only at maximum power, and still suffers with power backoff. As radio standards move to modulation schemes with higher crest factors (most notably OFDM), backoff to accommodate these crest factors adversely impacts power efficiency. More work is needed on PA designs that produce good power efficiency at all power levels rather than only at peak power.

The Doherty and Chireix amplifiers are approaches based on using two PAs that modulate each other's apparent load impedance to enhance the drain efficiency when operating at reduced output power. There may be merit in using such multiple-PA approaches, and they will likely require some form of linearization to be useful for tomorrow's modulation schemes.

Another approach that seems to have gained interest in recent years is the use of supply modulation, originally proposed for modulating the output amplitude in envelope-elimination and restoration (EE&R) polar modulators. While supply modulation pushes the problem of drain efficiency from the PA to the supply modulator, the supply operates at baseband rather than RF, and is perhaps more receptive to switch-mode techniques for improving efficiency. The challenges of EE&R modulators are in: dealing with phase discontinuity of the envelope-eliminated signal at the origin; being able synthesize an accurate supply voltage that generates the intended output amplitude; and in synchronizing the two signal paths. Perhaps the answer to these challenges would be to modulate the PA supply only with the intent of improving power efficiency and not relying on it to accurately control the output amplitude, and then operating the PA as a 'linear' amplifier near but not in clipping, relying on cartesian feedback to clean up whatever distortion gets introduced.

Such hybrid approaches aside, there are still some open questions to be solved for using cartesian feedback. The PA in the prototype was loaded with essentially an ideal 50Ω provided by the spectrum analyzer, but in a real transceiver, the load provided by the antenna can vary significantly with the operating environment. Antenna load aside, the sharp frequency response of

SAW filters (as may be used for transmit/receive diplexing) could also affect feedback stability if phase shifts from the filter band-edge get included inside the linearizing loop bandwidth. While it is possible to isolate the PA from these with a passive RF circulator, the circulator adds insertion loss, and other ways to achieve robustness against load variation deserve study. Performing phase-angle feedback within the cartesian feedback loop to actively perform upconverter/ downconverter phase alignment would be one step in this direction, but there may be other approaches too.

Another challenge for cartesian feedback is the trend towards wider channel bandwidths. With wider bandwidths for newer standards, the assumption of a memoryless channel may no longer be reasonable, and ensuring stability of the feedback will be much more difficult than it was in this work.

One approach which would relax the feedback bandwidths somewhat, would be to not rely so heavily on the feedback for linearization: an amplifier that produces less distortion at its output will require less feedback gain to suppress it. The output distortion could perhaps be reduced with some form of adaptive table look-up based predistortion, and cartesian feedback relied upon to only clean up what distortion is left over.

References

[1] J.C. Maxwell, “A Dynamical Theory of the Electromagnetic Field”, *Philosophical Transactions of the Royal Society of London*, vol. 155, pp. 459-512, 1865

[2] M. Faulkner, T. Mattsson, W. Yates, “Automatic adjustment of quadrature modulators”, *IEE Electronics Letters,* vol. 27, (3), pp. 214-216, Jan. 1991

[3] A. Lohtia, P.A. Goud, C.G. Englefield, “An adaptive digital technique for compensating for analog quadrature modulator/demodulator impairments,” in *IEEE Pac. Rim Conf. on Communications, Computers and Signal Processing,* 1993, pp. 447-450.

[4] V. Volterra, *Theory of Functionals and of Integral and Integro-Differential Equations*, New York: Dover, 1959.

[5] S. Narayanan, “Transistor distortion analysis using Volterra series representations,” *Bell Syst. Tech. J.*, vol. 46, (5), pp. 991-1024, May/ June 1967.

[6] R.G. Meyer, Class notes for EECS 242, University of California Berkeley, Fall 1995.

[7] B. Razavi, RF Microelectronics, Upper Saddle River NJ: Prentice-Hall, 1998.

[8] S.C. Cripps, *RF Power Amplifiers for Wireless Communications*, Boston MA: Artech House, 2000.

[9] P. Kenington, *High-Linearity RF Amplifier Design*, Artech House Publishers, 2000.

[10] T.T. Ha, "Chapter 6: Signal Distortion Characterizations and Microwave Power Combining Techniques," in *Solid-State Microwave Amplifier Design*, New York, NY: John Wiley and Sons, 1981.

[11] S.A. Maas, *Nonlinear Microwave and RF Circuits*, Boston MA: Artech House, 1988.

[12] T.A.H. Wilkinson and P.A. Matthews, "Assessment of UHF Power Amplifier Linearisation by Measurement and Simulation", in *IEE Proc. 5th Int. Conf. Mobile Radio,* Warwick, U.K., Dec. 1989, pp. 60-64.

[13] M. Faulkner and T. Mattsson, "Spectral Sensitivity of Power Amplifiers to Quadrature Modulator Misalignment," *IEEE Trans. on Vehicular Technology*, vol. 41, (4), pp. 516-525, Nov. 1992.

[14] S.C. Cripps, "Advanced Techniques in RF Power Amplifier Design for Wireless Communications", Boston, MA: Artech House Incorporated, 1999.

[15] D. Su, W. McFarland, "A 2.5-V, 1-W Monolithic CMOS RF Power Amplifier," in *Proc. of the IEEE 1997 Custom Integrated CIrcuits Conference*, May 1997, pp. 189-92.

[16] N.O. Sokal, A.D. Sokal, "Class E - A New Class of High Efficiency Tuned Single-Ended Power Amplifiers," *IEEE J. Solid State Circuits*, vol. 10, (3), pp. 168-176, June 1975.

[17] M.R. Elliott, T. Montalvo, B.P. Jeffries, F. Murden, J. Strange, A. Hill, S. Nandipaku, and J. Harrebek, "A Polar Modulator Transmitter for GSM/ EDGE," *IEEE J. Solid-State Circuits*, vol 39, (12), pp. 2190-2199, Dec. 2004.

[18] P. Reynaert and M.S.J. Steyaert, "A 1.75GHz Polar Modulated CMOS RF Power Amplifier for GSM-EDGE," *IEEE J. Solid-State Circuits*, vol 40, (12), pp. 2598-2608, Dec. 2005.

[19] D. C. Cox, "Linear Amplification using Non-linear Components", *IEEE Transactions on Communications*, vol. COM-22, pp. 1942-1945, 1974.

[20] S.A. Hetzel, A. Bateman, J.P. McGeehan, "A LINC transmitter," in *41st IEEE Veh. Technol. Conf.*, May 1991, pp. 133-137.

[21] L. Sundstrom, “Effects of reconstruction filters and sampling rate for a digital signal component separator on LINC transmitter performance,” *IEE Electronics Letters*, vol. 31, (14), pp. 1124-1125, July 1995.

[22] F.J. Casadevall. “The LINC transmitter,” in *R.F. Design*, vol. 13, (2), pp. 41-48, Feb. 1990.

[23] K.Y. Chan, A. Bateman, M. Li, “Analysis and realisation of the LINC transmitter using the combined analogue locked loop universal modulator (CALLUM)” in *44th IEEE Veh. Technol. Conf.*, June 1994, pp 484-488.

[24] K. Yamauchi, K. Mori, M. Nakayama, Y. Itoh, Y. Mitsui, and O. Ishida, “A novel series diode linearizer for mobile radio power amplifiers,” in *IEEE MTT-S Dig.*, June 1996, pp. 831-834.

[25] J.K. Cavers, “A linearising predistorter with fast adaptation,” in *Proc. 38th IEEE Veh. Technol. Conf.*, May 1990, Orlando, FL, pp. 44-47.

[26] M. Faulkner, M. Johansson, “Adaptive Linearization Using Predistortion - Experimental Results,” in *Proc. 43rd IEEE Veh. Technol. Conf.*, May 1994, pp. 323-332.

[27] A. Mansell, A. Bateman, “Practical Implementation Issues for Adaptive Presdistortion Transmitter Linearisation,” in *IEE Colloq. Linear RF Amplifiers and Transmitters*, Apr. 1994, pp. 5/1-5/7.

[28] M.A. Briffa, M. Faulkner, “Stability Considerations for Dynamically Biased Cartesian Feedback Linearization,” in *44th IEEE Veh. Technol. Conf.*, Jun. 1994, pp. 1321-1325.

[29] M.A. Briffa, M. Faulkner, “Stability analysis of Cartesian feedback linearisation for amplifiers with weak nonlinearities,” in *IEE Proc.-Commun.*, vol. 143, (4), pp. 212-218, Aug. 1996.

[30] L. Perraud, M. Recouly, et. al., “A direct-conversion CMOS transceiver for the 802.11a/b/g WLAN standard utilizing a Cartesian feedback transmitter,” *IEEE J. Solid-State Circuits*, vol. 39, (12), pp. 2226 2238, Dec. 2004.

[31] F. Carrara, A. Scuderi, and G. Palmisano, “Wide-bandwidth Fully Integrated Cartesian Feedback Transmitter,” in *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 451-454, Sept. 2003.

[32] S. Mann, M. Beach, "A Hybrid Cartesian Loop and Envelope Modulated PA Linear Transmitter Architecture," in *Proc. IEEE Symposium on Personal, Indoor and Mobile Radio Communications*, Sept. 2003 pp. 2721-2725.

[33] A.R. Mansell, A. Bateman, "Transmitter linearisation using composite modulation feedback," in *IEE Electronics Letters*, vol. 32, (23), pp. 2120-2121, Nov. 1996.

[34] C.A. Desoer, "A General Formulation of the Nyquist Criterion," *IEEE Trans. Circuit Theory*, vol. 12, (2), pp. 230-234, June 1965.

[35] C.A. Desoer and Y.T. Wang, "On the Generalized Nyquist Stability Criterion," *IEEE Trans. Autom. Control*, vol. AC-25, (2), pp. 187-196, Apr. 1980.

[36] A.N. Brown, V. Petrovic, "Phase delay compensation in HF Cartesian-Loop Transmitters," in *Fourth International Conference on HF Radio Systems and Techniques,* Apr. 1988, pp. 200-204.

[37] J.L. Dawson, T.H. Lee, "Automatic Phase Alignment for a Fully Integrated Cartesian Feedback Power Amplifier System," *IEEE J. Solid-State Circuits*, vol. 38, (12), pp. 2269-2279, Dec. 2003.

[38] S. Sastry, *Nonlinear Systems: Analysis, Stability and Control*, Reading Materials for EECS 222, University of California Berkeley, Spring 1998.

[39] J.E. Slotine, and W. Li, *Applied Nonlinear Control*. New Jersey: Prentice Hall, 1991.

[40] M. Boloorian and J.P. McGeehan, "Phase-lag compensated Cartesian feedback transmitter," *IEE Electronics Letters*, Vol. 32, (17), pp. 1547-1548, Aug. 1996.

[41] R.S. Narayanaswami, "RF CMOS Class C Power Amplifiers for Wireless Communications," Ph.D. Dissertation, Department of EECS, University of California, 2001.

[42] N. Wongkomet, "Efficiency Enhancement Techniques for CMOS RF Power Amplifiers," Ph.D. Dissertation, Department of EECS, University of California, Berkeley, 2006.

[43] R.M. Bocock, “Upconversion Mixers in 0.18µm CMOS for GSM-EDGE Transmitter Linearization Scheme,” M.S. Report, Department of EECS, University of California, Berkeley, 2001.

[44] M.J. Gingell, “The Synthesis and Application of Polyphase Networks with Sequence Asymmetric Properties,” Ph.D. Thesis, Faculty of Engineering, University of London, 1975.

[45] J. Crols, M. Steyaert, “A single-chip 900 MHz CMOS receiver front-end with a high-performance low-IF topology,” *IEEE J. Solid-State Circuits*, vol. 30, (12), pp. 1483-1492, Dec. 1995.

[46] J. Crols, M. Steyaert, “An analog integrated polyphase filter for a high performance low-IF receiver,” in *Symposium on VLSI Circuits.*, 1995, pp. 87-88.

[47] A. Abidi, “Direct-Conversion Radio Transceivers for Digital Communications,” *IEEE J. Solid-State Circuits*, vol. 30, (12), pp. 1399-1410, Dec. 1995.

[48] F. Behbahani, Y. Kishigami, J. Leete, A. Abidi, “CMOS mixers and polyphase filters for large image rejection,” *IEEE J. Solid-State Circuits*, vol. 36, (6), pp. 873-887, Jun. 2001.

[49] C. DeRanter, M. Borremans, M. Steyaert, “A wideband linearisation technique for non-linear oscillators using a multi-stage polyphase filter,” in *Proc. IEEE European Solid-State Circuits Conf.,* Sept 1999, pp. 214-217.

[50] A.C. Davies, “Digital Generation of Low-Frequency Sine Waves” in *IEEE Trans., Instrum. Meas*. IM-18, (2), pp. 97-105, Jun. 1969.

[51] J. Weldon, R.S. Narayanaswami, J.C. Rudell, L. Lin; M. Otsuka, S. Dedieu, L. Tee; K.-C. Tsai, C.-W. Lee, P.R. Gray, P.R., “A 1.75-GHz highly integrated narrow-band CMOS transmitter with harmonic-rejection mixers,” *IEEE J. Solid-State Circuits*, vol 36, (12), pp. 2003-2015, Dec. 2001.

[52] J.C. Rudell, “Frequency Translation Techniques for High-Integration High-Selectivity Multi-Standard Wireless Communication Systems,” Ph.D. Dissertation, Department of EECS, University of California, Berkeley, 2000.

[53] D.K. Shaeffer, et al., “A 115-mW, 0.5-um CMOS GPS Receiver with Wide Dynamic-Range Active Filters,” *IEEE J. Solid-State Circuits*, vol. 33, (12), pp. 2219-2231, Dec. 1998.

[54] E. Sacchi, I. Bietti, S. Erba, L. Tee, P. Vilmercati, R. Castello, “A 15mW, 70kHz 1/*f* Corner Direct Conversion CMOS Receiver,” in *Proc. IEEE Custom Integrated Circuits Conf.*, Sep. 2003, pp. 459-462.

[55] J. Crols, M. Steyaert, “A 1.5 GHz highly linear CMOS downconversion mixer,” *IEEE J. Solid-State Circuits*, vol. 30, (7), pp. 736-742, Jul. 1995.

[56] D. M. W. Leenaerts, W. Redman-White, “1/*f* noise in passive CMOS mixers for low and zero IF integrated receivers,” in *Proc. IEEE European Solid-State Circuits Conf.,* Sept 2001, pp. 41-44.

[57] D. Bohn et al., *Audio Handbook*. National Semiconductor Corporation, 1976.

[58] P. Horowitz, W. Hill, *The Art of Electronics* Cambridge, MA: Cambridge University Press, 1989.

[59] W. M. Leach, “Impedance Compensation Networks for the Lossy Voice-Coil Inductance of Loudspeaker Drivers,” *J. Audio Eng. Soc.*, vol. 52, (4), pp. 358-365, Apr. 2004.

[60] F. Boesch, J. Hagopian, “Minimum total capacitance RC realizations,” *IEEE Trans. Circuits Syst.* vol. 18, (2), pp. 286-288, Mar. 1971.

Appendix A

Volterra Kernels and Intermodulation Intercept Points

The third-order intermodulation intercept point, or IP_3 is a common measure of linearity. This measure is found from the two-tone test, where two sinusoids of equal amplitude at closely spaced frequencies are passed through a circuit.

$$x(t) = E\cos(\omega_1 t) + E\cos(\omega_2 t) \qquad \text{(Eq A-1)}$$

In the phasor domain, based on a centre frequency of ω_c, this is:

$$\tilde{x}(t) = Ee^{j(\omega_1 - \omega_c)t} + Ee^{j(\omega_2 - \omega_c)t} \qquad \text{(Eq A-2)}$$

From Section 2.4.3, the output of the system has a linear component of:

$$\tilde{y}_1(t) = \overline{A_1}\tilde{x}(t) = \overline{A_1}Ee^{j(\omega_1 - \omega_c)t} + \overline{A_1}Ee^{j(\omega_2 - \omega_c)t} \qquad \text{(Eq A-3)}$$

The third order product is:

$$\tilde{y}_3 = \overline{A_3}\tilde{x}\tilde{x}^*\tilde{x} = \overline{A_3}E^3(e^{j(\omega_1-\omega_c)t} + e^{j(\omega_2-\omega_c)t})^*(e^{j(\omega_1-\omega_c)t} + e^{j(\omega_2-\omega_c)t})^2 \quad \text{(Eq A-4)}$$
$$= 3\overline{A_3}E^3(e^{j(\omega_1-\omega_c)t} + e^{j(\omega_2-\omega_c)t}) + \overline{A_3}E^3(e^{j(2\omega_1-\omega_2-\omega_c)t} + e^{j(2\omega_2-\omega_1-\omega_c)t})$$

The first term of this is at the same frequencies as the linear component and contributes to gain compression, or possibly gain expansion. The second term consists of new tones that appear as though they were the linear products of tones at $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$ - these tones are known as third-order intermod, or IM_3.

The IM_3 intercept point, IP_3, is defined as the signal level for which these intermodulation products are the same power as the linear products of the input (not considering gain compression), that is:

$$\overline{A_1}E = \overline{A_3}E^3 \quad \text{(Eq A-5)}$$

$$\text{or: } E^2 = \frac{\overline{A_1}}{\overline{A_3}} \quad \text{(Eq A-6)}$$

Thus a system with an output of $\tilde{x}(t) + \tilde{x}(t)|\tilde{x}(t)|^2$, that is, $\overline{A_1} = \overline{A_3} = 1$, has an IP_3 of unity. A modulated signal with an RMS power of unity (equal to IP_3) will then have linear and third-order products of $\tilde{x}(t)$ and $\tilde{x}(t)|\tilde{x}(t)|^2$ respectively.

The same reasoning is easily applied to higher order distortion. The two-tone test with unity-magnitude input tones will produce tones at

$(i+1)\omega_1 - i\omega_2$ and $(i+1)\omega_2 - i\omega_1$ from $(2i+1)$'th order product $\tilde{x}(t)|\tilde{x}(t)|^{2i}$ of equal magnitude to the linear product from $\tilde{x}(t)$.

Note that (Eq A-6) is in terms of phasor-domain coefficients. Substituting (Eq 2-43) for the coefficients gives the more familiar result of:

$$E^2 = \frac{H_3}{\frac{3}{4}H_3} \qquad \text{(Eq A-7)}$$

Appendix B

Loop-Filter Synthesis

The desired loop filter transfer function, from (Eq 4-2) is:

$$L(s) = \frac{2\pi 10^6 (s + 2\pi 10^5)(s + 2\pi 10^6)(s + 2\pi 10^7)}{(s + 2\pi 100)(s + 2\pi 10^{4.5})(s + 2\pi 10^{5.5})(s + 2\pi 10^{6.5})} \quad \text{(Eq B-1)}$$

This can be broken into two portions, the dominant pole (integrator) and lag-compensation network:

$$L(s) = L_{dom}(s)L_{lag}(s) \quad \text{(Eq B-2)}$$

where the dominant pole is:

$$L_{dom}(s) = \frac{2\pi 10^6 10^{1.5}}{(s + 2\pi 100)} \quad \text{(Eq B-3)}$$

and the lag compensation transfer function is three decades of a half-pole rolloff:

$$L_{lag}(s) = \frac{(s + 2\pi 10^5)(s + 2\pi 10^6)(s + 2\pi 10^7)}{10^{1.5}(s + 2\pi 10^{4.5})(s + 2\pi 10^{5.5})(s + 2\pi 10^{6.5})} \quad \text{(Eq B-4)}$$

or, in more generation notation:

$$L_{lag}(s) = \frac{p_1 p_2 p_3 (s - z_1)(s - z_2)(s - z_3)}{z_1 z_2 z_3 (s - p_1)(s - p_2)(s - p_3)} \qquad \text{(Eq B-5)}$$

The dominant pole of $L_{dom}(s)$ represents a practical approximation of a real integrator $L_{int}(s) = \frac{2\pi f_{unity}}{s}$. The exact frequency of the pole typically is of limited interest, but the unity-gain frequency f_{unity} is important. The integrator is readily implemented with a miller integrator, and the value of the integrating capacitor controls where this frequency ends up.

A2.1 Foster-Network Lag Compensator Component Values

The lag compensator has a DC gain of unity, losing gain with increasing frequency, and can be implemented as a passive voltage divider network, as shown in Figure B.1

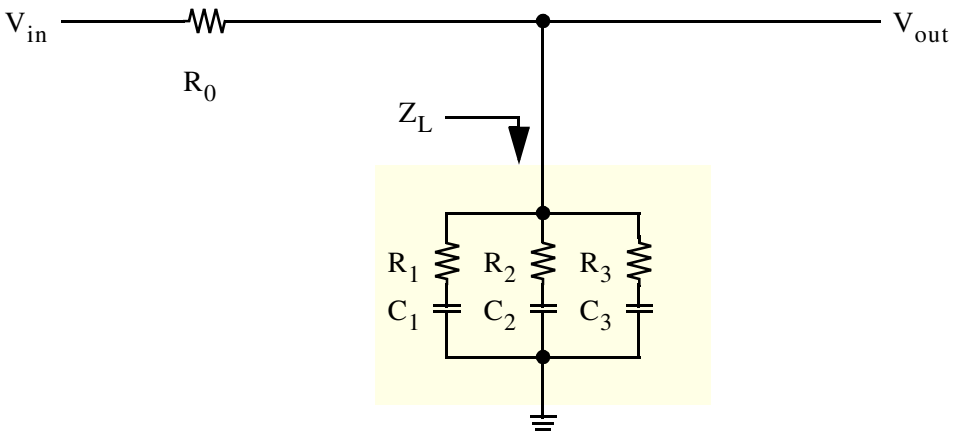


Fig. B.1: Passive Lag Compensator

The same components can also be arranged as a current divider as shown in Figure B.2. This topology is known as a Foster-II network.

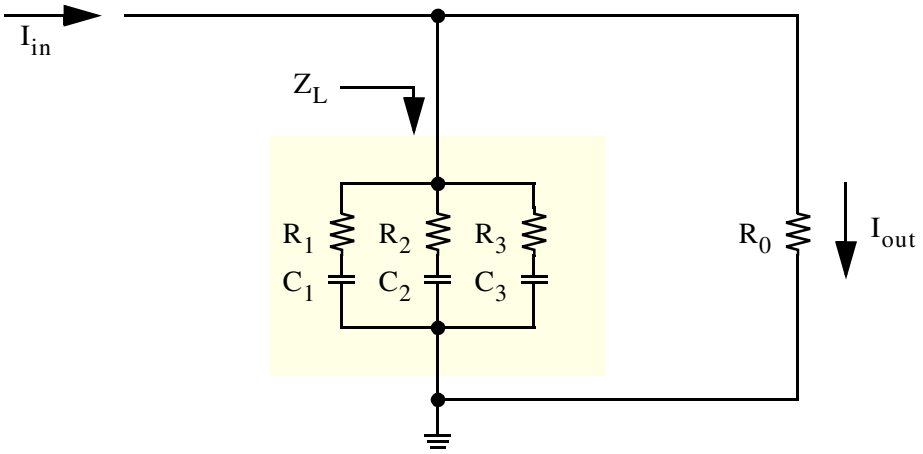


Fig. B.2: Passive Current-Mode Lag Compensator

The transfer function of either network is:

$$L_{lag}(s) = \frac{Z_L}{Z_L + R_0} \qquad \text{(Eq B-6)}$$

where:

$$Z_L = \left(R_1 + \frac{1}{sC_1}\right) \| \left(R_2 + \frac{1}{sC_2}\right) \| \left(R_3 + \frac{1}{sC_3}\right) \qquad \text{(Eq B-7)}$$

(Eq B-6) can be written in an alternate form for convenience:

$$\frac{1}{L_{lag}(s)} - 1 = \frac{R_0}{Z_L} = \frac{R_0}{\left(R_1 + \frac{1}{sC_1}\right)} + \frac{R_0}{\left(R_2 + \frac{1}{sC_2}\right)} + \frac{R_0}{\left(R_3 + \frac{1}{sC_3}\right)} \qquad \text{(Eq B-8)}$$

Rearranging slightly gives:

$$\frac{sR_0/R_1}{\left(s+\frac{1}{R_1C_1}\right)}+\frac{sR_0/R_2}{\left(s+\frac{1}{R_2C_2}\right)}+\frac{sR_0/R_3}{\left(s+\frac{1}{R_3C_3}\right)}=\frac{1}{L_{lag}(s)}-1 \qquad \text{(Eq B-9)}$$
$$=\frac{z_1z_2z_3(s-p_1)(s-p_2)(s-p_3)}{p_1p_2p_3(s-z_1)(s-z_2)(s-z_3)}-1$$

This equation goes to infinity as s tends to the zeros of $L_{lag}(s)$. From this, it is easily seen that:

$$z_1=\frac{-1}{R_1C_1},\ z_2=\frac{-1}{R_2C_2}\ \text{and}\ z_3=\frac{-1}{R_3C_3} \qquad \text{(Eq B-10)}$$

Thus, if R_1, R_2 and R_3 are known, the capacitances are easily found from the zero frequencies. These resistances can be found from the ratios R_0/R_1, R_0/R_2 and R_0/R_3 which in turn can be found by using Heaviside's method. First, for R_1, multiply (Eq B-9) by $\frac{(s-z_1)}{z_1}$:

$$\frac{s}{z_1}\frac{R_0}{R_1}+\frac{s}{z_1}\frac{R_0}{R_2}\frac{(s-z_1)}{(s-z_2)}+\frac{s}{z_1}\frac{R_0}{R_3}\frac{(s-z_1)}{(s-z_3)}=\frac{z_2z_3(s-p_1)(s-p_2)(s-p_3)}{p_1p_2p_3(s-z_2)(s-z_3)}-\frac{(s-z_1)}{z_1} \qquad \text{(Eq B-11)}$$

Then taking the limit of (Eq B-11) as s approaches z_1,

$$\frac{R_0}{R_1}=\frac{z_2z_3(z_1-p_1)(z_1-p_2)(z_1-p_3)}{p_1p_2p_3(z_1-z_2)(z_1-z_3)} \qquad \text{(Eq B-12)}$$

From this, given R_0, R_1 can be found, and R_2 and R_3 can be found similarly. Values for the capacitances then follow from (Eq B-10).

It is easily seen that at arbitrarily high frequencies, Z_L is simply $\frac{1}{\sum\frac{1}{R_i}}$.

This value depends only on the frequency range over which the 1/2-pole rolloff occurs, and would be the same if a larger number of more closely spaced pole/ zero pairs were used in the same range, as might be done to further reduce ripple of the phase response. Finer spacing would require greater total resistance, requiring a greater number of larger resistances combined in parallel to achieve the same net conductance. If resistor area is proportional to resistance (resistances limited by minimum resistor width), then the resistor area would increase with roughly the square of the number of pole/zero pairs. This puts a practical limit on how finely the pole/zero pairs can be spaced with this topology.

At arbitrarily low frequencies, Z_L is $\frac{1}{s\sum C_i}$, and it can be shown that

$$\sum C_i = \lim_{s \to 0} \frac{1}{sZ_L} = \frac{1}{R_0}\sum\left(\frac{1}{p_i} - \frac{1}{z_i}\right) \qquad \text{(Eq B-13)}$$

For uniformly spaced poles and zeros across a given frequency range, this total capacitance increases somewhat with the number of pole/zero pairs. The sum is always less than $\frac{1}{R_0 p_1}$ however, and this bound does not increase with the number of pole/zero pairs, (the sum approaches a limit that is roughly half of this bound), thus as long as capacitor area is proportional to total

capacitance (this is a reasonable assumption until the individual capacitances become excessively small), capacitor area does not limit increasing the number of pole/zero pairs.

A2.2 Cauer Topology (not used)

The need to combine resistances in parallel can be avoided with a change of circuit topology. Figure B.3 shows another topology which can give the same transfer function. This topology is known as a Cauer-I network.

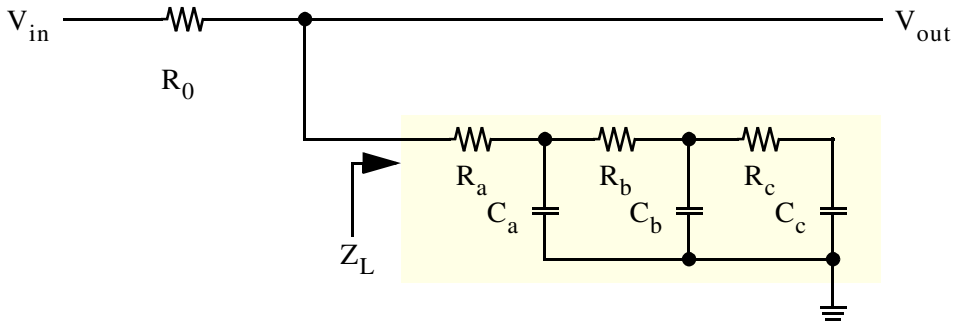


Fig. B.3: Cauer-Network Lag Compensator

The component values required for this topology are different from before, but are still straightforward to compute. The impedance of the network can be expressed in terms of the transfer function. For the transfer function of (Eq B-5), (Eq B-6) can be written as:

$$\frac{R_0}{Z_L} = \frac{1}{L_{lag}(s)} - 1 = \frac{A(s)}{B(s)} \qquad \text{(Eq B-14)}$$

where

$$A(s) = \left(\frac{s}{z_1} - 1\right)\left(\frac{s}{z_2} - 1\right)\left(\frac{s}{z_3} - 1\right) - \left(\frac{s}{p_1} - 1\right)\left(\frac{s}{p_2} - 1\right)\left(\frac{s}{p_3} - 1\right) \qquad \text{(Eq B-15)}$$

and

$$B(s) = \left(\frac{s}{p_1} - 1\right)\left(\frac{s}{p_2} - 1\right)\left(\frac{s}{p_3} - 1\right) \qquad \text{(Eq B-16)}$$

For convenience, these polynomials can be written as

$$A(s) = \sum_{i=1}^{3} a_i s^i \text{ and } B(s) = \sum_{i=0}^{3} b_i s^i \qquad \text{(Eq B-17)}$$

Note that the series $A(s)$ has no constant term - this corresponds with the load network being an open-circuit (no conductance) at DC.

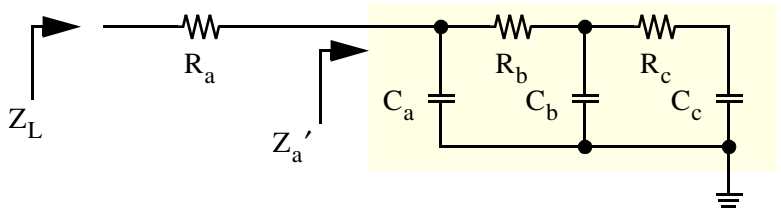


Fig. B.4: Removal of leading resistor from a Cauer network

The impedance Z_L is the series combination of the first resistor R_a and the impedance Z_a' of the rest of the network after it, or:

$$\frac{Z_L}{R_0} = \frac{B(s)}{A(s)} = \frac{R_a + Z_a'}{R_0} = \frac{R_a}{R_0} + \frac{B(s) - \frac{R_a}{R_0}A(s)}{A(s)} \qquad \text{(Eq B-18)}$$

The numerator for $\frac{Z_a'}{R_0}$ reduces in order when:

$$\frac{R_a}{R_0} = \frac{b_3}{a_3} \qquad \text{(Eq B-19)}$$

leaving

$$\frac{Z_a'}{R_0} = \frac{B'(s)}{A(s)} \qquad \text{(Eq B-20)}$$

where

$$B'(s) = B(s) - \frac{R_a}{R_0}A(s) = \sum_{i=0}^{2} b'_i s^i \qquad \text{(Eq B-21)}$$

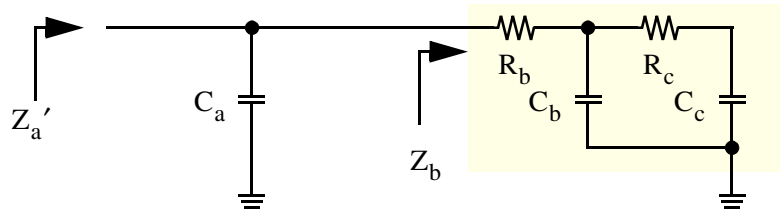


Fig. B.5: Removal of leading capacitor from a Cauer network

Similarly, the impedance Z_a' is the parallel combination of capacitor C_a and the impedance Z_b of the rest of the network after it.

$$\frac{R_0}{Z_a'} = \frac{A(s)}{B'(s)} = R_0\left(sC_a + \frac{1}{Z_b}\right) = R_0 sC_a + \frac{A(s) - sC_a B'(s)}{B'(s)} \qquad \text{(Eq B-22)}$$

The numerator of $\frac{R_0}{Z_b}$ reduces in order when:

$$C_a = \frac{1}{R_0}\frac{a_3}{b'_2} \qquad \text{(Eq B-23)}$$

leaving

$$\frac{R_0}{Z_b} = \frac{A'(s)}{B'(s)} \qquad \text{(Eq B-24)}$$

where

$$A'(s) = A(s) - sC_aB'(s) \qquad \text{(Eq B-25)}$$

The values of R_b and C_b can be extracted from this $A'(s)$ and $B'(s)$ just as R_a and C_a were extracted from $A(s)$ and $B(s)$, and the polynomials that remain after that would give R_c and C_c.

A2.3 Comparison of Topologies

For the third-order transfer function of (Eq B-4), the total resistance of the Cauer network is indeed slightly less (about 7%) than for the Foster topology. However, for fourth and higher order transfer functions spanning the same three decades of frequency, the total resistance required for the Cauer network is actually greater than for the Foster network - a sixth order transfer function needs about 64% more resistance in total, and the difference grows worse with higher order. While the Foster topology has a total resistance that

grows with the square of the order, the total resistance of the Cauer network appears to grows exponentially with the order.

An intuition was not found for why the total resistance of the Cauer network grows exponentially, but likely exists in existing circuit theory literature, with Foster/Cauer networks having being thoroughly studied since the 1920's. More important though, would be the question of whether the Foster network has the minimum possible total resistance. The total capacitance of the Foster network is known to be minimal [60] and perhaps a dual of the proof would show the total resistance is also minimal.

In practical implementation, the Cauer topology does have a minor advantage in that the impedance at arbitrarily high frequency is defined by just the first resistor, whereas all the resistances of the Foster network are active at high frequency. These resistances have parasitic capacitances associated with them, and at high frequencies where these parasitics would have an effect, the Cauer network is less affected. The difference is moot however, as the difference is only seen at frequencies of hundreds of megahertz which is well past the unity-gain loop bandwidth.