ANALYSIS AND DESIGN
OF CURRENT-COMMUTATING
CMOS MIXERS

by

Emmanouil Terrovitis

Memorandum No. UCB/ERL M01/25

18 May 2001

ANALYSIS AND DESIGN
OF CURRENT-COMMUTATING
CMOS MIXERS

by

Emmanouil Terrovitis

Memorandum No. UCB/ERL M01/25

18 May 2001

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Analysis and Design of Current-Commutating CMOS Mixers

by

Emmanouil Terrovitis

Grad. (National Technical University of Athens, Greece) 1992
M.S. (University of California, Berkeley) 1996

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 at BERKELEY

Committee in charge:

Professor Robert G. Meyer, Chair
Professor Borivoje Nikolic
Professor Keith Miller

Spring 2001

The dissertation of Emmanouil T. Terrovitis is approved:

Robert G Meyer   _5/15/01_
Chair   Date

C. Keith Miller   _May 15, 2001_
                                        Date

                         May 17, 2001
                                        Date

University of California at Berkeley

Spring 2001

Analysis and Design of Current-Commutating CMOS Mixers

Copyright Spring 2001

by

Emmanouil Terrovitis

Abstract

Analysis and Design of Current-Commutating CMOS Mixers

by

Emmanouil Terrovitis

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences
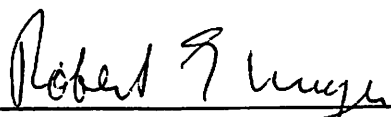
University of California at Berkeley

Professor Robert G. Meyer, Chair

The recent advances in low-cost CMOS fabrication processes have rendered them appropriate for the realization of high-frequency analog communication circuits, traditionally implemented in more expensive technologies such as bipolar or Gallium Arsenide. These CMOS implementations have the significant advantage that they can be more easily integrated with the low frequency analog and digital circuitry. The demand for short design cycles imposes the need for fast optimization of the high-frequency analog circuit blocks. Such a circuit, present in the front end of any communication system, is the mixer which performs frequency translation of the carrier signals. Because one of its inputs is the strong local oscillator signal, its operating point is periodically-time-varying. As a result, the analysis of its operation is considerably more complicated than that of the linear time-invariant blocks. The subject of this thesis is to analyze the operation of one commonly used class of mixers, those which employ a switching transistor pair to commutate the signal current.

The objective is to provide results in a form that can be applied by Radio Frequency circuit designers to systematically optimize their designs.

In the first part of this thesis the mixing operation is described and practical mixer nonidealities and related performance metrics are introduced. Several mixer topologies in CMOS technology are discussed and the current-commutating CMOS mixers, for which the results of this research apply, are emphasized. In the second part an analysis of the nonidealities which define the mixer dynamic range, namely the noise and the nonlinearity, is performed. The contribution of every internal and external noise source to the output noise is calculated and the mixer noise performance is predicted. The noise performance of a CMOS inductively degenerated transconductance stage is investigated in depth. Consequently, the nonlinearities of the CMOS transconductance stages are analyzed. These results are applicable besides mixers to other blocks that employ transconductance stages, such as low noise amplifiers and power amplifiers. Finally, the nonlinearity of the switching pair is investigated. In all cases the results are provided in terms of simplified analytical expressions or graphs of normalized parameters. A simple transistor model with continuous derivatives of any order in all operating regions is adopted and compared with more sophisticated simulator models. In the third part, the design of some single -balanced active mixers is presented as a demonstration of the application of the theoretical results derived in this thesis.

Professor Robert G. Meyer
Dissertation Committee Chair

To my parents

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my deepest appreciation and gratitude to my research advisor Robert G. Meyer, without whose guidance and support this research would have not been possible. In particular I am grateful to him for giving me the freedom to define this project in a way that best fit my abilities and background. I would also like to thank professors Borivoje Nikolic and Keith Miller for being readers of this thesis, and professor Robert Brodersen for serving as chair of my qualifying exam committee.

My research in integrated circuits has greatly benefited from the interaction with several persons from the industry. Ranjit Gharpurey of Texas Instruments contributed to this thesis by providing helpful advice during my summer internship at Texas Instruments, when a significant part of the noise analysis was accomplished. Ken Kundert of Cadence Design Systems triggered my interest in several interesting research topics such as the flicker noise of devices with time-varying operating point, phase noise in oscillators and noise simulation of large mixed-signal circuits. I am grateful to all of my fellow employees at Texas Instruments, Cadence Design Systems, and Philips Semiconductors where I held summer jobs, for transferring their experience to me. Similarly, I am thankful to my fellow employees at Atheros Communications, where I started working several months before I completed this dissertation, for creating such a friendly and technically stimulating workplace. Special thanks to Suni Mendis of Atheros Communications for proofreading part of this thesis. Kevin Mc Adams of Phillips semiconductors provided technical support during some of the measurements needed in this research. Many thanks are owed to Philips Semiconductors and National Semiconductor for fabricating the test chips used in this dissertation. Philips

Part I

Background

Chapter 1

Introduction

1.1 Motivation

The large market demand for new communication services motivates research towards a higher degree of integration in communication devices. Examples of these services include wireless applications such as cellular and cordless phones, pagers, wireless computer networks, (LANs and WANs), satellite communications, GPS systems, and wired communications such as cable TV, cable modems and XDSL modems. A high degree of integration of mass-produced communication devices is desirable because integration increases portability, functionality and reliability. It also reduces production cost and can lead to significant power savings which translates to longer operation time for portable devices. The integration of the logic circuitry and the analog low-frequency circuitry is routinely implemented today and imposes few challenges. A significant research effort is currently being conducted in industry and universities to increase the degree of integration of the high frequency — or radio frequency (RF) — analog circuitry of the wireless systems which has been tradition-

ally realized with high performance discrete active and passive components. The transition from discrete to integrated solutions involves significant changes in the system architecture and the structure of the circuit blocks.

Continuous scaling of CMOS technologies, mainly driven by the computer and the digital circuit industry, has enabled the realization of high-frequency analog circuits which have previously been realized mostly in GaAs or bipolar technologies. CMOS processes require fewer processing steps and are inexpensive when compared to other technologies. They offer the potential for a higher degree of integration since they are the technologies of choice for the digital and baseband analog circuitry. It remains a challenge for system engineers and circuit designers to integrate the analog RF part using CMOS technologies together with the baseband circuitry, while eliminating as many as possible of the high-quality discrete components. In particular, it is a challenge to accomplish this task in a power efficient way.

The multitude of different applications imposes different specifications on the analog circuit blocks while the fabrication technologies continuously improve. Circuit designers are often called on to redesign circuit blocks in new technologies and for a different set of specifications. The need for rapid time-to-market calls for a short design cycle. Therefore a better understanding of the RF block operation which will facilitate a systematic optimization is desirable. The mixer is an essential RF block of every communication system. In this thesis we investigate the operation and the nonidealities of a commonly used mixer topology implemented in CMOS technology.
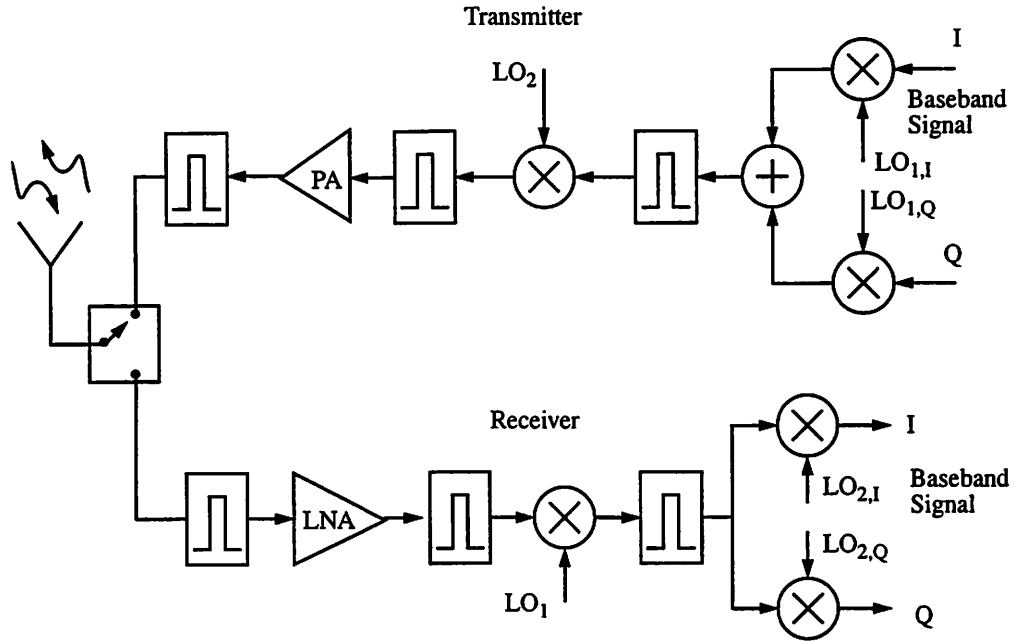
Figure 1.1: Simplified heterodyne transceiver architecture.

1.2 Mixers

In a communication system we desire to transfer a low-frequency signal from one geographic location to an other. This task is accomplished more conveniently if the low-frequency signal is translated to some higher frequency, called the carrier frequency. The circuit block that performs this task is the mixer, which essentially multiplies its input with a periodic signal supplied by another circuit block, the local oscillator (LO).

Fig. 1.1 shows a typical transceiver of a wireless system. In the transmitter side, the baseband signal is modulated to the carrier frequency in two mixer stages and after filtering it is transmitted to the air by the antenna. The signal is received by the antenna of the receiver, amplified by a low-noise amplifier (LNA), and translated to baseband again in two stages. Notice that many filters are used in several locations in both the receiver
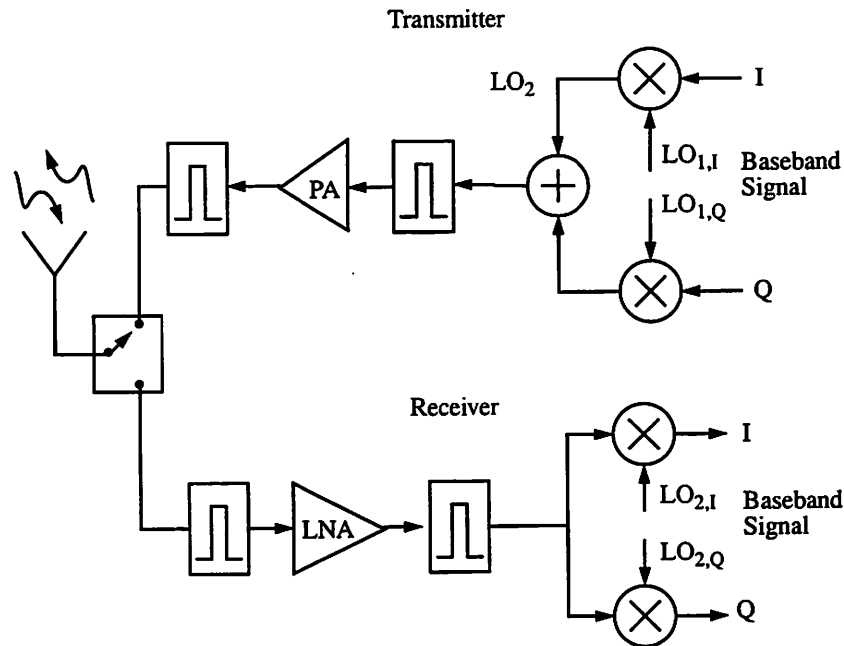
Figure 1.2: Simplified homodyne transceiver architecture.

and the transmitter, to reject undesired out-of-band signals and the image signals before

every mixer. The concept of the image will be explained further in the next chapter. The

receiver architecture of Fig. 1.1 is known as *heterodyne* or *superheterodyne* and because of

the high-quality passive image-rejection filters required it is mostly appropriate for discrete

component implementation. Several other architectures are investigated for implementation

in integrated technology. For example in the direct-conversion transceivers shown in Fig. 1.2

a single mixer performs the frequency translation, but these receivers face other challenges

such DC offsets [27, 68]. The Weaver architecture [100, 72] and low-IF receivers which

employ complex polyphase filters at low frequencies are other alternatives [12, 11].

Since mixers perform frequency translation, they are not linear-time-invariant

(LTI) systems. As we shall see they do not operate with a fixed operating point and

for this reason they are not as easy to analyze as other analog circuits such as amplifiers. For this reason designers usually have limited insight into the mixer operation and rely heavily on simulation or empirical approaches.

Mixer simulation on the other hand can be a difficult task. Traditional SPICE-type circuit simulators can only simulate some mixer performance characteristics such as gain and nonlinearity indirectly through transient analysis. *SpectreRF* [88] is a new simulator which is capable of efficiently simulating gain, noise and nonlinearity of periodically-driven circuits such as mixers. Although this simulator greatly facilitates the task of the designer, the simulations can still be time and memory consuming. In addition, simulations (especially of nonlinearity) can be unreliable if the transistor model is non physical (if for example it contains discontinuities in the transition from one region of operation to an other) as we shall see in chapter 6.

1.3 Research Goals

This thesis intends to facilitate the design of current - commutating CMOS mixers. It presents an analysis of the mixer operation and concentrates on the nonidealities that limit its performance. It aims to provide intuition by identifying the dependence of the performance on the design parameters. Simplified expressions and graphs of normalized parameters are derived from which one can obtain an approximate quantitative prediction of the performance. Using the results provided here, a designer can obtain a close to optimal design, without having to perform a large number of simulations. Furthermore the designer can easily decide on the fitness of a particular CMOS process for a particular set

of specifications.

1.4 Previously Published Related Work

Noise analysis of single-transistor bipolar mixers has been presented in [53, 54, 55]. The problem of noise prediction in mixers in general and bipolar current-commutating mixers in particular has been examined in [32, 31]. The intermodulation generated by a bipolar mixer switching pair has been examined in [56, 31]. Passive CMOS mixers on SOI technology have been examined in [39]. Several papers such as [83, 48, 49] and books [50] have been written by the microwave design community on single-device high-frequency mixers. The above publications contain interesting techniques some of which were adopted by the analysis of active CMOS mixers presented here. Some of the results described in this dissertation have also been published in [89, 90]. A more recent paper [52] concentrates on the flicker noise of low-power and low-voltage CMOS mixers.

1.5 Thesis Organization

Chapter 2 describes in abstract and theoretical terms the mixing operation and the nonidealities of a practical mixer. It introduces the performance metrics used to characterize mixers.

Chapter 3 presents several mixer topologies that can be implemented in CMOS processes and discusses their advantages and disadvantages. It introduces the class of current-commutating CMOS mixers for which the analysis presented in the following chapters applies.

Chapter 4 presents a noise analysis of current-commutating CMOS mixers. The contribution of all internal and external noise sources to the output noise is calculated. The noise figure of some simple mixer structures is estimated by computing only a few parameters or by reading them from provided normalized graphs. Simple explicit formulas for the thermal and flicker noise introduced by a switching pair are derived, and the upper frequency limit of validity of the analysis is examined. Although capacitive effects are neglected, the results are applicable up to the GHz frequency range for modern submicron CMOS technologies. The deviation of the device characteristics from the ideal square law is taken into account and the analysis is verified with measurements.

Chapter 5 is a noise analysis of a CMOS inductively degenerated, conjugately matched transconductance stage. These stages can be used in active mixers when the input is provided off chip. A more sophisticated transistor noise model is adopted here than used in chapter 4. The results of this chapter can be applied directly to the design of LNAs.

Chapter 6 presents an analysis of the nonlinearity of CMOS transconductance stages used in active mixers. Expressions are derived for the degenerated single-ended common-source, common-gate and differential-pair stages. The single-ended common-source stage is examined in depth and several approximations are made in order to provide both quantitative and intuitive results. The body-effect nonlinearity is shown to be significant for large degeneration and the degeneration impedance above which the body-effect nonlinearity dominates, is derived. The output third-order intercept

point is investigated for inductive and resistive degeneration, with and without a source matching restriction.

In Chapter 7 the nonlinearity behavior of the CMOS current-switching transistor pair is investigated. By treating the mixer as a periodically-time-varying weakly-nonlinear circuit we study the distortion-causing mechanisms and we predict the mixer distortion performance. Normalized graphs are provided from which the designer can readily estimate the mixer nonlinearity for particular process and design parameters. A simple CMOS transistor model appropriate for our calculations, which also takes into account deviation from the square law, is adopted. The significance of a physical transistor model for reliable distortion simulation is demonstrated. The predictions of our analysis are compared with simulation results and with experimental data.

Chapter 8 presents four different single-balanced active mixer designs. Implementation details about the mixer core and the on-chip LO buffer are discussed. It is demonstrated how the design parameters affect the conversion gain, the noise figure and the nonlinearity. Predicted performance characteristics are compared with measurements.

Chapter 9 concludes the thesis and identifies topics for future research.

Appendix A is a qualitative discussion on the implications of the time-varying – or cyclostationary – nature of the noise generated by mixers. The usual noise performance metric, the noise figure considers only the time-average of the noise power and discards the time variation. We show that in the majority of the practical cases the mixer noise figure is sufficient to provide an accurate noise performance prediction of

the overall system, but we identify cases in which the time variation is significant.

Appendix B is a derivation from first principles of the concept of the time-varying (and time-invariant) Volterra series. It is the basis for the analysis performed in chapters 6 and 7.

Appendix C is a discussion on flicker noise generated by CMOS devices. In chapter 4 we calculate the transfer function of the flicker noise from where it is generated to the output, but the flicker noise device model is not well established. The generating mechanisms and a model of flicker noise in MOSFETS has been a subject of controversy. We present an overview of the theories in this field and we identify the need for theoretical and experimental study of flicker noise of CMOS devices with time varying operating point.

Chapter 2

Mixer Fundamentals

2.1 Introduction

In this chapter we describe in abstract terms the operation and the nonidealities of mixers and we introduce the metrics which characterize their performance. The concepts presented in this chapter are general, independent of topology.

A mixer is a circuit block with three ports, the input port, the local-oscillator (LO) port, and the output port as shown in Fig. 2.1. This symbol for the mixer has already been used in Fig. 1.1 and Fig. 1.2. In the literature an ideal mixer is usually a multiplier whose output equals the product of the input with a sinusoidal LO waveform. The output spectrum in this case is the input spectrum shifted by the LO frequency.
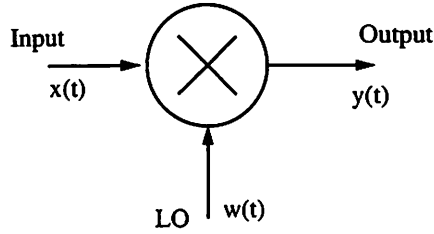
Figure 2.1: Mixer symbol.

2.2 Mixer Linear Operation

In a practical mixer, the input signal is usually small, such that it causes only a small perturbation to the current and voltage waveforms of the circuit – similarly to the input signal of an amplifier which causes only a small perturbation around the operating point. The LO signal is a strong periodic waveform such that it largely determines the current and voltage waveforms. In this thesis we call the voltages and the currents in the mixer in the absence of input signal the *periodically time-varying operating point*. We shall see next that for small input signals the input-output relation of a mixer is linear. A mixer is a *linear-periodically-time-varying* (LPTV) system and such systems provide frequency translation.

2.2.1 Low Frequency

Let us consider first the low-frequency case in which the mixer is a memoryless system. The output is a function of the instantaneous values of the input signal $x(t)$ and LO signal $w(t)$

$$y(t) = F(x(t), w(t)) \tag{2.1}$$

and since $x(t)$ is small a first-order Taylor expansion provides

$$y(t) = p_o(t) + p_1(t)x(t). \tag{2.2}$$

Waveform $p_o(t)$ is the output without an input signal present, does not contain signal information and will be omitted below. The mixer performs frequency translation by multiplying its input signal $x(t)$ with a periodic waveform $p_1(t)$. Waveform $p_1(t)$ is determined by the LO waveform and the mixer implementation. Since multiplication in the time domain is convolution in the frequency domain, the spectrum of the output signal $y(t)$ is given by

$$Y(f) = \sum_{n=-\infty}^{+\infty} p_{1,n} X(f + nf_{LO}) \tag{2.3}$$

where $X(f)$ is the spectrum of the input signal and $p_{1,n}$ are the Fourier coefficients of $p_1(t)$. The output spectrum consists of copies of the input spectrum shifted in frequency to integer multiples of the LO frequency and weighted by different coefficients. These coefficients are called *conversion gains*.

2.2.2   High Frequency

Let us now examine frequency conversion at high frequencies. We will assume for simplicity that the input signal is a single tone in complex representation $e^{j2\pi ft}$. Such a tone is not a physical waveform, but the response of the system to a physical input such as a sinusoid can be found as the response to a sum of single tones, and we will follow this approach even when we consider nonlinearities below. The mixing operation is ideally linear and is accurately described by the *periodically time-varying transfer function* $P_1(t, f)$, which is similar to the periodic gain $p_1(t)$ at low frequencies, but also depends on the frequency

of the input signal. For a proof see [107], or appendix B and discard the nonlinear terms. The transfer function of a linear-time-invariant (LTI) system is defined as

$$e^{j2\pi ft} \rightarrow \boxed{\text{LTI}} \rightarrow P_1(f)e^{j2\pi ft} \qquad (2.4)$$

meaning that the response of an LTI system to $e^{j2\pi ft}$ is $P_1(f)e^{j2\pi ft}$. Similarly the transfer function of an LPTV system such as the mixer is defined as follows

$$e^{j2\pi ft} \rightarrow \boxed{\text{LPTV}} \rightarrow P(t,f)e^{j2\pi ft} \qquad (2.5)$$

As in the low frequency case, in the frequency domain convolution provides the output spectrum

$$Y(f) = \sum_{n=-\infty}^{+\infty} P_{1,n}(f)e^{j2\pi(f+nf_{LO})t} \qquad (2.6)$$

which is similar to relation (2.3), but here the conversion gains $p_{1,n}$ have been replaced by the *conversion transfer functions* $P_{1,n}(f)$, the Fourier components of $P_1(t,f)$.

Observe that a mixer transfers a signal from a single frequency band to several output frequency bands. Similarly the signal at one output band originates from multiple input frequency bands. These multiple frequency bands at the input and the output are called *sidebands*, *image bands* or *images*.

Practical mixers are usually used for frequency translation by one LO frequency multiple because in this mode they usually provide the highest conversion gain and the rest of the output components are often removed by filtering. If the output signal frequency is lower than the input signal frequency, the mixer is said to perform *downconversion* while in the opposite case it is said to perform *upconversion*. Transmitters generally employ upconverters while receivers employ downconverters, (although there are some receivers,

such as broadband cable TV tuners and instrumentation equipment front-ends that first upconvert the signal and then downconvert it).

2.2.3 Conversion Gain

As introduced in the previous section, the conversion gain can be defined as the ratio of the magnitude of the output signal in the output frequency band over the magnitude of the input signal. So far we have not specified what physical quantities the input and output signals represent. If they are both voltages one can define the *voltage conversion gain* similarly to the voltage gain of an amplifier

$$G_c = \frac{V_{out}}{V_{in}} \quad and \quad G_c(dB) = 20 \log_{10}(\frac{V_{out}}{V_{in}}) \tag{2.7}$$

where V_{in} is the amplitude of the input signal considered sinusoidal and V_{out} is the amplitude of the output signal when all the out-of-band components are eliminated. Other similar definitions are possible, for example if the input is voltage and the output is current, the *conversion transconductance* can be used.

Microwave and RF designers usually consider the input and output signals to be power rather than voltage or current. The mixer input and output ports have some finite impedance, the source providing the signal to the input port has a finite impedance and the mixer delivers power to a load impedance. We will consider here that the source and load impedances are specified by the application, but the designer can use impedance transformation networks between the mixer ports and these impedances, and in fact the transformation networks can be considered part of the mixer. Similarly to the amplifier case [62, p. 606] several power gain definitions are possible, but the most commonly used
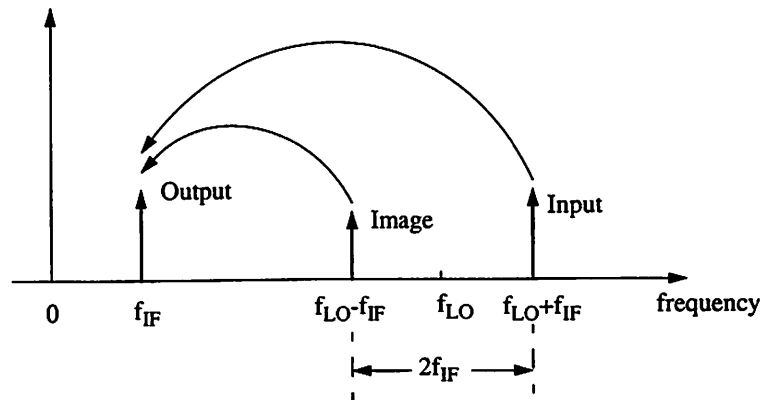
Figure 2.2: The input signal and its image appear at the mixer output.

is the following.

The *Power Conversion Gain* is the ratio of the power delivered to the load when the power of all out of band components is excluded, over the available power of the source. It depends on both the input and the output impedance transformation networks. This definition is also fitted to the way the mixer gain is measured in the lab, where the power delivered to the load (i.e. the 50Ω input impedance of the spectrum analyzer) is measured and the available power of the source (i.e. signal generator) is known. The Power Conversion Gain is maximized when both the input and the output are conjugately matched (but section 2.7 contains a more accurate discussion of this issue).

Very often the output signal is weaker than the input signal which translates to a negative conversion gain when expressed in dB. The term *Conversion Loss* is then used which equals the conversion gain in dB without the negative sign.

2.3 Image Problem

We saw that frequency translation from several input bands or image bands appears at a single output frequency. This effect usually causes problems. Images that are located far away from the input signal band in the frequency domain are easily filtered and do not represent a problem. However, in a downconverter employed in a heterodyne receiver, where the output is obtained at frequency f_{IF} and the input signal is frequency $f_{RF} = f_{LO} + f_{IF}$, the signal in the image band $f_{IM} = f_{LO} - f_{IF}$ is also translated to f_{IF}, as shown in Fig. 2.2 (we are considering real, not complex signals here). Either one of the input frequency bands can be used for the useful frequency translation. When the input signal is at $f_{LO} + f_{IF}$ it is said that *low-side injection* is used, while when it is at $f_{LO} - f_{IF}$ the term *high-side injection* is used. The spectrum usually contains undesired information or noise at the image frequency which contaminates the output.

Filters are often used before the mixer to eliminate this image. The input signal and its image are in distance $2f_{IF}$ and filtering becomes harder when a low f_{IF} is used. The image problem disappears when direct conversion or *zero IF* architecture is used. Finally, the image problem can be eliminated with image rejection mixers which employ phase shifters as shown in Fig. 2.3, with appropriate system architectures which employ an arrangement of several mixers, such as the Weaver architecture [100, 72], shown in Fig. 2.4. A different architecture that solves the image problem employs low IF frequency and complex polyphase filters [13, 12]. Since image cancellation is in practice limited by imperfect device matching, these techniques are often used in combination with some mild filtering which enhances the image rejection.
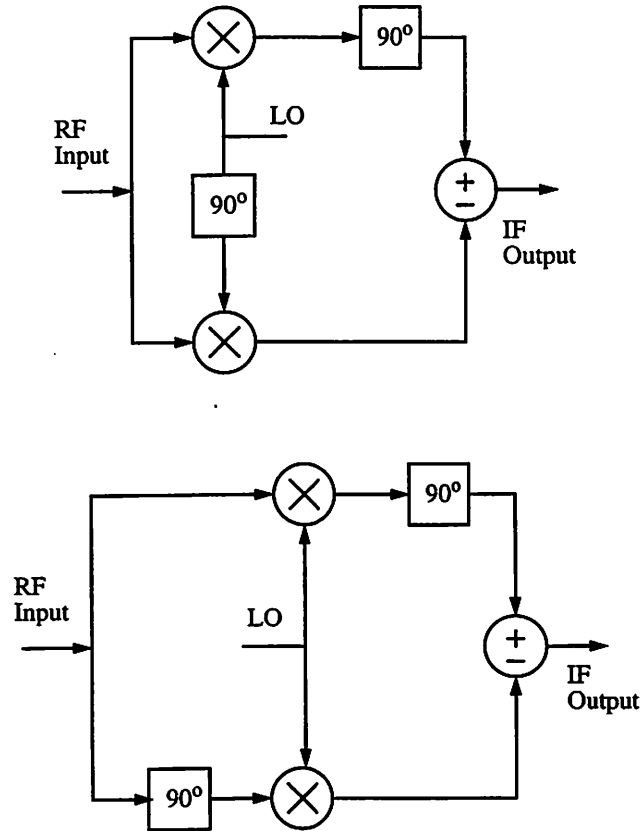
Figure 2.3: Mixer configurations which reject the image.

2.4 Port-to-Port Isolation

The only useful signal transfer in the mixer is from the input signal band of the input port to the output signal band of the output port. Any other power transfer is parasitic and is possibly detrimental. An example of a mixer with bad port-to-port isolation is the single-diode mixer in which all three ports share the two diode terminals. In this case isolation is provided exclusively by filtering.

Leakage of even a small amount of the strong LO signal to the output acts as an interferer for the circuits following the mixer, while LO leakage to the input port in a receiver
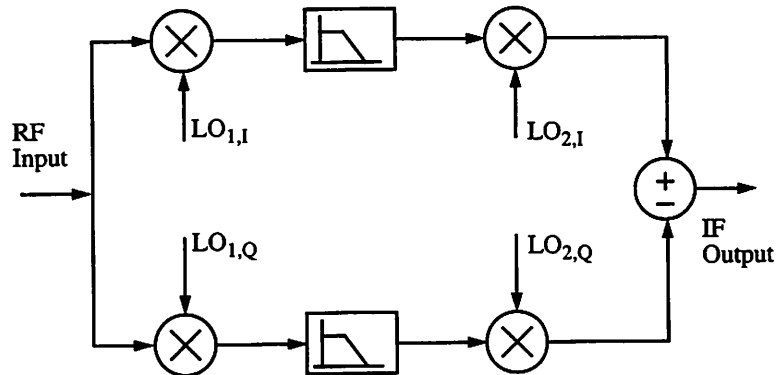
Figure 2.4: The Weaver architecture.

can reach the antenna and be transmitted causing problems to other users. Furthermore LO leakage to the input port, essentially squares the LO signal and creates DC offsets at the output which are a significant problem in direct conversion applications. Leakage of the input or output to the LO port can disturb the LO operation. Direct leakage from the input to the output port at the input signal band can pass DC components generated by nonlinearities before the mixer to the output which as mentioned cause problems in direct conversion receivers.

For these reasons, the specifications of the mixer often require that the isolation between two specified ports is higher than a certain ratio expressed in dB. In traditional discrete-component transceiver implementations, filters have been used to satisfy the isolation requirements. In modern integrated solutions however, high-frequency filtering is not easily implemented and besides the frequency spacing between the different frequency bands can be small or zero. For example in direct downconversion the LO and input signals are indistinguishable in frequency. Therefore balanced structures performing cancellation of the unwanted signals are extensively used (often in combination with some filtering).

2.5 Single-Balanced and Double-Balanced Mixers

Single-balanced mixers are mixers whose topology has the inherent ability to reject either one of the LO signal or input signal at the output. *Double-balanced* mixers reject both the LO and input component at the output. Port isolation through cancellation is desirable particularly in integrated solutions, as mentioned in the previous paragraph. In practice however the rejection of the undesired signals in balanced structures is limited by the several factors such as imperfect device matching, parasitic capacitances, coupling through the substrate and the package parasitics. One significant advantage of the balanced structures which reject the LO component at the output is that they also cancel or heavily suppress any noise coming from the LO port, which can otherwise significantly deteriorate the mixer noise performance.

2.6 LO Power Requirement

The amount of power required to the LO port of the mixer is often a concern and usually it is desirable that the mixer can operate properly with a relatively low amount of power delivered to the LO port.

2.7 Input, Output and LO Port Impedances

Since the mixer is a periodically time-varying circuit, in the general case the input, output and LO port impedances are periodically-time-varying, in a manner similar to the periodically-time-varying transfer function examined earlier in section 2.2. If a single-

frequency test current-signal $I(t) = I_o e^{j2\pi ft}$ is applied to such an impedance $Z(t, f)$, voltage

components at all frequencies $f + nf_{LO}$ will be generated

$$V(t) = Z(t, f) \cdot I(t) = I_o \sum_{n=-\infty}^{n=+\infty} Z_n(f) e^{j2\pi(f+nf_{LO})t}. \qquad (2.8)$$

Of these voltage components the one that is at the same frequency as the input test signal is

measured when we determine the input impedance with an instrument such as the network

analyzer. Then, the measured value is $Z_0(f)$ or the time-average of $Z(t, f)$.

If we model the input signal source with a sinusoidal voltage source at frequency f

in series with a constant source impedance, the current flowing through the voltage source

will have components at all frequencies $f + nf_{LO}$. However, only the current component at

the same frequency as the voltage excitation absorbs source power. This current component

is determined by the time-average input source impedance, and therefore the power delivered

from the source to the mixer is maximized when the source is conjugately matched to the

time-average mixer input impedance.

Let us now model the mixer output with a voltage source whose value equals the

output voltage without load, in series with a periodically-time-varying output impedance.

A time-invariant load is connected to this output. If the mixer input is at frequency f,

the mixer output voltage source contains components at all frequencies $f + nf_{LO}$. Assume

for example that the power delivered to the load at frequency $f + f_{LO}$ is desired. Now

power from the output voltage source at frequency $f + f_{LO}$ is maximized when the load

is conjugately matched to the time-average output impedance. However, power at the

frequency $f + f_{LO}$ is delivered to the load from other combinations of frequency components

of output voltage source and output impedance and it is not exactly accurate to consider

that matching the load to the time-average output impedance provides maximum power delivery to the load. In practical mixers however matching the load to the time-average output impedance usually provides very close to maximum power delivery.

Similar considerations hold for matching the LO port impedance. If the LO source is a single-frequency sinusoidal voltage source in series with a constant impedance, then matching the time-average LO port impedance to the LO source provides exactly maximum LO power transfer, otherwise the above matching provides close to optimal power transfer.

2.8   Input Return Loss

It is usually desirable that the impedance of the input port of the mixer is matched to the source, in order to eliminate reflections to the source. In a receiver, such reflections can reach the antenna and be retransmitted causing problems to other users. To avoid signal loss and noise, a lossless matching network is often used. Usually the input impedance of the mixer is made real such that the matching condition also satisfies the conjugate matching condition which provides maximum power transfer and maximum mixer gain. A metric for the reflection is the *input return loss* defined in *dB* as

$$RL = -20 \log |\Gamma| dB \qquad (2.9)$$

where Γ is the input reflection coefficient defined as

$$\Gamma = \frac{Z_{in} - Z_o}{Z_{in} + Z_o} \qquad (2.10)$$

and Z_{in} and Z_o are the mixer input impedance and source impedance (usually considered 50 Ω in discrete component implementations) respectively. In the general case that the

input impedance is time-varying, Z_{in} represents the time-average as mentioned in section 2.7.

A different metric to express the reflections from the input is the *standing wave ratio* (SWR) (or *voltage standing wave ratio*) (VSWR), which is defined as the ratio of the maximum to the minimum voltage amplitude on a transmission line with characteristic impedance Z_o terminated with Z_{in}. It can be shown to be equal to

$$SWR = \frac{1 + |\Gamma|}{1 - |\Gamma|} \tag{2.11}$$

Similar considerations hold for the LO port where reflections can disturb the LO operation and the reflection coefficient, input return loss, and standing wave ratio are defined similarly. Unlike the input port, in many cases a lossy matching network can be used at the LO port, since some LO signal loss and some noise can usually be tolerated (for example noise is rejected in double-balanced structures). If however a lossless matching network is used and the input impedance is made real, reflection elimination also provides maximum LO power transfer which is in many cases desirable.

2.9   Noise

It is a fundamental property of the electronic devices to generate noise of several kinds, such as thermal, shot and flicker [61]. The noise introduced by the mixer (and every block in a receiver) is a concern because it can mask a weak desired signal. The mixing function is inherently noisy, because while it transfers signal only from the input band to the output, it transfers noise from multiple frequency bands. The mixer is a significant noise contributor in most communication systems.

2.9.1 Mixer Noise Figure

The Noise Figure (NF) is a metric of the degradation of the signal to noise ratio (SNR) in the mixer, and is defined [18, 2] as

$$NF = \frac{SNR \ at \ the \ input \ in \ the \ input \ signal \ band}{SNR \ at \ the \ output \ in \ the \ output \ signal \ band} \qquad (2.12)$$

The input signal is considered provided by a source with an impedance containing a finite real part. The noise present at the input is then considered thermal noise of the source impedance. If the signal band is very narrow, the conversion gain and noise-power spectral density (PSD) in the input and output signal bands are almost flat. In this case the NF defined in (2.12) equals the following quantity which is also known as *spot NF*

$$NF = \frac{S_o}{S_{o1}}. \qquad (2.13)$$

In this expression

S_o is the total output noise per unit bandwidth including the contribution from the source impedance at all frequencies, and

S_{o1} is the part of the output noise per unit bandwidth due to the noise generated by the source impedance only in the input frequency band.

According to the IEEE definition the source impedance temperature is always $290^o K$ (even when the temperature of the circuit is different) [2]. The NF depends on the source impedance and a resistive impedance of 50 Ω is used by convention when RF block specifications are given.

2.9.2 Single-Sideband and Double-Sideband Mixer NF

In heterodyne receivers the mixer image band is not part of the input signal band and therefore the image does not contribute to S_{o1}. This is called the *Single-Sideband NF*. In direct-conversion receivers however the image band is part of the input signal band and contributes to S_{o1} and in this case the NF is called *Double-Sideband*. Assuming equal conversion gain from both sidebands and that source noise level is equal in the two sidebands, one can easily see that for the same mixer the single-sideband NF is $3dB$ higher than the double-sideband NF. This statement assumes that the source impedance contributes equal amount of output noise to S_o in both cases which is the case if filters are not present at the input. If however an input filter is employed which rejects the source noise outside the signal band, the difference is lower than 3 dB, but it approaches 3 dB if the noise generated inside the mixer dominates the output noise.

2.9.3 NF of an LTI Block

The NF is defined similarly for most communication circuit blocks such as LNAs and filters, and there the definition is even simpler since there is not a noise frequency-translation issue. A noisy linear time-invariant two-port (one input and one output port) can be represented with the equivalent input voltage and current noise generators preceding the noiseless network as shown in Fig. 2.5 [70]. The two generators are in general correlated and for this reason the voltage noise generator is partitioned into two voltage sources, one uncorrelated and one fully correlated with the current generator. Let us denote the current noise generator with I_n, the uncorrelated part of the voltage noise generator with V_n and its
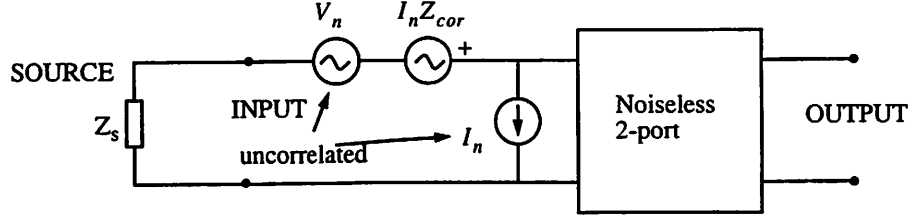
Figure 2.5: Equivalent representation of a noisy LTI twoport.

correlated part with $I_n Z_{cor}$ where $Z_{cor} = R_{cor} + j X_{cor}$ is called the correlation impedance.

Also let $Z_s = R_s + j X_s$ represent the source impedance and G_n, R_n be related with the power spectral densities of I_n and V_n as follows

$$\overline{I_n^2}/\Delta f = 4kTG_n \tag{2.14}$$

$$\overline{V_n^2}/\Delta f = 4kTR_n \tag{2.15}$$

where k is Boltzman's constant and T is the absolute temperature. It can be shown [4, 98] that the NF can be given as

$$NF_{Z_s} = 1 + \frac{R_n}{R_s} + \frac{G_n}{R_s} |Z_{cor} + Z_s|^2 \tag{2.16}$$

It can also be shown that the source impedance $Z_{s,opt} = R_{s,opt} + j X_{s,opt}$ which provides the minimum noise figure is

$$R_{s,opt} = \sqrt{\frac{R_n}{G_n} + R_{cor}^2} \tag{2.17}$$

$$X_{s,opt} = -X_{cor} \tag{2.18}$$

and the minimum noise figure is given by

$$NF_{min} = 1 + 2G_n R_{cor} + 2\sqrt{G_n R_n + (G_n R_{cor})^2} \tag{2.19}$$
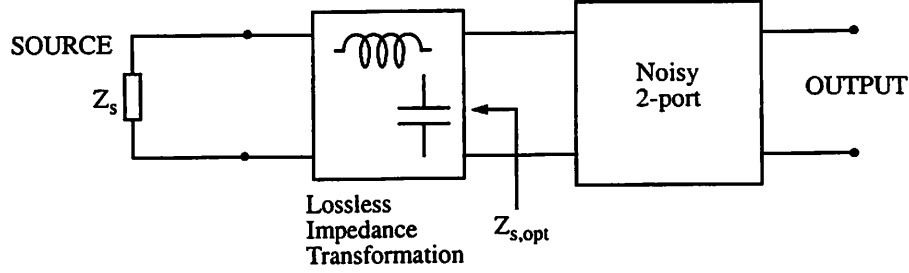
Figure 2.6: Transforming the source impedance to the optimal for noise figure value.

For a different source impedance the NF is

$$NF_{Z_s} = NF_{min} + \frac{G_n}{R_s} \cdot |Z_s - Z_{s,opt}|^2 \qquad (2.20)$$

From equation (2.16) we can see that if the NF is known for one source impedance value it is not generally possible to find its value for a different source impedance, as one has one equation and needs to solve for the quantities R_n, G_n and Z_{cor}. If however one of the two independent input noise generators is negligible it is possible to find the NF for a different source impedance. For example if G_n is negligible then $(NF_{Z_s} - 1)R_s$ is a constant, while if R_n is negligible $(NF_{Z_s} - 1)G_s$ is a constant where $G_s = R_s/|Z_s|^2$ is the source conductance.

Usually the designer has no control over the source impedance, for example the source impedance of the testing equipment is almost always 50Ω, and in discrete component implementation the implied input and output impedance of the RF blocks is also 50Ω. The discrete filters have their specified in-band and out-of-band attenuation only if they are terminated to a 50Ω impedance at the input and the output. The designer however can use a lossless impedance transformation network to transform the 50Ω (or any other lossy source impedance) to the optimal source impedance as shown in Fig. 2.6. Since the NF

is the ratio of the input over the output SNR and the SNR is the same before and after the lossless impedance transformation network, the NF of the noisy two-port for source impedance equal to the output impedance of the matching network, is equal to the NF of the combination of the impedance transformation network and the noisy two-port for the source impedance of 50Ω.

The mixer NF depends on the source impedance not only in the signal-band but also in all image bands that contribute output noise. Assuming that the noise behavior of the mixer does not depend heavily on the value of the out-of-band terminations, one could make similar considerations for the mixer as the ones made for the LTI blocks in this section. The latter assumption is often true if for example the mixer is preceded with a band-pass filter, or when the noise generated inside the mixer is much higher than the noise contribution of the source. One can then represent the mixer with a noiseless block preceded by two in-band input noise generators similarly to the LTI blocks. The role of the in-band source impedance would be similar and the optimal in-band source-impedance could be found. Under these assumptions the mixers can be treated indistinguishably from the LTI blocks in the next paragraph dealing with the NF of a system of cascaded blocks.

2.9.4 Noise Figure of a Cascade of RF Blocks

The NF of a system of cascaded blocks [18] is

$$NF_{tot} = NF_1 + \frac{NF_2 - 1}{G_1} + \frac{NF_3 - 1}{G_1 G_2} + \cdots \tag{2.21}$$

where NF_i and G_i are the NF and gain of the i-th block in the chain, and the input is connected to the 1-st block. This expression shows that the contribution of the NF of

each block to the total NF is essentially suppressed by the gain of all the previous blocks. Therefore, the LNA NF usually dominates the receiver NF, with the first mixer being the second highest contributor. Equation (2.21) assumes input-output matching of all blocks to a common impedance level (usually 50 Ω). This equation can be generalized as follows [68]

$$NF_{tot,Z_s} = NF_{1,Z_s} + \frac{NF_{2,Z_{out1}}}{G_{av,1}} + \frac{NF_{3,Z_{out2}}}{G_{av,1}G_{av,2}} + \cdots \quad (2.22)$$

where

$$G_{av,i} = \left| \frac{Z_{in,i}}{Z_{in,i} + Z_{out,i-1}} \right|^2 \frac{R_{out,i-1}}{R_{out,i}} A_{v,i} \quad (2.23)$$

is the available gain of the i-th block, the ratio of the available power at its output over the available output power at its input.

$A_{v,i}$ is the voltage gain of the $i - th$ block without output load,

$Z_{in,i}$ is the input impedance of the $i - th$ block and $R_{in,i}$ is its real part,

$Z_{out,i}$ is the output impedance of the $i - th$ block and $R_{out,i}$ is its real part,

$Z_s \equiv Z_{out,0}$ is the source impedance and $R_{out,0} = R_s$ is its real part, and

the second index in the NF denotes the source impedance for which it is defined.

It is useful to know that the NF of a passive network equals its loss. This is true even when the input and output impedances of the network are different and the input is not matched to the source impedance as long as the loss is defined similarly to the available power gain above, the ratio of the available power of the source over the available output

power. This property is useful for the noise figure of passive filters, and also passive mixers as we shall see in the next chapter.

2.9.5 Cyclostationarity

There are two reasons why the noise generated in a mixer has periodically time-varying statistics. First, the operating point of the devices changes periodically with time. Second, the processing of the signal from the point at which noise is generated to the output can be periodically time-varying [32]. A random process whose statistics are periodic functions of time is called cyclostationary, and a complete description of such a process requires a time-varying power spectral density (PSD) $S(f, t)$ [21]. This is different from the wide sense stationary (WSS) noise generated by a linear time-invariant circuit.

While introducing the NF as a metric for the noise performance of a mixer, we should mention that it is sufficient for most practical cases but it is not a complete characterization. Because of the cyclostationary nature of the noise, expressions (2.21) and (2.22) may not be accurate under certain conditions. The implications of this issue are examined in depth in appendix A.

2.10 Nonlinearity

The nonlinearity mechanisms and performance metrics of mixers are very similar to those of time-invariant systems. Therefore we will present the nonlinearity theory for the time-invariant case and we will extend the discussion to cover periodically time-varying systems.

2.10.1   Low-Frequency Nonlinearity

At low frequencies time-invariant systems are described in general by a nonlinear relation between the input signal x and the output signal y.

$$y = F(x) \tag{2.24}$$

For small input signals this relation can be linearized by taking a first-order Taylor expansion. For larger signals the system becomes weakly nonlinear and is described better with a power series which can be derived by keeping higher-order terms in the Taylor expansion. Thus

$$y = b_1 x + b_2 x^2 + b_3 x^3 + \cdots \tag{2.25}$$

where the signal independent term has been omitted.

2.10.2   High-Frequency Nonlinearity

At high frequencies the distortion characteristics of weakly nonlinear circuits can be derived from Volterra series, which are similar to the power series, but their coefficients depend on the frequency of the processed signal [104], [101]. They have the form

$$y = B_1(f_a) \circ x + B_2(f_a, f_b) \circ x^2 + B_3(f_a, f_b, f_c) \circ x^3 + \cdots \tag{2.26}$$

The operation $\circ$, is often called *Kronecker product* and acts as follows. If the signal x is a sum of k single complex tones at frequencies $f_1, f_2, \ldots, f_k$, the term x^n contains output tones at all frequencies $f'_1 + f'_2 + \cdots + f'_n$ where each f'_i can be any of $f_1, f_2, \ldots, f_k$. The contribution of each output tone is found by multiplying it with the n-th order Volterra coefficient evaluated at frequencies $f'_1, f'_2, \ldots, f'_n$. The single tones and the Volterra coefficients are

complex quantities, but when the sum of the complex tones x represents a real signal, the output complex tones always combine to give a real output signal. For example if the input is the sum of two cosines

$$x = \cos(2\pi f_1 t) + \cos(2\pi f_2 t) \tag{2.27}$$

the third-order nonlinear term x^3 contains

$$\frac{3}{4}\cos(2\pi(2f_1 - f_2)t) \tag{2.28}$$

whose output contribution is

$$\frac{3}{4}|B_3(f_1, f_1, -f_2)|\cos(2\pi(2f_1 - f_2)t + \angle B_3(f_1, f_1, -f_2)) \tag{2.29}$$

2.10.3 Weakly-Nonlinear Periodically-Time-Varying Systems

As we saw in section 2.2 a mixer is ideally an LPTV system. In practice however mixers exhibit nonlinearities and they behave rather as *weakly-nonlinear periodically-time-varying* systems. The input-output relation of such a system at low frequencies can be described with a power-series with periodically time-varying coefficients

$$y = p_o(t) + p_1(t)x + p_2(t)x^2 + p_3(t)x^3 + \cdots \tag{2.30}$$

This expression can be derived from (2.1) similarly to (2.2) by keeping higher-order terms in the Taylor expansion. Discarding the signal independent term of (2.30) and taking a Fourier expansion of the time-varying coefficients we obtain the following expression

$$y = \sum_{n=-\infty}^{+\infty} \left[p_{1,n}x + p_{2,n}x^2 + p_{3,n}x^3 + \cdots \right] e^{jn2\pi f_{LO}t} \tag{2.31}$$

where f_{LO} is the LO frequency. When the input signal is real, for every complex term its conjugate is also present in the sum, with which it combines to provide a real output. With real input signal (2.31) becomes

$$
\begin{aligned}
y \;=\;& p_{1,0}x + p_{2,0}x^2 + p_{3,0}x^3 + \cdots \\
&+ \sum_{n=1}^{+\infty} 2Re\left\{\left[p_{1,n}x + p_{2,n}x^2 + p_{3,n}x^3 + \cdots\right] e^{j2\pi n f_{LO}t}\right\}
\end{aligned}
\qquad (2.32)
$$

If the input signal x is a cosine of frequency f, the magnitude of all the *spurious responses* because of the nonlinearity and the frequency translation at frequencies $nf_{LO} \pm kf$ can be found. Similarly the response of the mixer to a sum of cosines, can be found. Since multiplication with $e^{j2\pi n f_{LO}t}$ only performs frequency translation, most of the nonlinearity performance metrics that describe the mixer can be derived from a usual time-invariant power series similar to (2.25) which corresponds to the principal (or useful) frequency translation, usually by one LO multiple, i.e. $b_i = p_{i,1}$. Notice that it is in general possible for b_i to be complex.

In appendix B we show how a weakly-nonlinear periodically-time-varying system can be described at high frequency by a Volterra series with periodically time-varying coefficients

$$
y = P_1(t, f_a) \circ x + P_2(t, f_a, f_b) \circ x^2 + P_3(t, f_a, f_b, f_c) \circ x^3 + \cdots
\qquad (2.33)
$$

and taking again a Fourier expansion of the coefficients

$$
y = \sum_{n=-\infty}^{+\infty} \left[P_{1,n}(f_a) \circ x + P_{2,n}(f_a, f_b) \circ x^2 + P_{3,n}(f_a, f_b, f_c) \circ x^3 + \cdots\right] e^{j2\pi n f_{LO}t}
\qquad (2.34)
$$

As with low frequencies, time-invariant Volterra series similar to (2.26) that describe intermodulation in the frequency band of interest can be extracted by taking the first

Fourier coefficients of $P_1(t, f_a)$, $P_2(t, f_a, f_b)$, and $P_3(t, f_a, f_b, f_c)$ respectively. Assume that the nth time-varying Volterra coefficient has the following Fourier expansion

$$P_k(t, .) = \sum_{n=-\infty}^{\infty} P_{k,n}(.)e^{j2\pi n f_{LO} t} \qquad (2.35)$$

where a dot has replaced the frequency arguments. For a downconverter $P_{k,-1}(.)$ must be used if the input signal frequency f_s is higher than f_{LO}, or its conjugate if f_s is lower than f_{LO}. For an upconverter $P_{k,1}(.)$ must be used when the output frequency is higher than f_{LO} while the conjugate of $P_{k,-1}(.)$ must be used if the output frequency is lower than f_{LO}, but the result is essentially the same if the input signal frequency is low compared to f_{LO}.

1

2.10.4 Harmonic Distortion

When the input to a weakly nonlinear system described by a power series or a Volterra series is a cosine $x = \cos(2\pi f t)$ then tones at frequencies nf are generated at the output by the nonlinear term x^n. In particular if $n = 2k$ is even, x^n generates tones at all frequencies $(2k)f, (2(k-1))f, \ldots, 0$, while if $n = 2k - 1$ is odd, it contributes at frequencies $(2k-1)f, (2k-3)f, \ldots, f$. The DC contribution of the even-order terms creates DC offsets which are a problem in direct conversion applications, while the contribution at the fundamental of the odd-order distortion alters the gain of the linear input-output relation. The $n - th$ order *Harmonic Distortion* is defined as

$$HD_n = \frac{Ampl.\ of\ output\ tone\ at\ freq.\ nf}{Ampl.\ of\ output\ tone\ at\ fund.\ f} \qquad (2.36)$$

When HD_n is specified, the nonlinearites of order higher than n are neglected.

[1]We mention that $\overline{P_k(t, f_1, \ldots, f_k)} = P_k(t, -f_1, \ldots, -f_k)$ and $\overline{P_{k,n}(f_1, \ldots, f_k)} = P_{k,-n}(-f_1, \ldots, -f_k)$, and the overline here denotes the complex conjugate.

2.10.5 Third-Order Intermodulation

Consider first a weakly-nonlinear time-invariant circuit such as an amplifier described by (2.25) whose input consists of two cosines of equal amplitude at frequencies f_1 and f_2

$$x = A_o cos(2\pi f_1 t) + A_o cos(2\pi f_2 t). \tag{2.37}$$

Because of the third-order nonlinearity x^3 the output will contain undesired tones at frequencies $2f_1 - f_2$ and $2f_2 - f_1$ which are usually detrimental because they fall in the signal band.

$$x^3 = \cdots + \frac{3}{4} A_o^3 \cos(2\pi(2f_1 - f_2)t) + \frac{3}{4} A_o^3 \cos(2\pi(2f_2 - f_1)t) + \cdots \tag{2.38}$$

The third-order intermodulation is defined as the ratio of the magnitudes of the undesired output term over the desired linear term. Assuming low frequency, the power series (2.25) provides

$$IM_3 = \frac{3}{4} \left| \frac{b_3}{b_1} \right| A_o^2, \tag{2.39}$$

the same for both tones at frequencies $2f_1 - f_2$ and $2f_2 - f_1$. At high frequencies, the output tone at $2f_1 - f_2$ for example is given by (2.29). The third-order intermodulation is given by

$$IM_3 = \frac{3}{4} \left| \frac{B_3(f_1, f_1, -f_2)}{B_1(f_1)} \right| A_o^2 \tag{2.40}$$

It is worth noticing that at high frequencies the two intermodulation terms have in general different magnitudes while when there are no reactive effects they have always equal magnitude.

2.10.6   Third-Order Input Intercept Point (IIP_3)

The third-order input intercept point is defined as the magnitude of the input signal for which the intermodulation term at the output has the same magnitude as the linear term assuming that nonlinearities of higher than third-order are negligible. According to this definition, at low frequencies it is easy to find from the power series (2.25) that

$$x_{IIP_3} = \sqrt{\frac{4}{3}\frac{b_1}{b_3}} \tag{2.41}$$

and similarly at high frequencies the Volterra coefficients replace the power series coefficients. At high frequencies where the two intermodulation tones might have different magnitude, the larger of the two should be used in the specification of IIP_3.

Usually the IIP_3 is expressed in terms of available power of the source, in dBm units. If for example the variable x in power series (2.25) represents voltage and the input is matched to the source impedance R_s

$$P_{av,IIP_3} = \frac{1}{2 \cdot R_s}\frac{4}{3}\left|\frac{b_1}{b_3}\right| \quad IIP_3(dBm) = 10\log_{10}\left(\frac{P_{av,IIP_3}}{1mW}\right) \tag{2.42}$$

Fig. 2.7 shows how IIP_3 can be found from measurements in the region that nonlinearities higher than third-order are negligible. In this region for every $1dB$ of increase in the linear output term, the third-order term increases by $3dB$. It is a simple geometry exercise to show from this graph that

$$IIP_3 = P_{in1} + \frac{P_{out1} - P_{out3}}{2} \tag{2.43}$$

where P_{in1} is the available source power, P_{out1} is the power of the output linear term, and P_{out3} is the power of the output intermodulation term. The value of the output linear term when the input equals the IIP_3 is called *Output Third-Order Intercept Point* (OIP_3).
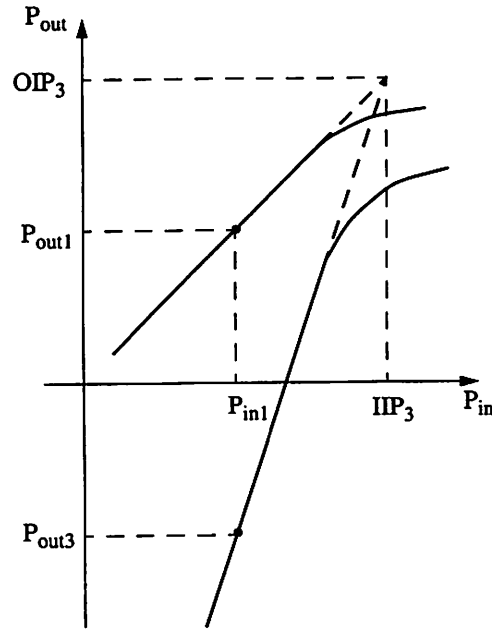
Figure 2.7: Graphical representation of IIP_3 and OIP_3.

In a weakly nonlinear time-varying system such as a mixer, IM_3 and the IIP_3 are defined similarly, only the output band where the output linear and intermodulation terms appear is different than the input frequency band.

2.10.7 Second-order Intermodulation

Second-order nonlinearity creates second-order intermodulation, i.e an input signal similar to (2.37) creates output tones at frequencies $f_1 - f_2$. The second-order intermodulation tones may fall in the output signal band, or create slowly time-varying offsets and disturb operation in direct conversion receivers. The second-order intermodulation and *Input Second-Order Intercept Point (IIP2)* are defined similarly to the third-order quantities examined in the previous paragraph. Second-order intermodulation specifications are

often provided to the mixer designers. Differential structures inherently reject the even-order nonlinearities. Such rejection however is in practice limited by the imperfect device matching.

2.10.8 IF/2 problem

It is possible that signal located at frequency $f_{LO} \pm f_{IF}/2$ appears at the output, and this effect is caused mainly by the second-order nonlinearity. Therefore this parasitic frequency translation is alleviated with differential structures.

One mechanism which is usually the dominant is through second-order harmonic distortion which transfers this signal to frequency $2f_{LO} \pm f_{IF}$ which in turn is translated to f_{IF}, by the parasitic frequency translation by two LO multiples. Another possible scenario is the following: a small amount of LO signal leaks to the input and through third-order intermodulation distortion it creates a tone at $f_{LO} \pm f_{IF}$ which is then transfered to f_{IF} together with the signal. A third possibility is that this tone is downconverted to $f_{IF}/2$ and then transfered to f_{IF} through second-order harmonic distortion of a subsequent non-linearity. Specifications for the rejection of the $f_{LO} \pm f_{IF}/2$ frequency components are often given to the designers.

2.10.9 1 dB Compression Point

As mentioned above, the odd-order nonlinearities can contribute components at the frequency of the output linear term and effectively change the gain. If the gain increases this phenomenon is called *gain expansion* while if it reduces it is called *gain compression*, with the latter being more common in practice. A related metric is the *1 dB compression*
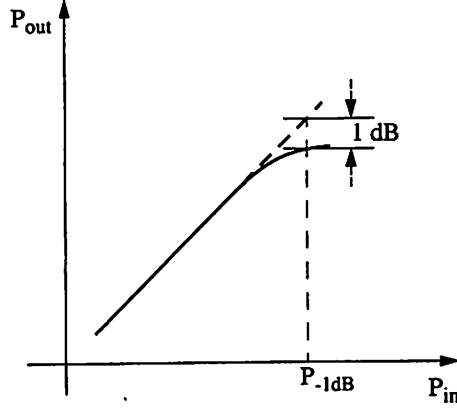
Figure 2.8: Graphical representation of the 1 dB compression point.

point defined as the value of the input signal for which the gain drops by $1dB$ and is depicted in Fig. 2.8. It is easy to show that if gain compression is caused exclusively by the third-order nonlinearity [68] the effective gain is given by

$$b_1 + \frac{3}{4}b_3 A^2 \qquad (2.44)$$

where A is the amplitude of the input sinusoidal tone. It is easy to find that the $1dB$ compression point is given by

$$x_{-1dB} = 0.33\sqrt{\frac{4}{3}\left|\frac{b_1}{b_3}\right|} \qquad (2.45)$$

Similarly to the IIP_3 the $1dB$ compression point is expressed in terms of the available power of the source (P_{-1dB}). It is easy to see from (2.45) and (2.41) that

$$P_{-1dB} = P_{IIP_3} - 9.6dB. \qquad (2.46)$$

More often than not however, the $1dB$ compression point is high, such that nonlinearities higher than third-order contribute significantly, and expressions (2.45) and (2.46) are not accurate.

2.10.10 Blocking and Desensitization

An other case in which nonlinearities can effectively change the gain of a circuit block is when the desired signal is weak but a strong undesirable signal is present at the input [58]. Assume that the input is

$$x(t) = A_1 \cos(2\pi f_1 t) + A_2 \cos(2\pi f_2 t) \qquad (2.47)$$

and that the signal with amplitude A_1 is the desired tone while the signal with amplitude A_2 is the out-of-band blocker. The third-order term x^3 contributes a component at frequency f_1 which together with the linear term provides the following tone at the fundamental frequency

$$y(t) = (b_1 + \frac{3}{2} b_3 A_2^2) A_1 \cos(2\pi f_1 t) \qquad (2.48)$$

If the blocker is large it can effectively alter and usually reduce the gain.

Chapter 3

CMOS Mixer Topologies

3.1 Introduction

Any circuit in which the input-output relation can become periodically time-varying by means of a periodically time-varying LO waveform can be used as a mixer. As an example, any electronic device whose characteristic equation demonstrates nonlinearity can be used as a mixer, if the sum of the input and LO signals is processed by the nonlinearity, since the intermodulation terms can be obtained as the desired output signal. In this chapter we present structures which can be used as mixers in CMOS technology and we introduce the class of the current-commutating CMOS mixers for which the analysis presented in the following chapters applies.

It is often desirable that the RF blocks provide gain. In a receiver for example, the signal at the antenna is possibly very weak and must be amplified significantly at the end of the receiver chain, typically by 100 dB. According to their ability to provide gain, mixers can be classified as *active* and *passive*. Active mixers are those whose topology can

$$I_{o1} = I_1 - I_2$$

$I_1 \downarrow$ $\downarrow I_2$

M1 M2

$+$

V_{LO}

$-$

$I_3 = I_B + i_s$

V_{in} M3
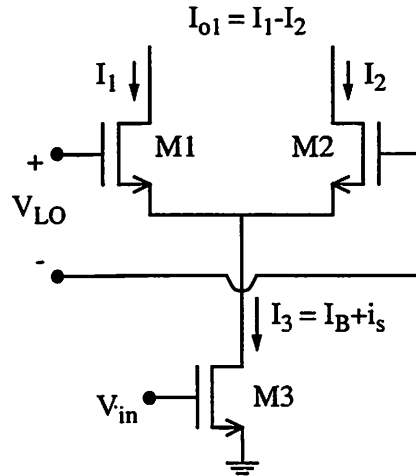
Figure 3.1: A typical single-balanced current-commutating CMOS mixer.

potentially amplify the signal, while passive mixers always have loss (at least in terms of power). A mixer which provides gain alleviates the gain requirements of the rest of the blocks in the chain, and the noise requirements of the blocks following the mixer.

3.2 Active Mixer Configurations

The most commonly used active mixers in integrated form are the current-commutating mixers. This section qualitatively discusses their operation and presents variations of these structures. It also presents some other active mixer topologies that can be found in the literature.

3.2.1 Current-Commutating Mixers

The current-commutating mixers [72, 37, 79, 105, 66, 60] employ switching pairs . A single-balanced active CMOS mixer is shown in Fig. 3.1. It consists of a transconductance

stage in this case a single transistor which transforms the input signal to current, and a switching pair driven by the strong LO waveform which commutates the current signal between the two output branches. If i_s is the small-signal at the output of the driver stage, assuming ideal switching, during the first half of the LO period the small signal output current is i_s while during the other half it is $-i_s$. This alternation in the sign of the output signal provides the desired mixing effect. When one of the devices of the switching pair is turned off, the other device is in the common-gate configuration and does not significantly contribute noise or distortion, at least when capacitive effects are negligible. The structure of Fig. 3.1 is single-balanced because ideally there is not direct feedthrough of the input signal to the output (of course in practice the input signal feedthrough cancellation is limited by device mismatch). If the output it taken single-ended neither the input nor the LO components are rejected and this configuration is unbalanced. Usually the output is obtained differentially because in addition to the balanced operation this way the conversion gain is higher.

A typical double-balanced active mixer or *Gilbert cell* is shown in Fig. 3.2. This circuit was originally designed with bipolar transistors [25] to operate as a precision multiplier, but it has been used widely as a mixer with the transistors driven by the strong LO signal acting as switches. The operation principle as a mixer is the same in CMOS technology. This mixer comprises a differential transconductance stage and two switching pairs. The output current is

$$I_o = I_{o1} - I_{o2} = (I_1 - I_2) - (I_5 - I_4) \tag{3.1}$$

where the above currents are defined in Fig. 3.2. The drains of the transistors of the two

$$I_o = I_{o1}\text{-}I_{o2} = (I_1\text{-}I_2)\text{-}(I_5\text{-}I_4)$$
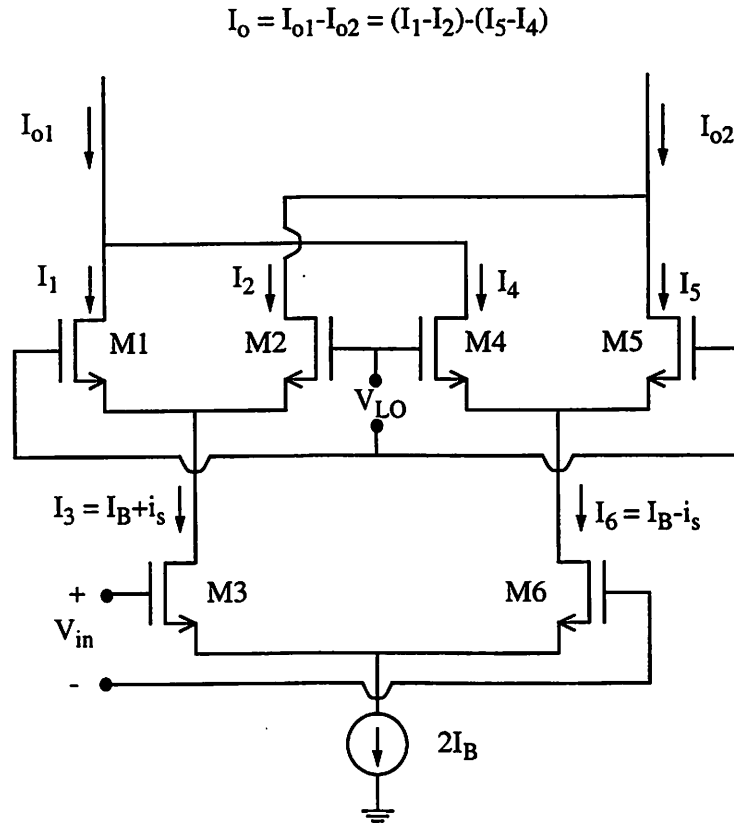
Figure 3.2: A typical double-balanced current-commutating CMOS mixer.

switching pairs are combined in such a way that the output ideally does not contain an LO component, but in practice the LO rejection is limited by the device mismatch. From (3.1), the output of the Gilbert cell is the difference of the output currents of two single-balanced mixers, and therefore the results of an analysis carry over easily from the single to the double-balanced case. The output is usually obtained differentially, if however it is obtained single-ended the mixer preserves the double-balanced operation but the conversion gain is only half.

An advantage of current-commutating mixers is the high port-to-port isolation. Consider for example the mixers of Fig. 3.1 and Fig. 3.2. Leakage from LO to the input

occurs through the gate-drain capacitance of the transconductance stage device. Assuming a balanced LO waveform and only DC common-mode LO voltage, the potential of the drain of the transconductance stage device moves at twice the LO frequency and the leakage to the input at this frequency is not quite as severe problem as the leakage at the LO frequency. Direct feedthrough from the input to the output is eliminated in both the single and double-balanced structures and leakage from the LO to the output is eliminated in the double-balanced. Finally, leakage from the output to the LO can occur through the gate-drain capacitances of the devices of the switching pair, but in the double-balanced structure this effect is also largely attenuated because of symmetry.

Let us now briefly compare the single-balanced mixer of Fig. 3.1 with the double-balanced mixer of Fig. 3.2 assuming that the bias current of each side of the double-balanced equals the bias current of the single-balanced, and let us assume that the transconductance stage of the double-balanced consists of two stages identical to that of the single-balanced. The two mixers have equal conversion gain, the Gilbert cell has better linearity since only half the signal is processed by each side, while the single-balanced has better noise performance since there are fewer devices to contribute noise. In addition, the Gilbert cell, being a differential structure rejects the even-order nonlinearities.

Active or passive loads can be used to transform the output current to voltage, or matching networks can be used to deliver maximum power to the mixer load.
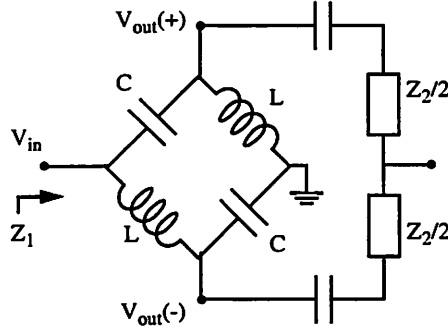
Figure 3.3: Narrow-band balun.

3.2.2 Passive Networks for Single-Ended to Differential (and inverse) Conversion

Baluns or equivalent narrowband structures such as those shown in Fig. 3.3 and Fig. 3.4 consisting of inductors and capacitors can be used to transform the input or LO signal from single-ended to differential, or the output signal from differential to single-ended.

For the narrow-band balun in Fig. 3.3 it is useful to know that at the resonant frequency $\omega = 1/\sqrt{LC}$

$$Z_1 Z_2 = \left(\frac{L}{C}\right)^2 \tag{3.2}$$

and therefore this arrangement can also be used for impedance transformation and impedance inversion. One should bear in mind however that the output voltage of this balun is generally unbalanced and a symmetric load with a center tap tied at the desirable voltage is needed to make it balanced, as shown in Fig. 3.3.

The current combiner of Fig. 3.4(a), at the resonant frequency $\omega = 1/\sqrt{2LC}$ provides the current difference $I_o = I_1 - I_2$. It is interesting to examine the operation of

the current combiner when a finite impedance Z is connected in parallel with capacitor C as shown in Fig. 3.4(b). It can be shown easily that the output impedance of the combiner at the resonance frequency is

$$Z_{out} = \frac{Z}{4} + \frac{j\omega L}{2} \tag{3.3}$$

and the current source at the output has the value

$$I_o = \left(\frac{Z - 2j\omega L}{Z + 2j\omega L}\right) I_1 - I_2 \tag{3.4}$$

The above relations suggest that for the proper operation of the current combiner, impedance Z must be much higher than $j\omega L$. According to these relations, of all the combinations of L and C that provide the correct resonance frequency, small values of L and large values of C must be chosen. However practical, finite Q inductors have losses and the lower the inductance value the lower the parallel resistance which represents these losses whose value is $QL\omega$. The optimal value of the inductance is chosen as a compromise between these two factors and usually relatively high inductance values provide a better combiner. It is worth noticing that the current combiner approximately steps down impedance Z by a factor of 4, similarly to a lossless transformer balun.

An other interesting calculation is shown in Fig. 3.4(c). When an impedance exists in parallel with the inductor on the left the equivalent output impedance is

$$Z_{out} = Z + 2j\omega L \tag{3.5}$$

and the equivalent output current source has the value

$$I_o = \left(\frac{Z}{Z + 2j\omega L}\right) I_1 - I_2 \tag{3.6}$$
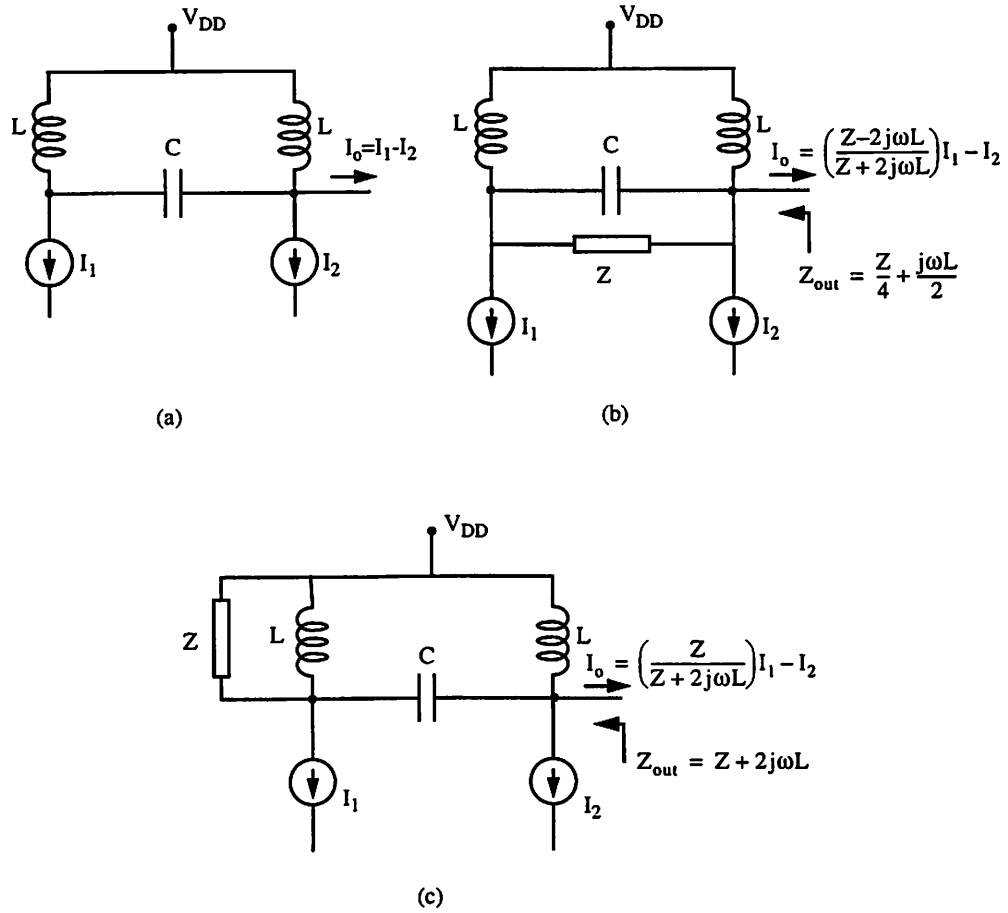
Figure 3.4: Current combiner.

Usually in integrated implementations all on-chip signals and circuits are differential in order to achieve common-mode noise rejection and minimize even-order nonlinearities. Often however these circuits need to interact with external discrete components that are single-ended, such as filters. In these cases the above structures as well as transformer baluns can be useful.
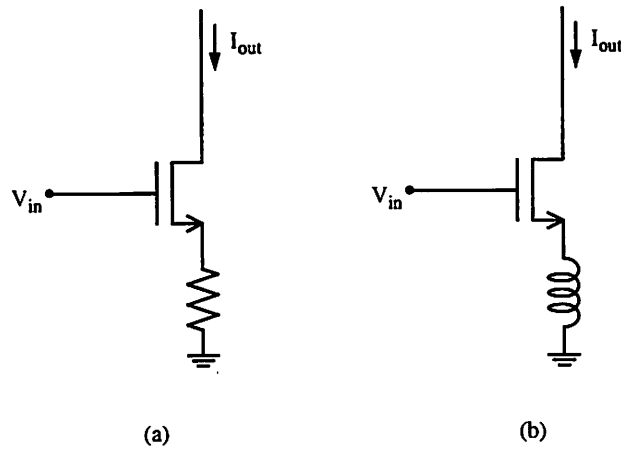
Figure 3.5: Single-ended common-source transconductance stages.

3.2.3 Single-Ended Transconductance Stages

Besides the single-transistor common-source single-ended transconductance stage shown in Fig. 3.1, several other can be used. Degeneration (which is series-series negative feedback) can be used to linearize this stage at the price of gain reduction as shown in Fig. 3.5. Resistive degeneration is appropriate for broadband applications but has the disadvantage that it introduces noise and that it consumes some voltage headroom. Inductive degeneration has the advantages of being noiseless and not requiring voltage headroom. It requires however large chip area, and is frequency dependent and therefore more appropriate for narrow-band applications.

We must notice that a CMOS device which closely follows the square law does not have odd-order nonlinearities and therefore has excellent IM_3 and gain compression performance. A device fabricated in a modern sub-micron technology however exhibits short channel effects which generate some odd-order nonlinearity and in this case degeneration can possibly improve the IM_3 and the compression gain as we shall see in chapter 6. A bond-
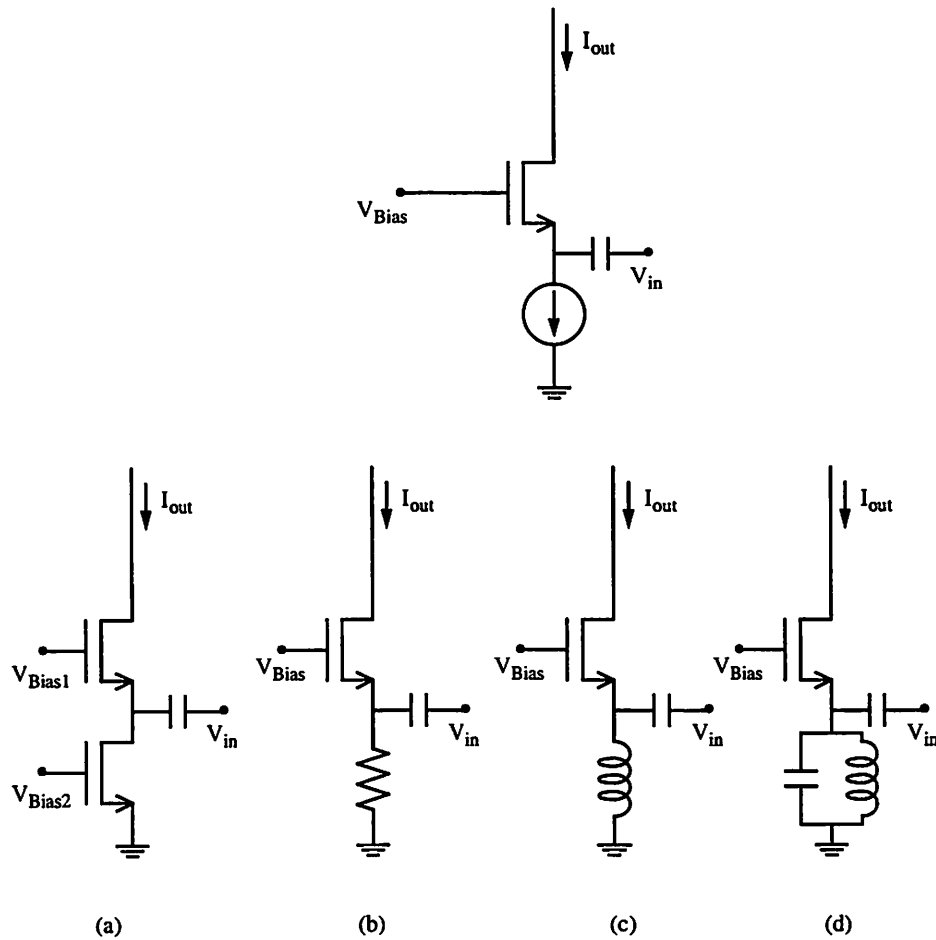
Figure 3.6: Single-ended common-gate transconductance stages.

wire inductance often represents inductive degeneration that cannot be avoided. Finally, degeneration facilitates impedance matching to the signal source, since, as we shall see in chapters 5 and 6, it creates a real part in the input impedance of this transconductance stage.

A common-gate transconductance stage shown on the top of Fig. 3.6 can also be used and has the advantage of easy and broadband matching to a real signal source impedance. It requires a current source which, as shown in this figure, can be implemented
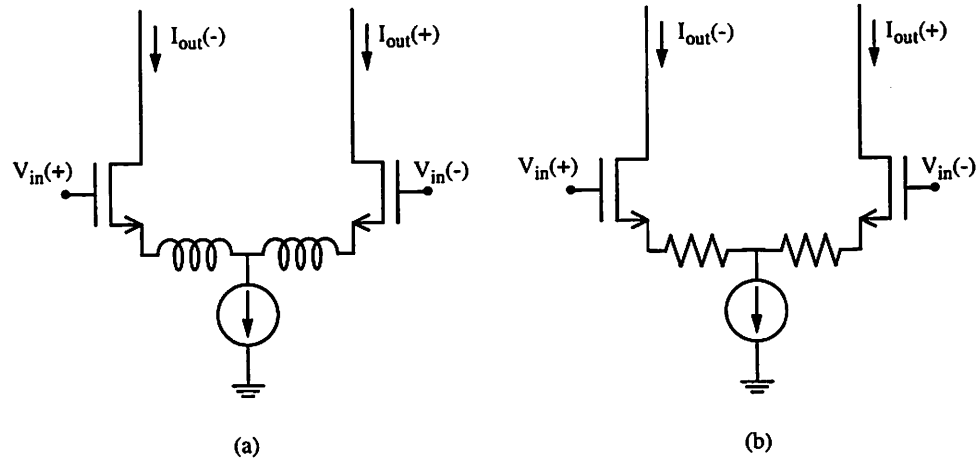
Figure 3.7: Differential transconductance stages.

in several ways: (a) with an other CMOS transistor in saturation, (b) with a resistor of higher value than the input impedance $1/g_m$, (c) with an inductor which exhibits high impedance and operates as a choke, and (d) with a parallel LC tank at resonance. The first two current sources have the disadvantage that they require some DC headroom, but they are broadband. The resistive implementation offers low parasitic capacitance and at high frequencies can be a better current source than the transistor in (a). Implementations (c) and (d) do not consume headroom but they require large area for the on-chip inductor. Finally (d) can be a good current source with high output impedance but, depending on the quality factor (Q) of the inductor, is rather narrow-band.

3.2.4 Differential Transconductance Stages

Double-balanced mixers can use two single-ended transconductance stages such as those described in the previous section. These stages require that the available input signal is perfectly differential, as is often the case in integrated implementations. Differential
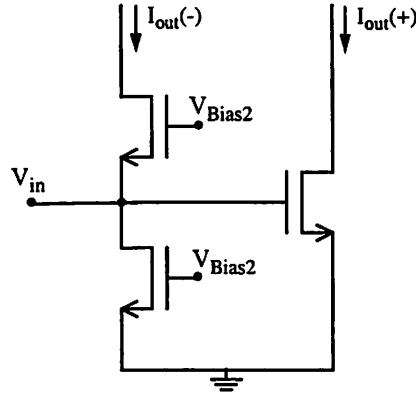
Figure 3.8: Single-ended to differential conversion with one common-source and one common-gate device.

common-source stages such as this shown in the Gilbert cell of Fig. 3.2 and also in degenerated form in Fig. 3.7 offer common-mode signal rejection. The current source can be implemented in all four ways demonstrated in Fig. 3.6 and discussed in the previous paragraph. These stages can be used for single-ended to differential implementation if one side is ac grounded, the parasitic impedance from the common-source node to ground however makes the conversion imperfect, particularly at frequencies that the parasitic capacitances are significant. An other advantage of these differential transconductance stages is that they reject noise from the current source since it is a common mode signal.

Another circuit that performs single-ended to differential conversion is shown in Fig. 3.8 which consists of one device in common-source and one in common-gate configuration. Because of the body effect, the different bias scheme and the different parasitics, no real matching exists between the two sides and one cannot expect to obtain fully differential signal. This stage has the advantage of easy and broadband source impedance matching. It was used in bipolar from in [26] and in CMOS in [60]. In both cases it operates as a class AB stage, and has high 1 dB compression point.
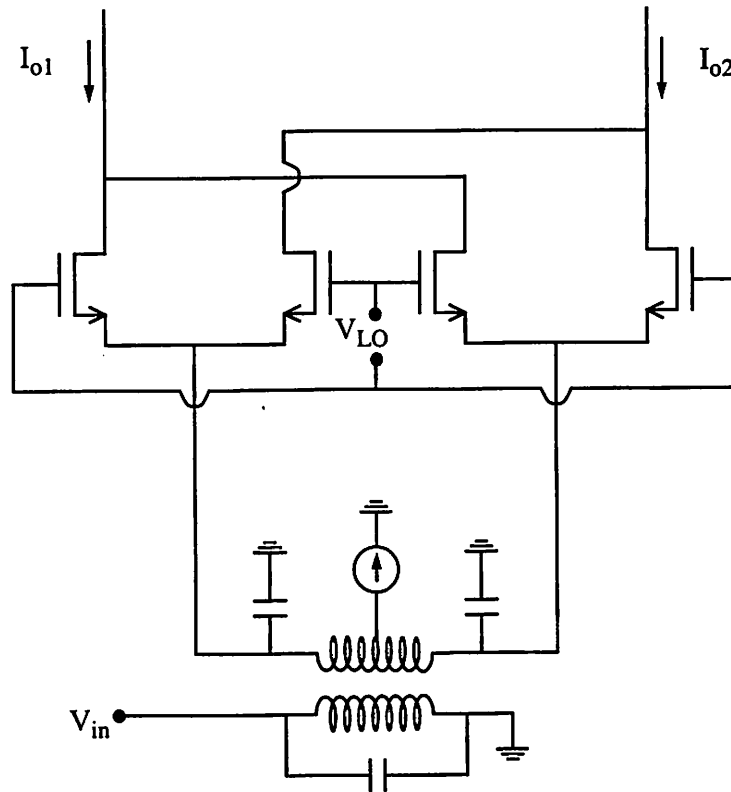
Figure 3.9: A transformer replaces the transconductance stage.

3.2.5 Balun Use in the Transconductance Stage

As explained above, the differential transconductance stages presented can be used for single-ended to differential conversion if one side is grounded, but the output signal is not fully differential. For example one cannot expect to completely cancel the even-order nonlinearity with these stages when the input signal is single-ended.

An on-chip balun is an efficient way to create a fully differential signal from a single-ended one, which comes at the price of large chip area and significant signal power loss, with 3 dB being a typical number.

A related structure is shown in Fig. 3.9 in which a balun does not just precede
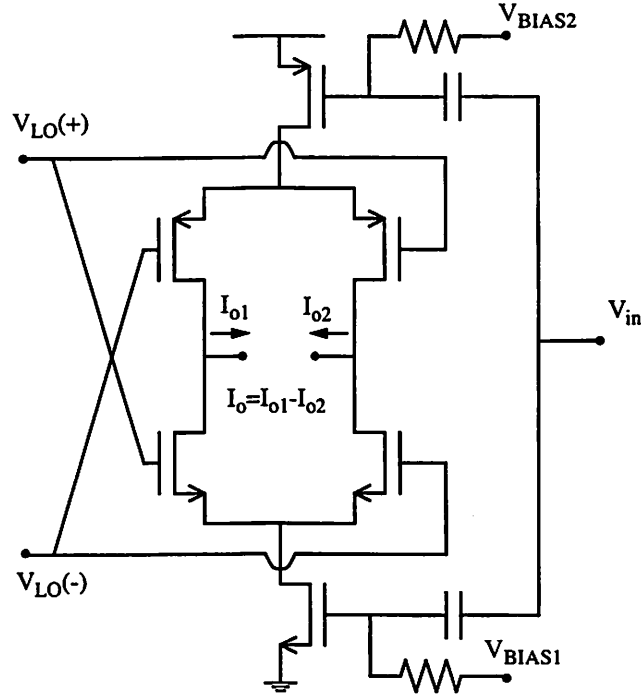
Figure 3.10: A current-reuse configuration.

a transconductance stage, but completely replaces it. Capacitors are used to tune out the inductances of the balun. Since the balun also offers impedance transformation, it can imitate the current sources that provide the signal for the switching pair. In addition, this technique has a significant DC headroom advantage.

3.2.6 A Current-Reuse Configuration

A current reuse, Gilbert cell type mixer has been presented in [37] and is shown in Fig. 3.10. It is essentially a Gilbert cell in which one of the two single-balanced mixers has been implemented with PMOS devices. The bias voltage V_{BIAS1} is set by a feedback loop in order for the NMOS current source to match the PMOS one. It is a double-balanced mixer since it rejects both the direct feedthrough of the input signal and the LO component
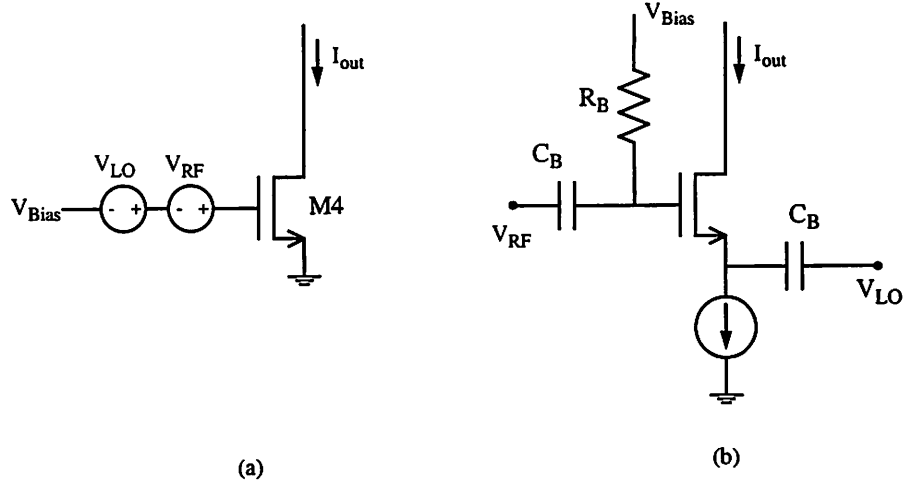
Figure 3.11: Single transistor mixers.

to the output. However, because it is not a differential structure it is not free of second-order nonlinearities. Its advantage is that it offers higher gain than a single-balanced mixer without increasing the power consumption. Its disadvantage is that because of the many stacked devices the signal headroom is limited.

3.2.7 Single Transistor Active Mixers

This is the simplest form of an active mixer. The mixer of Fig. 3.11(a) has also been used and analyzed in bipolar technologies [53, 54, 55]. The LO and input signals are both injected to the gate of the transistor and the input and LO ports are isolated by filtering. The output is the drain current. Due to the simplicity of this mixer the noise figure, gain, linearity and frequency response are probably good. Its disadvantage is that its port to port isolation is very bad and it requires high frequency filtering for isolation of the input and LO ports. It cannot be easily used as a downconverter if the IF frequency is low or zero.
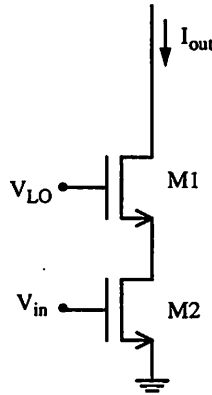
Figure 3.12: A dual-gate CMOS mixer.

Another single transistor mixer is shown in Fig. 3.11(b). Here the input signal is injected at the gate while the LO is injected at the source of the transistor, and therefore the LO to input port isolation is much better than for the mixer of Fig. 3.11(a).

3.2.8 Dual-Gate Mixers

Dual-gate transistors have been traditionally used as mixers in GaAs technology. They can also be implemented in CMOS technology as shown in Fig. 3.12. The source of the upper device (M1) and the drain of the lower device (M2) can easily share the same diffusion area in order to minimize the capacitance from that node to ground. The LO signal is connected to the gate of M1 which operates in saturation and the input signal is connected to the gate of M2 which primarily operates in the triode region. Mixing takes place in M2 whose current depends on both LO and input signals according to the IV relation of the CMOS transistor in saturation

$$I = k(V_{GS} - V_T - V_{DS}/2)V_{DS}.$$ (3.7)
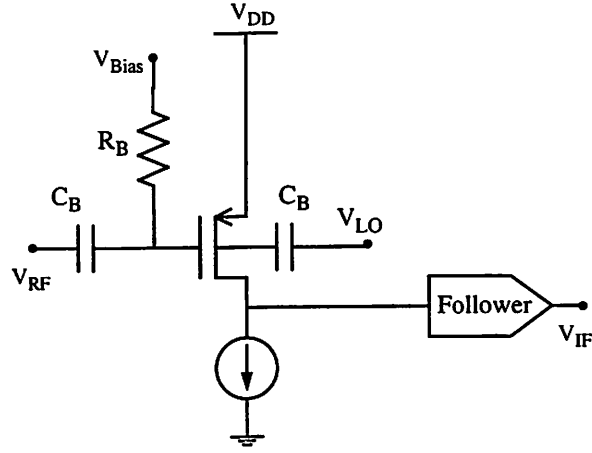
Figure 3.13: A back-gate mixer.

For large LO amplitude the lower device possibly enters the saturation region for a small part of the period. A balanced version of this mixer has been presented in [77, 78]. The LO to RF and LO to IF leakage of this structure happens at twice the LO frequency and therefore the problem is mitigated. However, the input component at the output is not rejected and this is not a double-balanced structure. The performance of these circuits does not appear to have significant performance advantages over the active current commutating mixers [77, 78], they require higher LO power and they probably have worse port-to-port isolation.

3.2.9 Back-Gate Mixer

Such a mixer has been presented in [103] and is shown in Fig. 3.13. The LO signal is injected from the body terminal. It achieves good performance in terms of gain, linearity and noise. It has several advantages such that it operates with very low supply voltage, it consumes very low power, it requires low LO power and it preserves a high conversion
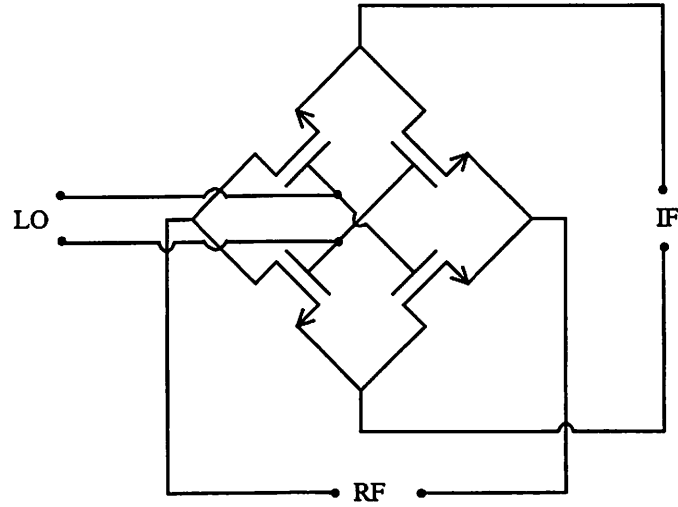
Figure 3.14: Double-balanced passive mixer.

gain even at very high frequencies. On the downside it injects the LO signal to the well of the PMOS devices and the LO component can easily couple to the substrate. Because the three ports are linked directly through the parasitic capacitances of the device, the port to port isolation is possibly a problem. In the balanced configuration presented in [103], the LO to the substrate, and LO to input port leakage happens at twice the LO frequency and is therefore alleviated.

3.3 Passive Mixers

3.3.1 Passive Switching Mixers

The most commonly used passive mixer is shown in Fig. 3.14 [74]. The transistors act as switches and when they conduct they are in the triode region. The circuit is shown in double-balanced form, but it can also be used in the unbalanced (single transistor passive
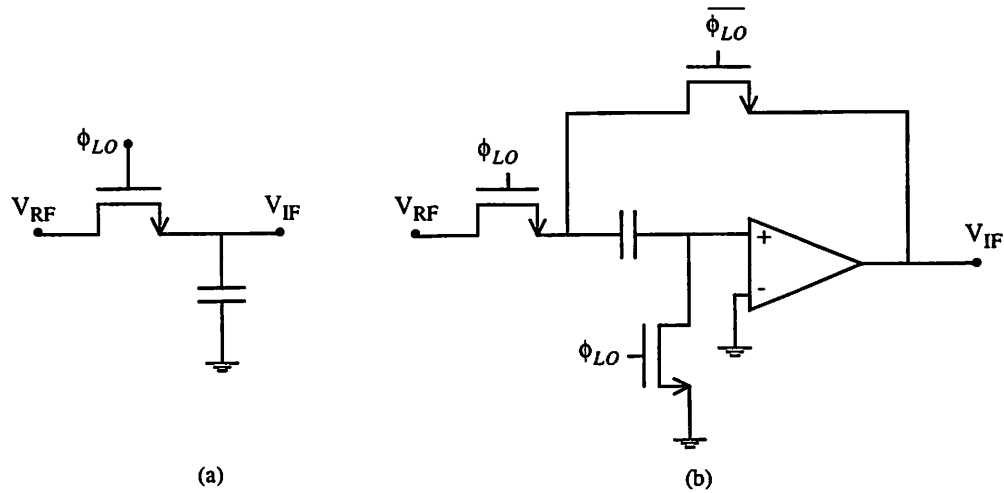
Figure 3.15: Subsampling mixers.

mixer) or single-balanced form. Its advantage is that it demonstrates exceptional linearity, while its disadvantage is the conversion loss which for ideal switching is $2/\pi$ or approximately $4dB$, but because of nonidealities a loss of 6 to 7 dB is a typical value. It generally requires higher LO power than active mixers. Its noise performance is good and its NF is almost equal to the loss as discussed about the NF of lossy blocks in chapter 2. Because of the conversion loss however, these mixers usually must be preceded or followed with amplifying stages which degrade the system linearity and noise.

3.3.2 Sub-Sampling Mixers

A sampler can be used as a downconversion mixer, as shown in its simplest form in Fig. 3.15(a). The circuit of Fig. 3.15(b) is similar, but provides output driving capability. Differential structures are more usual in practice and have the advantage of rejecting common mode noise, such as clock feedthrough. The high-frequency input signal must
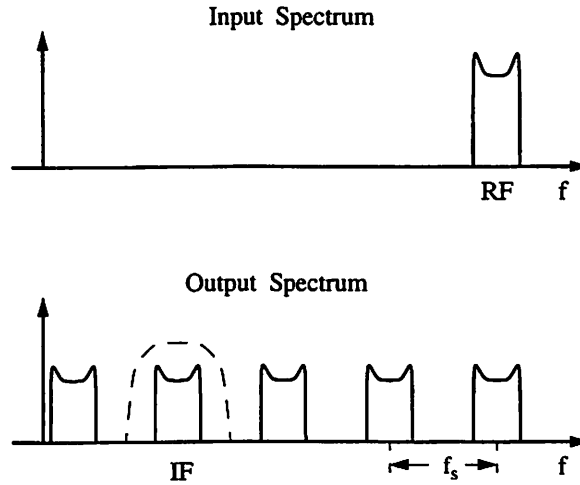
Figure 3.16: The spectrum at the input and the output of a sampling mixer.

be band-limited and the sampling frequency can be low, higher however than the signal

bandwidth. If the contents of the spectrum outside the signal band at the input have been

removed by filtering, the output contains non-overlapping copies of the input signal aliased

as shown in Fig. 3.16. With filtering one can obtain the original signal translated at low

frequencies. These mixers usually demonstrate excellent linearity but their noise perfor-

mance is poor since each sample is contaminated with noise folded from a large number

of sidebands and equals the sampling noise kT/C. Notice that the value of the sampling

capacitor cannot be made arbitrarily large, since the settling time of the sampling circuit

must be very fast and using a very wide transistor to reduce the resistance of the switch

will introduce parasitic nonlinear capacitors and clock feed-through problems. The noise

figure of a sampling mixer has been examined in [15]. The sampling clock can run at a

much lower frequency than the LO of a non-sub-sampling mixer, but its jitter must be only

a small part of the period of the sampled RF signal. Therefore clock jitter requirements are
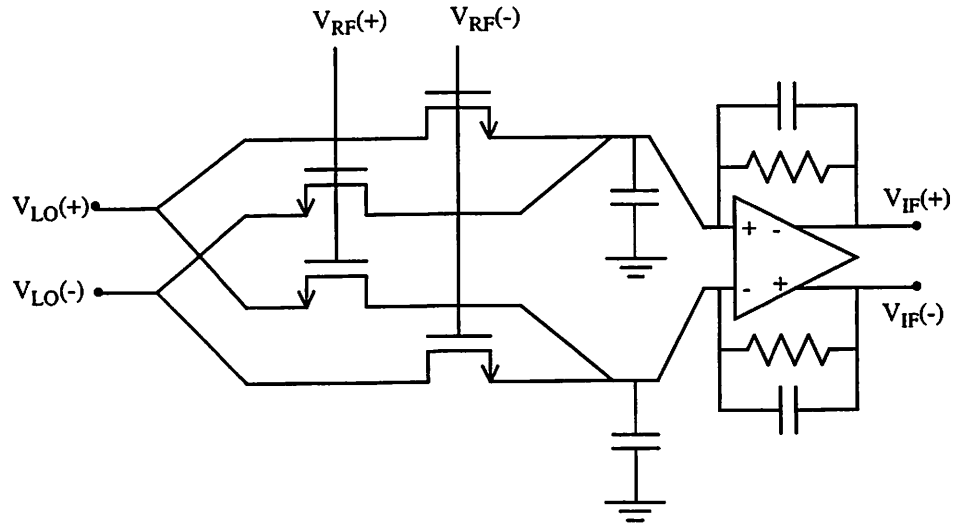
Figure 3.17: A downconverting triode-region mixer.

probably about as hard to satisfy as those of a high frequency LO of a non-subsampling

mixer.

3.3.3 Mixers Using the Transistor in the Triode Region

Such mixers take advantage of the IV relation of the MOS transistor in the triode

region (3.7) in which the product of V_{GS} and V_{DS} appears. They are also called potentio-

metric mixers in [45]. For example in [13] the circuit of Fig. 3.17 is used as a downconverter.

This is essentially the double-balanced passive mixer of Fig. 3.14, but the LO and signal

input have been exchanged. Despite the high LO amplitude that this circuit requires, the

conversion gain of the modulating part is much lower than one would obtain if one con-

nected the LO signal on the gates and completely switched on and off the transistors. The

output stage is used to amplify the low frequency output signal. However, as a result of

the fact that at the input of the amplifier the signal is heavily attenuated, the NF of this
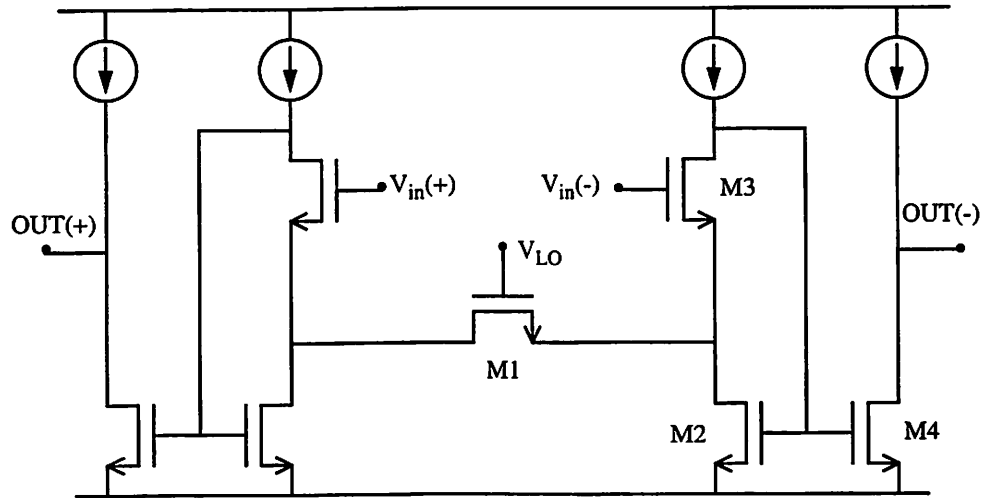
Figure 3.18: An upconverting triode-region mixer.

structure is very high.

An upconverter based on the same principle of operation has been presented in [40] and is shown in Fig. 3.18. Feedback from the drain of M3 to the gate of M2 is used to make M3 a very linear source-follower. Again the conversion loss is very high, the linearity is good while the NF is expected to be very poor.

Part II

Analysis

Chapter 4

Noise in Current-Commutating CMOS Mixers

4.1 Introduction

In this chapter we examine the noise performance of current-commutating active CMOS mixers, neglecting capacitive effects. The results are applicable when the mixer operates at moderate frequencies used at the IF stage of a receiver, or considering modern submicron technologies and high bias current, at higher frequencies used at the RF front end. A corresponding noise analysis of bipolar active mixers has been presented in [31]. The results of this chapter have also been presented in [89].

The simple single-balanced active mixer of Fig. 4.1(a) is examined, and the results are also presented for the double-balanced circuit, the Gilbert cell. The analysis can be readily adapted for variations of the above structures described in section 3.2.
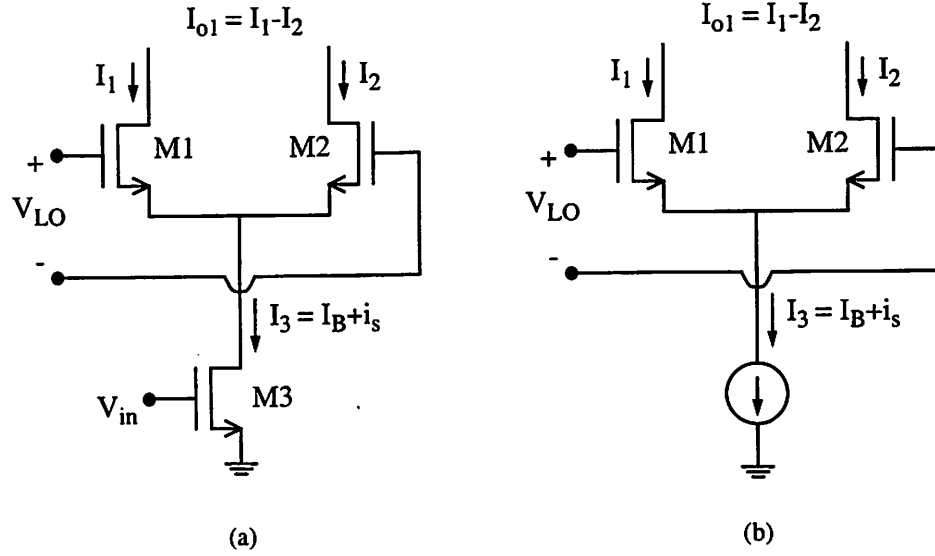
Figure 4.1: (a) A simple single-balanced active CMOS mixer, and (b) The basic model of a current commutating CMOS mixer

4.2 Transistor Model and Switching Pair Large-Signal Equations

The simple square-law MOSFET model is not accurate for modern short-channel technologies, and a better approximation for the I-V relation of a MOS transistor is [61]

$$I = K \frac{(V_{GS} - V_T)^2}{1 + \theta(V_{GS} - V_T)}. \tag{4.1}$$

In (4.1), I is the drain current, V_{GS} is the gate-source voltage, and V_T is the threshold voltage of the device. Parameter K depends on the technology and the size of the device, and is proportional to the channel width. Parameter θ models to a first order the source series resistance, mobility degradation due to the vertical field, and velocity saturation due to the lateral field in short-channel devices. It depends on the channel length and is independent of the body effect.

Fig. 4.1(b) shows the basic model of a current-commutating CMOS mixer. Since a large AC drive is applied to the switching pair, the bias of $M1$ and $M2$ is not fixed but varies periodically with time. When a differential voltage greater than a certain value V_x is applied between the gates of the two transistors, one of them switches off. When the absolute value of the instantaneous LO voltage V_{LO} is lower than V_x the current of the driver stage is shared between the two devices. In this case it is desirable to find the drain current of each transistor for a given LO voltage and driver stage bias current. We will assume that the output conductance of the devices can be neglected and therefore $M3$ can be modeled with an ideal current source I_B. We will also assume that the load of $M1$ and $M2$ is such that they remain in saturation during the part of the LO period that they are on.

The large-signal behavior of the switching pair is described by the system of two equations

$$K_1 \frac{(V_{GS1} - V_T)^2}{1 + \theta(V_{GS1} - V_T)} + K_1 \frac{(V_{GS2} - V_T)^2}{1 + \theta(V_{GS2} - V_T)} = I_B \tag{4.2}$$

$$V_{GS1} - V_{GS2} = V_{LO} \tag{4.3}$$

where K_1 is the K parameter of $M1$, $M2$ and V_{GS1}, V_{GS2} are the gate-source voltages of $M1$, $M2$. If we normalize I_B and V_{LO} as follows

$$J_B = \frac{\theta^2}{K_1} I_B \tag{4.4}$$

$$U_{LO} = \theta V_{LO} \tag{4.5}$$

and also let

$$U_1 = \theta(V_{GS1} - V_T) \quad U_2 = \theta(V_{GS1} - V_T) \tag{4.6}$$

(4.2) and (4.3) become:

$$\frac{U_1^2}{1+U_1} + \frac{U_2^2}{1+U_2} = J_B \qquad (4.7)$$

$$U_1 - U_2 = U_{LO}. \qquad (4.8)$$

Equations (4.7) and (4.8) can be transformed to one nonlinear equation with U_1 as

the unknown, which can be solved rapidly with an iterative numerical method. Considering

a positive V_{LO}, the desired value of U_1 lies between U_{LO} and

$$\theta V_x = \frac{J_B}{2} + \sqrt{\frac{J_B^2}{4} + J_B} \qquad (4.9)$$

which is the value of U_1 when the whole bias current passes through $M1$. With the trans-

formation of (4.4) and (4.5), the normalized current of each transistor can be found in terms

of J_B and U_{LO}, independent of the technology parameters. For $M1$ for example:

$$J_1 = \frac{\theta^2}{K_1} I_1 = \frac{U_1^2}{1+U_1}. \qquad (4.10)$$

The transconductance of each transistor will be needed below and can be calculated as the

derivative of I with respect to V_{GS} from (4.1), or in normalized form as the derivative of J_1

with U_1 from (4.10).

It is worth noticing that no specific value of V_T is needed to calculate the drain

current of $M1$ and $M2$. The behavior of the switching pair is independent of V_T and

therefore to a first order independent of the body effect and the common-mode LO voltage.

This observation allows us to omit the small-signal body transconductance below.

In the following analysis, some performance parameters of the switching pair will

be given in terms of the normalized bias current J_B and LO amplitude $U_o = \theta V_o$, V_o being

the real LO amplitude. The subthreshold conduction of the transistors has been neglected

and therefore if the devices operate at very low current density, the prediction will be inaccurate, especially for low LO amplitude where the transistors do not act as switches and their behavior depends on their I-V characteristics.

4.3 Deterministic Signal Processing

If capacitive effects are ignored, the single balanced mixer of Fig. 4.1 is a function of the instantaneous LO voltage $V_{LO}(t)$ and the current at the output of the driver stage $I_3 = I_B + i_s$, I_B being the bias current and is the small-signal current

$$I_{o1} = I_1 - I_2 = F(V_{LO}(t), I_B + i_s). \tag{4.11}$$

Since i_s is small, a first-order Taylor expansion gives:

$$I_{o1} = F(V_{LO}(t), I_B) + \frac{\partial}{\partial I_B} F(V_{LO}(t), I_B) i_s \tag{4.12}$$

or

$$I_{o1} = p_0(t) + p_1(t) \cdot i_s. \tag{4.13}$$

Both $p_0(t)$ and $p_1(t)$ are periodic waveforms, depicted in Fig. 4.2. As can be seen from (3.1) and (4.13), in a double-balanced structure with perfect device matching, $p_0(t)$ is eliminated. During the time interval Δ, when the LO voltage is between V_x and $-V_x$, and both transistors are on, $p_0(t)$ and $p_1(t)$ depend on $V_{LO}(t)$, I_B, and the I-V characteristics of the transistors. The small-signal current in each branch is determined by current division, and one can see that

$$p_1(t) = \frac{g_{m1}(t) - g_{m2}(t)}{g_{m1}(t) + g_{m2}(t)} \tag{4.14}$$
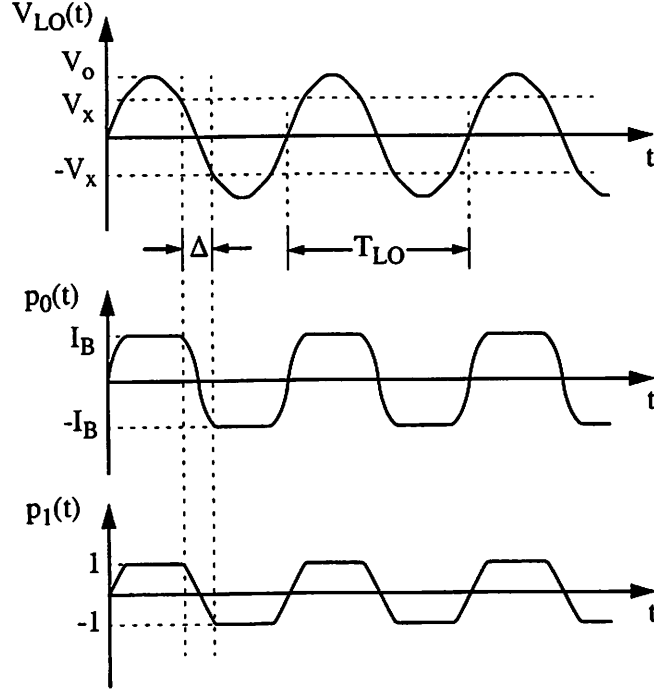
Figure 4.2: Waveforms $p_0(t)$ and $p_1(t)$

where $g_{m1}(t)$ and $g_{m2}(t)$ represent the instantaneous small-signal transconductances of $M1$ and $M2$. According to (4.13), a signal component $x(t)$ of $i_s(t)$ is multiplied by the waveform $p_1(t)$, and therefore the frequency spectrum of the corresponding output is

$$Y_x(f) = \sum_{n=-\infty}^{\infty} p_{1,n} \cdot X(f - nf_{LO}) \qquad (4.15)$$

where f_{LO} is the LO frequency, $p_{1,n}$ are the Fourier components of $p_1(t)$, and $X(f)$ is the frequency spectrum of $x(t)$.

It is worth noticing that with good device matching $p_1(t) = -p_1(t + T_{LO}/2)$, T_{LO} being the LO period, and hence $p_1(t)$ has only odd-order frequency components. The same observation can be made for $p_0(t)$. Usually the term for $n = 1$ or $n = -1$ is of interest, corresponding to shifting up or down the input signal in the frequency domain by one
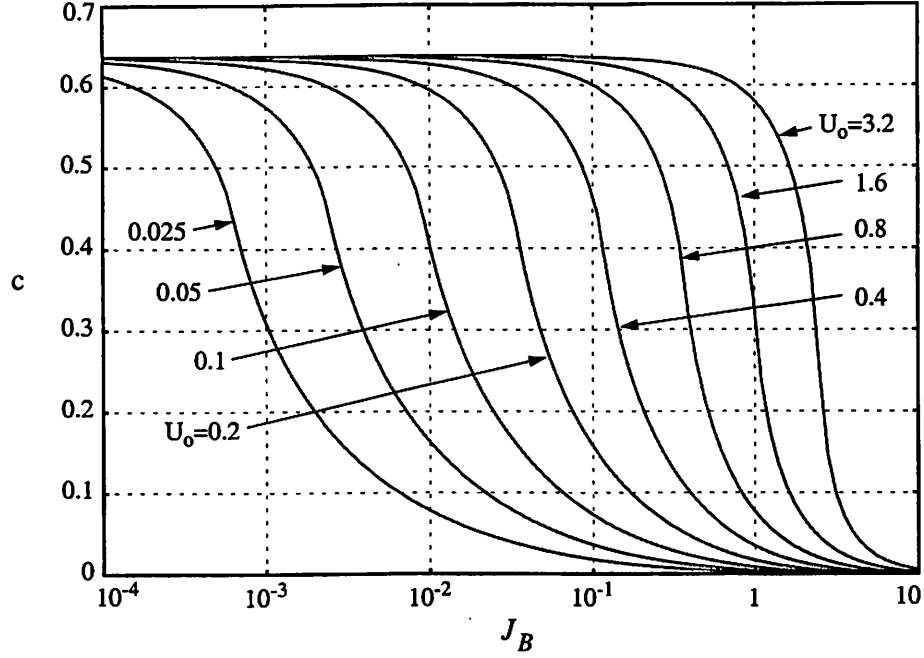
Figure 4.3: Numerically evaluated conversion gain of the switching pair c

multiple of the LO frequency, and in this case $c = |p_{1,1}| = |p_{1,-1}|$ represents the conversion gain of the switching pair alone. Since $x(t) = g_{m3} \cdot v_{in}(t)$ where $v_{in}(t)$ is the input voltage signal at the gate of $M3$, and g_{m3} is the transconductance of $M3$, the conversion gain of the single-balanced mixer in transconductance form is

$$g_c = c \cdot g_{m3} \tag{4.16}$$

For high LO amplitude, $p_1(t)$ approaches a square waveform and c approaches $2/pi$. Fig. 4.3 shows c, evaluated numerically as a function of the normalized bias current J_B and LO amplitude U_o, for a sinusoidal LO waveform. Assuming $V_o > V_x$, as it should be for proper mixer operation, an estimate for c can be obtained by approximating $p_1(t)$

with a straight line during Δ:

$$c = \frac{2}{pi} \left(\frac{sin(\pi \Delta f_{LO})}{(\pi \Delta f_{LO})} \right) \tag{4.17}$$

where for sinusoidal LO waveform

$$\pi \Delta f_{LO} = \arcsin \left(\frac{V_x}{V_o} \right) \tag{4.18}$$

and V_x is given by (4.9). Comparison with the numerically evaluated value of c for sinusoidal LO waveform shows that (4.17) is a good approximation for low values of U_o, introducing error below $1dB$ if $Uo < 0.7$, while it overestimates c for higher values of U_o, introducing error below $2dB$ if $U_o < 1.6$ and below $3dB$ if $U_o < 3.2$.

It is easy to observe that the conversion gain of the Gilbert cell is also given by (4.16). If degeneration or an input matching network is used, the transconductance of the driver stage is not g_{m3}, but can be calculated with linear circuit techniques and multiplied with c to provide the conversion gain.

4.4 Noise Analysis

Consider a device which with a fixed operating point produces shot or thermal white noise. It can be shown [14] that if the operating point changes with time, the resulting noise is still white, with a time- varying PSD given by the same formula as for the time-invariant case, if we replace the value of the fixed resistor with the time-varying one for thermal noise, and the value of the fixed current across the p-n junction by the time-varying one for the shot noise.

We will use the fact that the PSD of the drain current thermal noise generated by

a MOS transistor in saturation is

$$\overline{i_n^2} = 4kT\gamma g_m \qquad (4.19)$$

where g_m is the gate transconductance, k is Boltzman's constant, T is the absolute temper-

ature, and γ is 2/3 for long channel transistors, but can be higher for short channel devices,

can depend on bias and can be affected by hot electron phenomena [84, 1, 102, 91, 36, 85].

In the following analysis we will calculate the time average noise at the output of

the mixer, and based on that we will evaluate the noise figure.

4.4.1 Noise from the Transconductance Stage

Consider the noise component $n_3(t)$ of $i_s(t)$, in Fig. 4.1(b). This is considered

to be WSS with PSD $S_{n3}(f)$, and can represent noise generated in $M3$, or noise present

at the input of the mixer, and amplified by $M3$. The output noise component that $n_3(t)$

contributes

$$y_{n_3} = n_3(t) \cdot p_1(t) \qquad (4.20)$$

is cyclostationary, and its time average PSD is:

$$S_{n3}^o(f) = \sum_{n=-\infty}^{\infty} |p_{1,n}|^2 \, S_{n3}(f - nf_{LO}). \qquad (4.21)$$

Assuming that $n_3(t)$ is white over the bandwidth of interest, $S_{n3}(f)$ equals N_{n3}, a constant

and

$$S_{n3}^o(f) = N_{n3} \cdot \sum_{n=-\infty}^{\infty} |p_{1,n}|^2 = \alpha N_{n3} \qquad (4.22)$$

where

$$\alpha = \sum_{n=-\infty}^{\infty} |p_{1,n}|^2 = \frac{1}{T_{LO}} \int_{-\infty}^{\infty} (p_1(t))^2 \, dt \qquad (4.23)$$
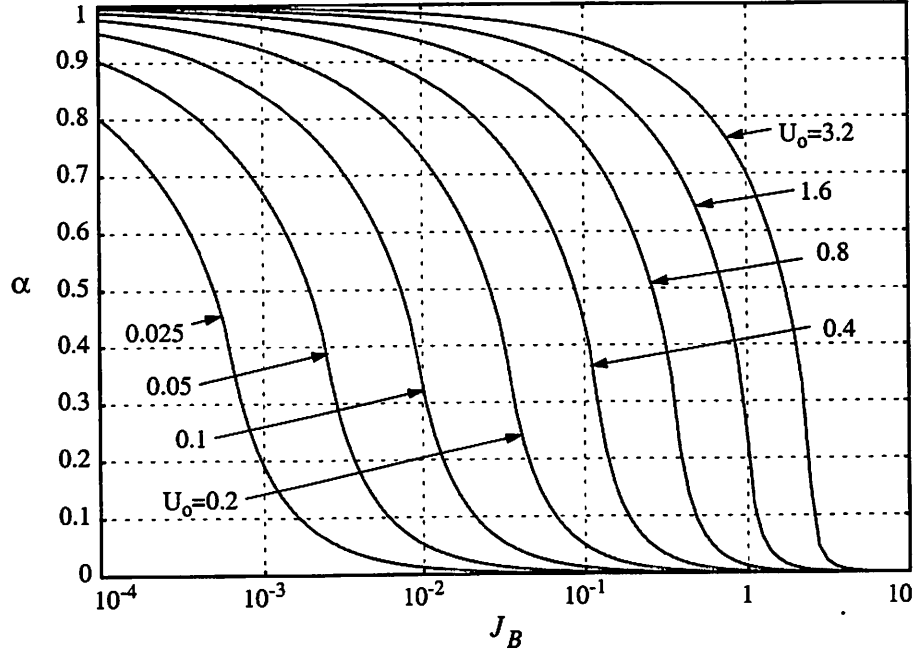
Figure 4.4: Numerically evaluated average power α of waveform $p_1(t)$

is the power of waveform $p_1(t)$. Equations (4.22) and (4.23) can be used to find the noise contribution to the output without any assumptions about the LO waveform or amplitude.

For large LO amplitude, $p_1(t)$ approaches a square waveform and its power α approaches 1. It is interesting to examine the contribution of every individual sideband in the case of a square waveform. Noise from $f_{LO} \pm f_{out}$, f_{out} being the output frequency, accounts for 81% of the noise transferred to the output, from $3f_{LO} \pm f_{out}$ for 9%, and from all higher order sidebands together for 10%. Parameter α is evaluated numerically and given in Fig. 4.4 as a function of the normalized bias current J_B and LO amplitude U_o, for sinusoidal LO waveform. Similarly to the conversion gain of the switching pair c, if $V_o > V_x$, an estimate for a can be obtained by approximating $p_1(t)$ with a straight line during Δ:

$$\alpha = 1 - \frac{4}{3}(\Delta f_{LO}) \tag{4.24}$$

where for sinusoidal LO waveform Δf_{LO} can be obtained from (4.18). Comparison with the numerically computed value of α in the case of sinusoidal LO waveform shows that (4.24) introduces error smaller than 25% if $U_o < 0.8$, while it overestimates a by less than 50% if $U_o < 1.6$, and by less than 70% if $U_o < 3.2$.

For the single-balanced mixer, assuming that N_{n3} consists of the thermal noise of $M3$, the input source resistance R_s, and the polysilicon gate resistance r_{g3}, the noise transferred to the output is

$$S_{n3}^o(f) = \alpha \cdot 4kT \left(R_s + r_{g3} + \frac{\gamma}{g_{m3}} \right) g_{m3}^2 \qquad (4.25)$$

while for the Gilbert cell

$$S_{n3}^o(f) = \alpha \cdot 4kT \left(R_s + 2r_{g3} + \frac{2\gamma}{g_{m3}} \right) g_{m3}^2. \qquad (4.26)$$

If resistive degeneration is used, the noise at the output of the driver stage is white and equation (4.21) applies. If inductive degeneration or an impedance matching network is used, the gain of the driver stage is frequency dependent. The PSD of the noise at the output of the driver stage at the frequencies of interest $-f_{LO} \pm f_{out}, 3f_{LO} \pm f_{out}$, etc.— can be calculated with linear circuit techniques and the output noise at f_{out} can be calculated from (4.21). Because of the frequency selective gain of the driver stage, possibly only a few sidebands need to be taken into account.

4.4.2 Thermal Noise Generated in the Switching Pair

We consider now thermal noise generated in $M1$ and $M2$ in Fig. 4.1, assuming that they remain in saturation during the part of the period that they are on. Neglecting

capacitive effects and the output conductance of the transistors, when $M1$ or $M2$ is off the output current is determined by I_3, and the switching pair does not contribute to the output noise. For this reason when the LO amplitude is high, the noise contribution of the switching pair is usually lower than that of the driver stage. During the time interval Δ, both $M1$ and $M2$ are on, and contribute to the output noise. The instantaneous noise PSD at I_1 is

$$4kT\gamma\left(\frac{1}{g_{m1}}\left(\frac{g_{m1}}{1+g_{m1}/g_{m2}}\right)^2 + \frac{1}{g_{m2}}\left(\frac{g_{m2}}{1+g_{m2}/g_{m1}}\right)^2\right) = 4kT\gamma\left(\frac{g_{m1}\cdot g_{m2}}{g_{m1}+g_{m2}}\right). \quad (4.27)$$

Since the sum of I_1 and I_2 equals I_3, the amplitude of the noise component at the output I_{o1} is twice that at I_1, and the corresponding output noise PSD is:

$$S^o_{n12}(f,t) = 16kT\gamma\left(\frac{g_{m1}\cdot g_{m2}}{g_{m1}+g_{m2}}\right) = 8kT\gamma G(t) \quad (4.28)$$

where

$$G(t) = 2\left(\frac{g_{m1}\cdot g_{m2}}{g_{m1}+g_{m2}}\right) \quad (4.29)$$

is the small-signal transconductance of the whole differential pair, from V_{LO} to I_{o1}. This time-varying PSD is flat in frequency since it represents white noise and is shown in Fig. 4.5. The peak of $S^o_{n12}(f,t)$ appears for $V_{LO} = 0$ and is independent of the LO amplitude. The higher the LO amplitude, the smaller the time interval Δ, and the lower the noise contribution to the output. From (4.28) we obtain the time-average PSD at the output:

$$S^o_{n12}(f) = 8kT\gamma\left(\frac{1}{T_{LO}}\int_0^{T_{LO}} G(t)dt\right) = 8kT\gamma\overline{G} \quad (4.30)$$

where $\overline{G}$ is the time average of $G(t)$. This expression can be used to calculate $S^o_{n12}(f)$ without any assumptions about the LO waveform or amplitude. However, the LO amplitude is
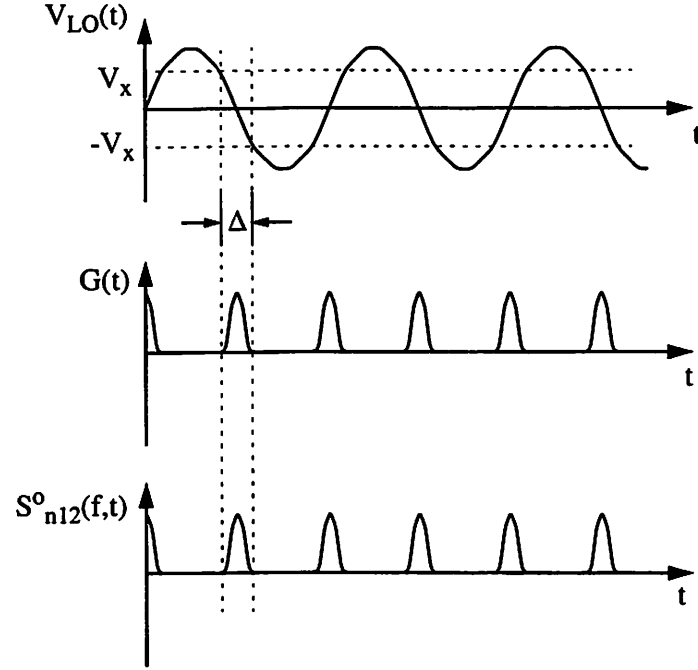
Figure 4.5: Time-varying transconductance of the switching pair, and time-varying PSD of the generated thermal noise

usually large and a further simplification is possible. Assuming sinusoidal V_{LO} and changing

the variable of integration from t to V_{LO} we obtain:

$$\overline{G} = \frac{1}{\pi V_o} \int_{-V_x}^{V_x} G(V_{LO}) \frac{1}{\sqrt{1 - (V_{LO}/V_o)^2}} dV_{LO}. \tag{4.31}$$

If the LO amplitude V_o is high, in the interval of integration V_{LO} is much smaller than V_o,

and $1/\sqrt{1 - (V_{LO}/V_o)^2} \simeq 1$. In this case, since $G(V_{LO}) = dI_{o1}/dV_{LO}$, (4.31) provides:

$$\overline{G} = \frac{1}{\pi V_o} \int_{-V_x}^{V_x} \left(\frac{dI_{o1}}{dV_{LO}} \right) dV_{LO} = \frac{2I_B}{\pi V_o}. \tag{4.32}$$

From (4.30) and (4.31) we obtain the contribution of the switching pair to the output noise

$$S_{n12}^o(f) = \frac{16kT\gamma}{\pi} \cdot \frac{I_B}{V_o}. \tag{4.33}$$

A sinusoidal LO signal was assumed above, but such a restriction is not necessary. A

relation similar to (4.32) can be obtained directly from the definition of $\overline{G}$), only with the
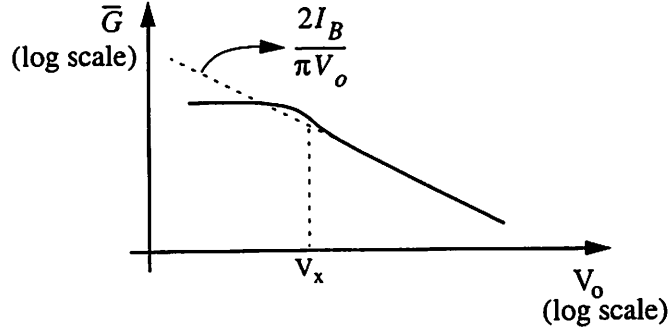
Figure 4.6: Time-average transconductance of the switching pair versus LO amplitude

assumption of linear dependence of V_{LO} in t during the time interval Δ, with slope λ:

$$\overline{G} = \frac{4}{T_{LO}} \cdot \frac{I_B}{\lambda}. \tag{4.34}$$

For sinusoidal LO waveform $\lambda = 2\pi V_o/T_{LO}$ and (4.32) results. Notice that no assumption was made about the I-V characteristics of the transistors, and that (4.32)-(4.34) are independent of the transistor dimensions. These expressions, with $\gamma = 1/2$, can also be used for the time-average transconductance and the collector shot noise of a bipolar switching pair.

We observe that the PSD at the output is proportional to the bias current I_B and inversely proportional to the zero crossing slope of V_{LO}. As can be seen in (4.31), if for moderate V_o the slope of $V_{LO}(t)$ (proportional to $\sqrt{1 - (V_{LO}/V_o)^2}$ drops close to the ends of Δ, equations (4.32)-(4.34) slightly underestimate the output noise. For smaller V_o, $\overline{G}$ and $S^o_{n12}(f)$ approach the values that correspond to the fixed operating point of $V_{LO} = 0$, and the above expressions overestimate the output noise, as demonstrated in Fig. 4.6. A graph of $(\theta/K_1)\overline{G}$ evaluated numerically as a function of J_B and U_o is given in Fig. 4.7 for a sinusoidal LO waveform. Comparison of the prediction of equation (4.32) with the computed value of $\overline{G}$ for a sinusoidal LO waveform, shows that if $V_o > V_x$ (4.32) underestimates $\overline{G}$
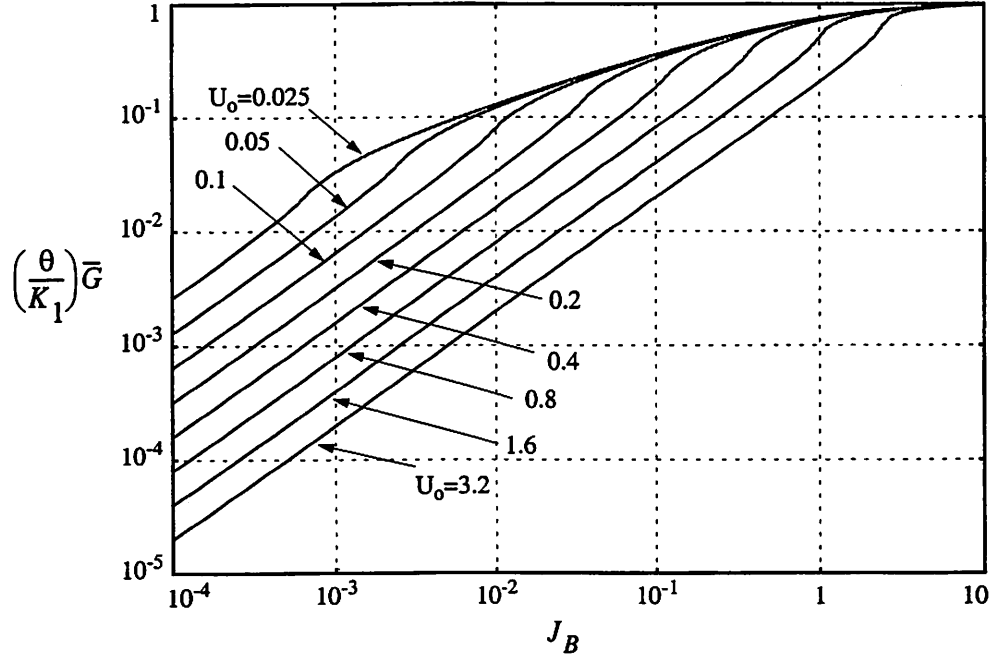
Figure 4.7: Numerically evaluated time-average transconductance of the switching pair

by less than 25% for all the values of U_o shown in Fig. 4.7, with the error growing as V_o approaches V_x. The output noise contribution of the two switching pairs of a Gilbert cell is twice that calculated for the single- balanced mixer.

Lacking a commonly accepted expression for γ as a function of bias, a fixed value was used above. In practice, the equations derived in this section can be used with the value of γ which corresponds to the bias condition of $V_{LO} = 0$. This is a reasonable approximation since the devices of the switching pair contribute the most noise for zero LO voltage.

4.4.3 Noise from the LO Port

Since the LO is a periodically time-varying circuit it is possible that the noise at its output contains a cyclostationary component. It is inaccurate to time average its PSD

and use it as if it were a WSS process, since the time-varying processing of this signal by the mixer tracks exactly the time variation of the noise statistics. Except for the case of white cyclostationary noise where time dependence of the PSD can be incorporated to the system [32], the treatment of such a problem is complicated and described in [21]. Below we will consider the simplified case at wh 198 ich the noise present at the LO port is stationary. The results also apply to intrinsic noise of $M1$ and $M2$ which can be modeled with a time-invariant stationary voltage-noise source in series with the gates, such as thermal noise of the gate resistances and flicker noise discussed in section 5.5.

We assume that the LO voltage has a noise component $n_{LO}(t)$. This contributes output noise:

$$y_{nLO} = G(t)n_{LO}(t) \tag{4.35}$$

where $G(t)$ is the time-varying transconductance of the switching pair defined in (4.29). If $n_{LO}(t)$ is WSS with PSD $S_{nLO}(f)$, $y_{nLO}(t)$ is a cyclostationary process with time average PSD

$$S^o_{nLO}(f) = \sum_{n=-\infty}^{\infty} |G_n|^2 S_{nLO}(f - nf_{LO}) \tag{4.36}$$

where G_n are the Fourier coefficients of the waveform $G(t)$.

If $n_{LO}(t)$ is also white with PSD $S_{nLO}(f) = N_{LO}$, the noise contribution to the output becomes

$$S^o_{nLO}(f) = N_{LO} \cdot \sum_{n=-\infty}^{\infty} |G_n|^2 = \overline{G^2} \cdot N_{LO} \tag{4.37}$$

where

$$\overline{G^2} = \frac{1}{T_{LO}} \int_0^{T_{LO}} G(t)^2 dt. \tag{4.38}$$
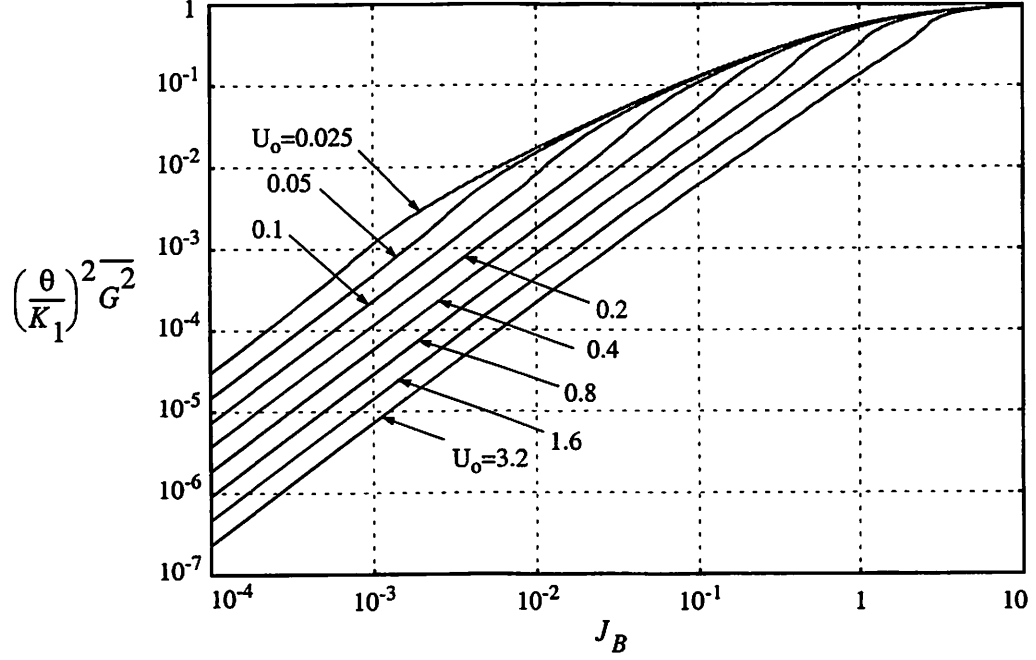
Figure 4.8: Numerically evaluated time-average square transconductance of the switching pair

With some manipulation it can be shown that for LO amplitude fairly larger than Vx and square law equations ($\theta = 0$)

$$\overline{G^2} \simeq 16 \left(\frac{ln(\sqrt{2}+1)}{\sqrt{2}} - \frac{1}{3} \right) \cdot \frac{K_1^{1/2} I_B^{3/2}}{\lambda T_{LO}} = 4.64 \cdot \frac{K_1^{1/2} I_B^{3/2}}{\lambda T_{LO}} \qquad (4.39)$$

where λ is as before the zero crossing slope of $V_{LO}(t)$. A plot of $(\theta/K_1)^2 \overline{G^2}$ as a function of the normalized bias current J_B and LO amplitude U_o, calculated numerically from (4.38) is shown in Fig. 4.8, assuming a sinusoidal LO waveform. Comparison of this computed value with the prediction of (4.39) derived for square law equations, shows that for $V_o > V_x$, (4.39) introduces error lower than 25% if $U_o < 0.8$, while it overestimates $\overline{G^2}$ by less than 50% if $U_o < 1.6$, and less than 90% if $U_o < 3.2$.

For the single balanced mixer, the white noise N_{LO} consists of the noise floor of the LO output spectrum, represented by an equivalent noise resistance R_{LO}, and the thermal

noise of the polysilicon gate resistance r_{g1} of the transistors:

$$S_{nLO}^o(f) = 4kT(R_{LO} + 2r_{g1})\overline{G^2}. \qquad (4.40)$$

The noise floor of the LO can significantly increase the noise figure of the mixer, and filters can be used to limit its effect. In a Gilbert cell the external noise present at the LO port is rejected, and only the gate resistances contribute noise:

$$S_{nLO}^o(f) = 4kT(4r_{g1})\overline{G^2}. \qquad (4.41)$$

4.4.4 Mixer Noise Figure

Having calculated the noise contribution from the various sources to the output, the noise figure of the mixer can be estimated. Consider that the load introduces output noise which can be represented by an equivalent noise resistance R_L. The single sideband (SSB) noise figure for the single-balanced mixer is

$$(NF)_{SSB} = \frac{\alpha}{c^2} + \frac{(\gamma_3 + r_{g3}g_{m3})g_{m3}\alpha + 2\gamma_1\overline{G} + (R_{LO} + 2r_{g1})\overline{G^2} + \frac{1}{R_L}}{c^2 g_{m3}^2 R_s}. \qquad (4.42)$$

and for the Gilbert cell is

$$(NF)_{SSB} = \frac{\alpha}{c^2} + \frac{2(\gamma_3 + r_{g3}g_{m3})g_{m3}\alpha + 4\gamma_1\overline{G} + (4r_{g1})\overline{G^2} + \frac{1}{R_L}}{c^2 g_{m3}^2 R_s}, \qquad (4.43)$$

where the quantities α, c, $\overline{G}$, $\overline{G}$ are evaluated with the bias current of each switching pair, and the symbols γ_1 and γ_3 have been used for the noise factor γ of $M1$ and $M3$ respectively. If a band-pass filter is used at the input (which filters out noise from the source resistor at frequencies outside the input signal band) the term α/c^2 in formulas (4.42) and (4.43) becomes 1. If the useful signal is present in both sidebands around the LO frequency, the

double sideband (DSB) noise figure is the appropriate noise performance metric. For the

single-balanced mixer and the Gilbert cell, this is half of the SSB noise figure given by

(4.42) and (4.43) respectively. As in the SSB case, if a band-pass filter is used at the input

to reject noise from the source resistor at frequencies outside the two input signal bands,

the first term $\alpha/(2c^2)$ becomes 1. Comparing the above equations and neglecting the noise

from the LO port, we observe that for equal conversion gain, the double-balanced structure

consumes twice the power of the single-balanced one and has a higher noise figure.

4.4.5   Flicker Noise Effects

In the above analysis the effect of flicker noise was neglected, but if the system

employs direct conversion this can be a limiting factor. Flicker noise from the driver stage

appears at the output around f_{LO} and all the odd-order harmonics, since, as discussed in

section 4.3, $p_1(t)$ has only odd-order frequency components. If the PSD of flicker noise

is known at the output of the driver stage, the PSD at the output around f_{LO} can be

easily found from (4.21), since the conversion gain of the switching pair $c = |p_{1,1}|$ has been

calculated in section 4.3.

To estimate the flicker noise contribution from the switching pair we need to know

the flicker noise behavior of MOS devices with time-varying operating point. A discussion

about the flicker noise generating mechanisms in MOSFETs and the case that the operating

points changes strongly with time is given in appendix C. Assuming that a usual time-

invariant flicker-noise voltage source in series with the gate is an appropriate model, from

(4.35) this noise is transferred to the output by multiplication with $G(t)$. It is easy to

see in Fig. 4.5 that the period of $G(t)$ is $T_{LO}/2$, and therefore it contains only even-order

Figure 4.9: Measurement setup of a single-balanced mixer

harmonics of the LO frequency. This means that flicker noise from the switching pair will appear at the output around DC, but not around f_{LO}. The PSD of the noise contribution of each transistor to the output around DC can be easily found from (4.36) since G_0 is the time-average transconductance of the switching pair $\overline{G}$, which has been calculated in section 4.4.2.

4.5 Measurements

The SSB noise figure of a single-balanced mixer shown in Fig. 4.9, fabricated in the Philips Qubic2 process, with minimum drawn length $0.8\mu m$, was measured at low

frequencies. The drains of $M1$ and $M2$ were brought off chip. No attempt was made to optimize its performance, the goal being to compare predictions with measurements. No input matching was used that would improve conversion gain and lower the noise figure. The measurements were taken with the noise figure meter HP8970A [82].

Baluns with a center tap were used to transform the differential output signal to single-ended and the single-ended LO signal to differential. The series $L_1 - C_1$ trap was used to null the strong LO component at the output, that could saturate the noise figure meter input and drive $M1$ and $M2$ to the triode region. A band-pass filter reduced the noise floor of the LO signal. Care was taken to avoid introducing noise from the bias circuit.

The noise figure meter measures its own noise figure with a 50Ω source impedance during calibration and it uses this measurement to extract the noise figure of the device under test (DUT). Therefore, the output impedance of the DUT must also be matched to 50Ω, and inductor L_2 and resistor R_1 were used for this purpose. The board and balun parasitics significantly affect the behavior of the output load. It was measured that the trap resonance frequency is $72MHz$, used as LO frequency, and that an output parallel RLC resonance appears at $19MHz$, used as IF, with an impedance close to 50Ω across the $4MHz$ bandwidth that HP8970A measures noise. The output impedance of $M1$ and $M2$ is high in the bias condition of Fig. 4.9 and does not significantly affect the output impedance of the circuit.

During the measurements, the need to characterize individual components of the circuit arose. By connecting the gate of $M2$ to ground and the gate of $M1$ to a fixed bias, a cascode linear amplifier was formed. Its gain and noise figure were measured with a 50Ω AC
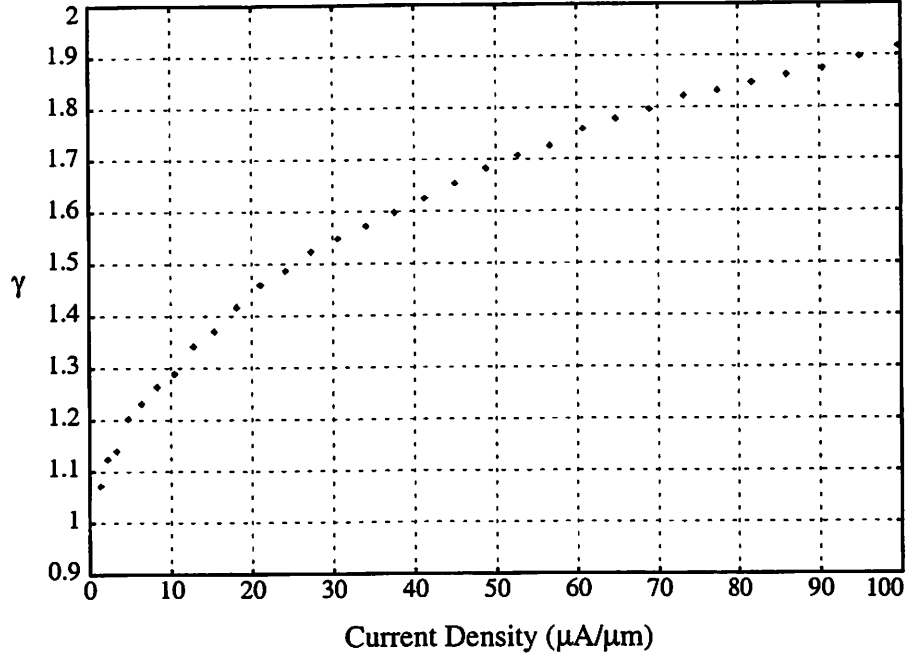
Figure 4.10: Extracted γ versus current density for a minimum channel length Qubic2 MOS transistor

load at $19MHz$, and g_m and γ of $M3$ as a function of bias were extracted. To characterize

the output load of Fig. 4.9, which was affected by the parasitics, this was used as a load of

the linear amplifier, and the gain and noise figure measurements were repeated. A second

estimate for γ was obtained, which essentially coincided with the previous one. The effect

of gate resistance noise [67] was removed, and the result is shown in Fig. 4.10. Parameter γ

was found to depend on the bias current, but not significantly on the drain or body voltage,

and therefore this measured value of γ versus bias current density was also used for the

transistors of the switching pair.

The I-V curve of $M3$ was measured and the parameters $\theta = 0.669V^{-1}$ and $K = 7.97mA/V^2$ were extracted with curve fitting. These values were used to calculate the bias

condition and small-signal transconductances of the transistors. For the transistors of the
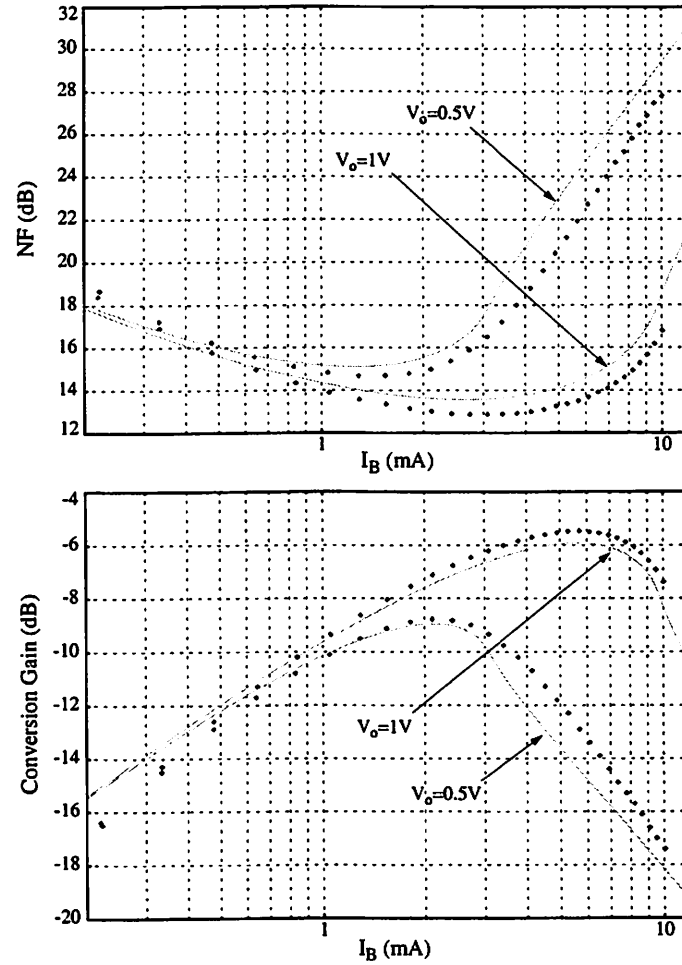
Figure 4.11: Measured (dots) and predicted (solid line) noise figure and conversion gain of the single-balanced mixer of Fig. 4.9, versus bias current

switching pair the value of γ which corresponds to zero LO voltage was used. The predicted (computed numerically) and measured values for the noise figure and conversion gain are shown in Fig. 4.11, in which fairly good agreement is observed. It is worth noticing that because of the noise of the switching pair, the optimum noise figure appears for lower current than the optimum gain. The discrepancy is mainly because of the conversion gain prediction and can be attributed to the fact that the LO amplitude applied to the switching pair can be estimated but is not exactly known because of the losses in the band-pass filter, the

Table 4.1: Parameters used in the calculation of noise figure for the mixer of Fig. 4.9. $V_o = 1V$ and $I_B = 5.6mA$.

Parameter	Numerically Evaluated	Closed-form Expression	
c	0.539	0.570	from (4.17), Vx=0.722V
a	0.590	0.658	from (4.24)
G	3.84 (mA/V)	3.57 (mA/V)	from (4.32)
G^2	32.7 $(mA/V)^2$	39.1 $(mA/V)^2$	from (4.39)

Table 4.2: Noise contribution from individual components of the mixer of Fig. 4.9. $V_o = 1V$ and $I_B = 5.6mA$.

Noise Contributors	Additional Information	Output Noise Power (pA^2/Hz)
$R_s = 50\Omega$	$g_{m3} = 7.98mA/V$	15.2
$R_L = 276\Omega$		59.6
$M3$, $r_{g3} = 3.3\Omega$	$\gamma_3 = 1.72$	135.6
$M1-M2$	$\gamma_1 = 1.36$ (for $V_{LO} = 0$)	172.1
$R_{LO} = 50\Omega$, $r_{g1} = r_{g2} = 6.6\Omega$		34.1

balun and the connections, and also to the fact that the transistor model used is inaccurate for low current density.

We will now elaborate on the calculation of the noise figure for one point of Fig. 4.11, namely for $V_o = 1V$ and $I_B = 5.6mA$ which corresponds to maximum conversion gain. Table 4.1 shows the numerically computed value of the parameters needed in the evaluation of the noise figure, together with the value resulting from approximate closed-form expressions derived in this chapter. Table 4.2 shows the contribution of individual components of the circuit to the output noise.

4.6 Upper Frequency Limit of the Analysis

To estimate the frequency range of validity of this noise analysis, it is necessary to consider the most significant of the transistor capacitances. Let C_1 and C_2 represent the gate-source capacitances of $M1$ and $M2$, and C_b denote the total capacitance from the

common source node to the ground, consisting of the source-body capacitances of $M1$ and $M2$, and the drain-body capacitance of $M3$.

For this analysis to hold, reactive effects must not significantly alter the periodically varying operating point considered in section 4.2. It is shown next in section 4.6.1 that assuming a sinusoidal LO waveform and DC common LO voltage, an approximate upper LO frequency limit for this to hold is

$$f_{LO1} = \epsilon_1 \frac{I_B}{2\pi(C_1 + C_2 + C_b)\left[\frac{V_o}{2} - (V_{gs1} - V_{gs2})\right]} \tag{4.44}$$

where ϵ_1 is a small number (e.g. 0.2 or 0.3), V_{gs1} and V_{gs0} is the low-frequency gate-source voltage of $M1$ for peak and zero LO voltage respectively, and the sum of the capacitances is evaluated for zero LO voltage. It is interesting to observe that high LO amplitude lowers this limit. Simulation with SpectreRF shows that (4.44) correctly predicts the LO frequency at which the operating point departs from the low frequency behavior. Fig. 4.12 shows simulation results for the drain current of transistor $M1$ of the mixer of Fig. 4.9, for $I_B = 2.3mA$ and for three different LO amplitudes, at the LO frequency $f_{LO1}(\epsilon_1 = 0.3)$. The sum of the three capacitances was estimated from the available SPICE model to be $0.936pF$. To avoid reactive effects at the output, the drains of $M1$ and $M2$ were connected directly to the positive supply. In the three cases we observe approximately equal overshoot above $2.3mA$ which is the peak value of I_1 at low frequencies, and therefore about equal deviation from the corresponding low frequency waveforms.

In addition, the small-signal conductance represented by the capacitors C_1, C_2 and C_b must be much lower than the sum of the conductances g_{m1} and g_{m2}. Otherwise, the signal is lost in these capacitors while the switching pair contributes noise even if one
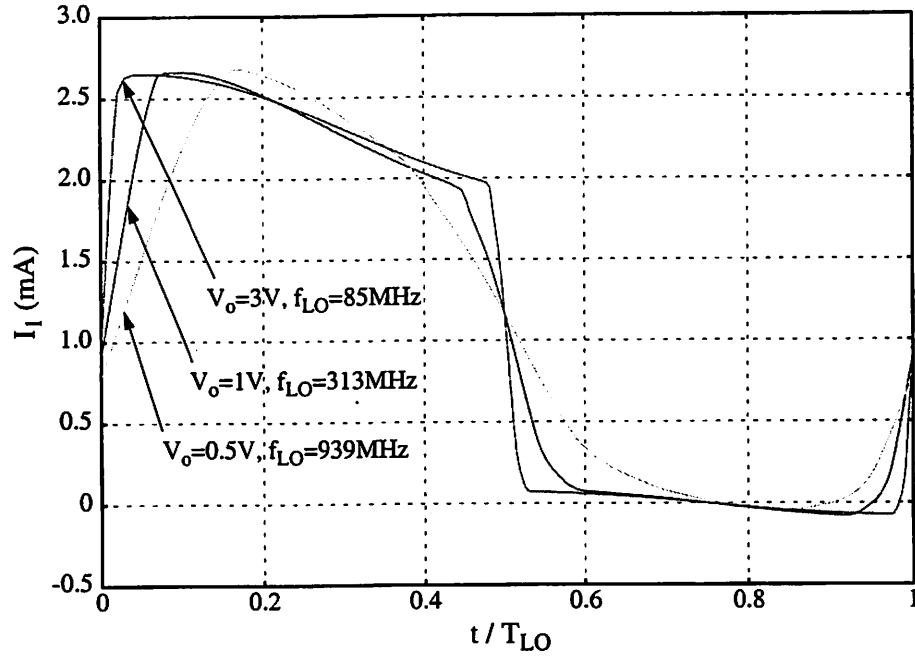
Figure 4.12: Simulated drain current of $M1$ of the mixer of Fig. 4.9 over one LO period, for three different LO amplitudes, and LO frequencies given by (4.44) with $\epsilon_1 = 0.3$. The bias current is $I_B = 2.3mA$.

of the transistors is off. For a down-conversion mixer in which the signal and the image frequencies are close to f_{LO}, a second approximate upper LO frequency limit is

$$f_{LO2} = \epsilon_2 \frac{g_{m1} + g_{m2}}{2\pi(C_1 + C_2 + C_b)} \tag{4.45}$$

where the sums of the capacitances and the conductances are considered constant and equal to their values for $V_{LO}(t) = 0$, and ϵ_2 is again a small number (e.g. 0.2).

Simulation shows that for LO frequency below f_{LO2}, the conversion gain and the noise figure are not significantly deteriorated by the change in operating point that occurs after f_{LO1}, and that in some cases they improve. For f_{LO} higher than f_{LO2}, the conversion gain and noise figure gradually degrade. Fig. 4.13 shows simulation results versus f_{LO}, for the conversion gain, noise figure and noise contribution of the switching pair of the mixer
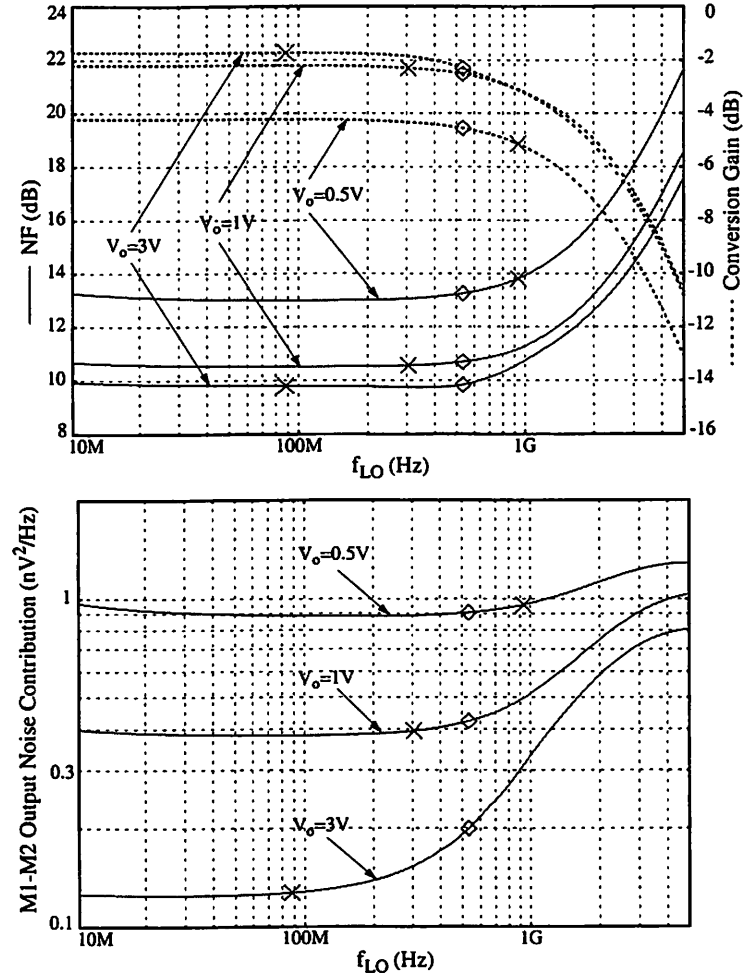
Figure 4.13: Noise figure, conversion gain and switching pair output noise contribution versus frequency for the mixer of Fig. 4.9. Frequencies f_{LO1} (X) with $\epsilon_1 = 0.3$, and f_{LO2} (diamonds) with $\epsilon_2 = 0.2$ are shown.

of Fig. 4.9, for three different LO amplitudes, together with the frequencies f_{LO1} ($\epsilon_1 = 0.3$) and f_{LO2} ($\epsilon_2 = 0.2$). The bias current is $2.3mA$, the input signal frequency is $1.1f_{LO}$ and the output signal frequency is $0.1f_{LO}$. For simplicity the filters shown in Fig. 4.9 were not included in simulation. Ideal baluns where employed and the output stage consisted only of the balun. Since the output is obtained at low frequencies, reactive effects at the output do not affect the conversion gain. We observe that for large LO amplitude the noise
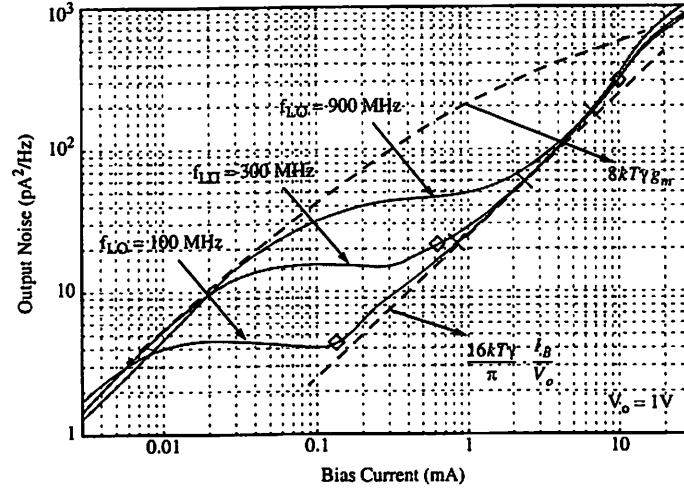
Figure 4.14: Switching pair noise versus bias current at high frequencies for the mixer of Fig. 4.9. The bias currents related to the frequency limits f_{LO1} (X) with $\epsilon_1 = 0.3$, and f_{LO2} (diamonds) with $\epsilon_2 = 0.2$ are shown. The dashed line corresponds to the noise of the two transistors when their common source is ac grounded.

contribution of the switching pair does increase after f_{LO1}, but in this case the switching pair is a minor contributor to the noise figure, which remains approximately constant up to f_{LO2}. The slight noise increase at low frequencies is caused by flicker noise of the switching pair appearing at baseband. Fig. 4.14 shows the noise generated by the switching pair of the mixer of Fig. 4.9 versus bias current, for LO amplitude $1V$ and for three different LO frequencies, $100MHz$, $300MHz$ and $900MHz$. Observe that for high bias current the noise coincides with the prediction of (4.33) while for very low bias current because of the capacitance at the common-source node, this point becomes ac ground and each transistor contributes noise $4kT\gamma g_m$, depicted with the top dashed line. The values of the bias currents related to the frequency limits f_{LO1} and f_{LO2} are also shown.

Let us now examine the flicker noise contribution of devices M1 and M2 at high frequencies. Fig. 4.15 shows the the transfer function from a voltage source in series with
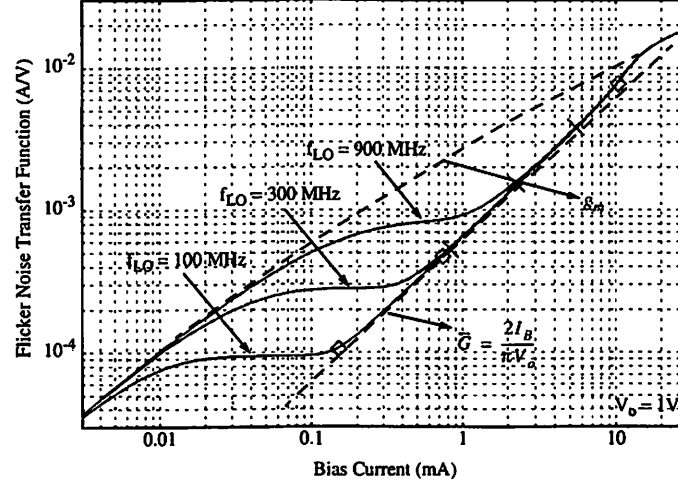
Figure 4.15: Transfer function for the flicker noise contribution of M1 and M2, versus bias current at high frequencies for the mixer of Fig. 4.9. The bias currents related to the frequency limits f_{LO1} (X) with $\epsilon_1 = 0.3$, and f_{LO2} (diamonds) with $\epsilon_2 = 0.2$ are shown.

the gate of M1 or M2 to the mixer output current, versus bias current and for LO frequency $100MHz$, $300MHz$ and $900MHz$. At high bias current we obtain the low-frequency prediction of the time-average transconductance (4.32). At low bias current the capacitance dominates the impedance of the common-source node which becomes an AC ground. In this case the transfer function is the transconductance of the transistor g_m. The same transfer function versus frequency is shown in Fig. 4.16 for a bias current $I_B = 2.3mA$ and three LO amplitudes $0.5V$, $1V$ and $3V$. The frequency limits f_{LO1} and f_{LO2} are also shown. Similarly to Fig. 4.13 deviation from the low frequency behavior is observed approximately after f_{LO2}. We must note that when considering flicker noise there is no input signal close to f_{LO} as was the case when we considered a downconverter before. However f_{LO2} is related to time constant in the system and it is worth placing on these graphs. These graphs agree qualitatively with these presented in [52].
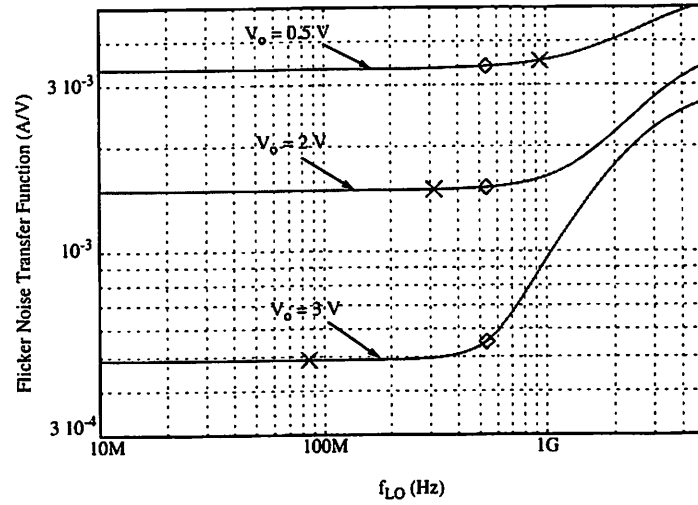
Figure 4.16: Flicker noise transfer function versus frequency for the mixer of Fig. 4.9. Frequencies f_{LO1} (X) with $\epsilon_1 = 0.3$, and f_{LO2} (diamonds) with $\epsilon_2 = 0.2$ are shown.

4.6.1 Proof of the Large-Signal High-Frequency Limit (4.44)

We derive here a limit for the frequency-independent operating-point assumption used in the analysis. The high-frequency large-signal equation for the switching pair is

$$I_B = I_1 + I_2 + C_1(V_1)\frac{dV_1}{dt} + C_2(V_2)\frac{dV_2}{dt} + C_B(V_{bs})\frac{dV_{bs}}{dt} \tag{4.46}$$

where V_1, V_2, V_{bs} are the voltages across the capacitors C_1, C_2, C_b respectively. We assume that the LO common voltage is constant with time, and equal to $V_{LO,c}$. The voltages V_1, V_2, V_{bs} can be expressed as

$$V_1 = V_{LO,c} + \frac{V_{LO}(t)}{2} - V_s \tag{4.47}$$

$$V_2 = V_{LO,c} - \frac{V_{LO}(t)}{2} - V_s \tag{4.48}$$

$$V_{bs} = -V_s \tag{4.49}$$

where $V_{LO}(t)$ is the LO voltage and V_s the potential of the common source of $M1$ and $M2$.

Using (4.47)-(4.49), (4.46) becomes

$$I_B = I_1 + I_2 + \frac{1}{2}(C_1 - C_2)\frac{V_{LO}}{dt} - (C_1 + C_2 + C_b)\frac{dV_s}{dt} \tag{4.50}$$

Of the two terms involving capacitances in this equation the last one is more significant and the other one is neglected. We will now estimate the maximum value of the derivative dV_s/dt, assuming a sinusoidal $V_{LO}(t)$. At frequencies that the reactive effects are negligible, $V_s(t)$ is a periodic waveform with frequency $2f_{LO}$, high voltage appearing when the LO voltage takes its peak value

$$V_{sh} = V_{LO,c} + \frac{V_o}{2} - V_{gs1} \tag{4.51}$$

and low voltage appearing when the LO voltage is zero

$$V_{sl} = V_{LO,c} - V_{gs0} \tag{4.52}$$

where voltages V_{gs1} and V_{gs0} are the low-frequency gate-source voltages of $M1$ for $V_{LO}(t) = V_o$, and $V_{LO}(t) = 0$ respectively. Approximating $V_s(t)$ with a sinusoid, its maximum derivative is

$$\begin{aligned} max\frac{dV_s}{dt} &= 2\pi(2f_{LO})\frac{1}{2}(V_{sh} - V_{sl}) \\ &= 2\pi f_{LO}\left[\frac{V_o}{2} - (V_{gs1} - V_{gs0})\right]. \end{aligned} \tag{4.53}$$

Capacitances C_1, C_2, are C_b are voltage dependent, but we will make the approximation that their sum is constant and equal to its value for $V_{LO} = 0$. From (4.50), for low-frequency behavior to hold, it must be

$$(C_1 + C_2 + C_b)\frac{dV_s}{dt} \ll I_B, \tag{4.54}$$

and using (4.53), (4.44) results.

4.7 Conclusions

A systematic study of the noise generating mechanisms in current commutating CMOS mixers has been completed, and analytical expressions for important parameters have been derived. We can now comment on the effect of the design parameters on the noise performance.

High bias current improves the driver stage transconductance and therefore the conversion gain and noise figure, provided that the LO amplitude and the size of the transistors of the switching pair are such that complete commutation is performed. As seen from (4.44) and (4.45) the use of high current density causes reactive effects to appear at higher frequencies.

Large LO amplitude increases the conversion gain and reduces the noise contribution of the switching pair and the LO port. After a certain value the conversion gain of the switching pair reaches its maximum value $2/\pi$, the noise contribution of the switching pair becomes negligible, and further increase does not reduce the noise figure considerably. Large LO amplitude also allows operation at higher frequencies because complete current-commutation can then be achieved with small channel width devices operating at high current density.

Increasing the channel width of $M1$ and $M2$ is desirable up to the point that for the given LO amplitude, c approaches $2/\pi$ and relations (4.33) and (4.39) hold. Further increase does not reduce the noise introduced by the switching pair as shown in (4.33), and it even increases the noise coming from the LO port as seen in (4.39). In addition, it introduces higher capacitances which cause high frequency deterioration in performance,

and represent a larger load for the LO. Increasing the channel width of $M3$ is desirable because this increases g_{m3} and therefore the conversion gain and reduces the noise figure. However, large channel width of $M3$ introduces parasitic capacitance which can degrade the performance at high frequencies and can represent a large load for the circuit driving the mixer.

Minimum channel length is preferred for the switching pair because increasing this reduces the conversion gain. Longer channel length requires larger channel width for operation with similar LO amplitude and bias current, which introduces higher parasitic capacitances. Minimum channel length is also appropriate for the driver stage since this maximizes the driver stage transconductance. However, for longer channel devices the noise factor γ of the transistors is closer to the ideal value of 2/3. Without an expression of γ as a function of channel length it is difficult to quantify this benefit.

Chapter 5

Noise in Inductively Degenerated, Conjugately Matched, Transconductance Stages

5.1 Introduction

The input of the mixer is often provided off-chip, for example when an image rejection filter is used after the LNA. In these cases the input of the mixer must be matched to 50Ω to guarantee maximum power transfer and proper filter operation. This chapter examines the noise performance of inductively degenerated conjugately matched transconductance stages, and also the noise performance of mixers which use such stages. We will adopt a more accurate transistor noise model than we did before in chapter 4. Results of the noise analysis of the transconductance stage can be directly applied to the design of
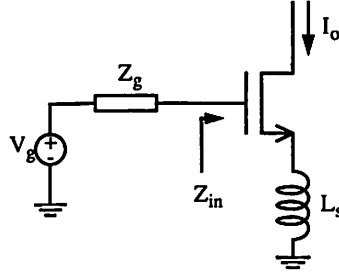
Figure 5.1: A CMOS inductively degenerated common-source transconductance stage.

LNAs.

5.2 Input Impedance

Consider the inductively degenerated CMOS transconductance stage shown in Fig. 5.1. The voltage source V_g and the impedance Z_g represent the Thevenin equivalent of the circuitry connected to the gate of the transistor. Symbol g_m represents the transconductance of the transistor, and C_{gs} its gate-source capacitance. Neglecting the body effect and the effect of the gate-drain capacitance C_{gd}, we find (see also chapter 6) that the input impedance at the transistor gate Z_{in} is given by

$$Z_{in} = \omega_T L_s + j\omega L_s + \frac{1}{j\omega C_{gs}} \qquad (5.1)$$

where $\omega_T = g_m/C_{gs}$ is the angular unity gain frequency of the transistor, and ω is the angular frequency of operation.

A more accurate expression which takes into account the body effect is

$$Z_{in} = Z_s + \frac{1}{j\omega C_{gs}} + \frac{g_m Z_s(1 - j\omega C_{gs}\chi Z_s)}{j\omega C_{gs}(1 + \chi g_m Z_s)}. \qquad (5.2)$$

where $Z_s = j\omega L_s$ is the degenerating impedance, $\chi = g_{mb}/g_m$ and g_{mb} is the body transconductance. A typical value for χ is 0.2.

Yet another expression which neglects the body effect but takes into account C_{gd} is

$$Z_{in} = \frac{\omega_T L_s + j\omega L_s + \frac{1}{j\omega C_{gs}}}{1 + \left(\frac{1}{j\omega C_{gd}} + Z_L\right)^{-1} \left(\omega_T L_s + j\omega L_s + \frac{1}{j\omega C_{gs}} + \frac{g_m Z_L}{j\omega C_{gs}}\right)} \qquad (5.3)$$

where Z_L is the load impedance connected to the drain. In the mixer case $Z_L \approx 1/(g_{m,c} + g_{mb,c})$ where $g_{m,c}$ and $g_{mb,c}$ are approximately the average gate and body transconductances of one of the devices of the switching pair.

For low values of degeneration most commonly used in practice the body effect has a negligible effect, and (5.3) provides a more accurate expression for Z_{in}. However, for simplicity, the first approximation (5.1) will be adopted below.

5.3 Transconductance Gain

Let us now define the following gain for the transconductance stage of Fig. 5.1 (also used in chapter 6) which relates the output current amplitude to the available power of the source

$$GAIN_{transc} = \frac{(RMS\ output\ current)^2}{available\ power\ of\ the\ source} \qquad (5.4)$$

Observe that $GAIN_{transc}$ as defined above has dimensions of conductance. The transconductance stage utilizes all the available power of the source when the input is conjugately matched or, neglecting the losses in the bias circuit, the matching network and the connections, when $Z_g = \overline{Z_{in}}$. Neglecting C_{gd} and using expression (5.1) we can show (see chapter
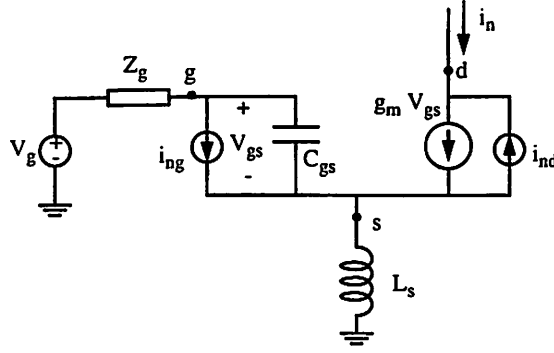
Figure 5.2: The model used in the noise analysis of the transconductance stage

6 for the derivation) that

$$GAIN_{transc} \approx \frac{\omega_T}{\omega} \frac{1}{\omega L_s} \tag{5.5}$$

Observe that $GAIN_{transc}$ depends on the ω_T but not on the device size. Taking into account

the finite quality factor Q_{L_s} of the degenerating inductor L_s, this expression becomes

$$GAIN_{transc} \approx \frac{\omega_T}{\omega} \frac{1}{\omega L_s} \frac{1}{\left(\frac{\omega}{\omega_T} \frac{1}{Q_{L_s}} + 1\right)^2} \tag{5.6}$$

which shows that even a low inductor Q does not appreciably change the prediction of (5.5).

5.4 Noise Analysis of the Transconductance Stage

We will assume that the transconductance stage of Fig. 5.1 is matched at the input.
We will adopt the model shown in Fig. 5.2. We model the noise behavior of the transistor
with the two noise current sources i_{nd} and i_{ng} [108, 41, 73, 45]. The noise current i_{nd}
represents thermal noise of the channel while i_{ng} represents noise of the channel coupled to
the gate through the distributed channel capacitance. Up to moderately high frequencies of
operation i_{ng} has negligible impact on the performance and for this reason it is not included

in more traditional transistor noise models [61]. The two noise sources are correlated. Their power spectral densities are given by

$$\frac{\overline{i_{nd}^2}}{\Delta f} = 4KT\gamma g_{do} \tag{5.7}$$

$$\frac{\overline{i_{ng}^2}}{\Delta f} = 4KT\delta g_g. \tag{5.8}$$

Above g_{do} is the gate-drain conductance in triode, for the same V_{GS} and zero V_{DS}. Quantity g_{do} is equal to g_m for square law devices, but when short channel phenomena are present it is higher than g_m, that is $g_{do} = g_m/\alpha$ where $\alpha = 0.8$ is a typical value. Quantity g_g represents the gate conductance of the transistor, which in a more complete transistor model, is connected between the gate and the source in parallel with C_{gs}. In the frequency of operation of our circuit this gate conductance has a negligible effect on the frequency response. Therefore it is omitted from our model, but the noise source associated with it is taken into account. It is given by expression [108, 41, 73, 45]

$$g_g = \frac{(\omega C_{gs})^2}{5g_{do}} \tag{5.9}$$

Parameters γ and δ have the values 2/3 and 4/3 respectively for long channel devices but are significantly higher for short channel devices. However δ remains approximately twice as high as γ [45]. The correlation between the two noise sources is expressed in terms of their correlation coefficient

$$\rho = \frac{\overline{i_{ng}^* i_{nd}}}{\sqrt{\overline{i_{ng}^2} \cdot \overline{i_{nd}^2}}}. \tag{5.10}$$

When the noise current sources have the direction shown in Fig. 5.2 for long channel devices it is shown in [108] that $\rho = j0.395$. Lacking information about its value for short

channel devices we will use the same value. Nevertheless, it is commented in [108] that the correlation terms have only a small effect in the noise performance of the circuit.

Using expression (5.1) and the fact that the input is conjugately matched, a linear analysis of the circuit of Fig. 5.2 shows that the output noise current is given by

$$i_n = x i_{nd} + y i_{ng},$$ (5.11)

where

$$x = -\frac{1}{2}$$ (5.12)

and

$$y = -\frac{1}{2}\left(\frac{Z_g}{j\omega L_s} + 1\right) = \frac{1}{2}\frac{1}{j\omega C_{gs}}\left(\frac{1}{j\omega L_s} - g_m\right)$$ (5.13)

Therefore,

$$|i_n|^2 = |x|^2|i_{nd}|^2 + |y|^2|i_{ng}|^2 + Re\left\{2xy^*i_{nd}i_{ng}^*\right\}$$ (5.14)

or

$$\overline{|i_n|^2} = |x|^2\overline{|i_{nd}|^2} + |y|^2\overline{|i_{ng}|^2} + Re\left\{2xy^*\rho\right\}\sqrt{\overline{|i_{nd}|^2|i_{ng}|^2}}$$ (5.15)

Using (5.12) and (5.13) we find that

$$Re\left\{2xy^*\rho\right\} = -\frac{0.395}{2}\frac{g_m}{\omega C_{gs}}$$ (5.16)

Substituting now (5.7), (5.8), (5.12), (5.13) and (5.16) in (5.15) we obtain

$$\frac{1}{KT}\frac{\overline{|i_n|^2}}{\Delta f} = \gamma g_{do} + \frac{\delta}{5 g_{do}}\left(\frac{1}{\omega^2 L_s^2} + g_m^2\right) - 0.79 g_m\sqrt{\frac{\gamma\delta}{5}}$$ (5.17)

The available noise power of the source is KT and the transconductance stage gain is given by (5.5). Therefore the output noise current due to the source impedance denoted

by $\overline{|i_{ns}|^2}$ is given by

$$\frac{1}{KT}\frac{\overline{|i_{ns}|^2}}{\Delta f} = GAIN_{transc} = \frac{\omega_T}{\omega}\frac{1}{\omega L_s} \qquad (5.18)$$

The noise figure of the transconductance stage is

$$
\begin{aligned}
NF_{transc} &= 1 + \frac{\overline{|i_n|^2}}{\overline{|i_{ns}|^2}} \\
&= 1 + \frac{\omega^2 L_s}{\omega_T} \cdot \left[\gamma g_{do} + \frac{\delta}{5g_{do}}\left(\frac{1}{\omega^2 L_s^2} + g_m^2\right) - 0.79 g_m\sqrt{\frac{\gamma\delta}{5}}\right] \qquad (5.19)
\end{aligned}
$$

Now substituting $g_{do} = g_m/\alpha$ we obtain

$$NF_{transc} = 1 + \frac{\omega}{\omega_T}\gamma z\left[\frac{1}{\alpha} - 0.79\sqrt{\frac{\delta}{5\gamma}} + \frac{\delta\alpha}{5\gamma}\left(1 + \frac{1}{z^2}\right)\right] \qquad (5.20)$$

or

$$NF_{transc} = 1 + \frac{\omega}{\omega_T}\gamma f(z) \qquad (5.21)$$

where $z = g_m\omega L_s$ is the feedback factor and

$$f(z) = z\left[\frac{1}{\alpha} - 0.79\sqrt{\frac{\delta}{5\gamma}} + \frac{\delta\alpha}{5\gamma}\left(1 + \frac{1}{z^2}\right)\right]. \qquad (5.22)$$

Fig. 5.3 shows quantity $f(z)$ as a function of the feedback factor z for $\delta/\gamma = 2$ and several values of α, and for $\alpha = 0.75$ and several values of δ/γ. We observe that this quantity has a flat minimum at approximately $z = 0.5$ for almost all the values of the parameters shown. The optimal value of $f(z)$ is always between 1 and 1.5. We can observe that without the gate referred noise (i.e. $\delta = 0$) it is $f(z) = z/\alpha$. An expression for the noise figure equivalent to (5.21) is

$$NF_{transc} = 1 + \frac{1}{\sqrt{GAIN_{transc} \cdot R_{in}}}\gamma f(z) \qquad (5.23)$$

where $R_{in} = \omega_T L_s$ is the real part of the input impedance Z_{in}. This expression shows that NF_{transc} is minimized for high $GAIN_{transc}$ and high R_{in}.
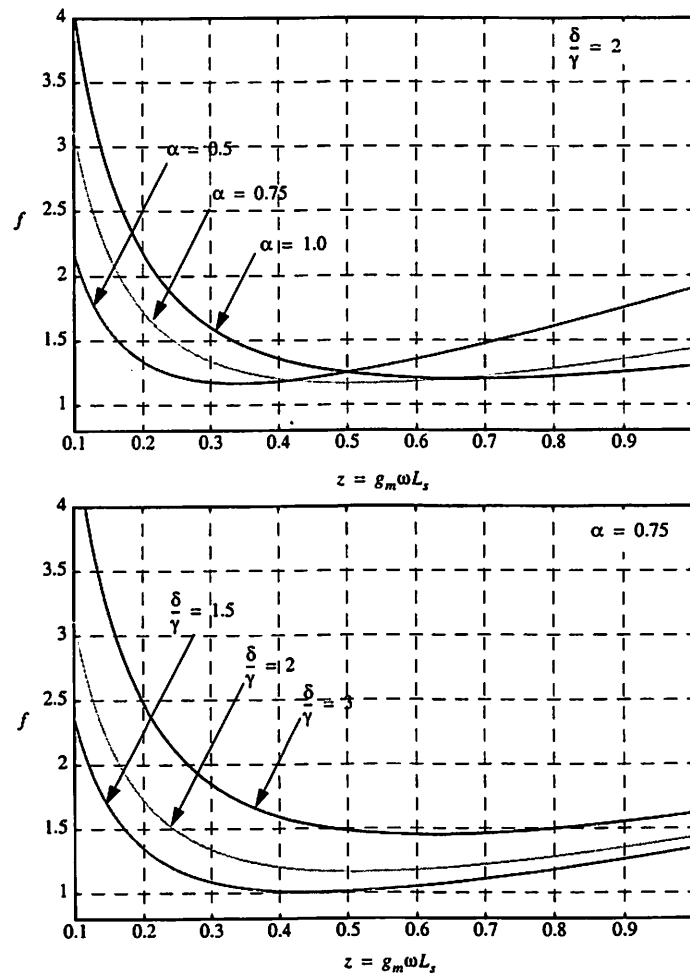
Figure 5.3: Quantity f(z) as a function of the feedback factor z, for several values of the parameters α and δ/γ.

It is now clear from (5.21) that in order to minimize the noise figure one must adhere to the following simple guidelines: a) Maximize the unity gain frequency of the input device by choosing small width devices biased at a high current density, or equivalently a high $V_{GS} - V_T$ value. b) Pick a degenerating inductor such that the feedback factor $g_m\omega L_s$ is approximately 0.5 (or close, such that the value of $f(z)$ is close to its minimum).

Although using very small width device increases ω_T and improves the noise figure, it also increases the input impedance of the transconductance stage and makes it hard to

match to the source. A high Q matching network is then required, which is harder to implement, sensitive to the component tolerances, and tends to be lossy, thereby increasing the noise figure. In particular, if ESD is used, more input signal is wasted on the lossy components of the ESD when the input impedance of the transconductance stage is high. From expression (5.1) we can find a parallel representation for Z_{in}, that is to calculate the value of a resistance R_p and a capacitance C_p whose parallel combination equals Z_{in}. We can then easily calculate that the gain loss and corresponding increase in the noise figure, expressed in dB is equal to

$$10 \log_{10} \left(\frac{R_{ESD} + R_p}{R_{ESD}} \right) \qquad (5.24)$$

where R_{ESD} is the corresponding resistance in the parallel represenation of the ESD loading. To minimize losses on the ESD it is desirable that R_{ESD} is much higher than R_p. Under the assumption

$$\frac{1}{\omega C_{gs}} >> \omega L_s \qquad (5.25)$$

which is usually satisfied, the quality factor of the input impedance is

$$Q_{in} = \frac{Im\{Z_{in}\}}{Re\{Z_{in}\}} \approx \frac{1}{z} \qquad (5.26)$$

where z is as defined before the feedback factor $g_m \omega L_s$. Then

$$R_p = Re\{Z_{in}\} \cdot (1 + Q_{in}^2) \approx \omega_T L_s \left(1 + \frac{1}{z^2} \right) \approx \frac{\omega_T}{g_m \omega} \left(z + \frac{1}{z} \right) \qquad (5.27)$$

and

$$C_p \approx C_{gs} \frac{Q_{in}^2}{1 + Q_{in}^2} \approx C_{gs} \left(\frac{1}{1 + z^2} \right). \qquad (5.28)$$

It is worth noticing that for small values of z increasing L_s reduces R_p. Minimizing the expression of R_p we see that a minimum is obtained for a feedback factor $z = 1$, as also
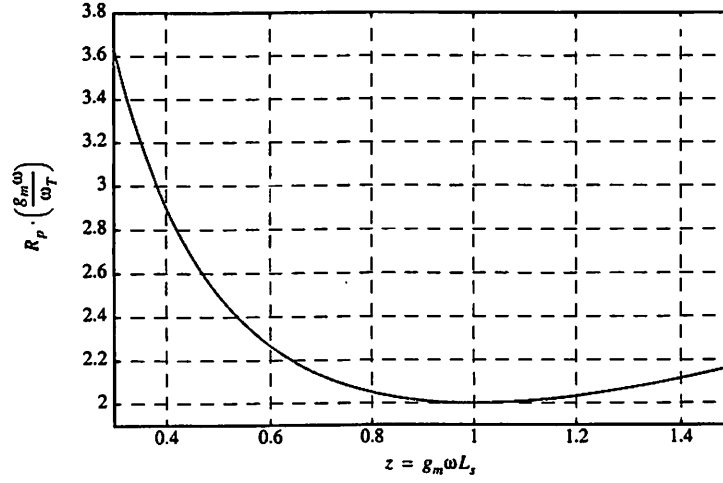
Figure 5.4: Normalized resistance R_p of the parallel representation of Z_{in} as a function of the feedback factor z.

shown in Fig. 5.4.

The input impedance R_p can be reduced without harming the noise figure at the expense of power consumption. Indeed by using a wider device biased at a higher bias current such that ω_T remains the same, and reducing L_s such that the feedback factor z remains the same, from (5.27) R_p will be lower because g_m will be larger. However, there is a limit to how small on-chip inductance can be realized reliably which is usually close to $1nH$. If a bond wire is used to realize this inductance the designer has no control over its value.

Besides the high input impedance, when using a very small transistor, the substrate resistance and the resistance of the gate polysilicon and the related contacts is divided by only a small number of fingers and can have an impact on the noise figure.

A capacitor C_A can be connected between the gate and the source of the transistor as in the transconductance stage of Fig. 5.5 to facilitate the input matching at the expense of gain. In this case the gate noise source does not increase as would be the case if we attempted
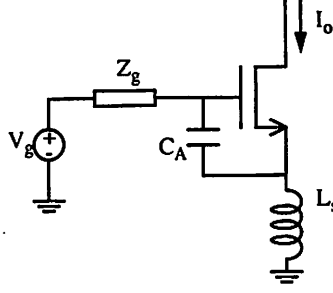
Figure 5.5: An inductively degenerated transconductance stage with added capacitance to facilitate input matching.

to improve the input matching by increasing the size of the input device. However, the introduction of C_A harms the noise figure by reducing ω_T. Similar analysis shows that the output noise current is given by the above equations if we replace parameter δ with

$$\delta' = \delta \left(\frac{C_{gs}}{C_{gs} + C_A} \right) \tag{5.29}$$

In this case the optimal value of the feedback factor is not 0.5 but can be found similarly. From Fig. 5.3 we observe that lower values of δ indeed reduce the value of $f(z)$ and also tend to lower the optimal value of the feedback factor z.

Longer than minimum length devices have a lower ω_T but also lower noise factors γ and δ. Therefore, it is possible that longer devices can provide better noise figures, or similar noise figures with lower input impedance. Since the values of the transistor noise factors as a function of channel length are unknown we can address this question only experimentally.

Some observations can be made on the above expressions of the noise figure. a) When inductive degeneration is used, the correlation between the gate and drain noise sources is such that it improves the noise figure. b) The noise figure depends almost pro-

portionally on the ratio ω/ω_T. c) The noise figure depends on ω_T but not on C_{gs}. d) Several noise sources that in practice degrade noise figure have been neglected in our analysis, for example the noise generated by the substrate resistance and amplified by the body transconductance, the noise generated by the substrate resistance and directly coupled to the output through the junction capacitance, the noise generated by the polysilicon resistance and the contacts connected to the gate. e) Finally, we should mention that we have adopted the power matching condition and not the noise matching. Therefore, in general there exists a different source impedance for which the noise figure is lower than predicted by the above analysis.

5.5   Effect of a Cascode Device

Usually a cascode device in common-gate configuration is connected to the output of the transconductance stage as shown in Fig. 5.6. In the case of an LNA this device provides reverse isolation and guarantees stability, while in the case of an active mixer this device represents the transistors of the switching pair. A pole is formed at the frequency

$$\omega_p = \frac{g_{mc} + g_{mbc}}{C_p} \tag{5.30}$$

where g_{mc} and g_{mbc} are the gate and body transconductance of device M_c and C_p is the total parasitic capacitance to ground (or some other low impedance node). Assuming that the gate of M_c is connected to a low impedance node, it is easy to see that the gate referred noise of transistor M_c adds directly to the output noise current of the transconductance stage. However, unlike the transconductance stage examined in the previous section, the effect of the gate referred current noise of the cascode transistor is usually small and will be neglected
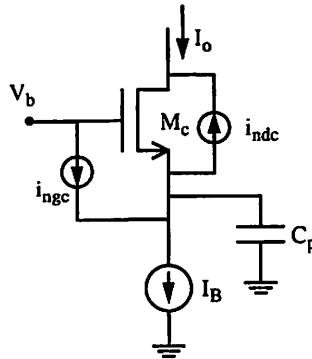
Figure 5.6: The cascode device of the transconductance stage.

here. (Indeed, observe in (5.13) that the gate referred noise of the transconductance stage is multiplied with $g_m/\omega C_{gs} = \omega_T/\omega$, which is usually large.) It is now easy to see that the contribution of the drain current of M_c to the output current is equal to

$$\frac{(\omega/\omega_p)^2}{1 + (\omega/\omega_p)^2} 4kT\gamma g_{do_c} \tag{5.31}$$

where g_{do_c} is the transistor conductance in triode, for the same V_{gs} and zero V_{ds}.

5.6 Mixer Noise Figure

An accurate calculation of the noise figure of a mixer which uses an inductively degenerated, conjugately matched transconductance stage would require consideration of the noise contribution of the transconductance stage from all possible sidebands. However, besides the complexity of the calculations required, the result depends on the out of band characteristics of the input matching network which depends on the package and board parasitics and is usually unknown during the design stage. In addition, the uncertainty in the noise model of the MOS devices renders meaningless any attempt to very accurately

predict the mixer noise figure. As an approximation we will take into account only the image band and we will assume that it contributes an equal amount of noise as the signal band.

We will adopt the low frequency expressions for the noise introduced by the switching pair derived in chapter 4. Indeed, we showed that if the conversion gain of the switching pair is not degraded by frequency effects, these expressions are valid. High frequency noise from the switching pair from the signal and the image band could be approximately included in the calculations as described in the previous section in the discussion about the cascode device.

Under these assumptions, similarly to equation (4.42), the mixer NF of a single balanced mixer is given by

$$
\begin{aligned}
NF \ &\approx\ \frac{2(\overline{i_n^2}c^2 + \overline{i_{ns}^2}c^2) + \overline{i_{sp}^2} + \overline{i_{LO}^2} + \overline{i_{R_L}^2}}{\overline{i_{ns}^2}c^2} \\
&\approx\ 2NF_{transc} + 4\frac{2\gamma_1\overline{G} + R_{LO}\overline{G^2} + 4/R_L}{GAIN_{transc}c^2}
\end{aligned}
\tag{5.32}
$$

where $\overline{i_n^2}$ is the noise generated by the transconductance stage at its output, $\overline{i_{ns}^2}$ is the noise generated by the source at the output of the transconductance stage, $\overline{i_{sp}^2}$ is the noise generated by the switching pair at the mixer output, $\overline{i_{LO}^2}$ is the noise introduced by the LO port at the mixer output, and $\overline{i_{R_L}^2}$ is the noise introduced by the mixer load of equivalent noise resistance R_L at the mixer output. Resistor R_{LO} represents the equivalent noise resistor at the LO port. Parameter γ_1 represents the γ noise factor of the transistors of the switching pair, and parameter c is the conversion gain of the switching pair calculated in chapter 4, usually slightly lower than $2/\pi$.

An observation can be made on the above expression of the noise figure. A

transconductance stage designed to have optimal noise figure when operating as an amplifier does not necessarily provide optimal mixer noise figure because of the noise introduced by the switching pair. To suppress this, $GAIN_{transc}$ must also be kept high. Therefore a smaller value for the degenerating inductor is favored in the case of the mixer. Again, a small size transconductance stage device operating at high ω_T is beneficial.

Let us now discuss the issue of the source impedance which provides optimal noise performance in a mixer with a given transconductance stage. First, the mixer output noise is a function of the source impedance at all the input bands, not only the input signal band. For the sake of simplicity let us assume that the source termination impedance at frequencies other than the signal band is insignificant, or as we considered above that only the signal and its image band are significant and they are close in frequency such that the source impedance in these two bands and the effect of source impedance in these two bands to the mixer output noise is identical. In this case we observe that the optimal source impedance for the mixer is in general different than that for the transconductance stage.

Chapter 6

Intermodulation Distortion in CMOS Transconductance Stages

6.1 Introduction

The linearity performance of the CMOS transconductance stages is of major concern in the design of transceiver blocks such as low-noise-amplifiers, active mixers and power amplifiers. In this chapter we analyze the linearity performance of such stages.

The CMOS transistors used in the transconductance stage of active mixers demonstrate fairly good linearity and are used often with little or no degeneration, in contrast with the bipolar transconductance stages which often require significant degeneration. Intermodulation distortion analysis of bipolar transconductance stages has been presented before in [17],[3]. Although the weakly nonlinear I-V relation of a bipolar device can be relatively accurately approximated by a single one-dimensional power series, in the MOS

case the drain current is generally strongly dependent not only on the gate-source voltage but also on the source-body voltage through the body effect. In several practical cases, the body-effect can dominate the nonlinearity and cannot be neglected. In [101] a two-dimensional power series is suggested for the nonlinearity analysis of a MOS transistor, which involves nine coefficients if nonlinearities up to third-order are considered. In our analysis we will consider two one-dimensional power series, one for the drain-current as a function of the effective gate-source voltage (the difference of the gate-source voltage minus the threshold voltage), and one for the threshold voltage as a function of the source-body voltage. As a result only six coefficients are involved and the final expressions are simpler (although still quite involved). The nine coefficients of the two-dimensional power series are not independent, but can be expressed in terms of the six coefficients of the two individual one-dimensional power series we employ.

We first analyze the common-source transconductance stage, and then we apply the same methodology to the common-gate transconductance stage and the differential pair.

6.2 Background and Analysis

Neglecting channel-length modulation, the I-V relation of the MOS transistor can be approximately expressed as a function of the effective gate-source voltage $V_{GST} = V_{GS} - V_T$, where V_{GS} is the gate-source voltage and V_T the threshold voltage of the device. Let V_{gst}, V_{gs} and V_t represent incremental values of V_{GST}, V_{GS} and V_T respectively around the operating point. If these perturbations are small enough that the I-V relation of the transistor remains weakly nonlinear, the incremental value of the drain current I_d can be

approximated by a third-order power series

$$I_d = g_1 V_{gst} + g_2 V_{gst}^2 + g_3 V_{gst}^3. \tag{6.1}$$

The distortion characteristics of a voltage-driven common-source stage without degeneration can be derived from the coefficients of this power series. The threshold voltage V_T depends on the source voltage through the body effect and the incremental value V_t can also be approximated by a third-order power series as function of the variation of the source voltage V_s

$$V_t = b_1 V_s + b_2 V_s^2 + b_3 V_s^3. \tag{6.2}$$

By inverting (6.1) we obtain

$$V_{gst} = V_{gs} - V_t = r_1 I_d + r_2 I_d^2 + r_3 I_d^3 \tag{6.3}$$

and by using (6.2) in (6.3) we find

$$V_{gs} = r_1 I_d + r_2 I_d^2 + r_3 I_d^3 + b_1 V_s + b_2 V_s^2 + b_3 V_s^3 \tag{6.4}$$

Expressions for g_i, r_i, and b_i, $i = 1, 2, 3$ are given in the Appendix. In the following analysis (6.4) will be used to describe the nonlinearities of the transistor which generate intermodulation distortion.

The following assumptions are adopted a) The gate-source capacitor C_{gs} is constant for small perturbations of the gate-source voltage; this is a good approximation if the device remains in strong inversion b) The nonlinear dependence on the source-body junction capacitance weak and this capacitance can be incorporated in a constant impedance connected from the source to ground c) Second-order effects that modify the threshold voltage such as Drain-Induced Barrier-Lowering can be neglected d) The output resistance of
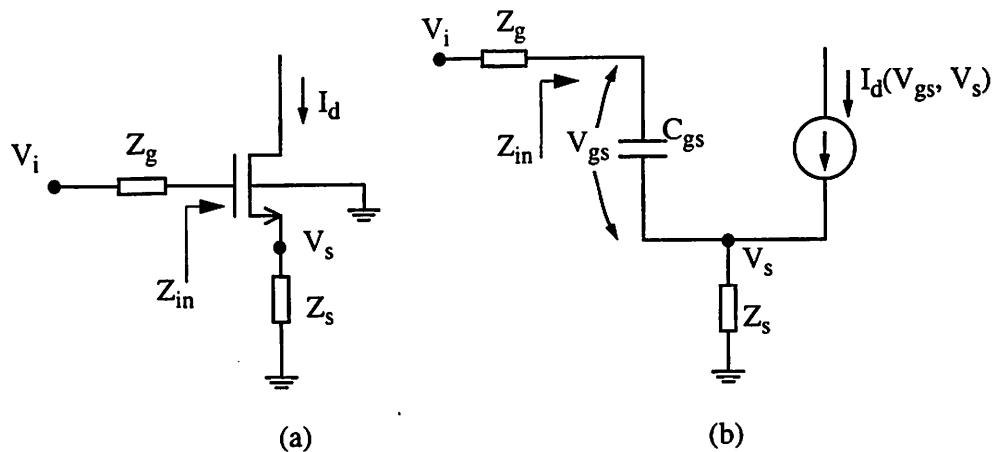
Figure 6.1: (a) A common-source transconductance stage, and (b) Equivalent circuit

the device is very high and can be neglected e) The gate-drain capacitance is small and although its effect is possibly enhanced via the Miller effect, it can be neglected. The last two assumptions are particularly valid when the transistor drain is a low-impedance node, as is the case when a low impedance load is used, a cascode device is used in common-gate configuration to provide isolation from the output to the input, or a switching pair is used to commutate the output current of the transconductance stage in an active mixer.

6.3 Common-Source Transconductance Stage

A common-source transconductance stage is shown in Fig.6.1. Impedance Z_s represents degeneration together with the source-body capacitance of the transistor and any stray capacitance to ground. Impedance Z_g and voltage V_i represent the Thevenin equivalent of the passive network in front of the transistor gate, including the input matching network, the impedance of the bias circuitry, and the signal-source impedance. The follow-

ing equations can be derived by inspection

$$V_i = K(\omega)V_{gs} + I_d Z_s \tag{6.5}$$

$$V_s = (j\omega C_{gs}V_{gs} + I_d)Z_s \tag{6.6}$$

where V_i is the incremental value of the input voltage, and

$$K(\omega) = 1 + j\omega C_{gs}(Z_s + Z_g). \tag{6.7}$$

We desire to express the output current I_d as a Volterra series of the input voltage V_i

$$I_d = G_1(\omega_a) \circ V_i + G_2(\omega_a, \omega_b) \circ V_i^2 + G_3(\omega_a, \omega_b, \omega_c) \circ V_i^3 \cdots \tag{6.8}$$

From (6.5), (6.6) and (6.8) we can obtain expressions for V_{gs} and V_s as Volterra series of V_i. Substituting these and (6.8) in (6.4) and equating terms of equal power of V_i we obtain

$$G_1(\omega) = \frac{A(\omega)}{N(\omega)} \tag{6.9}$$

$$G_2(\omega_a, \omega_b) = -\frac{1}{D(\omega_a + \omega_b)}\left[r_2 G_1(\omega_a)G_1(\omega_b) + b_2 H(\omega_a)H(\omega_b)\right] \tag{6.10}$$

$$\begin{aligned} G_3(\omega_a, \omega_b, \omega_c) = & -\frac{1}{D(\omega_a + \omega_b + \omega_c)}\left[r_3 G_1(\omega_a)G_1(\omega_b)G_1(\omega_c) + b_3 H(\omega_a)H(\omega_b)H(\omega_c)\right. \\ & \left. + 2r_2\overline{G_1(\omega_a)G_2(\omega_b, \omega_c)} + 2b_2\overline{H(\omega_a)P(\omega_b + \omega_c)G_2(\omega_b, \omega_c)}\right] \end{aligned} \tag{6.11}$$

where

$$A(\omega) = 1 - b_1 j\omega C_{gs}Z_s \tag{6.12}$$

$$B(\omega) = (1 + \frac{j\omega}{\omega_T})Z_s \tag{6.13}$$

$$P(\omega) = \left(1 - \frac{j\omega C_{gs}Z_s}{K(\omega)}\right)Z_s \tag{6.14}$$

$$D(\omega) = r_1 + b_1 Z_s + \frac{Z_s}{K(\omega)}(1 - b_1 j\omega C_{gs}Z_s) \tag{6.15}$$

$$N(\omega) = D(\omega)K(\omega) = (r_1 + (b_1 + 1)Z_s) + j\omega C_{gs}(r_1(Z_s + Z_g) + b_1 Z_s Z_g) \qquad (6.16)$$

$$H(\omega) = \frac{B(\omega)}{N(\omega)} \qquad (6.17)$$

and $\omega_T = g_1/C_{gs}$. Function $H(\omega)$ is the transfer function from V_i to V_s. If the input signal consists of two tones of equal amplitude V_{io} at frequencies ω_1 and ω_2,

$$V_i = V_{io}cos(\omega_1 t) + V_{io}cos(\omega_2 t) \qquad (6.18)$$

intermodulation products will be generated at frequencies $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$. The magnitude of the tone at $2\omega_1 - \omega_2$ is proportional to $G_3(\omega_1, \omega_1, -\omega_2)$. Usually in an intermodulation test $\omega_1 \simeq \omega_2 \simeq \omega_o$ and by letting $\Delta\omega = \omega_1 - \omega_2$ in this case we obtain

$$\begin{aligned}
G_3(\omega_1, \omega_1, -\omega_2) &= G_3(\omega_o, \Delta\omega) \\
&= \frac{1}{D(\omega_o)N(\omega_o)^2 N(-\omega_o)}[F_{r3}(\omega_o) + F_{b3}(\omega_o) \\
&\quad + F_{r2}(\omega_o, \Delta\omega) + F_{b2}(\omega_o, \Delta\omega) + F_{rb}(\omega_o, \Delta\omega)] \qquad (6.19)
\end{aligned}$$

where

$$F_{r3}(\omega_o) = -r_3 A(\omega_o)^2 A(-\omega_o) \qquad (6.20)$$

$$F_{b3}(\omega_o) = -b_3 B(\omega_o)^2 B(-\omega_o) \qquad (6.21)$$

$$F_{r2}(\omega_o, \Delta\omega) = \frac{2}{3}r_2^2 A(\omega_o)^2 A(-\omega_o)\left[\frac{2}{D(\Delta\omega)} + \frac{1}{D(2\omega_o)}\right] \qquad (6.22)$$

$$F_{b2}(\omega_o, \Delta\omega) = \frac{2}{3}b_2^2 B(\omega_o)^2 B(-\omega_o)\left[\frac{2P(\Delta\omega)}{D(\Delta\omega)} + \frac{P(2\omega_o)}{D(2\omega_o)}\right] \qquad (6.23)$$

$$\begin{aligned}
F_{rb}(\omega_o, \Delta\omega) &= \frac{2}{3}r_2 b_2\left[\frac{2A(\omega_o)B(\omega_o)}{D(\Delta\omega)}(B(-\omega_o) + A(-\omega_o)P(\Delta\omega_o))\right. \\
&\quad \left. + \frac{1}{D(2\omega_o)}\left(B(\omega_o)^2 A(-\omega_o) + A(\omega_o)^2 B(-\omega_o)P(2\omega_o)\right)\right] \qquad (6.24)
\end{aligned}$$

Evaluations of these expressions with typical parameters shows that the intermodulation is usually dominated by F_{r_3} and F_{b_3}, but F_{r_2} can also be significant. Considering only these three terms, the intermodulation is given by

$$
\begin{aligned}
IM_3 &= \frac{3}{4}\frac{|G_3(\omega_o, \Delta\omega)|}{|G_1(\omega_o)|}V_{io}^2 \\
&\simeq \frac{3}{4}\frac{\left|\left[r_3'(\Delta\omega, 2\omega_o)A(\omega_o)|A(\omega_o)|^2 + b_3 B(\omega_o)|B(\omega_o)|^2\right]\right|}{|D(\omega_o)||N(\omega_o)|^2|A(\omega_o)|}V_{io}^2
\end{aligned}
\tag{6.25}
$$

where

$$
r_3'(\Delta\omega, 2\omega_o) = r_3 - \frac{2r_2^2}{3}\left[\frac{2}{D(\Delta\omega)} + \frac{1}{D(2\omega_o)}\right]
\tag{6.26}
$$

The quantity $r_3'(\Delta\omega, 2\omega_o)$ depends on Z_s and Z_g at frequencies $\Delta\omega$ and $2\omega_o$ through the second order interaction terms. In [3] it was demonstrated that in bipolar transconductance stages it is possible to achieve cancellation of the third-order intermodulation without affecting the gain and noise performance by selecting appropriate out-of-band terminations. Such cancellation however is sensitive to process variations and requires nontrivial arrangements of passive networks in order to set the out of band terminations to the desired value. In many practical cases the input impedance transformation network and therefore impedance Z_g, depends on the package and board parasitics and is not exactly known during the design face. Often however, the value of the out-of-band terminations does not significantly affect the intermodulation and simply approximating $r_3'(\Delta\omega, 2\omega_o)$ with $r_3 = -g_3/g_1^4 + 2\frac{g_2^2}{g_1^5}$ or $-g_3/g_1^4$ introduces only a small inaccuracy in the intermodulation prediction.

Better approximations can be derived from the following expression which results from (6.26) by neglecting b_1,

$$
r_3'(\Delta\omega, 2\omega_o) \simeq -\frac{g_3}{g_1^4} + \frac{2}{3}\frac{g_2^2}{g_1^5}\left(\frac{2g_1 Z_s(\Delta\omega)}{K(\Delta\omega) + g_1 Z_s(\Delta\omega)} + \frac{g_1 Z_s(2\omega_o)}{K(2\omega_o) + g_1 Z_s(2\omega_o)}\right).
\tag{6.27}
$$

Assuming inductive degeneration $Z_s = j\omega L_s$, and that capacitor C_{gs} does not have a significant effect on the denominator of the terms in the parenthesis we obtain

$$r_3' \simeq -\frac{g_3}{g_1^4} + \frac{2}{3}\frac{g_2^2}{g_1^5}\frac{2g_1 L_s \omega_o}{1 + 2g_1 L_s \omega_o} \simeq -\frac{g_3}{g_1^4} + \frac{2}{3}\frac{g_2^2}{g_1^5} \tag{6.28}$$

where in the second expression we have used the fact that $Z_s(\Delta\omega) \approx 0$ and in the last we have assumed that $2g_1 L_s \omega_o \gg 1$. Assuming now resistive degeneration $Z_s = R_s$ and neglecting again C_{gs} we obtain

$$r_3' \approx -\frac{g_3}{g_1^4} + \frac{2g_2^2}{g_1^5}\frac{g_1 R_s}{1 + g_1 R_s} \tag{6.29}$$

It is worth noticing here, that from (6.111), g_3 for a MOS device is negative and therefore there is no value of degeneration in a resistively degenerated common-source stage for which the third-order intermodulation is nulled, as happens for a similar bipolar-transistor common-emitter stage [57].

We mention that for modern technologies and minimum channel length devices expressions (6.28) and (6.29) are usually dominated by $-\frac{g_3}{g_1^4}$. For example, using (6.109)-(6.111) we can see that $-g_3$ is higher than $\frac{2g_2^2}{g_1}$ if $\theta V_{GST} > \sqrt{(3)} - 1 = 0.73$.

The following issues deserve a further discussion.

6.3.1 Body-Effect Nonlinearity

For small values of degeneration the term proportional to b_3 in (6.25) is insignificant, while for very high values of degeneration it dominates the nonlinearity. The role of the body effect is easier to visualize at frequencies where the effect of C_{gs} is negligible. In this case

$$A(\omega) = 1, \quad B(\omega) = Z_s, \quad D(\omega) = N(\omega) = r_1 + (b_1 + 1)Z_s. \tag{6.30}$$

Then the transconductance is

$$G_1(\omega) = \frac{1}{r_1 + (b_1 + 1)Z_s} \tag{6.31}$$

and the magnitudes of the two terms in the intermodulation expression (6.25) caused by the nonlinearity in the gate and the body transconductance respectively are given by

$$IM_{3,gate} = \frac{3}{4} \frac{|r_3'(\Delta\omega, 2\omega_o)|}{|r_1 + (b_1 + 1)Z_s|^3} V_{io}^2 \tag{6.32}$$

$$IM_{3,body} = \frac{3}{4} \frac{|b_3 Z_s^3|}{|r_1 + (b_1 + 1)Z_s|^3} V_{io}^2 \tag{6.33}$$

In (6.32), $r_3'(\Delta\omega, 2\omega_o)$ is often approximately independent of the impedances Z_g and Z_s as discussed above after equation (6.26). Increasing the value of the degeneration impedance, we find that $IM_{3,gate}$ decreases asymptotically with the cube of $Z_s(\omega)$, while $IM_{3,body}$ asymptotically approaches a constant value. For high values of degeneration $IM_{3,gate}$ vanishes compared to $IM_{3,body}$, and further increasing $Z_s(\omega)$ reduces the gain but not the intermodulation which remains constant

$$IM_{3,Z_s \to \infty} = \frac{3}{4} \frac{b_3}{(b_1 + 1)^3} V_{io}^2 \tag{6.34}$$

The value of Z_s above which intermodulation is not reduced any more is given approximately by equating $IM_{3,gate}$ and $IM_{3,body}$

$$|Z_{s,lim}| \simeq \left(\frac{r_3'}{b_3}\right)^{1/3} \tag{6.35}$$

where r_3' can be approximated by r_3, $-g_3/g_1^4$, (6.28) or (6.29). Considering $r_3' = r_3$ and using the expressions (6.109)-(6.114) given in section 6.6, one can calculate the feedback factor

$$|g_1 Z_{s,lim}| \simeq 2(\phi + V_S) \left(\frac{2}{\gamma\sqrt{\phi + V_S}} \frac{(1 + \theta V_{GST})^2 + 1}{V_{GST}^2 (2 + \theta V_{GST})^2 (1 + \theta V_{GST})^2} \right)^{1/3} \tag{6.36}$$

Common values for this quantity are between 2 and 6. This value is higher for technologies with high values of ϕ, and it increases strongly as V_S increases or V_{GST} decreases.

At frequencies that the effect of C_{gs} is not negligible, (but still assuming that $b_1 \omega C_{gs} Z_{s,lim} \ll 1$ or $g_1 Z_{s,lim} \ll (1/b_1)(\omega_T/\omega_o)$ which is usually satisfied), similar considerations show that the value of $Z_s(\omega)$ at which the two intermodulation terms proportional to r_3' and b_3 contribute equally is given by

$$|Z_{s,lim}| \simeq \left(\frac{r_3'}{b_3}\right)^{1/3} \cdot \frac{1}{\sqrt{1 + (\omega_o/\omega_T)^2}} \qquad (6.37)$$

We observe that an operating point with high source voltage V_S is beneficial when the body-effect nonlinearity dominates, because this reduces the value of the coefficients b_1, b_2 and b_3 as can be seen from relations (6.116)-(6.118).

6.3.2 Input Impedance

The input impedance looking towards the transistor gate Z_{in} as shown in Fig. 6.1 will be needed below. A small-signal analysis provides

$$Z_{in} = Z_s + \frac{1}{j\omega C_{gs}} + \frac{g_1 Z_s (1 - j\omega C_{gs} b_1 Z_s)}{j\omega C_{gs}(1 + b_1 g_1 Z_s)}. \qquad (6.38)$$

Assuming that the degeneration impedance in the signal band is small enough to satisfy

$$|Z_s(\omega_o)| \ll \frac{1}{b_1 g_1}, \quad |Z_s(\omega_o)| \ll \frac{1}{b_1 \omega_o C_{gs}} = \frac{1}{b_1 g_1}\frac{\omega_T}{\omega_o} \qquad (6.39)$$

the input impedance $Z_{in}(\omega_o)$ becomes

$$Z_{in}(\omega_o) \simeq Z_s(\omega_o) + \frac{1}{j\omega_o C_{gs}} + \frac{g_1 Z_s(\omega_o)}{j\omega_o C_{gs}}. \qquad (6.40)$$

Assuming inductive degeneration $Z_s = j\omega L_s$ we have

$$Z_{in}(\omega_o) \simeq \omega_T L_s + j\omega_o L_s + \frac{1}{j\omega_o C_{gs}} \qquad (6.41)$$

while if the degeneration is resistive $Z_s = R_s$ we have

$$Z_{in}(\omega_o) \simeq R_s + \frac{1 + g_1 R_s}{j\omega_o C_{gs}}. \tag{6.42}$$

The input of the transconductance stage is often matched to the source using a lossless passive network in order to achieve maximum power transfer or to eliminate reflections. Neglecting the signal power loss in the bias circuit, impedance $Z_g(\omega_o)$ in this case should equal the conjugate of $Z_{in}(\omega_o)$ in the input signal band

$$Z_g(\omega_o) = \overline{Z_{in}(\omega_o)}. \tag{6.43}$$

6.3.3 Third-Order Input and Output Intercept Points

The third-order input intercept point (IIP_3) has been defined in chapter 2 as the available power of the source at which the third-order intermodulation product at the output equals the output linear term, assuming that nonlinearities higher than third-order are negligible. Letting $Z_g = R_g + jX_g$, considering a lossless input passive network and that no significant signal power dissipation occurs in the bias circuitry, the IIP_3 is given by

$$IIP_3 = \frac{1}{6R_g} \left| \frac{G_1}{G_3} \right|. \tag{6.44}$$

If the input is matched to the source, R_g equals the real part of Z_{in} which we have found before in section 6.3.2.

A different metric for the nonlinearity is the Output Third-Order Intercept Point I_{OIP_3}, defined here for the transconductance stage as the amplitude of the linear term of the output current when the input signal power equals IIP_3, assuming that nonlinearities

higher than third-order are negligible. It is given by

$$I_{OIP3} = \sqrt{\frac{4}{3}\left|\frac{G_1^3}{G_3}\right|} \qquad (6.45)$$

Let us now define the following gain for the transconductance stage

$$GAIN_{transc} = \frac{(RMS\ output\ current)^2}{Available\ power\ of\ the\ source} \qquad (6.46)$$

which has dimensions of conductance. Apparently the $GAIN_{transc}$, IIP_3 and I_{OIP3} are related with the equation

$$I_{OIP3}^2 = 2 \cdot GAIN_{transc} \cdot IIP_3 \qquad (6.47)$$

Assuming that the input matching network and bias circuitry are lossless

$$GAIN_{transc} = 4|G_1|^2 R_g. \qquad (6.48)$$

We established in the previous section that no linearity benefit is introduced by increasing the degeneration impedance above a certain value at which the body-effect non-linearity starts to dominate. We will assume here that the degeneration impedance is such that the body-effect nonlinearity does not dominate. Then from (6.25), neglecting the body effect nonlinearity, and assuming that conditions (6.39) are satisfied and therefore b_1 can be neglected

$$IIP_3 = \frac{1}{6R_g}\frac{|D(\omega_o)| \cdot |N(\omega_o)|^2}{r_3'(\Delta\omega, 2\omega_o)} \qquad (6.49)$$

Under the same assumptions we can find easily

$$I_{OIP3}^2 \simeq \frac{4}{3}\frac{|D(\omega_o)|}{r_3'(\Delta\omega, 2\omega_o)} \qquad (6.50)$$

$$G_1 = \frac{1}{N(\omega_o)} \qquad (6.51)$$

$$GAIN_{transc} = 4\frac{R_g}{|N(\omega_o)|^2} \tag{6.52}$$

$$D(\omega_o) \simeq r_1 + \frac{Z_s}{1 + j\omega_o C_{gs}(Z_s + Z_g)} \tag{6.53}$$

$$N(\omega_o) = r_1(1 + j\omega_o C_{gs}(Z_s + Z_g)) + Z_s \tag{6.54}$$

In the considered case that the body-effect nonlinearity does not dominate, there exists a tradeoff between the IIP_3 and the transconductance value. That is, increasing degeneration improves the IIP_3 but reduces the $GAIN_{transc}$. The I_{OIP_3} however has no obvious dependence on the degeneration value. Let us now examine the following individual cases.

Inductive Degeneration and Conjugate Matching

We assume here that the degenerating impedance is a lossy inductor $Z_s = j\omega_o L_s + R_s$. Assuming a lossless matching network and conjugate matching, using (6.41) and (6.43) in (6.53) and (6.54) we obtain

$$D(\omega_o) \simeq r_1 \left(1 + \frac{jQ + 1}{jQ + 2j\omega_o/\omega_T - 1}\right) \tag{6.55}$$

and

$$N(\omega_o) \simeq 2j\omega_o L_s \left(\frac{\omega_o}{\omega_T}\frac{1}{Q} + 1\right) \tag{6.56}$$

where $Q = \omega_o L_s/R_s$ is the quality factor of the degenerating inductor. If Q is relatively high, $D(\omega_o) \simeq 2r_1$, and $N(\omega_o) \simeq 2j\omega_o L_s$. Now from (6.49)-(6.52) we obtain

$$IIP_3 \simeq \frac{4}{3}\frac{r_1}{r_3'(\Delta\omega, 2\omega_o)}\frac{\omega_o}{\omega_T}(\omega_o L_s) \tag{6.57}$$

$$I_{OIP_3}^2 \simeq \frac{4}{3}\frac{2r_1}{r_3'(\Delta\omega, 2\omega_o)} \tag{6.58}$$

$$GAIN_{transc} \approx \frac{\omega_T}{\omega_o} \frac{1}{\omega_o L_s} \tag{6.59}$$

Approximating $r_3'(\omega_o, \Delta\omega)$ with (6.28) and using (6.109)-(6.111) in (6.58) we obtain

$$I_{OIP3}^2 \simeq \frac{8}{3}I_B^2 \frac{(2 + \theta V_{GST})^4}{[(2 + \theta V_{GST})\theta V_{GST} + 2/3]} \tag{6.60}$$

Therefore, under conjugate matching conditions the I_{OIP3} is largely independent of L_s, C_{gs}, and the input matching network. It is strongly dependent on the bias current I_B and for a given bias current I_B it depends weakly on V_{GST}, having a minimum for $\theta V_{GST} \approx 0.46$.

Inductive Degeneration and No Input Matching

Let us assume now that there is no conjugate matching restriction at the input and that $R_s = 0$ that is, the inductor is ideal. We can show that both IIP_3 and I_{OIP3} are maximized when the imaginary part of Z_g cancels that of Z_{in}

$$X_g = \frac{1}{\omega_o C_{gs}} - \omega_o L_s \tag{6.61}$$

In this case

$$D(\omega_o) = r_1 + \frac{L_s}{C_{gs}R_g} \tag{6.62}$$

$$N(\omega_o) = j(\omega_o/\omega_T)(R_g + \omega_T L_s) \tag{6.63}$$

$$IIP_3 = \frac{1}{6}\frac{r_1}{r_3'(\Delta\omega, 2\omega_o)}\left(\frac{\omega_o}{\omega_T}\right)^2 \frac{(R_g + \omega_T L_s)^3}{R_g^2} \tag{6.64}$$

$$I_{OIP3}^2 = \frac{4}{3}\frac{r_1}{r_3'(\Delta\omega, 2\omega_o)}\left[1 + \frac{\omega_T L_s}{R_g}\right] \tag{6.65}$$

$$GAIN_{transc} = \frac{4R_g}{(\omega_o/\omega_T)^2(R_g + \omega_T L_s)^2} \tag{6.66}$$

Both IIP_3 and I_{OIP3} are maximized for small R_g and large L_s.

Resistive Degeneration and Conjugate Matching

Considering now resistive degeneration and input matching, using (6.42) and (6.43) in (6.53) we obtain

$$D(\omega_o) \simeq r_1 \left[\frac{2j\omega_o/\omega_T}{2j\omega_o/\omega_T - 1} \right] \tag{6.67}$$

$$N(\omega_o) = 2j\frac{\omega_o}{\omega_T}R_s \tag{6.68}$$

$$IIP_3 = \frac{4}{3}\frac{r_1}{r_3'(\Delta\omega, 2\omega_o)} \frac{(\omega_o/\omega_T)^3 R_s}{\sqrt{1 + 4(\omega_o/\omega_T)^2}} \tag{6.69}$$

$$I_{OIP_3}^2 \simeq \frac{4}{3}\frac{2r_1}{r_3'(\Delta\omega, 2\omega_o)} \frac{\omega_o/\omega_T}{\sqrt{1 + 4(\omega_o/\omega_T)^2}} \tag{6.70}$$

$$GAIN_{transc} = \frac{1}{(\omega_o/\omega_T)^2 R_s} \qquad . \tag{6.71}$$

The expression for IIP_3 is largely independent of the value of the degenerating resistor. Comparing (6.58) and (6.70) we conclude that inductive degeneration provides a much higher Output Third-Order Intercept Point than the resistive degeneration. A similar conclusion has been reached in [17] for bipolar common-emitter stages.

Observe that at DC, expressions (6.70)-(6.69) approach zero while (6.71) approaches infinity. These however do not correspond to practical cases since the input matching condition requires an infinitely high inductance in the input matching network.

6.3.4 Degenerating Quasi-square-law Devices

It is well known that a MOS device that closely follows the square low does not exhibit third-order nonlinearity and therefore in this case degeneration will be detrimental for both the gain and the third-order linearity. On the other hand it is also known that for modern short channel devices degeneration does provides linearization benefit. In particular

we expect that if the device exhibits low short-channel effects, as we increase the degeneration linearity will initially degrade, it will reach some worst value and it will eventually improve for high degeneration values. We will attempt to answer here the question, when is it beneficial to introduce degeneration.

We will consider the simple case of low frequency and resistive degeneration R_s. Assuming that the body effect nonlinearity does not dominate and can be neglected, using equations derived earlier in this section we obtain

$$G_1 = \frac{1}{1 + (b_1 + 1)g_1 R_s} \tag{6.72}$$

$$G_3 = \frac{1}{(1 + (b_1 + 1)g_1 R_s)^4} \left[g_3 - \frac{2g_2^2}{g_1} \frac{(b_1 + 1)g_1 R_s}{1 + (b_1 + 1)g_1 R_s} \right] \tag{6.73}$$

Let us consider first that our figure of merit is the third-order input-intercept point

$$V_{IIP_3}^2 = \frac{4}{3} \left| \frac{G_1}{G_3} \right|. \tag{6.74}$$

We can show with some manipulation that IIP_3 improves monotonically as degeneration increases only if $-g_1 g_3 / g_2^2 > 2/3$ or using (6.109)-(6.111), $\theta V_{GST} > \sqrt{5/3} - 1 = 0.29$. In the opposite case the IIP_3 will initially decrease before it starts increasing monotonically, obtaining its worse value when the feedback factor is

$$(b_1 + 1)g_1 R_s = \frac{2g_2^2 + 3g_1 g_3}{6g_2^2 - 3g_1 g_3} \tag{6.75}$$

This value of the feedback factor is always lower than 1/3.

When the figure of merit is the third-order output-intercept point

$$I_{OIP_3}^2 = \frac{4}{3} \frac{G_1^3}{G_3} \tag{6.76}$$

similarly degeneration introduces monotonic benefit only if $-g_1 g_3 / g_2^2 > 2$ or using (6.109)-(6.111), $\theta V_{GST} > \sqrt{3} - 1 = 0.73$. Otherwise I_{OIP_3} obtains its lowest value when the feedback factor is

$$(b_1 + 1)g_1 R_s = \frac{2g_2^2 + g_1 g_3}{2g_2^2 - g_1 g_3} \tag{6.77}$$

This feedback factor is always lower than 1.

Let us now consider low frequency and inductive degeneration. Then considering $\Delta \omega L_s \approx 0$

$$G_1 = \frac{1}{1 + j(b_1 + 1)g_1 \omega_o L_s} \tag{6.78}$$

$$G_3 = \frac{1}{(1 + j(b_1 + 1)g_1 \omega_o L_s)^4} \left[g_3 - \frac{2g_2^2}{g_1} \frac{j(b_1 + 1)g_1 2\omega_o L_s}{1 + j(b_1 + 1)g_1 2\omega_o L_s} \right] \tag{6.79}$$

The third-order input-intercept point improves monotonically with degeneration only when $-g_1 g_3 / g_2^2 > 2.06$ or $\theta V_{GST} > 0.75$. Otherwise V_{IIP_3} obtains its lowest value when the feedback factor is

$$(b_1 + 1)g_1 \omega_o L_s = \sqrt{\frac{-(a + 2) + \sqrt{(a + 2)^2 - 3a(7 - 4a)}}{12a}} \tag{6.80}$$

where

$$a = (1 - \frac{2g_2^2}{3g_1 g_3})^2. \tag{6.81}$$

This feedback factor is always lower than 0.47. Finally the third-order output intercept point decreases monotonically with degeneration only if $-g_1 g_3 / g_2^2 > 5.65$ or $\theta V_{GST} > 1.58$. In the opposite case a minimum appears when the degeneration factor is

$$(b_1 + 1)g_1 \omega L_s = \sqrt{\frac{-1 + \sqrt{1 - a(5 - 4a)}}{4a}} \tag{6.82}$$

This feedback factor is always lower than 0.71

6.3.5 Distortion of Single Device in Common-Source Configuration

Lets us consider a single device in common source configuration without degeneration. It is well known that for high $V_{GS} - V_T$, the short channel effects tend to linearize the device and therefore third-order nonlinearity linearity is good. It is also known that CMOS devices for low $V_{GS} - V_T$ do not exhibit high short channel phenomena. Therefore one might claim that in this case the third-order nonlinearity must also be good, or that as the value of $V_{GS} - V_T$ increases the third-order nonlinearity first increases, reaches the worst value and then it improves. Such reasoning however is false. For any given device, the third-order linearity improves monotonically with $V_{GS} - V_T$. Indeed, both the third-order input-intercept point

$$V_{IIP3}^2 = \frac{4g_1}{3g_3} = \frac{4V_{GST}(2 + \theta V_{GST})(1 + \theta V_{GST})^2}{3\theta} \tag{6.83}$$

and the output intercept point

$$I_{OIP3}^2 = \frac{4g_1^3}{3g_3} = \frac{4V_{GST}^3(2 + \theta V_{GST})^3}{3\theta(1 + \theta V_{GST})^2} \tag{6.84}$$

are monotonically increasing functions of V_{GST}.

6.4 Common-Gate stage

A common-gate transconductance stage is shown in Fig. 6.2. From the circuit we obtain the equations:

$$-V_i = K(\omega)V_{gs} + Z_s(\omega)I_d \tag{6.85}$$

$$V_s - V_i = (j\omega C_{gs}V_{gs} + I_d)Z_s \tag{6.86}$$
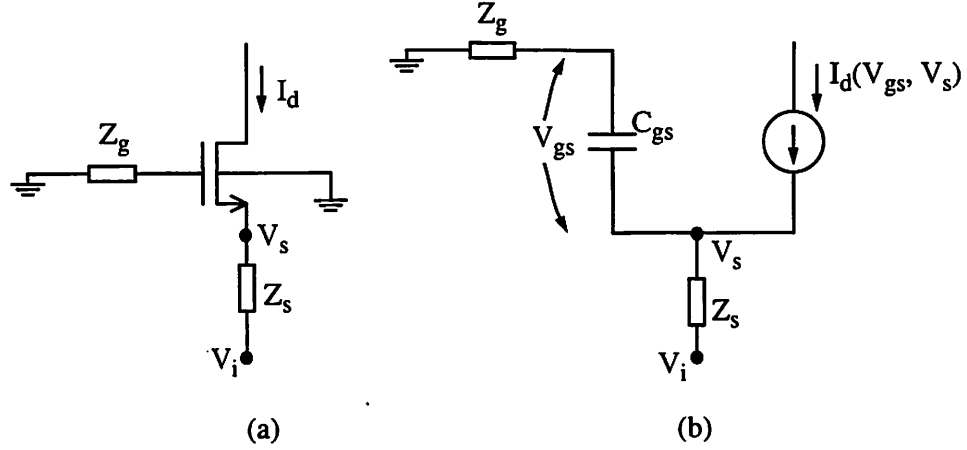
Figure 6.2: A common-gate transconductance stage

We desire to find a Volterra series relating I_d with V_i

$$I_d = G_1(\omega_a) \circ V_i + G_2(\omega_a, \omega_b) \circ V_i^2 + G_3(\omega_a, \omega_b, \omega_c) \circ V_i^3 \cdots \qquad (6.87)$$

From (6.85)-(6.87) V_{gs} and V_s can be expressed as a Volterra series of V_i and substituting

in (6.4) one obtains again the expressions (6.9) - (6.11) and (6.19) - (6.24) where functions

$A(\omega)$ and $B(\omega)$ are now given by

$$A(\omega) = -(1 + b_1(1 + j\omega C_{gs} Z_g)) \qquad (6.88)$$

$$B(\omega) = (1 + j\omega C_{gs} Z_g) r_1 \qquad (6.89)$$

and the rest of the quantities are as defined in (6.14) - (6.17).

6.5 Differential Transconductance Stage

We consider now the general case of a differential MOS transconductance stage

with degeneration at high frequency, shown in Fig. 6.3. Impedance $Z_c(\omega)$ can represent a
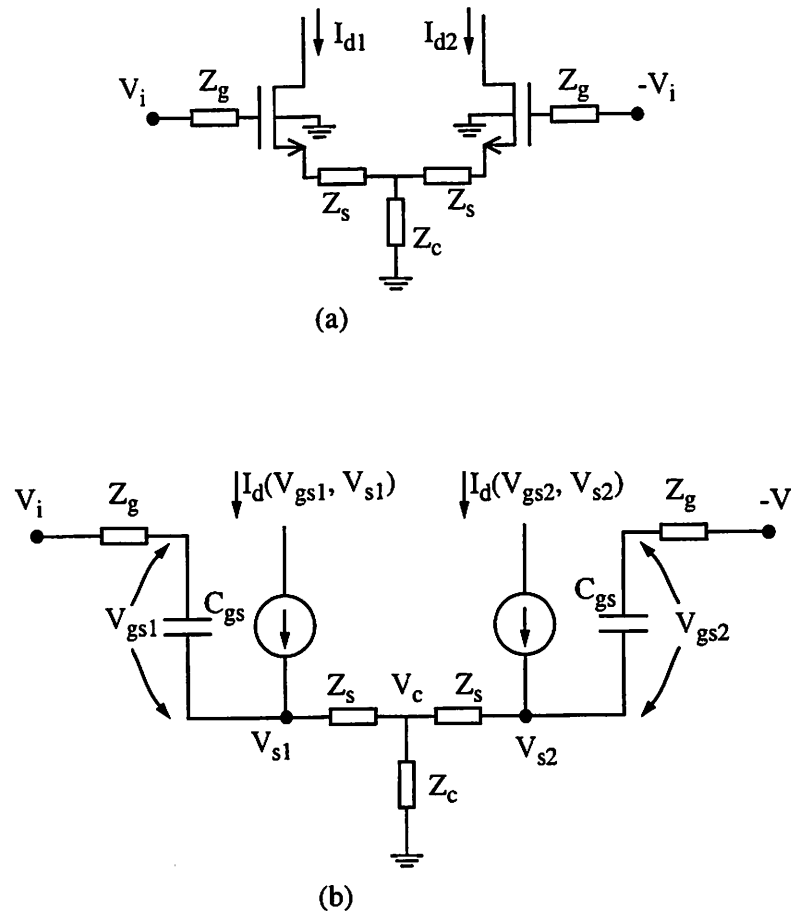
Figure 6.3: (a) A differential pair transconductance stage, and (b) Equivalent circuit

passive impedance (often used instead of a current source), the parasitic output impedance

of a current source, or stray capacitance from this point to ground such as this included in

the Π model of a spiral inductor. Finally, the network of Fig. 6.3 can include the parasitic

source-body capacitance and any stray capacitance form the transistor sources to ground,

if we apply a T to Π transformation in the T formed by impedances Z_s, Z_c and Z_s, combine

all the impedances from the sources to ground and then we apply a Π to T transformation.

From Fig. 6.3 we derive the equations

$$V_i - V_c = K(\omega)V_{gs1} + I_{d1}Z_s \tag{6.90}$$

$$-V_i - V_c = K(\omega)V_{gs2} + I_{d2}Z_s \tag{6.91}$$

$$V_{s1} - V_c = (j\omega C_{gs}V_{gs1} + I_{d1})Z_s \tag{6.92}$$

$$V_{s2} - V_c = (j\omega C_{gs}V_{gs2} + I_{d2})Z_s \tag{6.93}$$

$$V_c = (j\omega C_{gs}(V_{gs1} + V_{gs2}) + (I_{d1} + I_{d2}))\,Z_c \tag{6.94}$$

where all the symbols used are defined in Fig. 6.3, and $K(\omega)$ is defined in (6.7). It is desirable to express the output currents I_{d1} and I_{d2} as a Volterra series of the input voltage V_i:

$$I_{d1} = G_1(\omega_a) \circ V_i + G_2(\omega_a, \omega_b) \circ V_i^2 + G_3(\omega_a, \omega_b, \omega_c) \circ V_i^3 \cdots \tag{6.95}$$

$$I_{d2} = -G_1(\omega_a) \circ V_i + G_2(\omega_a, \omega_b) \circ V_i^2 - G_3(\omega_a, \omega_b, \omega_c) \circ V_i^3 \cdots \tag{6.96}$$

Combining equations (6.90) - (6.96), quantities V_c, V_{gs1}, V_{s1} can be expressed as Volterra series of V_i. For example

$$V_c = T(\omega_a + \omega_b)G_2(\omega_a, \omega_b) \circ V_i^2 \tag{6.97}$$

where

$$T(\omega) = 2Z_c\frac{K(\omega) - j\omega C_{gs}Z_s}{K(\omega) + 2j\omega C_{gs}Z_c} \tag{6.98}$$

Substituting these expressions in (6.4) we obtain:

$$G_1(\omega) = \frac{A(\omega)}{N(\omega)} \tag{6.99}$$

$$G_2(\omega_a, \omega_b) = -\frac{1}{D_d(\omega_a + \omega_b)}\left[r_2 G_1(\omega_a)G_1(\omega_b) + b_2 H(\omega_a)H(\omega_b)\right] \tag{6.100}$$

$$G_3(\omega_a, \omega_b, \omega_c) = -\frac{1}{D(\omega_a + \omega_b + \omega_c)} \left[r_3 G_1(\omega_a) G_1(\omega_b) G_1(\omega_c) + b_3 H(\omega_a) H(\omega_b) H(\omega_c) + \right.$$
$$\left. 2r_2 \overline{G_1(\omega_a) G_2(\omega_b, \omega_c)} + 2b_2 \overline{H(\omega_a) P_d(\omega_b + \omega_c) G_2(\omega_b, \omega_c)} \right] \qquad (6.101)$$

These equations are similar to (6.9)-(6.11), but $P(\omega)$ has been replaced by

$$P_d(\omega) = P(\omega) \left(1 + \frac{T(\omega)}{Z_s} \right) \qquad (6.102)$$

and $D(\omega)$ in $G_2(\omega_a, \omega_b)$ has been replaced by

$$D_d(\omega) = D(\omega) + T(\omega) \left[\frac{1}{K(\omega)} + b_1 \left(1 - \frac{j\omega C_{gs} Z_s}{K(\omega)} \right) \right] \qquad (6.103)$$

Functions $A(\omega)$, $B(\omega)$, $P(\omega)$, $D(\omega)$, $N(\omega)$ are as defined before in (6.12)-(6.17). For two closely spaced tones at frequencies ω_1 and ω_2, we obtain again $G_3(\omega_1, \omega_1, -\omega_2)$ from equations almost identical to (6.19) - (6.24), but now $D(\Delta\omega)$, $D(2\omega_o)$, $P(\Delta\omega)$, $P(2\omega_o)$ in (6.22)-(6.24) must be replaced by $D_d(\Delta\omega)$, $D_d(2\omega_o)$, $P_d(\Delta\omega)$, $P_d(2\omega_o)$ respectively.

If impedance Z_s is zero, it is easy to see from the above equations that G_1 and G_3 are independent of the body effect.

6.5.1 Differential Pair at Low Frequencies without Degeneration

Let as consider now the simple case of a differential pair without degeneration at low frequencies, biased with an ideal current source as shown in Fig. 6.4. Then the equations developed so far can be simplified as

$$G_1 = g_1 \qquad (6.104)$$

$$G_2 = 0 \qquad (6.105)$$

$$G_3 = \left(g_3 - 2\frac{g_2^2}{g_1} \right) \qquad (6.106)$$
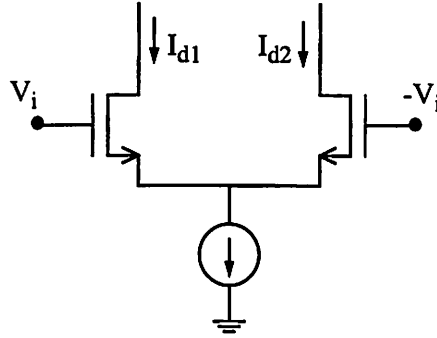
Figure 6.4: A differential pair transconductance stage without degeneration at low frequency

Let us now contrast this differential pair with with a differential transconductance stage consisting of two single transistors in common-source configuration with grounded sources. Apparently the corresponding coefficients in this case are $G_1 = g_1$, $G_2 = g_2$ and $G_3 = g_3$. Assuming that the same bias current is consumed in both cases, the transconductance value is the same. The third-order coefficient however has always lower magnitude for the transistors with grounded sources, since g_3 is negative, as we can see from (6.111). This is opposite from what happens in the corresponding comparison with bipolar devices, where the third order nonlinearity coefficient is positive.

Using (6.109) - (6.111) in (6.106) we find that for the differential pair

$$G_3 = -\frac{2K}{(1 + \theta V_{GST})^4} \left(\theta + \frac{2}{V_{GST}(2 + \theta V_{GST})} \right) \tag{6.107}$$

where V_{GST} is the effective gate-source voltage of each device.

6.6 Appendix: I – V curve power series

The I-V relation of a transistor in the strong inversion and saturation region can be approximated by [61]

$$I = K\frac{V_{GST}^2}{1 + \theta V_{GST}} \tag{6.108}$$

where $V_{GST} = V_{GS} - V_T$ represents the effective gate-source voltage. With differentiation we find the coefficients of the power series (6.1):

$$g_1 = K\frac{V_{GST} \cdot (2 + \theta V_{GST})}{(1 + \theta V_{GST})^2} \tag{6.109}$$

$$g_2 = K\frac{1}{(1 + \theta V_{GST})^3} \quad \cdot \tag{6.110}$$

$$g_3 = K\frac{-\theta}{(1 + \theta V_{GST})^4} \tag{6.111}$$

In the above expressions we have assumed that V_{GST} is at least 0.2V, such that the devices are in the strong inversion. Inverting (6.1) one obtains the coefficients of (6.3).

$$r_1 = \frac{1}{g_1} \tag{6.112}$$

$$r_2 = -\frac{g_2}{g_1^3} \tag{6.113}$$

$$r_3 = 2\frac{g_2^2}{g_1^5} - \frac{g_3}{g_1^4} \tag{6.114}$$

The threshold voltage dependence on the source voltage can be approximated by the well known expression [61]

$$V_T = V_{T0} + \gamma(\sqrt{\phi + V_S} - \sqrt{\phi}) \tag{6.115}$$

where V_{T0} is the threshold voltage when the source is tied to the body, γ is the body effect coefficient and ϕ is the surface potential. In fact, for small dimension devices the above

expression is not exact [94] but some experimentation shows that it is a good approximation if γ and ϕ are treated as fitting parameters. A Taylor expansion provides the coefficients of (6.2).

$$b_1 = \frac{1}{2}\frac{\gamma}{(\phi + V_S)^{1/2}} \tag{6.116}$$

$$b_2 = -\frac{1}{8}\frac{\gamma}{(\phi + V_S)^{3/2}} \tag{6.117}$$

$$b_3 = \frac{1}{16}\frac{\gamma}{(\phi + V_S)^{5/2}} \tag{6.118}$$

Coefficient b_1 is the ratio g_m/g_{m_b} of the small-signal body transconductance over the small-signal body transconductance and a typical value for it is 0.2.

Chapter 7

Intermodulation Distortion of the Switching Pair

7.1 Introduction

Having examined the nonlinearity of the CMOS transconductance stages in chapter 6 we examine here the nonlinearity of the switching pair. Its nonlinearity imposes problems, particularly when together with high linearity, high gain is required from the mixer, since then the signal at the output of the transconductance stage is large. While the exponential I-V characteristics of the bipolar transistor make the bipolar transistor switching pair arbitrarily linear at low frequency if the device base resistance is low [56][31], this is not true for the CMOS switching pair which demonstrates significant nonlinearity even at low frequencies.

The switching pair is treated as a weakly-nonlinear periodically-time-varying cir-

cuit and time-varying power series are employed in the analysis. It will be shown that if the capacitive effects are negligible, in the frequency band of interest the distortion behavior of the switching pair can be described by a time-invariant power series which can be cascaded with the power series of the transconductance stage to calculate the total distortion. The methodology also applies to high frequencies, where time-varying Volterra series replace the time-varying power series. Similar approaches have been used in [83] and [48] for the distortion of diode mixers, and in [49] for the distortion of passive MESFET mixers. Time-varying Volterra series have also been used in [106] for the analysis of MOS track and hold sampling mixers. Our approach identifies the characteristics of a transistor model for a reliable mixer distortion simulation.

A distortion study of a bipolar switching pair has been presented in [56] and [31], but the method used was different from the one employed here. The behavior of the bipolar switching pair was found to depend on only a few normalized variables, and transient analysis was used to find their effect on the distortion. Using the methods described here for fast evaluation of the CMOS switching pair distortion we also provide normalized graphs from which one can predict the intermodulation for any given technology and operating conditions. The analysis presented in this chapter has also been published in [90].

7.2 Transistor Model

The transistors of the CMOS switching pair operate in weak, moderate and strong inversion. It will become apparent below that a model which ignores the subthreshold region, or uses different equations to describe the different modes of operation, is inappro-

priate for a distortion analysis of the switching pair. A model which describes all three regions with a single analytical expression and therefore has continuous derivatives of any order is needed. Furthermore, it must be simple enough for analytical calculations.

Continuous MOS transistor models have been presented in [93] and [94]. We will use the same kind of smooth interpolation between the regions of operation. We will take into account to a first order the deviation from the square law in strong inversion while we will neglect second-order phenomena such as channel-length modulation, which complicate the transistor model. As a result of the latter assumption we will neglect the distortion introduced by the output impedance of the devices, assuming that a linear load dominates the mixer output. The drain current I as a function of the gate-source voltage V_{GS} is modeled in this chapter by

$$I = f(V_{GS} - V_T) = K \frac{X^2}{1 + \theta X} \tag{7.1}$$

where

$$X = 2\eta \phi_t \ln(1 + e^{\frac{(V_{GS} - V_T)}{2\eta \phi_t}}). \tag{7.2}$$

Above, V_T is the threshold voltage, ϕ_t is the thermal voltage kT/q, and K is a constant depending on the technology and the transistor dimensions, proportional to the transistor width. Parameter θ approximately models source series resistance, mobility degradation because of the vertical field, and short-channel effects such as velocity saturation [61]. For an existing 0.8 μm technology it was estimated to be $0.9\,V^{-1}$ and for a different 0.25 μm technology it was found to be approximately $2.5\,V^{-1}$. It is a function of the channel length and is independent of the body effect. Parameter η determines the rate of exponential increase of the drain current with the gate-source voltage in the subthreshold region, and

also the size of the moderate inversion region. It takes values approximately between 1 and 2 and it decreases (approaching 1) because of the body effect when the source-body voltage increases, while it tends to be higher when short-channel phenomena are present. It is shown in [93] to depend slightly on V_{GS}, but for simplicity η will be considered here a constant for a given channel length. The parameter corresponding to our parameter X is referred to in the BSIM3 version 3 manual [94] as the effective $V_{GS} - V_T$ voltage. Using this model the analysis presented in [89] can also be performed while taking into account the effect of weak and moderate inversion.

This model reduces to known expressions in strong and weak inversion. In strong inversion where the exponential term dominates the argument of the logarithm in (7.2), $X = V_{GS} - V_T$, and (7.1) becomes

$$I = K \frac{(V_{GS} - V_T)^2}{1 + \theta(V_{GS} - V_T)} \tag{7.3}$$

the common I-V relation in saturation. In weak inversion, using the approximation $\ln(1 + z) \approx z$ for small z, we obtain

$$X = 2\eta\phi_t e^{\frac{(V_{GS} - V_T)}{2\eta\phi_t}} \tag{7.4}$$

and since X is small, 1 dominates in the denominator of (1), and it provides

$$I = K(2\eta\phi_t)^2 e^{\frac{(V_{GS} - V_T)}{\eta\phi_t}} \tag{7.5}$$

which is the exponential I-V relation of the transistor in weak inversion. In moderate inversion equations (7.1) and (7.2) provide a smooth monotonic increase, interpolating between equations (7.3) and (7.5). However, the proportionality constant of (7.5) is probably inaccurate and the value of parameter $\eta\phi_t$ that provides the correct exponential increase

in weak inversion does not necessarily provide accurate moderate-inversion modeling. For distortion prediction of the switching pair the moderate-inversion region is more significant than the subthreshold and it is preferable to consider an η value that better models the moderate inversion. Nevertheless, it will be shown that the value of η has only a minor effect on the distortion prediction.

7.2.1 Derivatives of the Drain Current of a CMOS Transistor

If $I = f(V)$ is the I-V relation of a transistor in saturation as given by (7.1) and (7.2), with direct differentiation we find

$$f_V = f_X \cdot X_V \tag{7.6}$$

$$f_{VV} = f_{XX} \cdot X_V^2 + f_X \cdot X_{VV} \tag{7.7}$$

$$f_{VVV} = f_{XXX} \cdot X_V^3 + 3(f_{XX} \cdot X_V \cdot X_{VV}) + f_X \cdot X_{VVV} \tag{7.8}$$

where

$$f_X = K \frac{X \cdot (2 + \theta X)}{(1 + \theta X)^2} \tag{7.9}$$

$$f_{XX} = K \frac{2}{(1 + \theta X)^3} \tag{7.10}$$

$$f_{XXX} = K \frac{-6\theta}{(1 + \theta X)^4} \tag{7.11}$$

are the first three derivatives of f with respect to X,

$$X_V = \frac{1}{1 + s^{-2}} \tag{7.12}$$

$$X_{VV} = \frac{1}{2\eta\phi_t} \cdot \frac{1}{(s + s^{-1})^2} \tag{7.13}$$

$$X_{VVV} = -\frac{1}{(2\eta\phi_t)^2} \cdot \frac{(s - s^{-1})}{(s + s^{-1})^3} \tag{7.14}$$

are the first three derivatives of X with respect to V, and

$$s = e^{\frac{(V-V_T)}{4\eta\phi_t}}$$

(7.15)

7.2.2 Comparison of the Simple Model with Spice Models

The I-V transistor curve in saturation and its first three derivatives with respect to V_{GS} obtained from this simple model were compared with the corresponding curves obtained from the SPICE models BSIM3 version 3 and version 2, in Fig. 7.1 and Fig. 7.2 respectively. The two SPICE models describe different technologies of channel length $0.25\,\mu m$ and $0.8\,\mu m$ respectively. The parameters of the simple model were curve fitted to the I-V curves obtained from SPICE. The derivatives of the simple model were derived analytically (appendix), while those of the SPICE models were calculated numerically. As numerical noise imposes problems in the evaluation of the second and third derivative with successive differences, a more sophisticated method was used. For every value of V_{GS} a polynomial was fitted to a number of points around this value and then the derivatives of the polynomial were taken analytically [63].

As can be seen in Fig. 7.1, model BSIM3 version 3 provides smooth derivatives, as one would expect from a physical model and the simple model is in close agreement with it. In Fig. 7.2 we observe that the I-V curve and the first derivative generated with the BSIM3 version 2 model coincide with those of the simple model. However, the use of a different equation for the weak, moderate and strong inversion in the BSIM3 version 2 model becomes apparent in the second and third derivatives, where discontinuities appear at the transitions. We will see below the effect of these discontinuities on the distortion
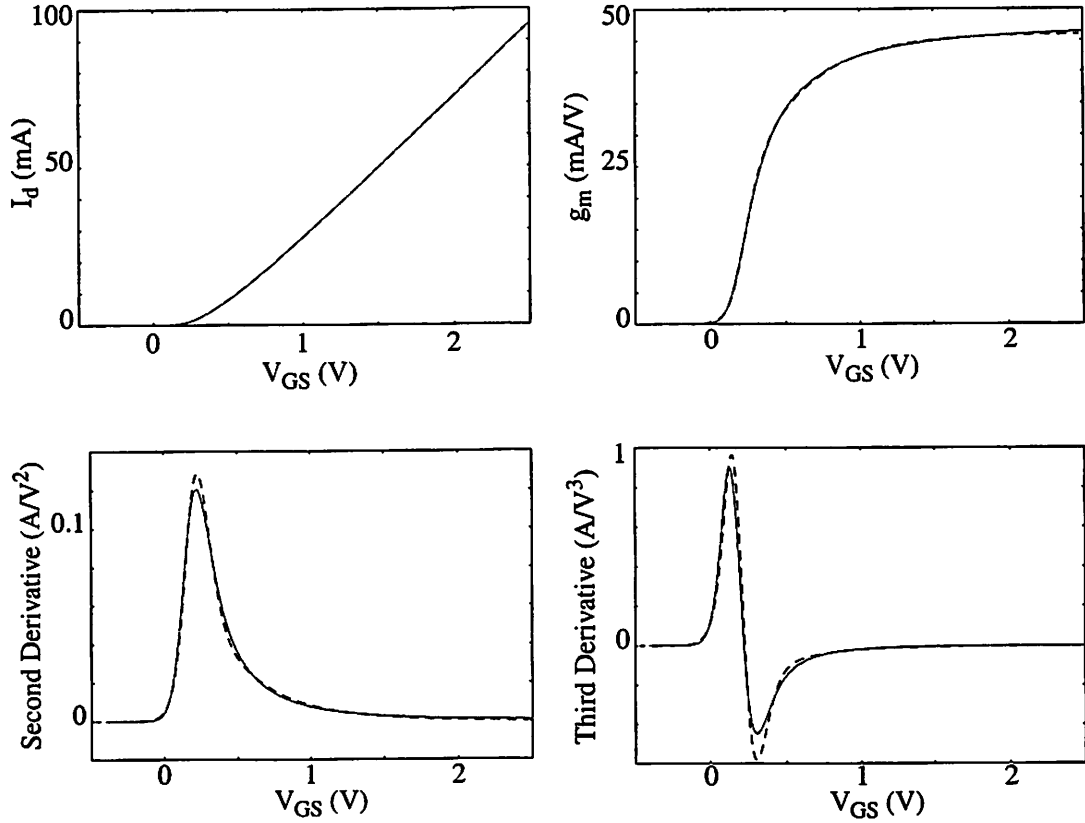
Figure 7.1: The I-V curve and the first three derivatives for a quarter-micron CMOS technology. The solid line is the simple model and the dashed line is obtained from the BSIM3 version 3 model.

simulation of the switching pair.

7.3 Switching Pair Distortion at Low Frequencies

7.3.1 Low-Frequency Large-Signal Equations

Consider the single-balanced mixer of Fig. 7.3. The operating point of the transistors of the switching pair varies periodically with time. In the following analysis we need to be able to find this operating point for a given bias current I_B and instantaneous local oscillator voltage $V_{LO}(t)$. The output conductance of the devices is neglected and the load
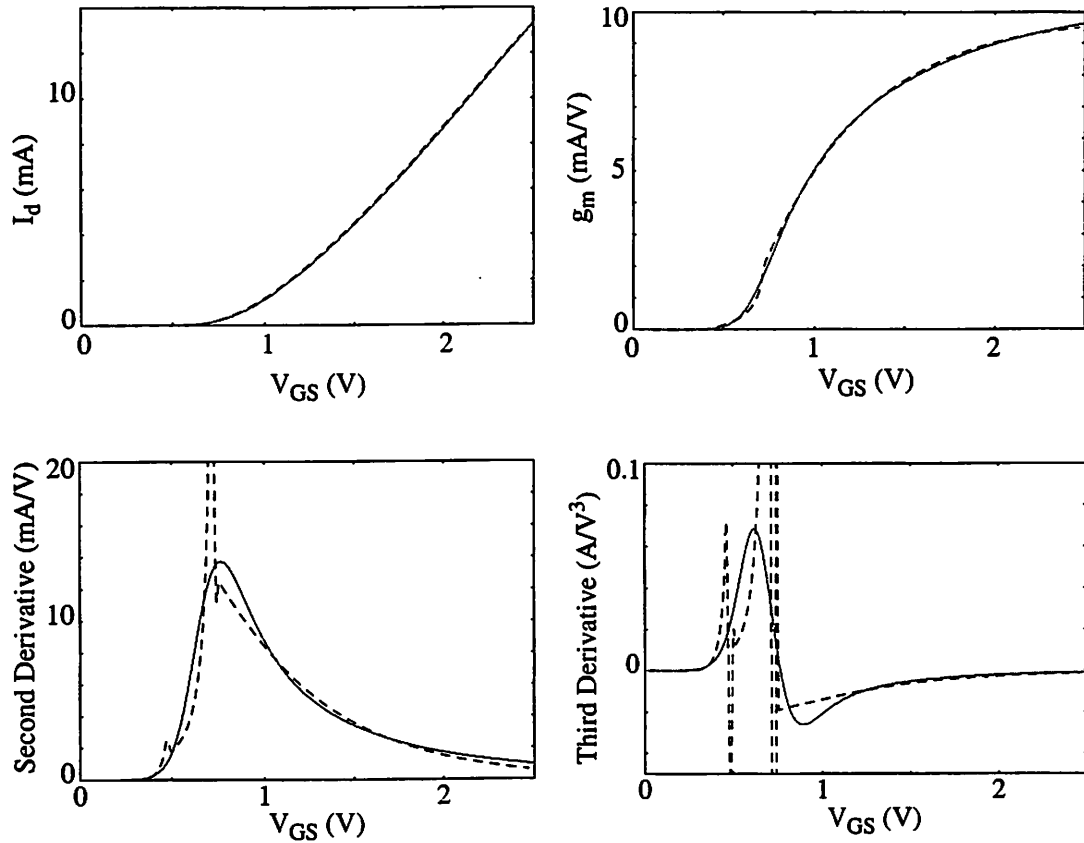
Figure 7.2: The I-V curve and the first three derivatives for a 0.8 μm CMOS technology. The solid line is the simple model and the dashed line is obtained from the BSIM3 version2 model.

at the drains of M1 and M2 is assumed such they remain in saturation during the whole LO period. This assumption is usually satisfied since if the transistors of the switching pair enter the triode region, the common source node becomes a high impedance point and performance is degraded because of reactive effects. If $I = f(V_{GS} - V_T)$ is the I-V relation of a transistor as given by (7.1) and (7.2), the large-signal behavior of the switching pair is described by the following equations

$$f(V_1) + f(V_2) = I_B \tag{7.16}$$

$$V_1 - V_2 = V_{LO} \tag{7.17}$$
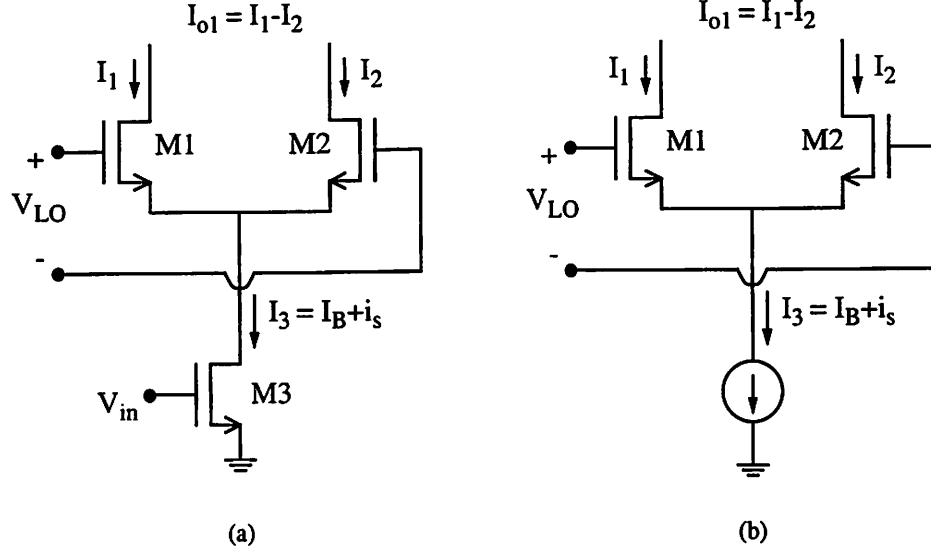
Figure 7.3: (a) A simple single-balanced active CMOS mixer, and (b) The basic model of a current commutating CMOS mixer

where $V_1 = V_{GS1} - V_T$, $V_2 = V_{GS2} - V_T$, and V_{GS1} and V_{GS2} are the gate-source voltage of M1 and M2 respectively. Substituting V_2 from (7.17) to (7.16), we obtain one nonlinear equation with V_1 as an unknown which can be solved rapidly with an iterative numerical method.

From (7.17) and (7.16) we observe that the drain current of each transistor does not depend on V_T, and therefore to the extent of validity of the transistor model used here, the behavior of the switching pair is independent of the body effect and the common-mode LO voltage. The same conclusion was reached in [89], but without taking into account the subthreshold region of operation.

7.3.2 Distortion Calculations

At low frequencies the switching pair is a memoryless system. Neglecting the output resistance of the transconductance stage, the output current defined as the difference

of the drain currents of M1 and M2, is a function of the instantaneous values of the output current of the transconductance stage and the LO voltage

$$I_{o1} + i_{o1} = F(V_{LO}(t), I_B + i_s) \tag{7.18}$$

where I_{o1}, I_B denote values without input signal present, and i_{o1}, i_s denote incremental values. Since i_s is small, a third-order Taylor expansion provides

$$i_{o1} = \frac{dF}{dI_B} \cdot i_s + \frac{1}{2} \frac{d^2 F}{dI_B{}^2} \cdot i_s{}^2 + \frac{1}{6} \frac{d^3 F}{dI_B{}^3} \cdot i_s{}^3 \tag{7.19}$$

or

$$i_{o1} = p_1(t) \cdot i_s + p_2(t) \cdot i_s{}^2 + p_3(t) \cdot i_s{}^3 \tag{7.20}$$

where $p_1(t)$, $p_2(t)$, $p_3(t)$ are periodic waveforms of which a typical shape is shown in Fig. 7.4. The value of these waveforms is easily determined when one of the transistors is off. For example, when M2 is off $p_1(t) = 1$, and $p_2(t) = p_3(t) = 0$. When instantaneously $V_{LO}(t) = 0$, $i_{o1} = 0$, and $p_1(t) = p_2(t) = p_3(t) = 0$, because of symmetry. When the conductance of both M1 and M2 is significant, $p_1(t)$, $p_2(t)$, $p_3(t)$ depend on the bias current I_B, the LO voltage V_{LO}, and the device characteristics.

With some manipulation, waveforms $p_1(t)$, $p_2(t)$, $p_3(t)$ can be expressed in terms of the derivatives of the I-V function f with respect to V_{GS} as follows

$$p_1(t) = \frac{f_{1V} - f_{2V}}{f_{1V} + f_{2V}} \tag{7.21}$$

$$p_2(t) = \frac{f_{2V} f_{1VV} - f_{1V} f_{2VV}}{(f_{1V} + f_{2V})^3} \tag{7.22}$$

$$p_3(t) = \frac{(f_{1VV} + f_{2VV})(f_{1V} f_{2VV} - f_{2V} f_{1VV})}{(f_{1V} + f_{2V})^5}$$
$$+ \frac{f_{2V} f_{1VVV} - f_{1V} f_{2VVV}}{3(f_{1V} + f_{2V})^4}. \tag{7.23}$$
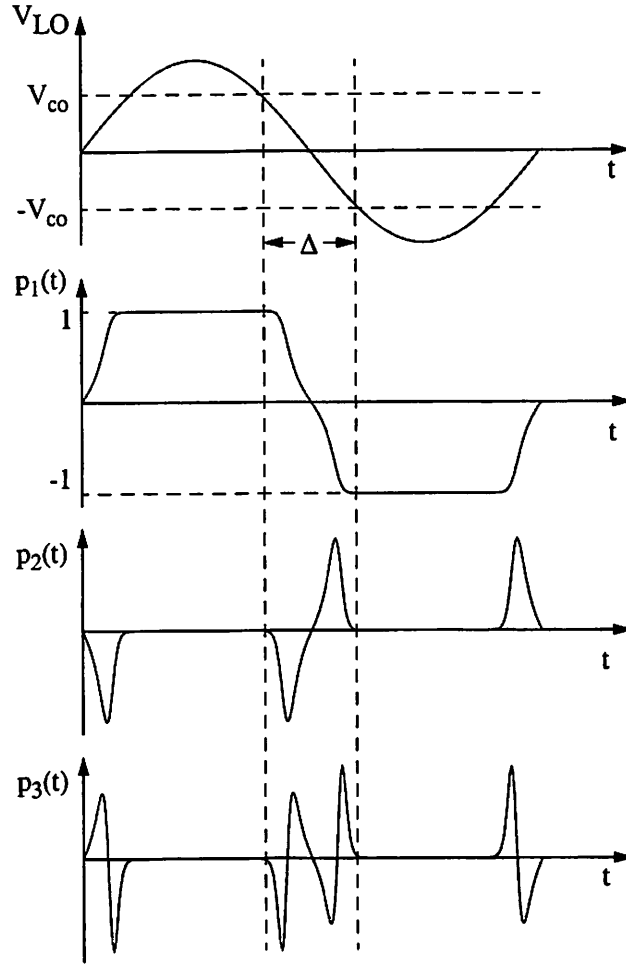
Figure 7.4: Typical shape of waveforms $p_1(t)$, $p_2(t)$ and $p_3(t)$.

The derivatives of f are denoted with the symbol f followed by an index, whose first character (1 or 2) denotes the transistor (M1 or M2 respectively) and the number of V's following denotes the order of the derivative.

Without loss of generality, $p_1(t)$, $p_2(t)$ and $p_3(t)$ can be considered odd functions of time and can be expanded in a series of sinusoids. In this case (7.20) provides

$$i_{o1} = \sum_{k=1}^{\infty} [p_{1,k} \cdot i_s + p_{2,k} \cdot i_s^2 + p_{3,k} \cdot i_s^3] \cdot \sin(2\pi k f_{LO} t) \qquad (7.24)$$

where $p_{i,k}$ is the kth coefficient of the waveform $p_i(t)$ in the series, and f_{LO} is the LO frequency. The mixer is usually used for upconversion or downconversion by one LO multiple and in this case the distortion behavior of the switching pair in the frequency band of interest can be described by a time-invariant power series[1]

$$i_{o1} = b_1 \cdot i_s + b_2 \cdot i_s{}^2 + b_3 \cdot i_s{}^3 \tag{7.25}$$

where

$$b_i = \frac{p_{i,1}}{2} = \frac{1}{T_{LO}} \int_0^{T_{LO}} p_i(t) \sin(2\pi f_{LO} t) dt \tag{7.26}$$

and T_{LO} is the LO period. If i_s consists of two tones of equal magnitude I_s at two closely spaced frequencies f_1 and f_2

$$i_s = I_s \cos(2\pi f_1 t) + I_s \cos(2\pi f_2 t) \tag{7.27}$$

the generated third-order intermodulation is

$$IM_3 = \frac{3}{4} \frac{b_3}{b_1} I_s^2. \tag{7.28}$$

For high LO amplitude $p_1(t)$ resembles a square waveform and b_1 approaches $2/\pi$. Assuming that the time interval Δ (see Fig. 7.4), during which $p_2(t)$ and $p_3(t)$ are non-zero, is small compared to the whole period and that during this time the LO voltage is a linear function of time with slope λ, it can be shown that the coefficients b_2 and b_3 decrease inversely proportional to the square of λ. Indeed, for b_3 for example, approximating the sinusoid with its argument, (7.26) provides

$$b_3 = \frac{4 \cdot 2\pi}{\lambda^2 T_{LO}{}^2} \int_0^{V_{co}} p_3(V_{LO}) V_{LO} dV_{LO} \tag{7.29}$$

---

[1]Equation (7.25) as well as equation (7.36) below are 'loose', in the sense that the left and right side refer to the output and input frequency bands respectively.
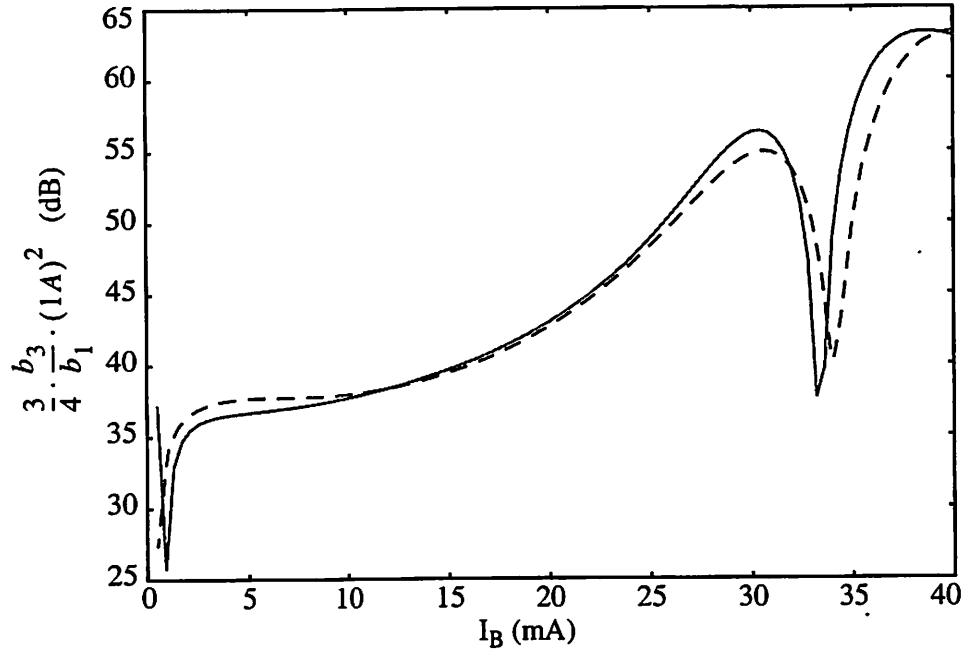
Figure 7.5: Comparison of prediction (solid line) and simulation (dashed line) using the BSIM3 version 3 model, of the low-frequency intermodulation versus bias current for a switching pair of the 0.25 μm technology and channel width 100 μm. The LO amplitude is 1 V.

where V_{co} is some LO cutoff voltage above which the conduction of one of the two devices is insignificant and $p_3(t)$ is zero. Simulation with sinusoidal LO waveform of amplitude V_o shows that indeed intermodulation asymptotically reduces proportionally to $1/V_o^2$ for high values of V_o while it drops at a higher rate for moderate values of V_o. For the rest of the chapter the LO waveform will be considered sinusoidal and V_o will denote its amplitude.

Fig. 7.5 and Fig. 7.6 show the quantity $0.75(b_3/b_1)(1\,A)^2$ (in dB, calculated as $20log_{10}(IM_3)$) versus bias current simulated with SpectreRF using the BSIM3 version 3 and BSIM3 version 2 model respectively, and also as obtained from the simple model. This quantity corresponds to the intermodulation value for I_s equal to 1A, assuming that the system remains weakly nonlinear. If for example I_s is 1 mA, one must subtract 120 dB from
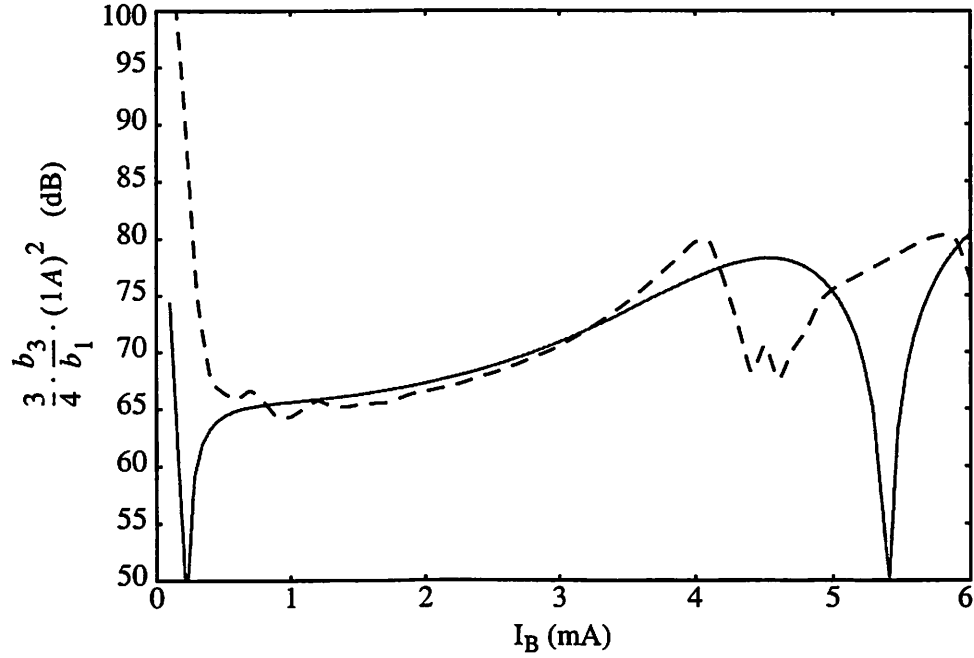
Figure 7.6: Comparison of prediction (solid line) and simulation (dashed line) using the BSIM3 version 2 model, of the low-frequency intermodulation versus bias current for a switching pair of the 0.8 μm technology and channel width 100 μm. The LO amplitude is 1 V.

the value read from these figures. The transistor width was in both cases equal to 100 μm and the LO amplitude was 1 V. The simulation result in Fig. 7.5 is a smooth curve, in very close agreement with the prediction of the simple model. However the BSIM3 version 2 model used in Fig. 7.6 is inappropriate for distortion simulation of the switching pair, as the discontinuities in the second and third derivatives of the I-V curve observed in Fig. 7.2 create large errors. Very high numerical accuracy is needed to reduce discontinuities in the intermodulation curve versus bias current and obtain the curve of Fig. 7.6, and even then the result shows large discrepancies from the simple model prediction. As we shall see, the latter agrees well with measurements.

The intermodulation prediction is largely insensitive to the value of parameter $\eta\phi_t$.
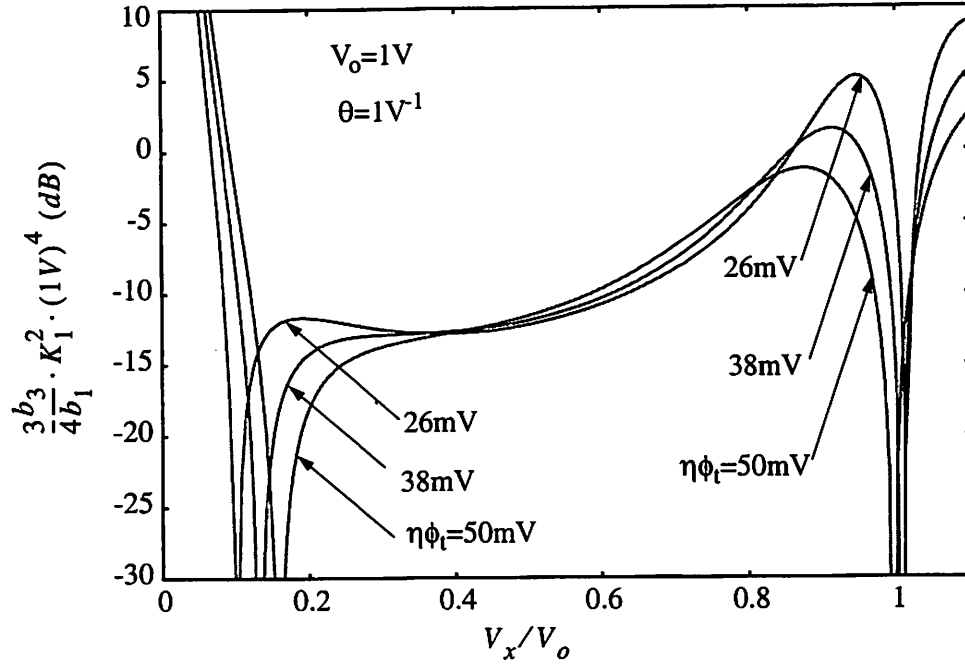
Figure 7.7: Low-frequency intermodulation of a switching pair for different values of the parameter $\eta\phi_t$. The intermodulation value is almost insensitive to this parameter.

Fig. 7.7 shows predicted intermodulation as a function of bias for three different values of $\eta\phi_t$, $26\,mV$, $38\,mV$, and $50\,mV$, versus bias. Instead of bias current we express here bias in terms of V_x/V_o where V_x is the LO voltage value sufficient to completely switch off one of the two devices (neglecting the subthreshold conduction[2]), and has been found in [89]

$$V_x = \frac{I_B\theta}{2K_1} + \sqrt{\left(\frac{I_B\theta}{2K_1}\right)^2 + \frac{I_B}{K_1}}.$$ (7.30)

The conversion gain of the switching pair has been found approximately equal to

$$b_1 \approx \frac{2}{\pi}\left(\frac{V_x/V_o}{Arcsin(V_x/V_o)}\right).$$ (7.31)

In [89] the subthreshold conduction was neglected and parameter normalization was used to express the performance of the switching pair as a function of fewer independent

[2]Voltage V_{co} is generally higher than V_x because of the subthreshold conduction
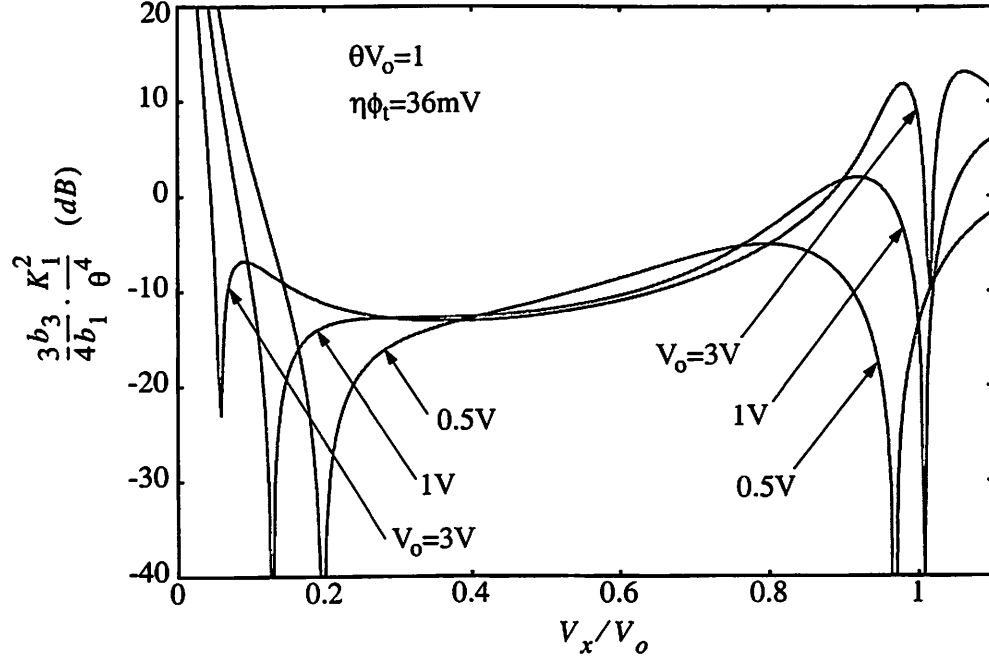
Figure 7.8: Normalized intermodulation for a fixed value of θV_o but three different values of V_o, 0.5 V, 1 V and 3 V.

parameters. For this purpose, all currents were multiplied with θ^2/K_1 and all voltages were multiplied with θ. Even though this normalization is not exact in the present analysis because the subthreshold region of operation is taken into account, we can conjecture that the intermodulation can be approximately expressed in terms of similarly normalized quantities. Evaluation of the intermodulation as described previously shows that indeed this is the case. Fig. 7.8 shows the value of the quantity $0.75(b_3/b_1)(K_1^2/\theta^4)$ versus V_x/V_o, for $\theta V_o = 1$ and for three different values of V_o. We observe that the three curves approximately coincide for moderate values of V_x/V_o which are most often used in practice, while they differ for very low and very high values of V_x/V_o. The agreement is better for higher LO amplitudes because these correspond to higher bias currents in this graph and the subthreshold region has a smaller effect. This observation allows us to give normalized intermodulation graphs.
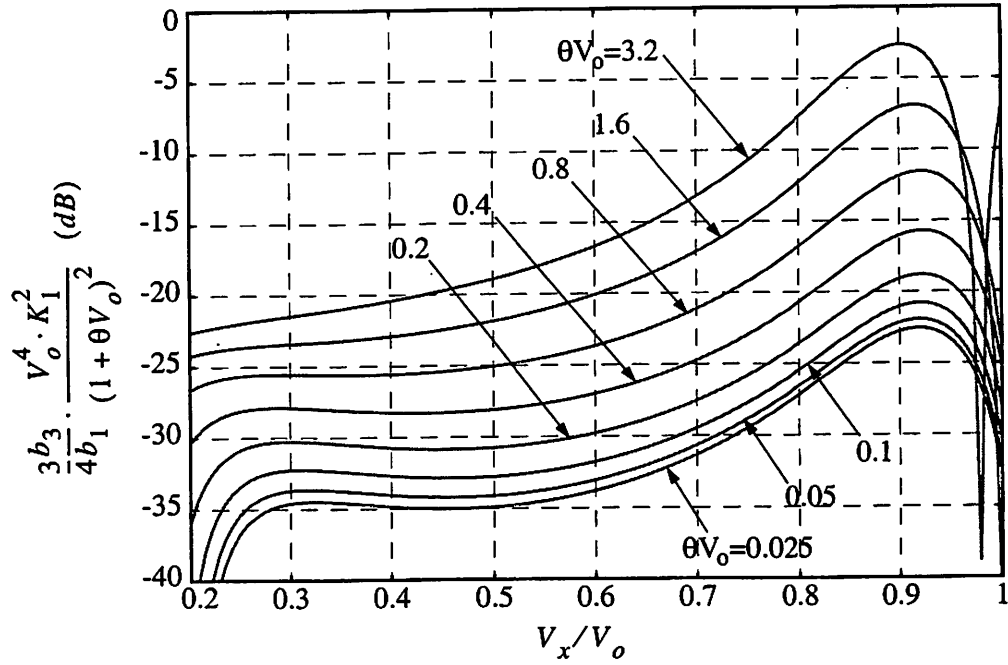
Figure 7.9: Normalized intermodulation versus V_x/V_o.

In order to reduce the range of the intermodulation values and improve the readability of the normalized graphs we express intermodulation in terms of the quantity

$$\frac{3}{4}\frac{b_3}{b_1}\frac{K_1^2}{\theta^4}\frac{(V_o\theta)^4}{(1+V_o\theta)^2}.$$

(7.32)

The result is shown in Fig. 7.9 which was generated with $\eta\phi_t = 36\,mV$ and $V_o = 1\,V$. The ratio b_3/b_1 can be calculated from this graph.

7.3.3 Cascading the Driver Stage and the Switching Pair

Let us assume now that the nonlinearity of the transconductance stage is described by a power series as follows

$$i_s = a_1 v_{in} + a_2 v_{in}^2 + a_3 v_{in}^3$$

(7.33)

where v_{in} is the input voltage. Cascading the power series of the transconductance stage with that of the switching pair, the output current can be related to the input voltage with a new time-varying power series. Substituting (7.33) in (7.20) we obtain

$$
\begin{aligned}
i_{o1} &= a_1 p_1(t) \cdot v_{in} \\
&+ (a_2 p_1(t) + a_1^2 p_2(t)) \cdot v_{in}^2 \\
&+ (a_3 p_1(t) + 2a_1 a_2 p_2(t) + a_1^3 p_3(t)) \cdot v_{in}^3.
\end{aligned}
\tag{7.34}
$$

Using the expansion of $p_1(t)$, $p_2(t)$ and $p_3(t)$ in a series of sinusoids as in (7.24) we obtain

$$
\begin{aligned}
i_{o1} &= \sum_{k=1}^{\infty} [a_1 p_{1,k} \cdot v_{in} \\
&+ (a_2 p_{1,k} + a_1^2 p_{2,k}) \cdot v_{in}^2 \\
&+ (a_3 p_{1,k} + 2a_1 a_2 p_{2,k} + a_1^3 p_{3,k}) \cdot v_{in}^3] \\
&\cdot \sin(2\pi k f_1).
\end{aligned}
\tag{7.35}
$$

If frequency translation by one LO multiple is of interest, the distortion performance can be described by a time-invariant power series

$$
i_{o1} = c_1 \cdot v_{in} + c_2 \cdot v_{in}^2 + c_3 \cdot v_{in}^3 + \ldots
\tag{7.36}
$$

where

$$
c_1 = a_1 b_1
\tag{7.37}
$$

$$
c_2 = a_2 b_1 + a_1^2 b_2
\tag{7.38}
$$

$$
c_3 = a_3 b_1 + 2a_1 a_2 b_2 + a_1^3 b_3.
\tag{7.39}
$$

Observe that these coefficients can be obtained directly by cascading the power series (7.33) and (7.25).

The total mixer third-order intermodulation is now given by

$$IM_3 = \frac{3}{4}\frac{c_3}{c_1}V_{in}^2 \approx \frac{3}{4}(\frac{a_3}{a_1}V_{in}^2 + \frac{b_3}{b_1}a_1^2V_{in}^2) \tag{7.40}$$

where in the last expression V_{in} is the amplitude of each input tone in the intermodulation test and the second term on the right side of (7.39) has been neglected as small. The total mixer intermodulation is approximately equal to the sum of the intermodulation values that the driver stage and the switching pair would generate if the other stage were ideal.

7.3.4 Distortion of Differential Versus Single-Ended Output

A differential output was considered above for the single-balanced mixer of Fig. 7.3 but it can be shown that if the LO waveform is symmetric around zero the distortion behavior is exactly identical if the output is taken single-ended. Assume that if I_1 is taken as output, the mixer distortion performance is described by a time-varying power series

$$i_1 = q_1(t) \cdot v_{in} + q_2(t) \cdot v_{in}^2 + q_3(t) \cdot v_{in}^3 + \ldots \tag{7.41}$$

and let us denote the Fourier coefficients of the waveform $q_k(t)$ by $q_{k,n}$. It is easy to see that if the LO waveform is symmetric the relevant power series for I_2 is

$$i_2 = q_1(t + \frac{T_{LO}}{2}) \cdot v_{in} + q_2(t + \frac{T_{LO}}{2}) \cdot v_{in}^2 + q_3(t + \frac{T_{LO}}{2}) \cdot v_{in}^3 + \ldots \tag{7.42}$$

and that the Fourier coefficients of the waveform $q_k(t+\frac{T}{2})$ are $(-1)^n q_{k,n}$, where for frequency translation by one LO multiple $n = 1$. The coefficients of the corresponding time-invariant power-series for differential output are twice these for single-ended output, and the generated distortion in the two cases is identical. A similar argument holds at high frequencies where Volterra series replace the power series. A similar approach shows that the distortion of
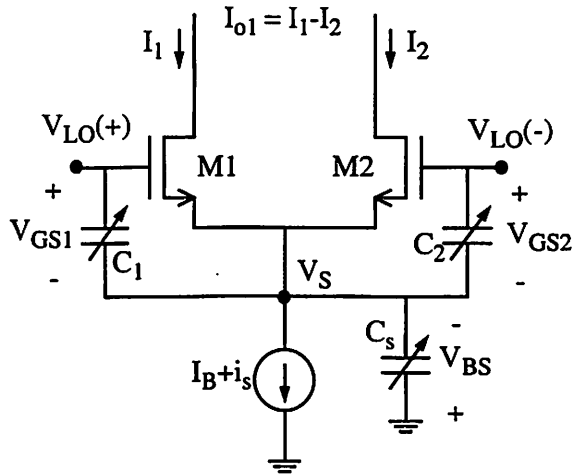
Figure 7.10: The switching pair as considered in the high frequency analysis.

the Gilbert cell is identical if single-ended or differential output is obtained. In fact for the Gilbert cell this statement can be shown true even if the LO waveform is not symmetric.

7.4 High Frequencies

7.4.1 Numerical Calculations

Time-varying Volterra series can be used at high frequencies to analyze the high-frequency intermodulation performance of the switching pair. We will consider now the effect of the gate-source capacitors of the transistors M1 and M2, C_1 and C_2 respectively, and the total capacitance from the common source node to ground C_b consisting of the source-body capacitance of M1 and M2 and the drain-body capacitance of M3, as shown in Fig 7.10. We will neglect the gate-drain capacitances and we will assume that the LO voltage is perfectly sinusoidal, generated by an ideal voltage source while the common LO voltage is a constant. Denoting again the I-V relation of a transistor with $I = f(V)$, with

no small signal present, the equation describing the switching pair at high frequencies is

$$I_B = f(V_1) + f(V_2) + \frac{d}{dt}Q_1 + \frac{d}{dt}Q_2 + \frac{d}{dt}Q_s \qquad (7.43)$$

where Q_1, Q_2, Q_s are the charges of the capacitors C_1, C_2, C_b respectively, and $V_1 = V_{GS1} - V_T$, $V_2 = V_{GS2} - V_T$ as defined before. The capacitances satisfy

$$C_1(V_1) = \frac{dQ_1}{dV_{GS1}} \qquad (7.44)$$

$$C_2(V_2) = \frac{dQ_2}{dV_{GS2}} \qquad (7.45)$$

$$C_b(V_{BS}) = \frac{dQ_s}{dV_{BS}} \qquad (7.46)$$

where V_{BS} is the body-source voltage of M1 and M2. Capacitance C_b is the sum of an area and a sidewall junction capacitance and each of its components is given by an expression of the form

$$C_j = \frac{C_{jo}}{\left(1 - \frac{V_{BS}}{\phi_j}\right)^{m_j}} \qquad (7.47)$$

where the symbols C_{jo}, ϕ_j and m_j have the usual meaning. Capacitances C_1 and C_2 are dependent on the region of operation and will be approximated with the expression

$$C_1(V_1) = Cgs_{ol} + \frac{Cgs_{sat} - Cgs_{ol}}{1 + \exp^{\left(-\frac{V_1}{n\phi_t}\right)}} \qquad (7.48)$$

which provides a smooth transition from the overlap capacitance in subthreshold Cgs_{ol} to the capacitance value in saturation Cgs_{sat}. From (7.47) and (7.48), analytical expressions for the derivatives of the capacitances with respect to their voltage arguments can be obtained. From (7.43), (7.44)-(7.46) and expressing V_{GS1}, V_{GS2} and V_{BS} in terms of $V_{LO}(t)$ and the common source potential V_S, we obtain

$$I_B \;=\; f(V_1) + f(V_2)$$

$$+ \frac{1}{2}(C_1 - C_2)\frac{dV_{LO}}{dt} - (C_1 + C_2 + C_b)\frac{dV_S}{dt} \qquad (7.49)$$

Substituting

$$V_1 = V_{LO,c} + \frac{V_{LO}}{2} - V_S - V_T \qquad (7.50)$$

$$V_2 = V_{LO,c} - \frac{V_{LO}}{2} - V_S - V_T \qquad (7.51)$$

where $V_{LO,c}$ is the common LO voltage considered constant, (7.49) becomes a nonlinear differential equation with periodic boundary conditions which must be solved to find the periodic steady-state operating point of the devices. This was accomplished by discretizing (7.49), and solving for the vector of the values of V_S over one period, using a Newton-Raphson method as described in [87]. A software package for sparse matrix manipulation [44] was employed. The body effect was neglected and the threshold voltage was considered a constant. We will adopt this approximation for the rest of the analysis, and will comment on the role of the body effect later.

When a small-signal current $i_s(t)$ is present at the transconductance stage output, voltages V_1, V_2 and V_{BS} will change to $V_1 + v$, $V_2 + v$ and $V_{BS} + v$ respectively. Taking a third-order Taylor expansion of (7.43) and removing the large-signal part of the equation we obtain

$$
\begin{aligned}
i_s &= Gv + G_v v^2 + G_{vv} v^3 \\
&\quad + \frac{d}{dt}[Cv + C_v v^2 + C_{vv} v^3]
\end{aligned} \qquad (7.52)
$$

where

$$G = f_{1V} + f_{2V} \qquad (7.53)$$

$$G_v = \frac{1}{2}(f_{1VV} + f_{2VV}) \qquad (7.54)$$

$$G_{vv} = \frac{1}{6}(f_1 vvv + f_2 vvv) \tag{7.55}$$

$$C = C_1 + C_2 + C_b \tag{7.56}$$

$$C_v = \frac{1}{2}(C_{1V} + C_{2V} + C_{bV}) \tag{7.57}$$

$$C_{vv} = \frac{1}{6}(C_{1VV} + C_{2VV} + C_{bVV}) \tag{7.58}$$

and C_{1V}, C_{1VV}, C_{2V}, C_{2VV}, C_{bV}, and C_{bVV}, denote the first and second derivatives of the capacitances C_1, C_2 and C_b, with respect to their voltage arguments in (7.44)-(7.46). The incremental voltage v is related to the incremental current i_s with a time-varying Volterra series:

$$v = H_1(t, f_a) \circ i_s + H_2(t, f_a, f_b) \circ i_s^2 + H_3(t, f_a, f_b, f_c) \circ i_s^3 + \ldots \tag{7.59}$$

Substituting (7.59) into (7.52), equating terms of similar power of i_s, and using the usual notation for i_s as a sum of sinusoids [104] we obtain:

$$G(t)H_1(t, f_a)e^{j2\pi f_a t} + \frac{d}{dt}[C(t)H_1(t, f_a)e^{j2\pi f_a t}] = e^{j2\pi f_a t} \tag{7.60}$$

$$G(t)H_2(t, f_a, f_b)e^{j2\pi(f_a+f_b)t} + \frac{d}{dt}[C(t)H_2(t, f_a, f_b)e^{j2\pi(f_a+f_b)t}] =$$
$$-G_v(t)H_1(t, f_a)H_1(t, f_b)e^{j2\pi(f_a+f_b)t} - \frac{d}{dt}[C_v(t)H_1(t, f_a)H_1(t, f_b)e^{j2\pi(f_a+f_b)t}] \tag{7.61}$$

$$G(t)H_3(t, f_a, f_b, f_c)e^{j2\pi(f_a+f_b+f_c)t} + \frac{d}{dt}[C(t)H_3(t, f_a, f_b, f_c)e^{j2\pi(f_a+f_b+f_c)t}] =$$
$$-[(2G_v(t)\overline{H_1(t, f_a)H_2(t, f_a, f_b)} + G_{vv}(t)H_1(t, f_a)H_1(t, f_b)H_1(t, f_c))e^{j2\pi(f_a+f_b+f_c)t}]$$
$$-\frac{d}{dt}[(2C_v(t)\overline{H_1(t, f_a)H_2(t, f_a, f_b)} + C_{vv}(t)H_1(t, f_a)H_1(t, f_b)H_1(t, f_c))e^{j2\pi(f_a+f_b+f_c)t}] \tag{7.62}$$

where the bar above certain terms denotes as usual the average over all the terms that result from all possible permutations of the frequency arguments [104]. These are linear

differential equations with periodic boundary conditions and were solved as described in [86] for the case of a periodic AC analysis, by discretizing them and solving one algebraic sparse linear system of equations. Once the time-varying Volterra series that relates i_s with v is known (coefficients H_1, H_2 and H_3), it can be cascaded with the time-varying power series which relates v with the output current i_{o1}, whose coefficients are

$$d_1(t) = f_{1V}(t) - f_{2V}(t) \tag{7.63}$$

$$d_2(t) = \frac{1}{2}\left(f_{1VV}(t) - f_{2VV}(t)\right) \tag{7.64}$$

$$d_3(t) = \frac{1}{6}\left(f_{1VVV}(t) - f_{2VVV}(t)\right) \tag{7.65}$$

in order to relate i_s with i_{o1} as follows:

$$i_{o1} = P_1(t, f_a) \circ i_s + P_2(t, f_a, f_b) \circ i_s^2 + P_3(t, f_a, f_b, f_c) \circ i_s^3 + \ldots \tag{7.66}$$

Above

$$P_1(t, f_a) = d_1(t)H_1(t, f_a) \tag{7.67}$$

$$P_2(t, f_a, f_b) = d_1(t)H_2(t, f_a, f_b) + d_2(t)H_1(t, f_a)H_1(t, f_b) \tag{7.68}$$

$$
\begin{aligned}
P_3(t, f_a, f_b, f_c) \;=\; & d_1(t)H_3(t, f_a, f_b, f_c) \\
& + 2d_2(t)\overline{H_1(t, f_a)H_2(t, f_b, f_c)} \\
& + d_3(t)H_1(t, f_a)H_1(t, f_b)H_1(t, f_c)
\end{aligned} \tag{7.69}
$$

As discussed in section 2.10.3, similarly with low frequencies, time-invariant Volterra coefficients $B_1(f_a)$, $B_2(f_a, f_b)$, and $B_3(f_a, f_b, f_c)$ that describe intermodulation in the frequency band of interest can be extracted by taking the first Fourier coefficients of $P_1(t, f_a)$, $P_2(t, f_a, f_b)$, and $P_3(t, f_a, f_b, f_c)$ respectively.

If i_s consists of two tones of equal magnitude as in (7.27) where $f_1 \approx f_2 \approx f_s$ and f_s is the input signal frequency, the third-order intermodulation generated by the switching pair can now be calculated as

$$IM_3 = \frac{3}{4} \cdot \left| \frac{B_3}{B_1} \right| \cdot I_s^2 \qquad (7.70)$$

where we have used the notation $B_1 = B_1(f_s)$ and $B_3 = B_3(f_s, f_s, -f_s)$. It is easy to see that to evaluate these coefficients we need $H_1(t, f_s)$, $H_2(t, f_s, f_s)$, $H_2(t, f_s, -f_s)$, and $H_3(t, f_s, f_s, -f_s)$. Hence we must solve (7.60) once, (7.61) twice and (7.62) once.

The total mixer distortion can be found by cascading the power series or Volterra series that describes the transconductance stage with the time-invariant Volterra series derived for the switching pair. Assuming that the Volterra series that describes the transconductance stage is

$$i_s = A_1(f_a) \circ v_{in} + A_2(f_a, f_b) \circ v_{in}^2 + A_3(f_a, f_b, f_c) \circ v_{in}^3 \qquad (7.71)$$

the Volterra coefficients that describe the total mixer distortion are

$$C_1(f_a) = B_1(f_a)A_1(f_a) \qquad (7.72)$$

$$C_2(f_a, f_b) = B_1(f_a + f_b)A_2(f_a, f_b) + B_2(f_a, f_b)A_1(f_a)A_1(f_b) \qquad (7.73)$$

$$C_3(f_a, f_b, f_c) = B_1(f_a + f_b + f_c)A_3(f_a, f_b, f_c) \qquad (7.74)$$

$$+ 2\overline{B_2(f_a, f_b + f_c)A_1(f_a)A_2(f_b, f_c)} \qquad (7.75)$$

$$+ B_3(f_a, f_b, f_c)A_1(f_a)A_1(f_b)A_1(f_c) \qquad (7.76)$$

To calculate $C_3(f_1, f_1, -f_2)$, besides $B_1(f_s)$ and $B_3(f_s, f_s, -f_s)$ we need $B_2(f_s, 0)$ and $B_2(-f_s, 2f_s)$ which implies that $H_2(f_s, 0)$, $H_2(-f_s, 2f_s)$, $H_1(0)$ and $H_1(2f_s)$ must also be calculated. Approximately, the second-order interaction can be neglected and the total mixer distortion
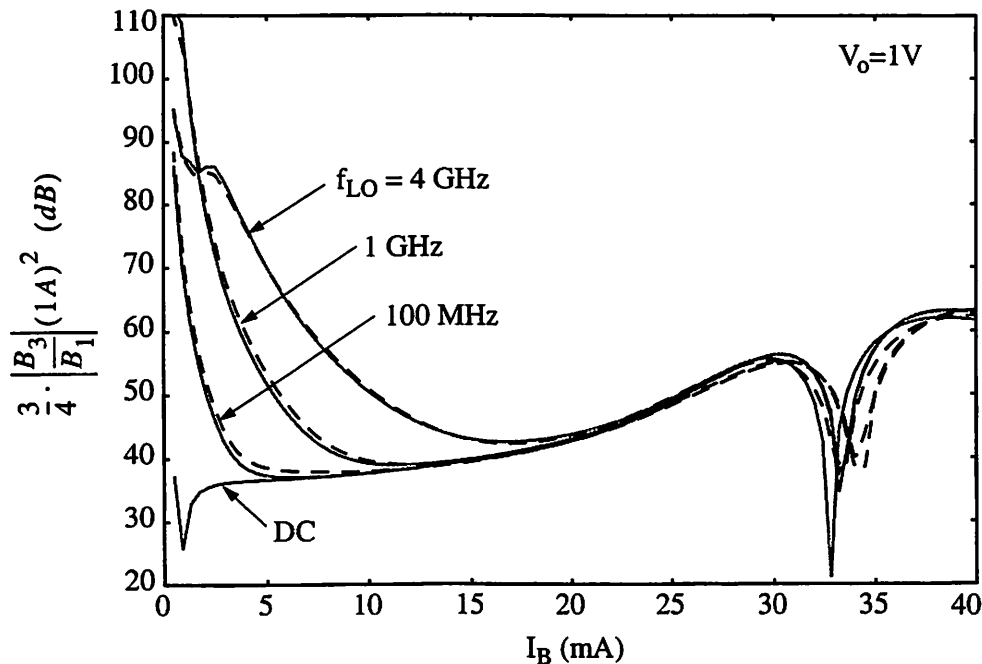
Figure 7.11: High-frequency intermodulation prediction (solid line) and simulation with spectreRF (dashed line) for a switching pair operating as a downconverter, versus bias current and for several LO frequencies. The model corresponds to a 0.25 μm technology, the channel width is 100 μm, and the the LO amplitude is 1 V.

can be given as the sum of the intermodulation generated by the transconductance stage and the switching pair separately, in a similar fashion to equation (7.40) for the low frequency case. It is worth noting however in the example of Fig. 7.20 below that when the high-frequency switching-pair nonlinearity dominates the mixer distortion, the interaction between the two stages partially improves linearity and the total mixer intermodulation is lower than that of the switching pair alone. This behavior is not observed at low frequencies.

7.4.2 Results and Comments

Fig. 7.11 shows intermodulation for downconversion operation, as predicted with the above method and as simulated with SpectreRF, versus bias current and for several
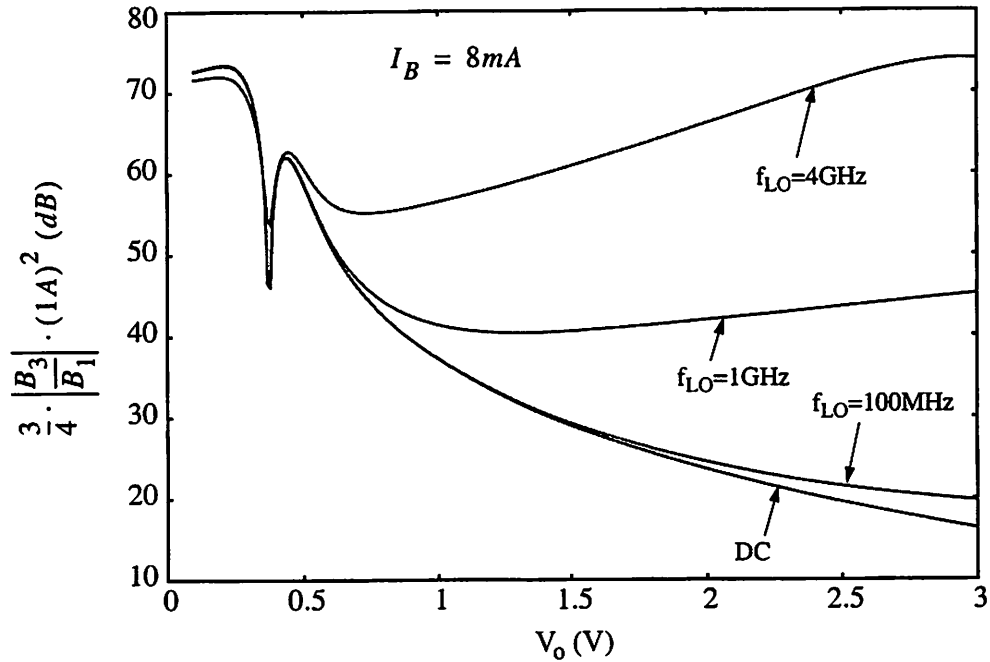
Figure 7.12: High-frequency intermodulation versus LO amplitude, for a fixed bias current and several LO frequencies.

values of the LO frequency. The quarter-micron technology whose BSIM3 version 3 model is available was used, the channel width was 100 μm and the LO amplitude was 1 V. Similar simulation with the BSIM3 version 2 model results in a high frequency intermodulation curve with large discontinuities, caused by discontinuities in the derivatives of the gate-source capacitance of this model.

Fig. 7.12 shows intermodulation for the same switching pair of the same 0.25 μm technology, performing downconversion as a function of the LO amplitude for a fixed bias current $I_B = 8\,mA$ and several LO frequencies. We observe that contrary to the low-frequency case where the intermodulation improves monotonically as the LO amplitude increases, at high frequencies there exists an optimal value after which the intermodulation increases. The same behavior is observed for upconversion and has also been reported for
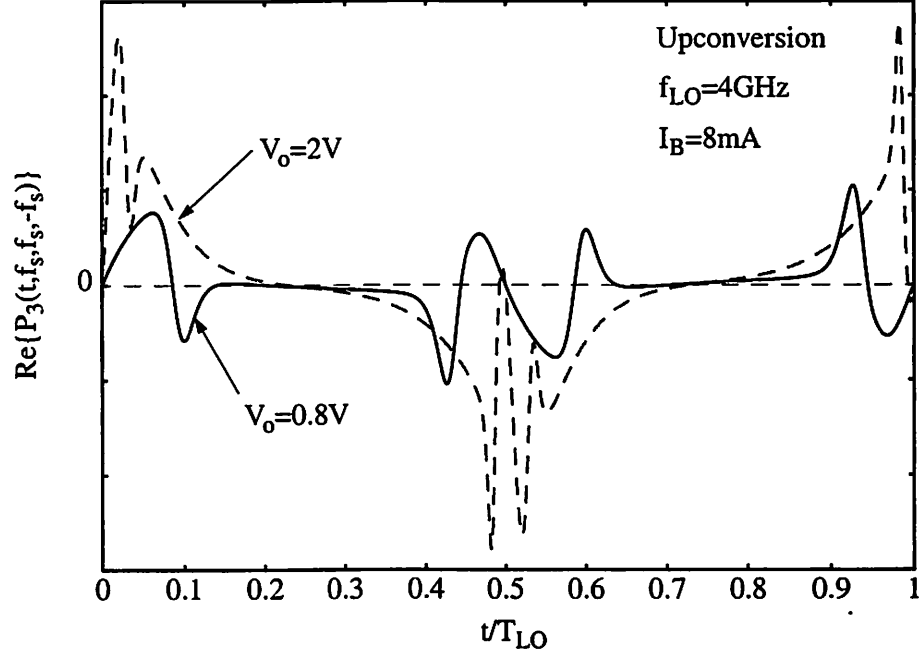
Figure 7.13: The third-order time-varying Volterra coefficient $P_3(t, f_s, f_s, -f_s)$ for fixed LO frequency $f_{LO} = 4\,GHz$, bias current $I_B = 8\,mA$, and two different LO amplitudes $V_o = 0.8\,V$ and $2\,V$.

bipolar transistor switching pairs in [56] and [31]. Because of the higher voltage swing of the common-source node when the LO amplitude is high, higher current is injected by the parasitic capacitances which accentuates the high frequency phenomena and alters the periodic operating point of the devices. Fig. 7.13 shows the real part of the third order time-varying Volterra coefficient $P_3(t, f_s, f_s, -f_s)$ of an upconverter over one LO period, for two different values of the LO amplitude $0.8\,V$ and $2\,V$, and for the same bias current $8\,mA$ and LO frequency $4\,GHz$ (for upconversion the complex exponentials of (7.60)-(7.62) approach one and the time-varying Volterra coefficients are mostly real).

Let us qualitatively comment on the role of the body effect. We have already established in section 7.3.1 that at low frequencies the behavior of the switching pair is largely insensitive to the body effect. Observe that in the I-V relation of the transistor, the sum

$V_S + V_T$ appears. Whenever because of a change in the LO voltage or the transconductance stage current a change in $V_S + V_T$ occurs, the change in V_S is smaller in the presence of body effect than in the absence, because part of the variation in $V_S + V_T$ is contributed by the change in V_T. Reduced V_S swing means that at high frequencies the body effect effectively reduces the value of the capacitance connected to the common-source node approximately by the quantity $1/(1 + b)$ where b is the ratio of the small-signal body transconductance over the small-signal gate transconductance (although b depends on V_S, for simplicity it is considered here a constant). A typical value for the quantity $1/(1 + b)$ is 0.9 and this modification to the capacitance value causes only a minor change to the distortion prediction. However, it has been taken into account in the predicted curves shown in this chapter.

Finally let us comment on some assumptions adopted about the local oscillator. The presence of a time-varying common LO voltage results in an additional voltage swing of the common source node and enhances the reactive effects which appear at lower frequencies. This was also verified experimentally. The assumption about the approximately sinusoidal shape of the LO voltage waveform is usually realistic at high frequencies. In the case that a tuned load LC buffer provides the LO signal, the LO waveform remains approximately sinusoidal as long as the load capacitance is not dominated by the time-varying gate-source capacitances. The presence of an output impedance of the local oscillator causes small error in the predicted intermodulation value if the actual applied LO waveform amplitude is considered in the prediction.

7.4.3 High-Frequency Intermodulation in terms of Normalized Parameters.

The calculation of the switching-pair intermodulation at high frequencies is very fast when compared with the performance of a circuit simulator. It requires, however, the use of numerical methods and is not easily applicable by a designer. For this reason we will attempt to capture the intermodulation performance of the switching pair in graphs of normalized variables.

Having the ability to rapidly evaluate the intermodulation of the switching pair, we can experiment with the related parameters. Neglecting the terms involving derivatives of the capacitances in (7.60)-(7.62) does not appreciably change the prediction, while replacing all the time-varying capacitors by one of constant value C_{tot} from the common-source node to ground causes only a small inaccuracy. Since for the largest part of the LO period one of M1 and M2 is cut-off, a reasonable value for C_{tot} is the sum of the total junction capacitance to ground, the gate-source capacitance of one of the two transistors in saturation, and the gate-source overlap capacitance of the other transistor. To generate the following graphs of normalized variables we will make the arbitrary but better than the constant capacitance approximation that $0.25C_{tot}$ is a gate-source capacitance dependent on the transistor region of operation as in (7.48) and the rest $0.75C_{tot}$ is a constant capacitance to ground. Fig. 7.14 shows intermodulation of a switching pair operating as a downconverter versus LO frequency, for a fixed capacitor connected to the common source node, for the simple model that arbitrarily breaks down C_{tot} as just described, and for the more complete model of voltage-dependent capacitors described in section 7.4.1. In this
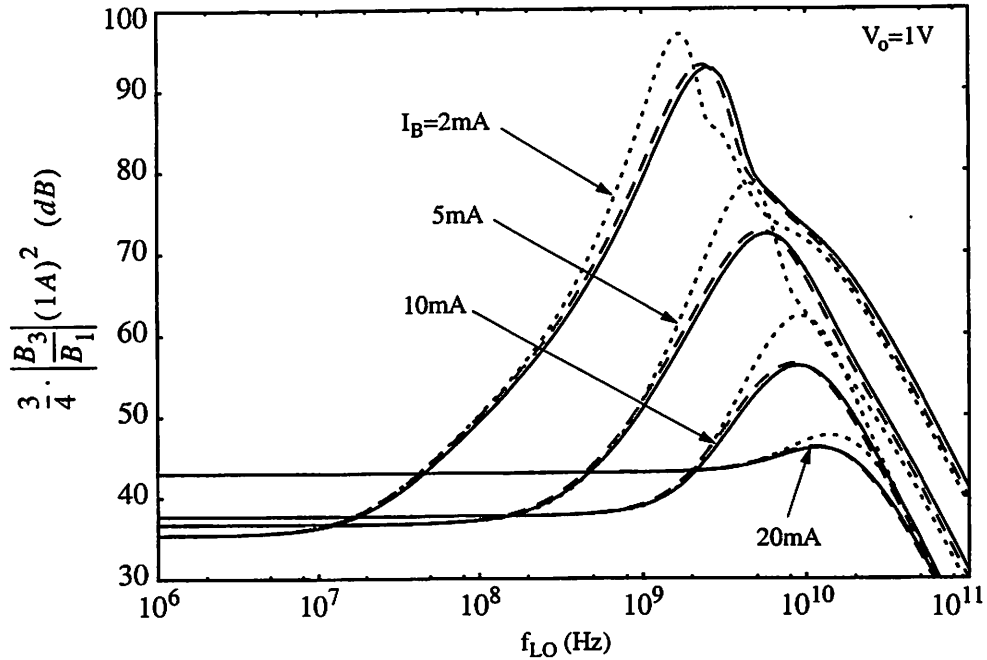
Figure 7.14: Comparison of the-high frequency intermodulation prediction using realistic voltage-dependent capacitor models (solid line), a fixed-value capacitor connected from the common-source node to ground (short dashed line), and the simple approximation in which $0.25C_{tot}$ is a gate-source capacitor and $0.75C_{tot}$ is a fixed capacitor to ground (long dashed line).

simulation the quarter-micron technology was used, the channel width was 100 μm, the LO amplitude was 1V, and several values of the bias current are shown. It is worth noticing in this graph that for relatively high bias current the high frequency deterioration up to very high frequencies is almost negligible.

Fig. 7.15 shows intermodulation of the switching pair operating as a downconverter versus bias current for a fixed LO amplitude of $1V$, for several values of the parameter θ, and for several values of the parameter $C_{tot}f_{LO}$, including the one that corresponds to DC. We observe that the intermodulation at any given frequency can be viewed as the sum of its value at low frequencies and a high-frequency component. In addition, the high-frequency component does not significantly depend on parameter θ, and can be

devices and neglecting the subthreshold conduction, we obtain the equations that determine

where $V_{LO,c}$ is the common LO voltage. Using the latter in (7.49), assuming square-law

$$V_S = V_{LO,c} - \frac{V_1 + V_2}{2} - V_r \qquad (7.77)$$

the mixer behavior. From (7.50) and (7.51) the common source potential equals

we can assert that the subthreshold conduction of the devices does not significantly affect

For reasonably high LO amplitudes, and when capacitive effects are significant,

approximately 6 dB.

For $C_{tot}f_{LO}/K_1 = 3 \cdot 10^{-2} V$ the curve for $\theta = 0$ stands higher than that for $\theta = 3 V^{-1}$ by

devices. For very high frequencies however, the curves for different θ values start to deviate.

approximately calculated from the particular case of $\theta = 0$ which corresponds to square-law

parameters $C_{tot}f_{LO}$ and for θ and for $V_o = 1V$.

Figure 7.15: High-frequency intermodulation versus bias current for several values of the
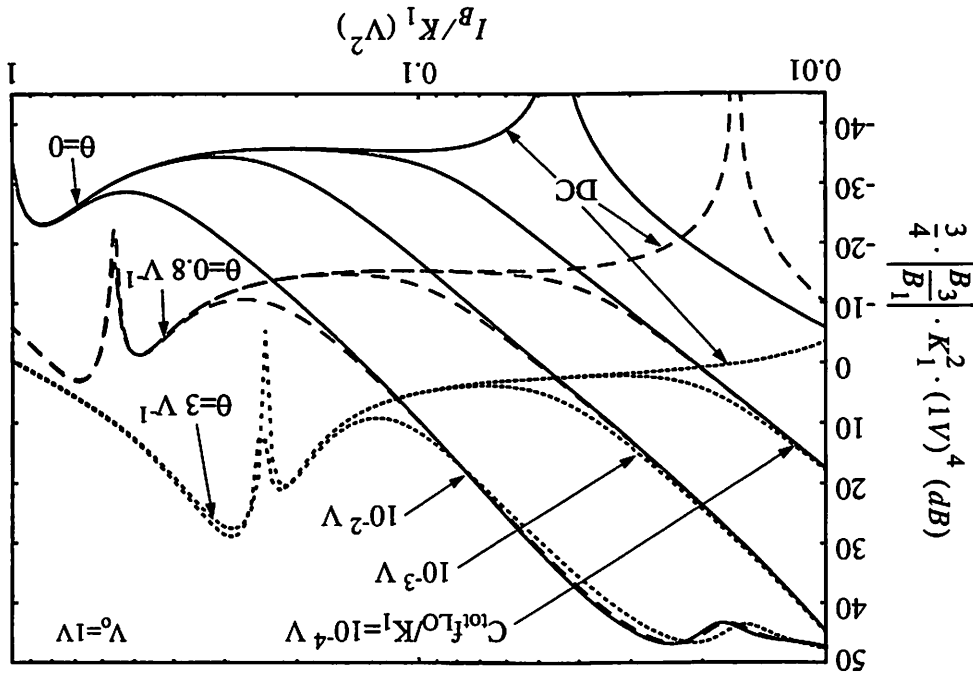
the high-frequency intermodulation

$$I_B = K_1 u(V_1)^2 + K_1 u(V_2)^2$$

$$+ (C_1 - C_2)V_o \pi f_{LO} \cos(2\pi f_{LO} t)$$

$$+ \frac{(C_1 + C_2 + C_b)}{2} \frac{d}{dt}(V_1 + V_2) \tag{7.78}$$

$$V_o \sin(2\pi f_{LO} t) = V_1 - V_2 \tag{7.79}$$

where $u(x)$ is the step function which equals its argument if it is positive and is zero otherwise. Equivalently

$$\frac{I_B}{K_1 V_o^2} = u\left(\frac{V_1}{V_o}\right)^2 + u\left(\frac{V_2}{V_o}\right)^2$$

$$+ \frac{(C_1 - C_2)f_{LO}\pi \cos(2\pi t')}{K_1 V_o}$$

$$+ \frac{(C_1 + C_2 + C_b)f_{LO}}{2K_1 V_o} \frac{d}{dt'}\left(\frac{V_1}{V_o} + \frac{V_2}{V_o}\right) \tag{7.80}$$

$$\sin(2\pi t') = \frac{V_1}{V_o} - \frac{V_2}{V_o} \tag{7.81}$$

where the normalized time variable $t' = tf_{LO}$ has also been introduced. It is now apparent that the high-frequency part of the intermodulation can approximately be expressed in terms of two parameters, $Z = I_B/(K_1 V_o^2)$ and $Y = C_{tot}f_{LO}/(V_o K_1)$. Fig.7.16 shows normalized intermodulation $3/4 (B_3/B_1) K_1^2 V_o^4$ for downconversion operation in terms of the two parameters. These particular curves were obtained for LO amplitude 1 V, but close agreement is observed if the calculations are repeated with a different one. The same normalization can be shown to be valid for an upconverter. Letting $f_s \simeq 0$ we obtain the corresponding graphs shown in Fig. 7.17. In these normalized graphs the body effect is neglected, but can be approximately accounted for by reducing the capacitance value as discussed previously in section 7.4.2.
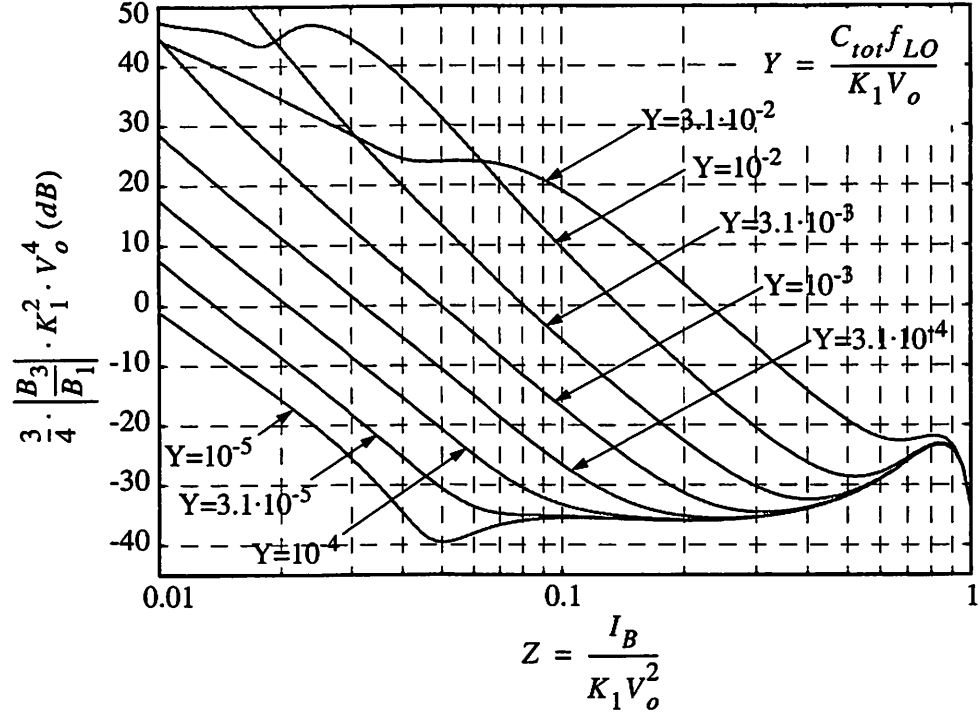
Figure 7.16: Normalized high-frequency intermodulation for a downconverter and square-law devices.

7.5 Measurements

To experimentally verify the validity of our results, third-order input intercept point (IIP_3) measurements were taken from a single-balanced active CMOS mixer fabricated in the Philips Qubic2 process. The measurement setup is shown in Fig. 7.18. The LO frequency was $375\,MHz$, the input signal consisted of two tones around $395\,MHz$ spaced $60\,KHz$ apart, and the output was obtained at $20\,MHz$. The input was resistively matched to 50Ω.

The width of the transconductance stage transistor was $100\,\mu m$, that of the switching pair devices was $200\,\mu m$ while minimum channel length $0.8\,\mu m$ was used for all the transistors. An I-V curve was obtained using the available SPICE model, and the param-
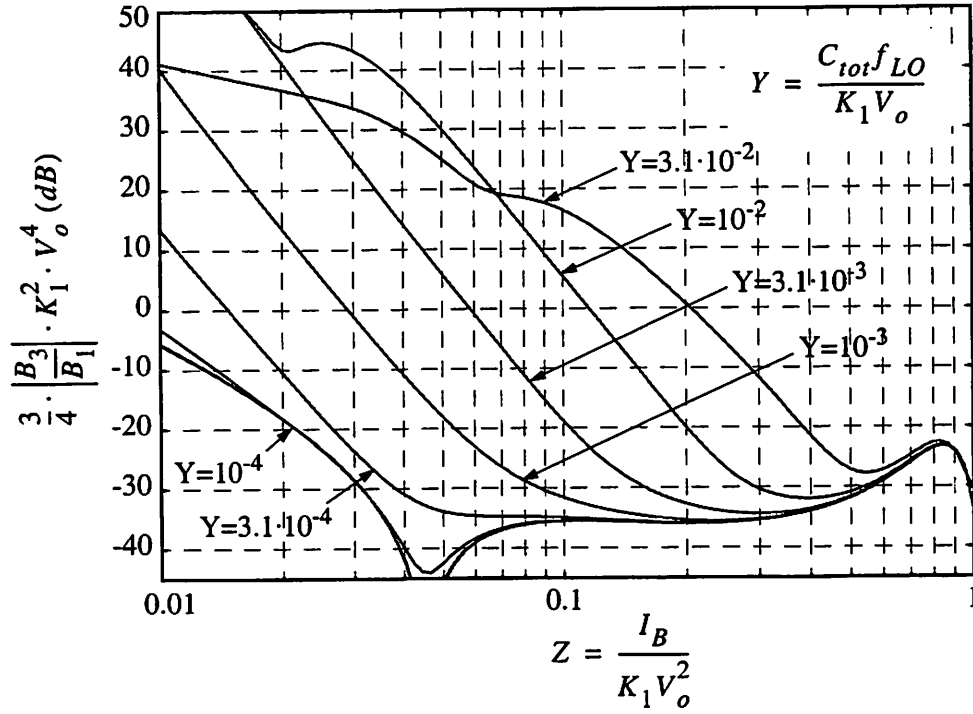
Figure 7.17: Normalized high-frequency intermodulation for an upconverter and square-law devices.

eters $\theta = 0.94\,V^{-1}$, $K = 10.4\,mA/V^2$, $\eta\phi_t = 44\,mV$ were extracted with curve fitting for a $100\,\mu m$ wide device. The capacitances were also estimated from the available SPICE model and provided a total effective capacitance of approximately $0.8\,pF$, connected to the common-source node. For the total mixer intermodulation prediction, the nonlinearity of the transconductance stage was provided by a power series which can be easily derived from the expressions given in the appendix.

Fig. 7.19 shows measurements and prediction of IIP_3 versus bias current for a fixed LO amplitude of $1\,V$. Very good agreement is observed. Also shown are the individual contributions of the transconductance stage and the switching pair, as well as the contribution that the switching pair would have at low frequencies. The switching pair
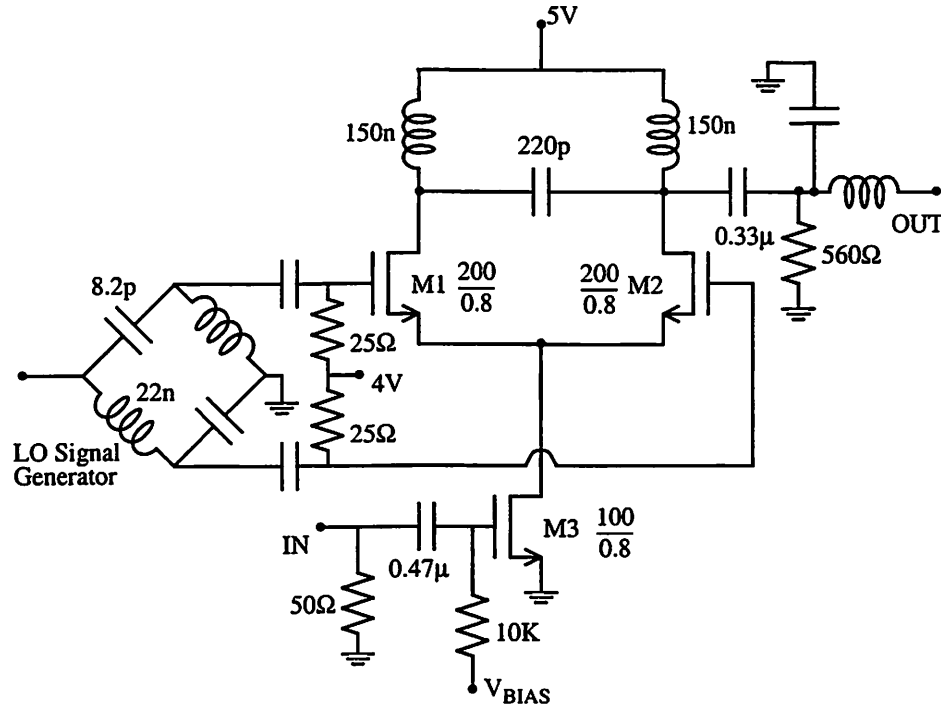
Figure 7.18: Measurement setup.

nonlinearity dominates at high bias current. At low bias current, where the switching-pair performance deteriorates compared to DC, the performance of the transconductance stage is also poor, and as a result the total mixer intermodulation prediction is almost identical with that at low frequencies.

The high frequency effects are better demonstrated in Fig. 7.20 where the bias current is fixed at the low value of $1.5\,mA$ and the LO amplitude is swept from $0.5\,V$ to $2\,V$. The individual contributions of the transconductance stage and the switching pair, together with the total mixer IIP_3 at low frequencies, are shown. Clearly the large LO amplitude causes high-frequency deterioration. Again very good agreement between prediction and measurement is observed.
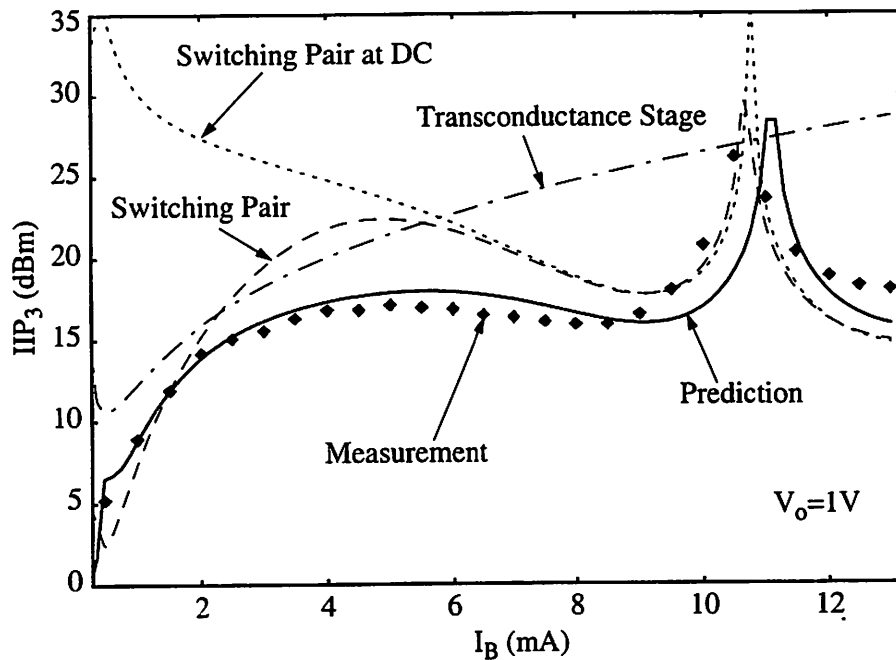
Figure 7.19: Intermodulation measurements versus bias current for a fixed LO amplitude $V_o = 1\,V$.

7.6 Conclusions

A nonlinearity analysis of the CMOS transistor switching pair has been performed. We demonstrated that in the frequency band of interest its nonlinearity can be accurately described by a regular power series, or Volterra series at high frequencies, and we described how the coefficients of these series can be calculated. As a result of our analysis we produced normalized graphs from which the active-mixer intermodulation can be predicted for any technology parameters and operating conditions. Using these, the designer can rapidly estimate the suitability of a given CMOS process for a given set of mixer specifications, and can accelerate the design cycle. Several useful results were derived in the course of the analysis. The importance of a physical CMOS transistor model, describing weak moderate and strong inversion with a single analytical equation was demonstrated and a simple
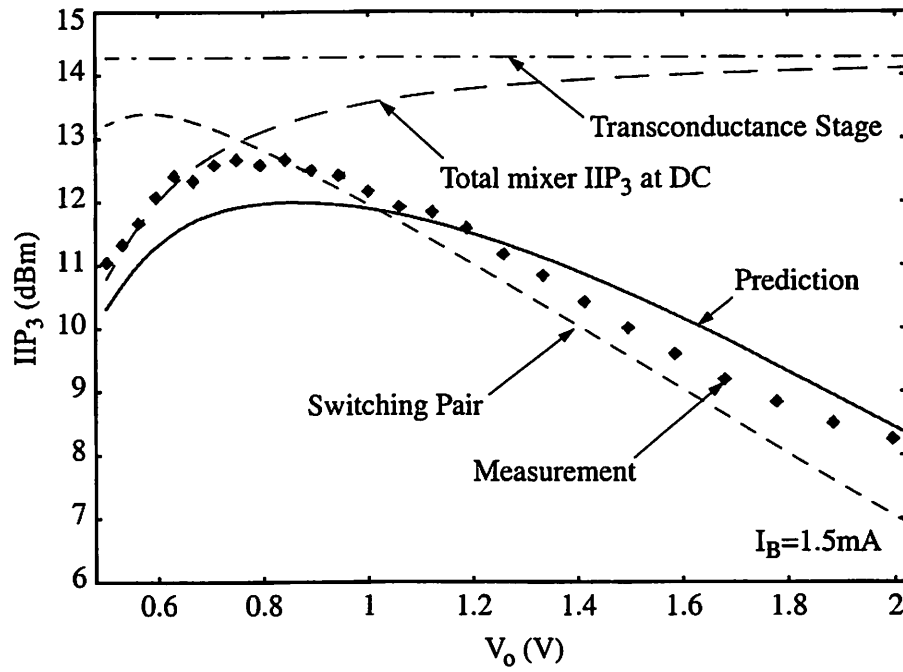
Figure 7.20: Intermodulation measurements versus LO amplitude for a fixed bias current $I_B = 1.5\,mA$.

appropriate model was discussed and used.

Part III

Application

Chapter 8

Design Examples

8.1 Introduction

This chapter describes some examples of single-balanced downconversion RF mixer designs. Although double-balanced mixers are more commonly used in integrated systems, these designs are sufficient to demonstrate the use of the theoretical results of the previous chapters. The available process is standard CMOS of minimum drawn gate length $0.24\mu m$, with two polysilicon layers.

8.2 Design Topology

The topology of the designs is shown in Fig.8.1. Care was taken to make this design flexible during testing. The mixer core employs an inductively degenerated common-source transconductance stage. The degeneration inductance L_s is implemented with the bond wire and the pin of the TSSOP package and is expected to be approximately $3nH$.
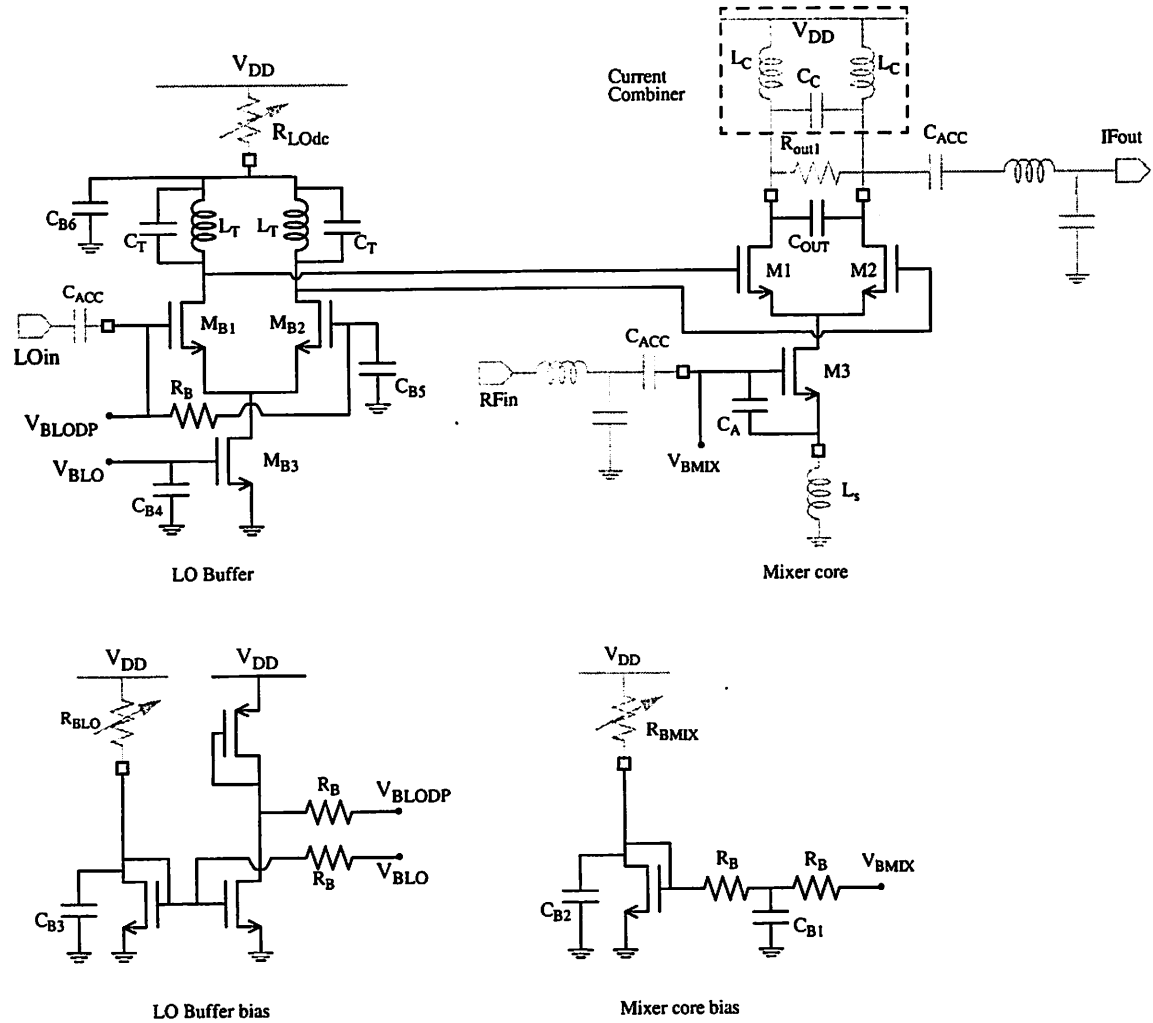
Figure 8.1: The topology of the implemented designs

Capacitor C_A employed in some of the designs is connected between the gate and the source

of the transconductance stage transistor to reduce the real and the imaginary part of the

impedance looking towards the gate of the transistor and therefore facilitate input matching.

The input and output matching networks and the ac coupling capacitors are external.

The current combiner described in chapter 3, section 3.2.2 is used at the output, also

implemented with discrete components. The capacitance of the current combiner $C_C + C_{OUT}$

shorts the LO feedthrough at the output. Since the bond wire inductance represents a significant impedance at the LO frequency which can give a significant voltage swing, part of the capacitance of the combiner $C_{OUT} = 6pF$ is implemented on chip. Resistor R_{out1} is intended to reduce the output impedance of the mixer, to reduce possible excess gain and reduce the nonlinearity generated by the nonlinear output resistance of the transistors of the mixer switching pair.

The LO buffer transforms the external single-ended LO waveform to differential. It employs a differential pair, one side of which is ac grounded and the other side ac connected to the external LO signal. It utilizes on chip tuned LC tanks. The resistance R_{LOdc} which sets the DC level of the LO waveforms supplied to the mixer is external. The bypass capacitor C_{B6} is used to reduce the common mode LO voltage. No matching network is used at the LO port since there is no matching requirement and, as we shall see, the necessary internal LO amplitude can be obtained without impedance matching at this port.

Independent bias networks for the mixer core and the LO buffer allow us to independently set the bias currents in these blocks. The dc level of the gates of the mixer switching pair can be adjusted though R_{LOdc}, and the dc level of the gates of the LO buffer differential pair can be set independently since V_{DD} is connected on chip only to the bias circuit of the LO. The bias resistors R_B are implemented with diffusion and all the on-chip capacitors are implemented with double polysilicon. The capacitors whose quality factors are critical are partitioned to small units such that the Q remains high.

High injection was employed since higher degeneration in the image band provides lower out-of-band noise contribution.

8.3 Mixer Core Design

In this section we use the theoretical results of the previous chapters to relate the performance with the design parameters.

8.3.1 Mixer Conversion Gain

Since the input is conjugately matched, the gain of the transconductance stage is given by expression (6.59)

$$GAIN_{transc} = \frac{(RMS\ current\ at\ transconductance\ stage\ output)^2}{available\ source\ power} = \frac{\omega_T}{\omega} \cdot \frac{1}{\omega L_s} \quad (8.1)$$

where $\omega_T = g_m/(C_{gs} + C_A)$ is the unity gain angular frequency of $M3$ and ω is the angular operation frequency. To obtain the power gain of the mixer we must calculate the amount of signal current that is delivered to the mixer load. The conversion loss in the switching pair must be subtracted from the gain of the transconductance stage. This, as explained in the previous chapters for perfect switching, is equal to $2/\pi$ or $-3.9dB$. In our case however $-5dB$ is a better prediction, as we shall see shortly in section 8.5. The output impedance of the switching pair R_{sp} is in fact time-varying and its time-average is approximately the effective value as discussed in chapter 2. When one of the two devices is turned off, the other device forms a cascode with the transconductance stage device and the output impedance of the switching pair is very high. When both devices conduct, the output impedance of the switching pair is significantly lower, almost equal to $2/g_{ds}$, where g_{ds} is the output conductance of $M1$ or $M2$. At the output of the combiner this impedance is stepped down by a factor of 4 as we saw in chapter 3, section 3.2.2 (or as would be the case if a transformer balun had been used instead of the combiner). Now, impedance R_{out1}, which we use to limit

the mixer output impedance if necessary, is in parallel with R_{sp}. Finally the total mixer output resistance is

$$R_{out} = \frac{R_{sp}//R_{out1}}{4} \tag{8.2}$$

and the mixer power conversion gain equals

$$CG = 10\log_{10}(GAIN_{transc} \cdot \frac{1}{4} \cdot R_{out}) - 5 \; dB \tag{8.3}$$

The factor of 1/4 in the logarithm accounts for the fact that since the mixer output is assumed matched, half of the output current will be delivered to the external load and half will be consumed on the mixer internal output impedance. In practice the real part of the internal mixer output impedance R_{out} is further reduced by the nonidealities of the current combiner as explained in chapter 3. Ohmic losses in the combiner inductors, the input and output matching networks and board traces further reduce the gain.

Most of the factors that determine the mixer output resistance are not easily predicted accurately. Therefore it is a good practice to design a mixer with a conservative value for R_{out}. If the measured gain is higher than desired it can always be reduced by introducing R_{out1}.

8.3.2 Linearity

Equation (7.40) approximately provides the intermodulation distortion of the mixer

$$IM_3 \approx \frac{3}{4}\left(\left|\frac{a_3}{a_1}\right| + |a_1|^2 \left|\frac{b_3}{b_1}\right|\right) \cdot V_{in}^2 \tag{8.4}$$

where V_{in} is the input signal voltage, a_1, a_3 are the power (or Volterra) series coefficients of the transconductance stage and b_1, b_3 are the corresponding coefficients of the switching

pair. Actually the output resistance of the switching pair also contributes nonlinearity and in this case an approximate expression for IM_3 is

$$IM_3 \approx \frac{3}{4}\left(\left|\frac{a_3}{a_1}\right| + |a_1|^2\left|\frac{b_3}{b_1}\right| + |a_1 b_1|\left|\frac{d_3}{d_1}\right|\right) \cdot V_{in}^2 \tag{8.5}$$

where d_1 and d_3 describe the nonlinearity of the output resistance. However, this nonlinearity can be reduced if a linear resistor dominates the mixer output. Then, the input-third order intercept point of the mixer in terms of voltage is given by

$$V_{in,IIP3}^2 = \frac{1}{|a_1|^2}\left(\frac{3}{4}\left|\frac{a_3}{a_1^3}\right| + \frac{3}{4}\left|\frac{b_3}{b_1}\right|\right)^{-1} \tag{8.6}$$

In terms of available power of the source, using the terminology of section 6.3.3 the third-order input intercept point is

$$IIP_3(available\ source\ power) \approx \frac{1}{2 \cdot GAIN_{transc}}\left(\frac{1}{I_{OIP3,transc}^2} + \left|\frac{3}{4}\frac{b_3}{b_1}\right|\right)^{-1} \tag{8.7}$$

where $I_{OIP3,transc}$ is the output third-order intercept point of the transconductance stage which has dimensions of current. We remind here that

$$IIP_3(available\ source\ power) = \frac{V_{in,IIP3}^2}{8R_g} \tag{8.8}$$

$$GAIN_{transc} = 4|a_1|^2 R_g \tag{8.9}$$

$$I_{OIP3,transc}^2 = \frac{4}{3}\left|\frac{a_3}{a_1^3}\right| \tag{8.10}$$

and R_g is the real part of the input impedance looking towards the transistor gate. From (8.7) we obtain the IIP_3 in dBm:

$$IIP_3(dBm) \approx 30 - 10\log(2 \cdot GAIN_{transc}) - 10\log\left(\frac{1}{I_{OIP3,transc}^2} + \left|\frac{3b_3}{4b_1}\right|\right) \tag{8.11}$$

or approximately

$$IIP_3(dBm) \approx 27 - 10\log(GAIN_{transc}) - \frac{1}{2}\max\left\{-20\log(I^2_{OIP3,transc}), 20\log(\left|\frac{3b_3}{4b_1}\right|)\right\}$$

(8.12)

Although we abusively take the logarithms of non unitless quantities, if we use the same system of units, the final result is apparently correct. The term 30 originates from the logarithm of $1mW$ which appears in the conversion of power to dBm.

The IIP_3 due to the nonlinearity of the transconductance stage and the switching pair alone respectively can be expressed as

$$(IIP_3)_{tr}(dBm) \approx 27 - 10\log(GAIN_{transc}) + 10\log(I^2_{OIP3,transc})$$

(8.13)

and

$$(IIP_3)_{sp}(dBm) \approx 27 - 10\log(GAIN_{transc}) - \frac{1}{2} \cdot 20\log(\left|\frac{3b_3}{4b_1}\right|)$$

(8.14)

Quantities $GAIN_{transc}$ and $I_{OIP3,transc}$ can be calculated from the expressions of chapter 6, while $20\log(3b_3/4b_1)$ can be read from the normalized graphs of chapter 7.

Transconductance Stage Nonlinearity

The output third-order intercept point of the transconductance stage used in equation (8.12), when the input is conjugately matched, is given by (6.58), repeated here

$$I^2_{OIP3,transc} \simeq \frac{4}{3}\frac{2r_1}{r'_3(\Delta\omega, 2\omega_o)}$$

(8.15)

where $r_1 = 1/g_1$, and r'_3 is given by (6.28)

$$r'_3 \simeq -\frac{g_3}{g_1^4} + \frac{2}{3}\frac{g_2^2}{g_1^5}$$

(8.16)

Above g_1, g_2, g_3 are the coefficients of the power series that describes the drain current of $M3$ versus its V_{gs}.

Switching Pair Nonlinearity

Experimenting with the methodology described in chapter 7 one can see that the intermodulation distortion of the switching pair always improves when the bias current or the LO amplitude increases, provided that the width of the transistors is adjusted to optimize the performance. Similarly the intermodulation always improves when the length of the transistors is reduced. Therefore we will use the maximum allowed bias current, optimize the LO buffer to obtain maximum LO amplitude and choose appropriate width for the switching pair devices. The corresponding graphs of intermodulation versus device size are shown in section 8.5.

8.3.3 Noise Figure

The noise figure of an active mixer with a conjugately matched transconductance stage has been discussed before in section 5.6. This has not been significantly degraded by high frequency effects, since as we shall see in section 8.5 the conversion gain of the switching pair is still very close to the maximum theoretical value. According to the discussion of chapter 4 the low frequency prediction of the switching pair noise is sufficient in our operation frequency and is used in eq. (5.32). However, this is only a coarse estimate which can vary by several dB from the measured value for the reasons discussed in section 5.6 and also later in section 8.6

8.4   LO Buffer Design

The size of the devices of the switching pair $MB1$ and $MB2$ of the LO buffer is chosen as a compromise between two factors. First the width of the devices must be large enough such that for the given input LO power the tail current is almost completely switched and therefore close to maximum output LO swing is obtained. Second the parasitic capacitance introduced from the common source node to ground must be small enough such that a good common mode rejection and therefore a good single-ended to differential conversion is achieved. Minimum channel length is used for these transistors since it benefits both of the above factors.

The current source $MB3$ is chosen to have a non-minimum gate length in order to increase its output resistance and improve the common mode rejection. Since the capacitance from the drain of this device to ground is dominated by the parasitic capacitances of $MB1$ and $MB2$, only a very small benefit is introduced by reducing the width of $MB3$ to a value lower than the width of $MB1$ and $MB2$.

The tank is implemented with on chip inductors and capacitors. Although a larger LO swing can be achieved with lower capacitance and larger inductance, some capacitance was introduced such that the time-varying capacitance C_{gs} of the transistors does not dominate. As we mentioned before in chapter 7, such capacitance is harmful since it reduces the zero crossing slope of the LO waveform and elongates the time interval during which the switching pair introduces noise and frequency distortion.

The on-chip inductors were implemented with metal 5, metal 4 and metal 3 in parallel to improve the quality factor (Q). However, the low resistivity of the epi substrate
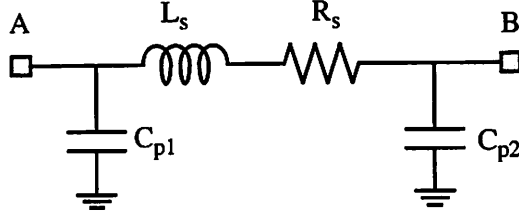
Figure 8.2: On-chip spiral inductor model used in simulation, provided by ASITIC

used allows eddy currents to flow and introduce losses which significantly reduce Q. The low Q translates to low impedance at resonance which means that relatively high current is needed to acquire the desired LO swing. However, the low Q makes the LO buffer broadband and less sensitive to tuning. The model of the inductors shown in Fig. 8.2 was derived with the program ASITIC [59].

8.5 Implementation

Our mixer designs use devices of two different gate lengths $0.24\mu m$ and $0.49\mu m$. The simplified technology parameters discussed in chapters 4 and 7 for these channel lengths as extracted from the available spice models are shown in Table 8.1. The specifications that

Table 8.1: Process Parameters

Parameter (W=100 μm)	L=0.24 μm	L=0.49 μm	Units
K	9.212e-2	2.67e-2	A/V
θ	2.46	0.778	V^{-1}
V_T	0.531	0.511	V
C_{gs1} (cutoff)	21	21	fF
C_{gs2} (saturation)	140	250	fF
$n\phi_t$	38	34	mV

all the designs satisfy are shown in Table 8.2.

Our LO buffer, optimized as described in the section 8.4 for maximum output LO

voltage, provides a differential LO voltage of amplitude approximately $1V$, and amplitudes

close to this value are used in the optimization of the switching pair.

Table 8.2: Design Specifications

Parameter	Value
Input signal frequency (f_{RF})	$1.9GHz$
Output signal frequency (f_{IF})	$100MHz$
LO frequency (f_{LO})	$2GHz$
Power Supply Voltage (V_{DD})	$2.5V$
Current Dissipation of mixer core (I_B)	$8mA$
Current Dissipation of LO Buffer (I_{BB})	$4mA$
Required external LO power	$< -8dBm$
Input Return Loss (in 50 Ω environment)	$< -12dB$
Output Return Loss (in 50 Ω environment)	$< -12dB$
LO input Return Loss	not specified

Fig. 8.3 and Fig. 8.4 show conversion gain and intermodulation distortion re-

spectively of the switching pair versus transistor width for channel length $0.25\mu m$. The

corresponding graphs for channel length $0.5\mu m$ are shown in Fig. 8.5 and Fig. 8.6. In both

cases the bias current is $8mA$, the LO amplitudes shown are $0.7V$, $1V$ and $1.3V$, the LO

frequency is $2GHz$ and the input signal frequency is $1.9GHz$. Some extra capacitance equal

to $60fF$ is included from the common source node to ground, which represents as a first

approximation the drain capacitance of M3 and the wiring capacitance. It is apparent from

these graphs that the optimal width values are approximately $100\mu m$ and $150\mu m$ for the

channel lengths $0.25\mu m$ and $0.5\mu m$ respectively. The transistor size changes the LO buffer

loading and affects the amplitude, but the capacitance of the tuned tank is adjusted such

that the amplitude at the transistor gates in all cases is $1V$.

Four designs were fabricated whose design parameters are shown in Table 8.3.
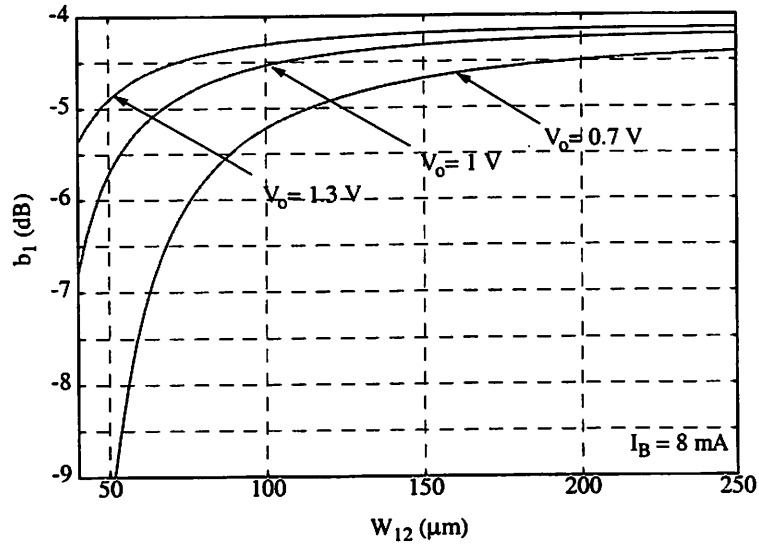
Figure 8.3: Conversion gain versus device width for minimum device length and several LO amplitudes.
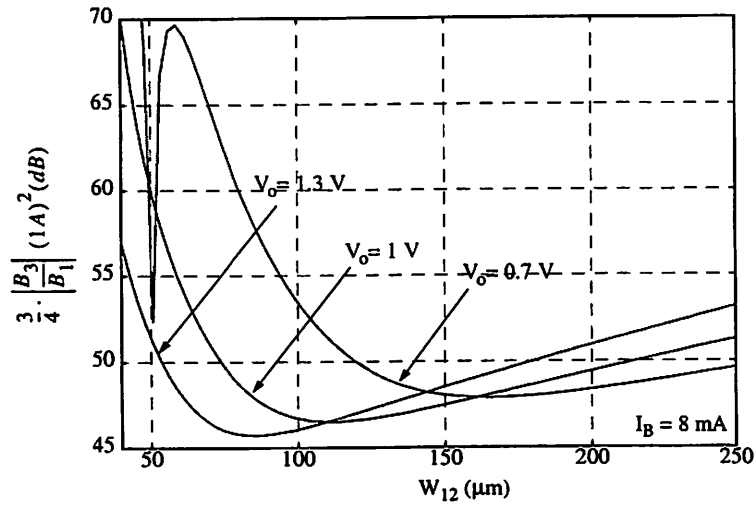
Figure 8.4: Intermodulation distortion versus device width for minimum device length and several LO amplitudes.

Their differences in the performance are discussed in the next section.
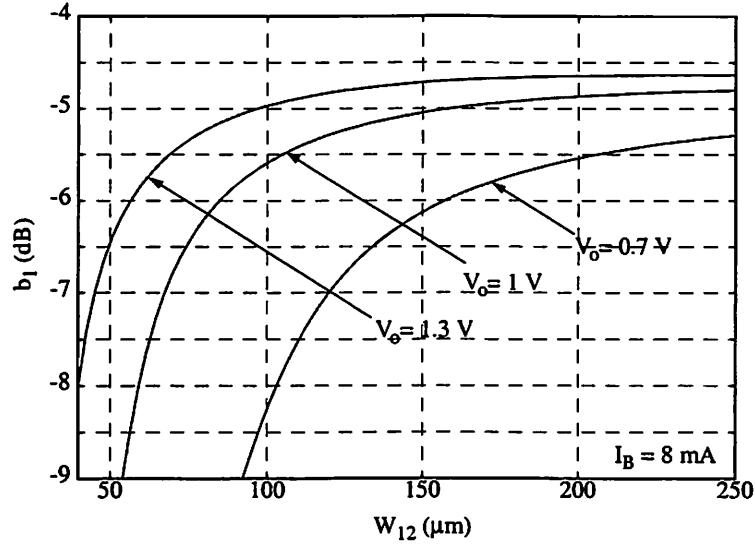
Figure 8.5: Conversion gain versus device width for device length 0.5 μm and several LO amplitudes.
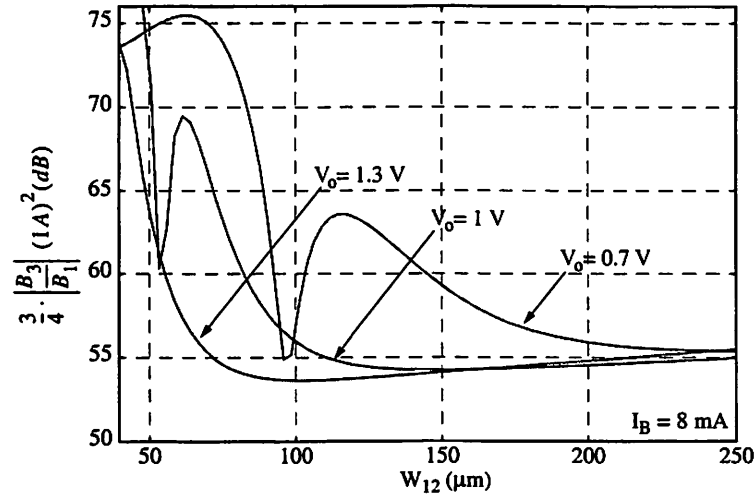
Figure 8.6: Intermodulation distortion versus device width for device length 0.5 μm and several LO amplitudes.

8.6 Predicted and Measured Results

Before we compare predicted and measured results we should mention that the prediction is not expected to be accurate since there are several unknown parameters for which we use only a rough approximation. For example the value of the degenerating

Table 8.3: Design Parameters

	Design A	Design B	Design C	Design D	
Parameter					Units
W/L of M1,M2	100/0.24	150/0.49	100/0.24	150/0.49	$\mu m/\mu m$
W/L of M3	40/0.24	150/0.24	50/0.24	100/0.49	$\mu m/\mu m$
W/L of MB1,MB2	200/0.24	200/0.24	220/0.24	220/0.24	$\mu m/\mu m$
W/L of MB3	100/0.49	100/0.49	100/0.49	100/0.49	$\mu m/\mu m$
C_A	0	0	200	100	fF
L_T	5	5	5	5	nH
C_T	600	400	600	400	fF

inductor $L_s = 3nH$ which is implemented with a bond wire is only a first order estimate. The input and output matching networks are formed with inductors and board traces which demonstrate ohmic losses that are difficult to estimate.

The output current combiner is also imperfect and according to the discussion in chapter 3, if the used inductors are lossless, it would approach perfect operation for small inductance values and high mixer output impedance. However, lab measurements showed that small inductance values induce high losses. Relatively high inductance values equal to $270nH$ and completely eliminating the external capacitor C_C were found to give the highest gain and were used in practice. As we shall see this gain is yet significantly lower than the theoretically achievable with lossless inductors.

From equation (3.4), the operation of the current combiner improves for high mixer output resistance. Also, high mixer output resistance provides higher power gain because it approaches the ideal current source which can provide infinite power. Long channel devices in the switching pair provide higher output resistance because of improved channel length modulation effect. It is simulated from the available spice model, that under the bias conditions that appear in the switching pair devices when they are on and

contribute some finite output resistance, the output resistance is $r_{out} \approx 1.2K\Omega$ and $r_{out} \approx 3K\Omega$ for $L = 0.24\mu m$ and $L = 0.49\mu m$ respectively. According to our discussion on the current combiner of chapter 3 the output resistance of each one of the transistors appears at the output in parallel with the current source which gives the difference of the two drain currents. In particular the time-average of each one of the resistors should be considered, but because calculating this time-average is not straightforward and because we would rather obtain a pessimistic estimate for the power gain prediction, we will assume that the parallel combination of the two impedances is half of the value quoted above.

The internal mixer output resistance depends on the drain voltage of the devices, is highly nonlinear and tends to dominate the linearity when an external linear resistor R_{out1} is not connected. The output resistance affects the linearity less when long channel devices are used for the switching pair, because then the internal nonlinear resistance is higher.

We should also note that there is some uncertainty with the measurements of the IIP_3. Several measurements that were repeated with matched input and output but with different matching networks were found to be as much as $2dB$ different. This probably has to do with the fact that the out of band behavior of the matching networks has a minor effect in the intermodulation as discussed in chapter 6.

The power gain and linearity improve when the DC level of the LO voltage is relatively low. This happens because as the drain-source voltage of the switching pair devices increases their output resistance increases. A DC LO voltage of $2V$ was used for all our measurements.

The input $1dB$ compression point P_{-1dB} is dominated by the output headroom when external resistor R_{out1} is not connected. As we shall see P_{-1dB} increases by a lot when a low R_{out1} is used to reduce the output swing, which also reduces the power gain.

Two different quality factors can be defined for the mixer input. One, denoted with Q_{in} is the ratio of the imaginary and the real part of the input impedance of the transconductance stage, including the bond wire of the gate. The second, denoted with Q_m is the quality factor of the matching network which matches the real part of the input impedance of the mixer to the impedance of the source 50Ω and is equal to $\sqrt{Re\{Z_{in}\}/50 - 1}$. The higher these quality factors are, the higher the losses introduced and the more sensitive matching is to the component value variation.

Estimating the noise figure is also a difficult task because of the uncertainty in the design parameter values and the complexity of the accurate calculations. We will attempt to obtain only a rough estimate and observe the trend as the design parameters change rather than predicting it accurately. First the noise parameters of the transistors γ and δ are unknown. Arbitrarily we will use $\gamma = 4$ and $\gamma = 2$ for channel lengths $L = 0.24\mu m$ and $L = 0.49\mu m$ respectively. Second, losses at the input matching network introduce an unknown amount of attenuation which increases the noise figure. Third, since these designs are single balanced, noise form the LO port appears at the output. The LO buffer generates its own noise and also amplifies the noise floor of the signal generator used as external LO source. This noise floor is in the order of $-130dBm/Hz$ to $-140dBm/Hz$, (which as a comparison is significantly higher than the available noise power of a resistor at room temperature $-174dBm/Hz$). Now from the discussion of chapter 4 we know that because

of symmetry, the transconductance of the switching pair $G(t)$ is periodic with period half of the LO period. Therefore, only the even order sidebands contribute noise, that is, only noise at frequencies f_{IF}, $2f_{LO} \pm f_{IF}$ at the output of the LO buffer will be transfered to the output. Fortunately the tuned tank LO buffer load attenuates the noise at these frequencies, but this attenuation is finite, since our inductor Q is relatively low (simulated with ASITIC approximately equal to 3.5). In addition, because of asymmetries introduced by the output current combiner which is an asymmetric load, and the fact that the LO buffer has a single-ended input and does not perform perfect single-ended to differential conversion, we do expect that some noise will also be transfered from the $f_{LO} \pm f_{IF}$ frequency bands where the LO buffer load is tuned. As a very coarse approximation we will assume that the noise at the LO port is white and is generated by an equivalent resistor equal to the resistance of the tuned tank at resonance $R_{LO} = 440\Omega$ (each tuned tank has a parallel resistance of approximately 220Ω). Finally, we use expression (5.32) which as explained in chapter 5 does not accurately take into account the out of band noise contribution of the transconductance stage.

Several other approximations have been made during the derivation of the expressions used to predict the mixer performance, for example the gate-drain capacitances have been neglected.

We must note that many of the factors that limit the accuracy of our prediction, also limit the accuracy of the prediction of a circuit simulator. For example the simulator gives overoptimistic results for gain if the board inductor losses are neglected.

Some critical parameters which have been found helpful in the performance predic-

Table 8.4: Some critical design quantities

Parameter	Design A	Design B	Design C	Design D	Units
$GAIN_{transc}$	0.555	0.425	0.140	0.131	Ω^{-1}
$b1$	0.59	0.55	0.59	0.55	
$-I^2_{OIP_3,transc}$	46	51.4	47	49.2	$20\log(A^2)$
$|3b_3/4b_1|$	46.5	54.1	46.5	54.1	$20\log(A^2)$
$(IIP_3)_{tr}$	6.5	5.0	12.0	11.2	dBm
$(IIP_3)_{sp}$ (Low R_{out1})	6.3	3.7	12.3	8.8	dBm
$\overline{G}$	5.1e-3	5.1e-3	5.1e-3	5.1e-3	Ω^{-1}
$\overline{G^2}$	1.6e-4	1.05e-4	1.6e-4	1.05e-4	Ω^{-2}
R_{in}	712	545	179	168	Ω
$Q_{in} = Im\{Z_{in}\}/Re\{Z_{in}\}$	2.00	0.60	1.33	1.00	
$Q_m = \sqrt{Re\{Z_{in}\}/50 - 1}$	3.64	3.15	1.61	1.53	

tion during the analysis performed so far are shown in Table 8.4. The predicted performance parameters are shown in Table 8.5, while the measured performance parameters are shown in Table 8.6. Some critical design quantities are given in table 8.4, while the prediction for the performance parameters are shown in table 8.5.

Table 8.5: Predicted Performance Parameters

Parameter	Design A	Design B	Design C	Design D	Unit
Power Gain (No R_{out1})	14.2	17.3	8.2	12.2	dB
IIP_3 (Low R_{out1})	6.3	3.7	12.0	8.8	dBm
$NF(dB)$	6.8	7.8	11.0	9.9	dB

8.6.1 Design A

The devices of both the switching pair and the transconductance stage are of minimum channel length. The small width of the transconductance stage device provides high ω_T and therefore high transconductance gain. The relatively high Q input matching network introduces losses at the input and makes the input matching network very sensitive to the component values. Also imperfect current combiner operation introduces losses at the

output and makes the power gain about $4dB$ lower than predicted. The transconductance stage and the switching pair contribute approximately equally to the nonlinearity when the mixer output is dominated by a linear impedance. The predicted IIP_3 due to the transconductance stage and switching pair alone is $6.5dBm$ and $6.3dBm$ respectively. Because of the interaction of the two nonlinearities we expect the measured IIP_3 to be a little lower than these values, but on the other hand the losses of the input matching network tend to increase this value. Therefore, the measured value of about $7dBm$ is in excellent agreement with the prediction.

Table 8.6: Measured Performance Parameters

Parameter	Design A	Design B	Design C	Design D	Unit
Power Gain (No R_{out1})	10.2	13.3	6.0	9.2	dB
IIP_3 (No R_{out1})	2.0	1.0	5.3	4.8	dBm
IIP_3 (Low R_{out1})	7.0	4.0	9.5	7.0	dBm
P_{-1dB} (No R_{out1})	-11.2	-14	-7	-11.3	dBm
P_{-1dB} (Low R_{out1})	-5.5	-7	-1.8	-2.2	dBm
$NF(dB)$	9.3	9.2	12.8	9.5	dB

Table 8.7 shows how the nonlinearity is dominated by the output resistance when R_{out1} is not present and how linearity improves as R_{out1} becomes low. The input P_{-1dB} is also shown in this table. When it is dominated by the output resistance nonlinearity it is $-11.2dBm$ while when a low R_{out1} is connected it is significantly higher, $-5.5dBm$. The difference of the P_{-1dB} and the IIP_3 is $12.5dB$, higher than $9.6dB$ that would have been the value if only third-order nonlinearity was present (see chapter 2). Therefore at this power level, 5-th or higher order nonlinearities are excited. The measured noise figure $9.3dB$ is about to $2.5dB$ higher than our coarse prediction. Such a discrepancy is expected because of the input matching network losses and the other reasons explained before.

Table 8.7: Gain, IIP3 and input P_{-1dB} as a function of the output resistance

$R_{out1}(\Omega)$	Power Gain (dB)	IIP3 (dBm)	P_{-1dB} (dBm)
∞	10.17	2	-11.2
$2.7K$	6.17	5.47	
$1.1K$	3.83	6.16	
510	1.00	6.25	
200	-2.67	6.66	-5.5
100	-5.67	7.41	

8.6.2 Design B

Here the devices of the switching pair are $0.5\mu m$ long. This limits the linearity of the switching pair which is less linear than the transconductance stage. Minimum length but large width equal to $150\mu m$ is used for the transconductance stage, which provides a little lower transconductance gain than in Design A because of the lower ω_T, but also lower Q input matching network and lower input losses.

The high output impedance of the long switching pair devices provides good operation of the current combiner and high power gain since the output current is delivered by an equivalent current source with a high output resistance. The predicted power gain $17.3dB$ is higher than the measured $13.3dB$. The IIP_3 for a low R_{out1} is limited by the switching pair, and the measured value $3.7dB$ agrees very well with the measured $4dB$. The long channel switching pair devices are beneficial in reducing the output resistance nonlinearity since a relatively high R_{out1} is sufficient to dominate the output resistance, without excessively reducing the power gain.

The measured noise figure $9.2dB$ is again higher than the rough estimate $7.8dB$. The long switching pair devices have now lower noise contribution than in design A because of their reduced γ factor and because they attenuate more the noise from the LO port since

their $\overline{G^2}$ value is lower.

8.6.3 Design C

Here a minimum channel length switching pair and minimum channel length and small width transconductance stage are used, similarly to design A. However a capacitor $C_A = 200fF$ is connected between the gate and the source of the input transistor to reduce the input impedance and facilitate input matching. This circuit is indeed very easy to match and not very sensitive to the matching component value variation. Capacitor C_A however reduces ω_T, which reduces the power gain. The output current combiner again introduces high loss which makes the measured power gain $6.0dB$, again lower than the predicted $8.2dB$.

The predicted IIP_3 is about $12dB$ from both the switching pair and the transconductance stage. The measured value $9.5dB$ is slightly lower but very close, within the limits of the interaction of the two kinds of nonlinearity. Besides, we should mention that the IIP_3 prediction of the transconductance stage is not accurate since in the derivation of expression (6.28) we had assumed that C_{gs} is small. This assumption does not hold very well here since capacitor C_A is used. Indeed such a capacitor would tend to increase r'_3 and reduce linearity.

The coarse estimate for the noise figure $11.0dB$ is close to the measured value $12.8dB$. The low ω_T of the transconductance stage gives a high noise figure transconductance stage and also a low transconductance gain which does not significantly suppress the noise generated by the switching pair and introduced by the LO port.

8.6.4 Design D

This design uses long channel devices for both the transconductance stage and the switching pair. It also uses capacitor $C_A = 100fF$ to provide a low input impedance. The long switching pair devices provide high output impedance which boosts the power gain. The predicted power gain 12.2 is higher than the measured $9.0dB$.

The IIP_3 is dominated again by the switching pair. The predicted is $8.8dBm$ and the measured is $7.0dBm$. Similarly to Design C, in reality the prediction of the transconductance stage nonlinearity is expected to be optimistic because of the large gate-source capacitance, assumed small in (6.28).

The coarsely predicted noise figure is close to the measured value. Here, besides the low gain of the transconductance stage, the noise figure remains below $10dB$ because of the low noise factor γ of the long channel devices and the low transconductance value of the devices of the switching pair which suppress the noise form the LO port.

Chapter 9

Conclusions

9.1 Thesis Summary

This thesis presented a systematic analysis of the operation of a commonly used class of CMOS mixers, those which employ a transistor switching pair to commutate the signal current. It was demonstrated that by carefully formulating the problems and by adopting appropriate transistor models and some approximations, intuition can be obtained and the performance can be qualitatively and quantitatively predicted. Some performance characteristics, for instance the thermal and flicker noise generated by the switching pair transistors, can be shown to have very simple analytical expressions, reported for the first time as a result of this research. Other characteristics such as the switching pair intermodulation require numerical solution of complicated equations, but in this case the performance can be captured in graphs of normalized parameters. The linearity and noise of CMOS transconductance stages were also studied and several useful results were derived. These results can be applied to several other RF blocks besides mixers, such as LNAs and power

amplifiers. Our theoretical results were validated with measurements of appropriate test structures. A designer can rapidly optimize the performance of an active mixer by using the results of this thesis without lengthy simulations. The designer can also easily judge if a given CMOS technology has the potential to satisfy the desired specifications. The use of the theoretical results was demonstrated with the design of some active mixers operating in the $2GHz$ frequency band.

Some specific contributions are listed below:

- In chapter 4 expression (4.33) which predicts the thermal noise contribution of the mixer switching pair was derived. In section 4.4.5 we showed how the estimate for parameter $\overline{G}$ derived in section 4.4.2 can be used to predict the flicker noise generated by the switching pair. The noise introduced by the LO port in the simplified case that noise present at this port is white was found in section 4.4.3, and an expression for the related gain (4.39) was derived. Simplified expressions for the switching pair conversion gain, the gain of the white noise at the output of the transconductance stage and the frequency limits where the prediction of our analysis is degraded by capacitive effects were derived in chapter 4.

- The noise behavior of the commonly used inductively degenerated, conjugately matched common-source CMOS transconductance stage was analyzed extensively in chapter 5. Expressions (5.21) and (5.22) were derived from which the noise figure can be predicted and optimized. In particular the problem of optimizing the noise figure in the presence of a lossy ESD was examined.

- In chapter 6 the intermodulation performance of several CMOS transconductance

stages was studied. Expressions (6.59), (6.58) and (6.28) were derived from which

the IIP_3 of an inductively-degenerated conjugately-matched transconductance stage

can be predicted. Similar results for resistively degenerated stages with and without

input matching restrictions were derived. The body effect nonlinearity was found to

impose a limit in the linearization benefit introduced by degeneration, and the value of

degeneration above which no linearization benefit is obtained was found in expression

(6.35).

- In chapter 7 the concept of the time-varying power series and Volterra series was

  used to find the intermodulation of the switching pair at low and high frequencies

  respectively. The results are expressed in form of graphs of normalized parameters

  shown in figures 7.9, 7.16 and 7.17. A simple transistor model whose IV curves

  have continuous derivatives of any order in weak, moderate and strong inversion was

  proposed. The continuity of the derivatives of the current and capacitance curves

  versus bias in a transistor model used for intermodulation simulation was shown to

  be critical.

9.2 Future Research Opportunities

One possible application of the research conducted in this thesis is generating

macromodels for mixers. The macromodels for example can take as parameters the design

parameters such as the bias current and the device sizes and model the gain, the output

noise and the nonlinearity effects. An accurate noise model could take into account the

cyclostationary nature of the noise.

Several mixer performance characteristics were not analyzed in this thesis and can be subject of future research, for example the port-to-port isolation and the 1 dB compression point. An other interesting problem is the relation between device mismatch and even-order nonlinearities. As mentioned in chapters 4 and appendix C the flicker noise behavior of devices with time-varying operating noise needs to be characterized theoretically and experimentally. The thermal noise characteristics of short channel CMOS devices remains a dark modeling area. In many modern applications the gate referred noise described in [108] is significant, and verification of this model for short channel devices with measurements is of very high importance. Theoretical and experimental work needs to be done, to quantify all the fudge factors that the current model includes (i.e factors γ, δ and ρ) described in chapter 5. Consideration of this more complete model for the characterization of the switching pair noise at high frequencies is one more open problem. There is room for analytical work on other mixer structures such as those discussed in chapter 3 and even other circuit blocks such as oscillators, low noise amplifiers and power amplifiers. Finally the problems analyzed here can be subject of future research using different approximations or approaches in order to provide possibly more accurate, more intuitive or simpler ways to predict the mixer performance.

Appendix A

Cyclostationary Noise in Radio-Frequency Communication Systems

A.1 Introduction

The concepts of noise figure and noise temperature have been introduced to describe the noise performance of circuits and receivers [18, 2]. They are convenient performance metrics because the noise figure and noise temperature of a system of cascaded blocks can be found easily from the corresponding quantities of the individual blocks. However, the simple formulas for a system of cascaded blocks assume that the noise at the input and the output of every block is a wide-sense-stationary (WSS) process. There are two reasons why the mixer output noise is in fact not WSS but has periodically time-varying statistics. First,

the operating points of the devices may vary with time, and second the transfer function of the noise signal from the point at which it is generated to the output can have time-varying characteristics [31]. The mixer output noise is a cyclostationary process and its complete description requires a periodically time-varying power-spectral-density (PSD) $S(f, t)$ [21]. An accurate evaluation of the output noise when cyclostationary noise is processed by a linear-periodically-time-varying (LPTV) system is considerably more complicated than the evaluation of the output noise of a linear-time-invariant (LTI) system processing WSS noise. The corresponding analysis and methodology is given in [21], and a related circuit simulator has been presented in [71].

Despite the fact that the mixer output noise is cyclostationary, the noise figure calculated using the time-average output noise PSD has been traditionally used to characterize mixers, and the simple formulas for the noise figure of a system of cascaded blocks have been used to find the noise figure of a receiver. We shall show here that this treatment provides the correct noise characterization of a communication system in most practical cases, but we will examine cases in which it could lead to an inaccurate prediction. The pitfalls of applying the stationary process theory to cyclostationary signals have been presented in mathematical terms in [20]. Here we discuss qualitatively some related results that can be useful in the design of radio-frequency communication systems.

A.2 Cyclostationary Noise and its Time Average

The complete description of a cyclostationary noise signal with its time-varying PSD $S(f, t)$, as opposed to its description with $S(f)$, the time-average of $S(f, t)$, is significant

only when the block to which the cyclostationary noise is input is synchronized to the variation of $S(f, t)$ with time. This statement will be explained on an intuitive basis, and it also gains support from the following theorem [19]

If a uniformly distributed random variable from zero to one cycle period is added to the time variable t of a cyclostationary process with PSD $S(f, t)$, (that is, the information about the phase of the periodically varying PSD is lost) the resulting process is stationary and its statistics are the time-average of the statistics of the cyclostationary process.

If the system to which the cyclostationary noise is input does not track the PSD variation with time, the phase of $S(f, t)$ for this system is unknown. In the absence of information about the phase of $S(f, t)$ the process becomes stationary, with PSD equal to the time-average of $S(f, t)$.

Usually, the noise performance of the analog part of a communication system consisting of a chain of radio-frequency circuit blocks, is characterized by measuring the time-average noise PSD at the output of the chain. Noise measuring equipment measures the noise PSD at a frequency f by measuring the noise signal power at the output of a very narrow-band filter around f, without tracking the time variation of the noise statistics and provides the time-average PSD.

When a cyclostationary noise signal passes through a LTI filter and the time-average PSD is measured at the output, the same result is obtained if only the time-average PSD is considered at the input of the filter [21]. However, when a cyclostationary noise signal is fed to a time-varying system, consideration of only the time-average PSD of the input noise can lead in the general case to wrong results [21]. For instance, if the time-
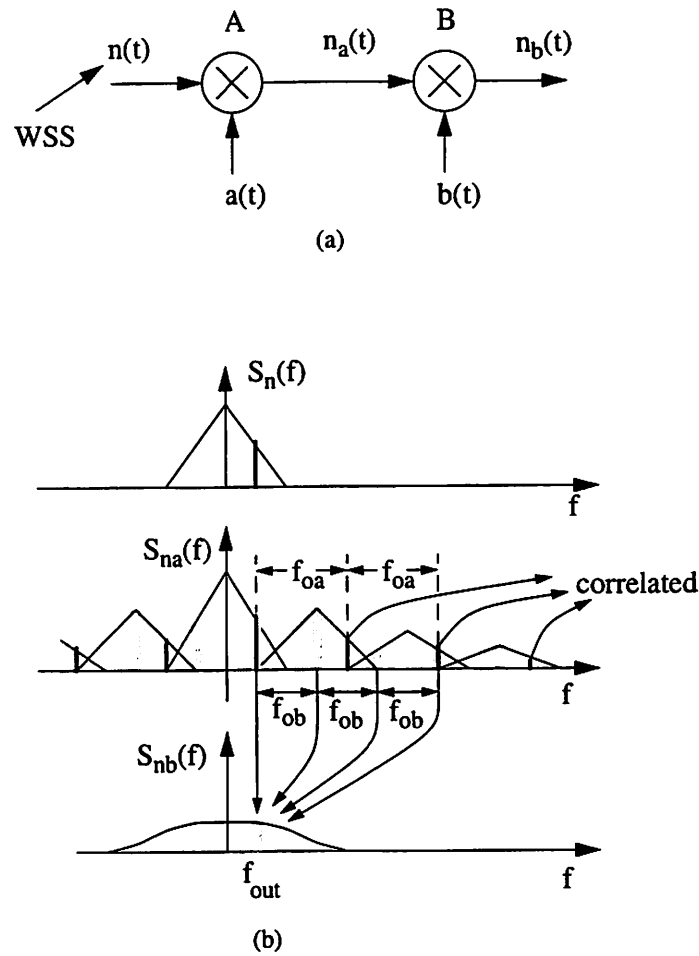
Figure A.1: a) A cascade of two mixers. (b) Time-average PSD of noise at the input, after the first mixer and the output.

varying gain and the power of the input noise obtain their peak values simultaneously, considering only the time-average input noise will underestimate the output noise. The following example will help clarify the situation.

Consider that a WSS signal $n(t)$ with PSD $S_n(f)$ is fed to a mixer A, and the output of this $n_a(t)$ is fed to a mixer B, as shown in Fig. A.1(a). The random signal $n(t)$ can represent noise present at the input of mixer A, or noise generated by its devices

[1]. The mixing operation is modeled by multiplication of the input signal $n(t)$ with a periodic waveform (time-varying gain) generated by a local oscillator, $a(t)$ with frequency f_{oa} and $b(t)$ with frequency f_{ob}, for mixers A and B respectively [2]. The output of mixer A is a cyclostationary process whose time-average PSD consists of copies of $S_n(f)$ shifted in frequency integer multiple of f_{oa}, and weighted by different coefficients. It is easy to see that frequency components of $n_a(t)$ in distance integer multiple of f_{oa} are correlated, since they contain the same frequency component of $n(t)$. A random process can be cyclostationary with cycle frequency f_{oa} only if there exists correlation between two different frequency components in distance f_{oa}. The spectral correlation can be expressed in terms of the cyclic spectra, the Fourier components of the time-varying PSD, and in fact the k-th cyclic spectrum for positive k is the correlation between frequency components in distance kf_{oa}, while the 0-th order cyclic spectrum is the time-average PSD. A random process can be WSS only if any two different frequency components are uncorrelated [21]. The output of mixer B is a cyclostationary process with two cycle frequencies f_{oa} and f_{ob}. If f_{oa} and f_{ob} are commensurate (their ratio is a rational number), $n_b(t)$ can be viewed as cyclostationary with one cycle frequency equal to the maximum common divider frequency of f_{oa} and f_{ob}.

---

[1]In the case of noise generated by devices with time-varying operating point, this noise is cyclostationary and white and its time variation can be incorporated to the system [31]. Therefore in any case the input noise $n(t)$ can be considered WSS. For every noise source inside the mixer the time-varying gain is a different function.

[2]At high frequencies where reactive effects are not negligible, the mixing operation also depends on the input-signal frequency and is better modeled with a periodically-time-varying transfer function $A(f, t)$ [107, 31], instead of a periodically-time-varying gain $a(t)$. Frequency translation is described with the Fourier components of $A(f, t)$, the conversion transfer functions instead of the conversion gains. For simplicity reactive effects are neglected in our mixer model. However the qualitative arguments presented here also apply at high frequencies.

A.2.1   Effect of LO frequency relation

Let us examine now the spectral content of the output of mixer B $n_b(t)$ at a frequency f_{out}. Frequency components of $n_a(t)$ at frequencies $f_{out} + k f_{ob}$, k being an integer, are folded on f_{out} as shown in Fig. A.1(b). If $n f_{oa} = m f_{ob}$ for some integers n and m, there exists correlation among these components, and it is incorrect to add their power, as we would do if $n_a(t)$ were WSS, since a valid addition would require correlation terms. However, if the ratio of f_{oa} and f_{ob} is not a rational number, such integers n and m do not exist and simply adding the different frequency components of the time-average PSD $S_{na}(f)$ provides the correct result, since the added terms are uncorrelated.

In practice, the ratio of two LO frequencies generated by different free running oscillators can always be considered an irrational number, since because of the random phase error they cannot track each other. The situation is different however if the two LOs are locked to a common reference frequency. In a superheterodyne receiver which employs two mixers, it is a common practice to generate the two LO signals from two PLLs with a common reference frequency, which means that f_{oa}/f_{ob} is a rational number m/n (we will assume below that m and n are such that a common integer divider greater than one does not exist). Despite this, a rational frequency ratio $f_{oa}/f_{ob} = m/n$ with m or n very large numbers is expected to have the same practical effect as an irrational frequency ratio. In fact, the LO frequencies in a receiver chain are often chosen such that they do not have a simple relation in order to avoid spurious responses.

Assuming a smooth $b(t)$ with low frequency content, we can see that the conversion gain of mixer B drops rapidly with the order of the sideband, and only the first few (for

example up to 3 or 4) contribute significantly. Therefore, considering again the integers m and n that satisfy $f_{oa}/f_{ob} = m/n$, if m is a large integer, in every set of correlated frequency components of $n_a(t)$ in distance integer multiple of $mf_{ob} = nf_{oa}$ that contribute to f_{out}, only one term contributes significantly and only a minor error is introduced by adding the power of all the components. If n is large, assuming a smooth $a(t)$, the effect of noise correlation is also attenuated for a similar reason: the copy of $n(t)$ around nf_{oa} has low power. Concluding, the effect of spectral correlation is insignificant if $a(t)$ is smooth and n is large, or if $b(t)$ is smooth and m is large, or both. Very often in practice, especially at high frequencies $a(t)$ and $b(t)$ are smooth functions, and unless the ratio of the two LO frequencies is a simple rational number m/n with m, n small integers, calculating the time-average at the first mixer output and treating it as if it were the PSD of WSS noise, does not introduce a significant error in the noise estimation at the second mixer output. Nevertheless, there are practical situations where the time-varying gain of a mixer is not a smooth waveform. An example is the sampling or subsampling mixer, in which case the time-varying gain is a pulse train which has high frequency content.

The above argument can be easily visualized in the time domain with the example of Fig. A.2. Assume that b(t) is an impulse train, so that mixer B is essentially a sampling mixer as shown in Fig. A.2(a) and that we desire to estimate the time-average power of the samples at the output of the sampler. Consider that the time-varying power $\sigma_a(t)$ of the cyclostationary noise $n_a(t)$ – the integral of the time-varying PSD over all frequencies – at the first mixer output is the periodic function of time shown in Fig. A.2(b). If $f_{oa} = f_{ob}$, or $f_{oa} = mf_{ob}$, we always sample $n_a(t)$ when $\sigma_a(t)$ is at the same point of the period as
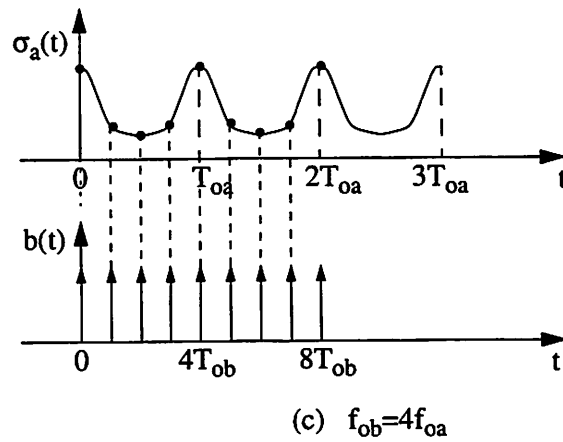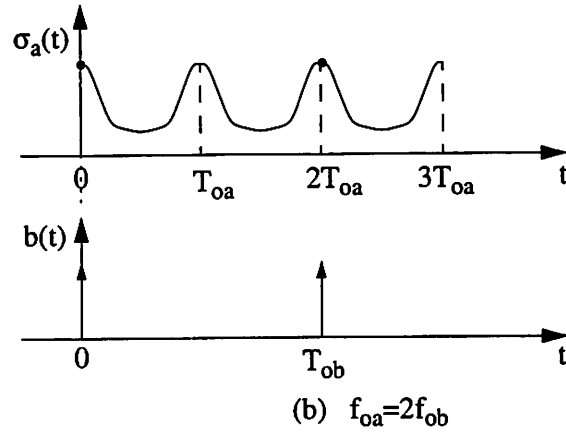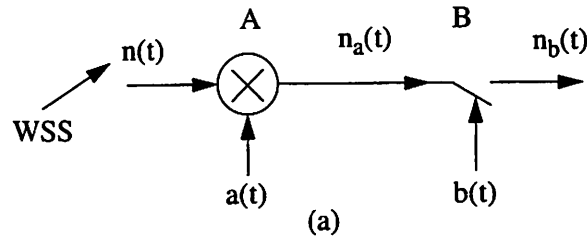
Figure A.2: Sampling cyclostationary noise.

shown in Fig. A.2(b), and if instead the time-average of $\sigma_a(t)$ is considered at the input of

the sampler, we probably significantly overestimate or underestimate the output noise. In

this case, since $b(t)$ is not a smooth function of time and its spectral content does not die

out at high frequencies, the effect of spectral correlation is not diminished if m is large. If

$f_{oa}/f_{ob} = m/n$ is a rational number and n is a small integer, we sample repeatedly only a few points in the period and it is possible that considering the time-average of $\sigma_a(t)$ at the sampler input will result again in an erroneous noise estimation. However, if n is a large number, the same points of the period are repeatedly sampled, but they are many and uniformly distributed across a period, as shown in Fig. A.2(c), so considering the time-average at the sampler input would give a practically correct result. When f_{oa}/f_{ob} is not a rational number, after long enough time the whole period is uniformly sampled and in fact the same point is never sampled twice. In this case time-averaging at the sampler input provides exactly the correct result.

Let us examine now the effect of the LO frequency relation in a more quantitative manner. Referring to Fig. A.1, we can see that $n_b(t)$ consists of scaled copies of $n(t)$ shifted in frequencies $k_a f_{oa} + k_b f_{ob}$, where k_a and k_b are the sidebands at which the conversion gain of mixers A and B respectively is significant, determined by the spectral content of the waveforms $a(t)$ and $b(t)$ and possibly as we shall see below by filtering the mixer outputs. If two of these frequencies coincide, the spectral correlation affects the output noise estimation. If k'_a and k'_b is a second set of mixer sidebands, the relation

$$k_a f_{oa} + k_b f_{ob} = k'_a f_{oa} + k'_b f_{ob} \tag{A.1}$$

or

$$\frac{k_b - k'_b}{k_a - k'_a} = -\frac{f_{oa}}{f_{ob}} \tag{A.2}$$

can only hold if f_{oa}/f_{ob} is a rational number, as we also concluded before. Furthermore, if $f_{oa}/f_{ob} = m/n$, spectral correlation has an effect only if there are integers k_a, k'_a, k_b, and

k_b' that represent mixer sidebands with significant conversion gain such that

$$\frac{k_b - k_b'}{k_a - k_a'} = -\frac{m}{n} \tag{A.3}$$

If for example, $a(t)$ and $b(t)$ are sinusoidal with frequencies mf_o and nf_o, where f_o is some reference frequency, k_a, k_a', k_b, and k_b' can only be $+1$ and -1, and spectral correlation can have an effect only if $n = m$.

Similarly, one can examine the effect of spectral correlation when a third mixer C follows the chain of A and B. Denoting the frequency of C by f_{oc} and the sidebands of C with some significant conversion gain by k_c and k_c', spectral correlation affects the noise estimation only when there are mixer sidebands with significant conversion gain, such that

$$k_a f_{oa} + k_b f_{ob} + k_c f_{oc} = k_a' f_{oa} + k_b' f_{ob} + k_c' f_{oc} \tag{A.4}$$

If the LO frequencies are related, i.e. $f_{oa} = mf_o$, $f_{ob} = nf_o$, $f_{oc} = pf_o$, where f_o is some reference frequency and m,n, and p integers with no common factors, (A.4) becomes

$$(k_a - k_a')m + (k_b - k_b')n + (k_c - k_c')p = 0 \tag{A.5}$$

In this case, it is possible that conditions (A.4) and (A.5) hold for low order sidebands, even if the LO frequency relation is not simple. For example, if $f_{oa} = 2000MHz$, $f_{ob} = 660MHz$ and $f_{oc} = 10MHz$ the above equations are satisfied for $k_a - k_a' = 1$, $k_b - k_b' = -3$, and $k_c - k_c' = -2$.

A.2.2 Filtering a Cyclostationary Noise Process

If filtering takes place at the output of a mixer, as in Fig. A.3(a), it is possible that the noise at the output of the filter is stationary, and no cyclostationary noise considerations
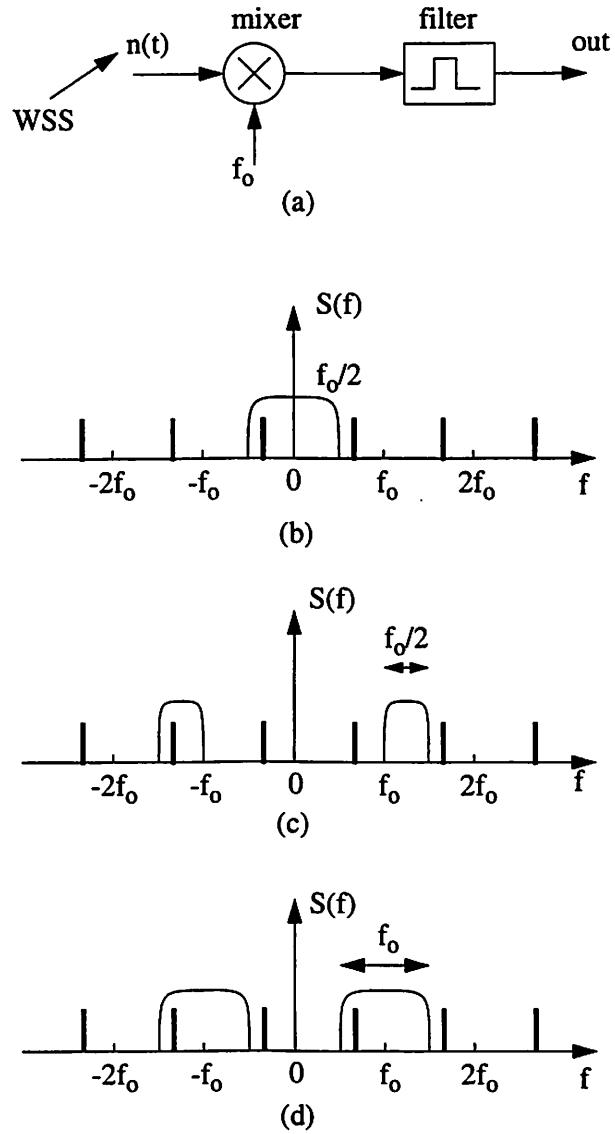
Figure A.3: Filtering cyclostationary noise

need to be made, or that the characteristics of the cyclostationary noise change. Some

relevant theorems have been presented in [71], where they were derived in a mathematical

way. Similar results can be found elsewhere [31, 76]. These theorems become intuitive

by examining filtering of a set of correlated frequency components. Let us consider a

cyclostationary noise process with cycle frequency f_o and a set of correlated frequency components in distance integer multiple of f_o. The results of [71] can be observed:

1. Consider a low-pass filter with cut-off frequency $f_o/2$ or lower, as in Fig. A.3(b). One can see that only one component of the set of correlated components can fall in the window $[-f_o/2, f_o/2]$ that the filter allows to pass. Therefore any frequency components at the output of the filter are uncorrelated and the output noise is stationary.

2. Consider a single-sided bandpass filter, either upper band or lower band with respect to f_o, and bandwidth $f_o/2$ or less, as in Fig. A.3(c). After filtering, only one frequency component of the correlated set remains, and the resulting noise is stationary.

3. Consider a bandpass filter with center frequency f_o and bandwidth f_o or less, as in Fig. A.3(d). One can easily see that after filtering, the remaining correlated frequency components can only be in distance $2f_o$, and therefore only the stationary and the second-order cyclic spectra can exist.

Many other similar results can be visualized in a similar manner. For example if the filter is a low-pass filter with a cut off frequency f_o, the resulting process can contain only the stationary and first-order cyclic spectrum. A possible application of such a result as well as of result 3 above is the following: If it is known that the random signal at the output of mixer A in Fig. A.1 does not contain the n-th order cyclic spectrum, $k_a - k'_a$ in (A.2)-(A.5) cannot be equal to n.

In a receiver chain the first mixer is typically followed by a bandpass IF filter. In this case one can apply the following theorem, which can also be verified easily by inspection:

If cyclostationary noise with cycle frequency f_o passes through a bandpass filter with bandwidth $f_o/2$ or less, and the frequencies $k(f_o/2)$ where k is an integer do not fall into the passband, the output noise is stationary.

The latter has been used in [31] but here we clearly define the necessary properties of the filter passband. Results 1 and 2 above can be seen as individual cases of this last theorem.

A.2.3   Mixing a Band-Limited Cyclostationary Noise Process

In the previous section the passband characteristics of a filter following a mixer were related to the frequency of the LO waveform driving the mixer in order for the output noise signal to have certain properties. Here we will examine the case of Fig. A.4(a) in which a general cyclostationary signal for which we have no information about the location of the correlated frequency components, passes through a filter and the filter output is fed to a mixer (or more generally a time-varying circuit). We will relate the filter characteristics with the frequency f_o of the LO signal driving the mixer, in order for the time-average noise at the mixer output to be unaffected by the spectral correlation.

If the filter is low-pass with cut-off frequency $f_o/2$ or lower as shown in Fig. A.4(b), no overlap will take place during mixing, and the average noise at the output will not be affected by spectral correlation. This situation appears often at the back-end of a receiver where sampling (for example performed by a switched capacitor filter or an analog to digital converter) is proceeded by an anti-alias filter.

If the filter is bandpass with center frequency f_c and bandwidth w, as in Fig. A.4(c),
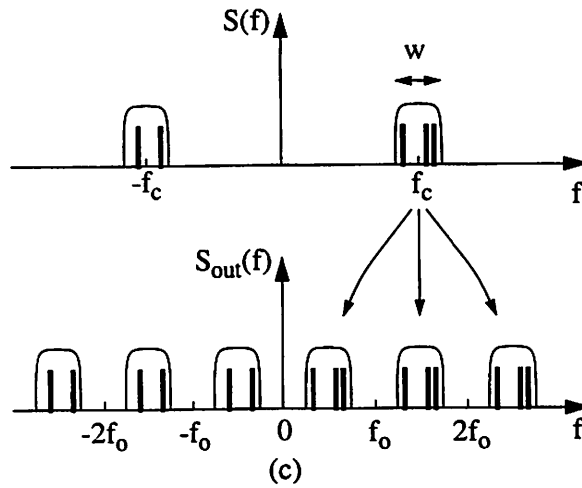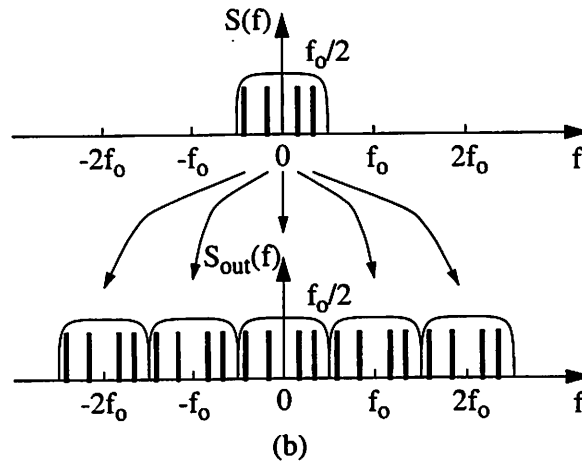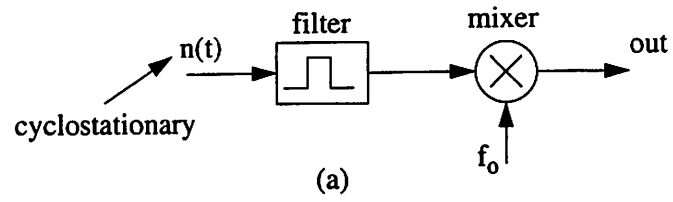
Figure A.4: Mixing band-limited cyclostationary noise.

one can see that overlap will not happen if

$$|(k - k')f_o + 2f_o| > w \qquad (\text{A.6})$$

for all mixer sidebands k and k' with some significant conversion gain. This results from the observation that the positive passband will be transferred to frequency bands with center $kf_o + f_c$ and width w, the negative passband will be transferred to frequency bands with center $k'f_o - f_c$ and width w, and to avoid overlap the centers of the two frequency bands must be in distance greater than w.

A.3 Two cases where spectral correlation is significant

A practical situation that deserves attention is when an interfering signal or blocker is present at the input of a receiver. If this signal is strong it can change the operating point of the devices and affect the circuit noise performance. The noise generated by the circuit will acquire cyclostationary characteristics with cycle equal to the blocker period, and if the blocker is not filtered or modulated to a different frequency, it acts as a common LO for successive cascaded blocks. In this case, although a block can still be characterized with the noise figure under the presence of a blocker, use of the formulas for cascaded blocks to estimate the noise figure of the whole receiver can lead to an inaccurate prediction. This situation could arise for example when an in-band blocker is processed together with the weak desirable signal by the LNA and the RF mixer of a receiver.

Let us consider now noise introduced to a mixer from the LO port. The LO is a periodically time-varying circuit and it is possible that the noise at its output contains some cyclostationary component. The time-varying processing of this signal by the mixer tracks exactly the time variation of the noise statistics since the mixer instantaneous operating point is determined by the LO drive. Therefore, it is not correct to time-average the noise

PSD at the LO output and use it as if it were a WSS process.

A.4   Conclusions

We examined qualitatively the significance of the cyclostationary nature of the noise generated in a communication system. We saw that cyclostationarity is equivalent to the presence of correlated components in the frequency spectrum. From the above discussion it results that in the majority of the practical cases, use of the concept of noise figure and considering only the time-average component of the cyclostationary noise at the input and the output of every block does not introduce significant inaccuracy in the noise characterization of the overall system for two reasons. First, the local oscillator frequencies used usually do not have a simple relation and the situation resembles the case at which the two frequencies are noncommensurate. Second, usually filtering takes place in several places in the receiver chain which converts the cyclostationary noise to stationary noise. However, we examined practical cases where cyclostationarity cannot be ignored, namely when the subsequent stage is time-varying synchronously with the cyclostationarity, as for example when the subsequent stage is driven nonlinearly by the stage generating cyclostationary noise. In these situations, if noise characterization is desirable by means of a circuit simulator which provides the time-average output noise the time-varying circuit blocks must be simulated together. Alternatively, the simulator of [71] can be used to calculate the cyclic spectra of every block separately and create appropriate macro-models which can then be used in a behavioral level simulation.

Appendix B

Time-varying Volterra series

Although the concept of the time-varying Volterra series have been used previously in the literature [83, 106, 49], the authors could not easily locate a proof. This appendix presents a derivation of this concept.

B.1 Taylor Expansion of a Functional

A functional is a function of a function. Consider the functional $F(y(t))$ which depends on the value of $y(t)$ over a time interval that we will assume to be $(-\infty, \infty)$. (For example $F(y(t)) = \int_a^b y(t)dt$ is a functional).

Theorem 1 *Consider $h(t)$ a small perturbation around $y(t)$. Then*

$$F(y(t) + h(t)) = F(y(t)) + \sum_{n=1}^{\infty} F_n(y(t), h(t)) \tag{B.1}$$

where

$$F_n(y(t), h(t)) = \frac{1}{n!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} k_n(\xi_1, \ldots, \xi_n) h(\xi_1) \ldots h(\xi_n) d\xi_1 \ldots d\xi_n \tag{B.2}$$

and

$$k_n(\xi_1, \ldots, \xi_n) \equiv \frac{\partial^n F(y(t))}{\partial y(\xi_1) \ldots \partial y(\xi_n)} \tag{B.3}$$

are the Volterra Kernels. *The concept of the Volterra kernels will be explained better below, during the proof of this theorem.*

Proof   Consider the function $f(\epsilon) = F(y(t) + \epsilon \cdot h(t))$. Taylor expansion in one dimension provides

$$f(\epsilon) = f(0) + f'(0)\epsilon + \frac{1}{2!}f''(0)\epsilon^2 + \frac{1}{3!}f'''(0)\epsilon^3 + \ldots \tag{B.4}$$

The Italian mathematician Vito Volterra (1860-1940) [99] calculated $f^{(k)}(0)$ by approximating the continuous-time functions $y(t)$ and $h(t)$ with the discrete-time functions, or vectors $\underline{y} = (y_1, \ldots, y_N)$ and $\underline{h} = (h_1, \ldots, h_N)$. If F is a function whose argument is a vector, and $f(\epsilon) = F(\underline{y} + \epsilon \underline{h})$ we know that

$$f'(\epsilon) = \sum_{i=1}^{N} \mathcal{D}_i \left[F(\underline{y} + \epsilon \underline{h}) \right] \cdot h_i \tag{B.5}$$

where $\mathcal{D}_i$ is the differentiation operator with respect to the $i - th$ argument of $F(\cdot)$. In analogy, in the continuous time case we have

$$f'(\epsilon) = \int_{-\infty}^{+\infty} D_\xi \left[F(y(t) + \epsilon h(t)) \right] h(\xi) d\xi \tag{B.6}$$

where $\mathcal{D}_\xi \left[F(y(t)) \right]$ represents differentiation of $F(y(t))$ with respect to $y(\xi)$:

$$D_\xi \left[F(y(t)) \right] = \frac{\partial F(y(t))}{\partial (y(\xi))} \equiv F^{(1)}(y(t); \xi) \tag{B.7}$$

Considering $F(y(t))$ as a function of infinitely many variables $y(t)$ indexed by t, $D_\xi \left[F(y(t)) \right]$ is the derivative with respect to one of those, $y(\xi)$. Since

$$f'(\epsilon) = \int_{-\infty}^{+\infty} F^{(1)}(y(t) + \epsilon h(t); \xi) h(\xi) d\xi \tag{B.8}$$

$$f'(0) = \int_{-\infty}^{+\infty} F^{(1)}(y(t); \xi) h(\xi) d\xi \qquad \text{(B.9)}$$

Now

$$f''(\epsilon) = \frac{d}{d\epsilon} f'(\epsilon) = \int_{-\infty}^{+\infty} \frac{d}{d\epsilon} F^{(1)}(y(t) + \epsilon h(t); \xi) h(\xi) d\xi \qquad \text{(B.10)}$$

but $F^{(1)}(y(t); \xi_1)$ is also a functional and using a similar definition for $F^{(2)}(y(t); \xi_1, \xi_2)$ we

obtain

$$f''(\epsilon) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F^{(2)}(y(t) + \epsilon h(t); \xi_1, \xi_2) h(\xi_1) h(\xi_2) d\xi_1 d\xi_2 \qquad \text{(B.11)}$$

hence

$$f''(0) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F^{(2)}(y(t); \xi_1, \xi_2) h(\xi_1) h(\xi_2) d\xi_1 d\xi_2 \qquad \text{(B.12)}$$

Working similarly we can find $f^{(k)}(0)$. Substituting in (B.4) and taking $\epsilon = 1$ we

obtain

$$
\begin{aligned}
F(y(t) + h(t)) &= F(y(t)) + \int_{-\infty}^{+\infty} F^{(1)}(y(t); \xi) h(\xi) d\xi \\
&+ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F^{(2)}(y(t); \xi_1, \xi_2) h(\xi_1) h(\xi_2) d\xi_1 d\xi_2 + \ldots \qquad \text{(B.13)}
\end{aligned}
$$

which is (B.1) and (B.2) for

$$k_n(\xi_1, \ldots, \xi_n) \equiv F^{(n)}(y(t); \xi_1, \ldots, \xi_n) \qquad \text{(B.14)}$$

The Volterra kernels remain unchanged after a permutation of their arguments. For example, $k_2(\xi_1, \xi_2) = k_2(\xi_2, \xi_1)$ because $F^{(2)}(y(t); \xi_1, \xi_2) = F^{(2)}(y(t); \xi_2, \xi_2)$ in the same way that

$$\frac{\partial^2 F(\underline{y})}{\partial y_i \partial y_j} = \frac{\partial^2 F(\underline{y})}{\partial y_j \partial y_i} \qquad \text{(B.15)}$$

where y_i and y_j are two components of the vector $\underline{y}$.

B.2 Volterra Series

The response of a time-varying system to an input $y(t)$ is a functional $F(t, y(t))$ which similar to this considered in the previous section, but it is also a function of time. The response $F(t, y(t))$ depends on t and on $y(t)$ for all $t \in (-\infty, +\infty)$. The results of the previous section apply, since dependence of the functional on t does not affect the proof of section B.1. The response of the system to a small perturbation $x(t)$ is

$$
\begin{aligned}
w(t) \equiv{}& F(t, y(t) + x(t)) - F(t, y(t)) \\
={}& \int_{-\infty}^{+\infty} k_1(t, \xi) x(\xi) d\xi \\
&+ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} k_2(t, \xi_1, \xi_2) x(\xi_1) x(\xi_2) d\xi_1 d\xi_2 \\
&+ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} k_3(t, \xi_1, \xi_2, \xi_3) x(\xi_1) x(\xi_2) x(\xi_3) d\xi_1 d\xi_2 d\xi_3 + \dots \quad (B.16)
\end{aligned}
$$

and let us denote the nth term of this sum by $w_n(t)$. Assume now that the perturbation is a sum of K single frequency tones

$$
x(t) = \sum_{i=1}^{K} a_i e^{j2\pi f_i t}. \tag{B.17}
$$

Then

$$
w_n(t) = \sum_{i_1=1}^{K} \dots \sum_{i_n=1}^{K} a_{i_1} \dots a_{i_n} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} k_n(t, \xi_1, \dots, \xi_n) e^{j2\pi(f_{i_1}\xi_1 + \dots + f_{i_n}\xi_n)} d\xi_1 \dots d\xi_n.
$$

$$
\tag{B.18}
$$

Consider now the transformation

$$
k_n(t, \xi_1, \dots, \xi_n) = b_n(t, t - \xi_1, \dots, t - \xi_n) \tag{B.19}
$$

Then

$$
w_n(t) = \sum_{i_1=1}^{K} \dots \sum_{i_n=1}^{K} a_{i_1} \dots a_{i_n} \cdot
$$

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} b_n(t, t - \xi_1, \ldots, t - \xi_n) e^{j2\pi(f_{i_1}\xi_1 + \ldots + f_{i_n}\xi_n)} d\xi_1 \ldots d\xi_n$$

$$= \sum_{i_1=1}^{K} \cdots \sum_{i_n=1}^{K} a_{i_1} \ldots a_{i_n} \cdot$$

$$\left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} b_n(t, u_1, \ldots, u_n) e^{-j2\pi(f_{i_1}u_1 + \ldots + f_{i_n}u_n)} du_1 \ldots du_n \right]$$

$$\cdot e^{j2\pi(f_{i_1} + \ldots + f_{i_n})t} \tag{B.20}$$

Finally

$$w_n(t) = \sum_{i_1=1}^{K} \cdots \sum_{i_n=1}^{K} a_{i_1} \ldots a_{i_n} B_n(t, f_{i_1}, \ldots, f_{i_n}) \cdot e^{j2\pi(f_{i_1} + \ldots + f_{i_n})t} \tag{B.21}$$

where $B_n(t, f_{i_1}, \ldots, f_{i_n})$ is the Fourier transform of $b_n(t, u_1, \ldots, u_n)$ with respect to $u_1, \ldots, u_n$.

Quantity $B_n(t, f_{i_1}, \ldots, f_{i_n})$ is the time-varying Volterra coefficient.

B.3   Time-invariant Systems

For the special case of a time-invariant system $F(t, y(t)) = F(t + \tau, y(t + \tau))$ for all values of τ. This means that

$$k_n(t, \xi_1, \ldots, \xi_n) = k_n(t + \tau, \xi_1 + \tau, \ldots, \xi_n + \tau) \tag{B.22}$$

or

$$b_n(t, t - \xi_1, \ldots, t - \xi_n) = b_n(t + \tau, t - \xi_1, \ldots, t - \xi_n) \tag{B.23}$$

for all τ and therefore $b_n(t, u_1, \ldots, u_n)$ does not depend on its first argument, and the Volterra coefficient is independent of time.

Appendix C

Flicker Noise

C.1 Introduction

Flicker noise is called a noise signal with frequency spectrum proportional to $1/f^\gamma$ with γ close to 1. Such a spectrum is present in many diverse signals, such as those which describe noise of active electronic devices and resistors, weather fluctuations, social effects, biological phenomena, music, etc. This observation led some researchers to seek a common underlying principle that generates flicker noise, but this attempt was unsuccessful [38]. The behavior of a signal with an $1/f$ spectrum is described as follows [38]: *$1/f$ noise combines the strong influence of the past events with the influence of the current events. The result is an overall context or pattern and somewhat predictable behavior but with the possibility of new trends developing and of occasional surprises.* The $1/f$ spectrum cannot be generated by filtering a white spectrum with a filter with a small number of poles. A system with infinite number of poles is required, such as the infinite RC transmission line whose input

impedance is

$$Z = \sqrt{\frac{R}{j2\pi fC}} \tag{C.1}$$

where R and C are the resistance and capacitance per unit length respectively.

A certain ambiguity arises about the fact that infinite energy is contained at the low frequency band of a random signal with an $1/f$ spectrum. Since the low frequencies are associated with long time constants of the generating system, some investigators have attempted to find an upper limit in the time constants that generate flicker noise in MOS-FETS, which would mean that the $1/f$ spectrum levels off at very low frequencies. No change in the $1/f$ shape was observed down to $10^{-6.3} Hz$ (1 cycle in 3 weeks) [38]. However, we always observe the signal over a finite window of time, and the observed signal has an $1/f$ spectrum down to the lowest frequency allowed by the limited observation time.

In this appendix we review the theories that have been developed to explain the flicker noise behavior of electronic devices and in particular of MOS transistors. These theories mainly cover the case that the operating point of the device is time-invariant. However in many practical analog circuits such as mixers, oscillators and switched capacitor filters the device operating point changes periodically with time. The last section of this appendix discusses this issue and summarizes the literature on this topic.

C.2 Theories for a Fixed Operating Point

Mainly two theories have been developed to explain and model flicker noise of MOS transistors. One of them is the *number fluctuation* (or *carrier fluctuation*, or *trapping*, or *Mc Whorter's*) theory and the other is the *mobility fluctuation* (or *Hooge's*) theory. Both

explain some measurement results but fail to explain some other. The two theories have been combined, and the resulting models seem to be able to match better experimental data.

C.2.1 Number Fluctuation Theory

The number fluctuation theory attributes flicker noise to random capturing and releasing of carriers from traps located in the $Si - SiO_2$ interface and in the oxide close to the interface [10, 5, 69, 23, 22, 109]. When the trapped charge changes, the channel charge which is responsible for the conduction changes also, and the drain current is affected. The current fluctuation caused by a single trapping-detrapping process, has a Lorentzian Power Spectral Density (PSD)

$$S_L(f) = \frac{c\tau}{1 + (2\pi f \tau)} \qquad (C.2)$$

where f is the frequency, c a constant associated with the amplitude of the fluctuation, and τ the mean time between two trapping events (time constant of the trap). The superposition of many Lorentzians with appropriately distributed time constants τ results in an $1/f$ power spectral density (PSD).

It can be proved (see [64] for a review of the number fluctuation theory) that if a) only oxide traps contribute to the flicker noise b) the oxide trap density N_t ($cm^{-3}ev^{-1}$) per volume unit and per energy unit is uniformly distributed in space and in energy, c) the mobility is assumed a constant, d) the transistor is biased in strong inversion and the linear region, e) trapping involves tunneling through the oxide and the time constant of an oxide trap increases exponentially with the distance from the interface, and f) the oxide

traps responsible for the flicker noise are located very close to the interface and therefore a change in the trapped charge causes an approximately equal change in the channel charge, then the PSD of the gate referred flicker noise is given by

$$S_{V_G} = \frac{\lambda T}{WLC_{ox}^2} \cdot \frac{N_t(E_{fn})}{f} \tag{C.3}$$

where λ is a constant, function of some physical constants, T the absolute temperature, W and L the effective dimensions of the transistor, C_{ox} the oxide capacitance per unit area, and E_{fn} the electron quasi-Fermi level in silicon. The PSD of the normalized current is

$$\frac{S_I}{I^2} = S_{V_G} \cdot \frac{g_m^2}{I^2} = \frac{\lambda T}{q^2 WLf} \cdot \frac{N_t(E_{fn})}{N^2} \tag{C.4}$$

where N is the number of carriers per unit area in the channel, and q the electron charge.

Since $N_t(E)$ is uniformly distributed in energy, it does not depend on the gate voltage and S_{V_G} is independent of bias. Such insensitivity to bias is actually observed mainly in NMOS transistors [95]. If there exists some nonuniformity in the distribution of the oxide traps with the distance from the interface, a dependence $1/f^\gamma$ on frequency is predicted by the above theory, with γ slightly different than 1. Indeed, such a spectrum is observed on MOSFET measurements with γ from 0.7 to 1.2 [8]. In addition, a nonuniformity in energy would result on a similar deviation on the noise spectrum because of oxide band bending even with a uniform distribution in space, and in this case γ is bias dependent.

Reimbold [69, 23] extended the trapping theory presented above to the weak inversion. He found that the ratio of the change δQ_n in the channel charge over the change δQ_t in the trapped charge that causes it, is given by

$$R = \left| \frac{\delta Q_n}{\delta Q_t} \right| = \frac{-\beta Q_n}{(C_D + C_{ox} + C_{it} - \beta Q_n)}, \tag{C.5}$$

where $\beta = q/kT$, k being Boltzman's constant, Q_n is the channel charge and C_D, C_{ox}, and C_{it} are the depletion region, oxide and interface trap capacitances per unit area respectively. This relation is true in the linear region for any gate bias. Using (C.5), (C.4) can be generalized as follows to give the normalized PSD of the drain current in the linear region of both strong and weak inversion

$$\frac{S_I}{I^2} = \frac{\beta^2 \lambda T}{WLf^\gamma} \cdot \frac{N_t(E_{fn})}{(C_D + C_{ox} + C_{it} - \beta Q_n)^2} \tag{C.6}$$

In strong inversion $|\beta Q_n| \gg |C_D + C_{ox} + C_{it}|$ and (C.4) results from (C.6) since $Q_n = Nq$. In weak inversion $|\beta Q n| \ll |C_D + C_{ox} + C_{it}|$ and (C.6) becomes

$$\frac{S_I}{I^2} = \frac{\beta^2 \lambda T}{WLf^\gamma} \cdot \frac{N_t(E_{fn})}{(C_D + C_{ox} + C_{it})^2}. \tag{C.7}$$

Since C_D varies very slowly with bias and C_{it} and $N_t(E_{fn})$ are generally weak functions of bias, (C.7) implies that S_I/I^2 is approximately bias independent. This plateau of S_I/I^2 versus bias is actually observed in measurements [69].

Equation (C.7) has been derived for the linear region in weak inversion but it can be shown [69] that it holds in the non linear region as well. Equation (C.4) is valid in linear region and strong inversion and can be modified in a manner described in [33] to model flicker noise in saturation.

A different formulation of the number fluctuation theory [109] relates the PSD of the fluctuation in flat-band voltage V_{fb} (or threshold voltage) with the PSD of the trapped charge fluctuation per unit area

$$S_{V_{fb}} = \frac{S_{N_t}}{WLC_{ox}^2}. \tag{C.8}$$
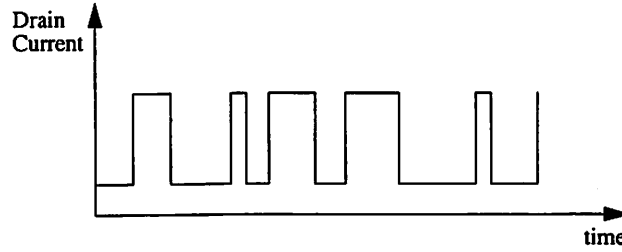
Figure C.1: A random telegraph signal.

The PSD of N_t is given by

$$S_{N_t} = \frac{\lambda N_t(E_{fn})}{f} \tag{C.9}$$

and relation (C.3) results.

The number fluctuation theory gains support from experiments with small dimension transistors ($area < 1mm^2$) with only one active trap [65, 95, 34]. In this case the resulting modulated drain current is a Random Telegraph Signal (RTS), such as this shown in Fig. C.1. The statistics of the RTS provide useful information about the trap characteristics.

Experiments with devices that have been stressed in order to change the trapping situation also attest to the number fluctuation theory. The stress is exercised by hot electron injection [6, 75, 9, 92], by a tunneling current through the oxide [47], or by radiation [16]. Measurements of flicker noise before and after stressing show that both the magnitude of the noise and the exponent γ are affected.

The trapping theory fails to explain the dependence of the gate referred noise on bias that is observed mainly on PMOS transistors [8].

There exists some confusion in the literature about whether the traps that cause

flicker noise are oxide or interface traps. All reported descriptions qualitatively agree with a picture of traps distributed continuously with the distance from the interface. The further a trap is located from the interface the harder it interacts with carriers and the longer its time constant is.

C.2.2 Mobility Fluctuation Theory

The mobility fluctuation theory was first developed for resistor flicker noise. Hooge [30] found empirically by examining homogeneous samples of semiconductors and metals that flicker noise is caused by fluctuation in the mobility of the material, independent of the current flowing. He found experimentally that the PSD of the mobility is given by

$$S_\mu(f) = \frac{\alpha_H}{f N_{tot}} \tag{C.10}$$

where N_{tot} is the total number of carriers that contribute to the conduction, and α_H is a constant with value approximately $2 \cdot 10^{-3}$ for all the materials he examined. The mobility fluctuations have been attributed to lattice scattering [30].

When this theory is applied to MOS transistors [29], predicts that the PSD of the normalized drain current variation in the linear region is given by

$$\frac{S_I}{I^2} = \frac{\alpha_H}{f W L N} \tag{C.11}$$

and the gate referred noise is

$$S_{V_G} = \frac{\alpha_H q^2}{f W L C_{ox}^2} N = \frac{\alpha_H}{f W L N} \cdot (V_{GS} - V_T)^2. \tag{C.12}$$

The effect of the bias dependent mobility has been taken into account in [42]. Equation (C.12) shows that according to the mobility fluctuation theory the gate referred noise is a

function of bias, since it is proportional to N. This contradicts the prediction of the number fluctuation theory expressed by equation (C.3).

A certain inconsistency arises about the value of Hooge's constant α_H when this theory is applied to MOSFETs. Experimental data indicate a value much smaller than $2 \cdot 10^{-3}$. In order to resolve this discrepancy, a modification in the value of this constant was suggested [30, 97, 96] as follows

$$\alpha'_H = \alpha_H \cdot \frac{\mu}{\mu_{ph}} \tag{C.13}$$

where α'_H is the new value of the constant and μ_{ph} is the mobility determined only by phonon scattering. Even this modification was not sufficient to always explain the value of this constant that results from measurements [28].

The mobility fluctuation theory explains the dependence of the gate referred flicker noise to bias that is observed mainly in PMOS transistors, but fails to explain the insensitivity of the gate referred noise to bias that is observed mainly in NMOS transistors [8]. If both mechanisms mobility fluctuation and trapping are responsible for flicker noise, it is reasonable to believe that trapping dominates in NMOS devices where the channel is in contact with the interface, while mobility fluctuation dominates in PMOS devices which are usually buried channel devices and an abundance of free carriers does not exist close to the interface for small gate bias.

The mobility fluctuation theory also fails to explain the weak inversion plateau of S_I/I^2 described in the number fluctuation theory. According to equation (C.11), S_I/I^2 should decrease strongly with increasing the gate voltage in the subthreshold region, since the number of carriers per unit area N increases rapidly.

C.2.3 Unified Theories

These theories [64, 35, 24] are based on the hypothesis that trapped charge fluctuation causes correlated mobility fluctuation. Experimental evidence for this effect comes from the RTS on the drain current of single trap transistors. It has been observed [65, 34] that the amplitude of the fluctuations cannot be explained only by the change of the channel charge by one carrier, and a correlated change in the mobility needs to be considered, which is attributed to oxide charge scattering. We will outline the derivation suggested by Hung et. al in [35]. The resulting model is simple and suitable for a circuit simulator.

The model starts with the assumption that the mobility is correlated with the number of trapped carriers per unit area, and is given by Matthiessen's rule [1] as

$$\frac{1}{\mu} = \frac{1}{\mu_n} + \frac{1}{\mu_{ox}} = \frac{1}{\mu_n} + \alpha N_t \qquad (C.15)$$

where μ is the total mobility, μ_{ox} is the mobility determined by oxide charge scattering, and μ_n is the mobility determined by other scattering mechanisms. The oxide charge scattering coefficient α is a function of the distance of the trap from the interface and an effective value is used in equation (C.15). Parameter α is also a function of the carrier density in the channel. Proceeding as in the number fluctuation theory, we finally obtain an expression for the PSD of the gate referred noise in strong inversion and linear region (low drain voltage)

$$S_{V_G} = \frac{\lambda T}{fWLC_{ox}^2} \cdot (1 + \alpha\mu N)^2 N_t(E_{fn}). \qquad (C.16)$$

[1]Matthienssen's rule states that if the mobility μ is determined by many independent mechanisms, it is given by

$$\frac{1}{\mu} = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \cdots + \frac{1}{\mu_n}, \qquad (C.14)$$

where $\mu_1, \mu_2, \ldots, \mu_n$ are the mobilities determined by each one of those mechanisms alone.

The PSD of the normalized drain current is

$$\frac{S_I}{I^2} = \frac{\lambda T}{q^2 W L f}(1 + \alpha \mu N)^2 \frac{N_t(E_{fn})}{N^2}.$$ (C.17)

Comparison of (C.17) and (C.16) with (C.3) and (C.3) shows that the expressions derived with the new theory are identical to those of the classical number fluctuation theory, if the trap density is replaced by

$$N_t^*(E_{fn}) = (1 + \alpha \mu N)^2 N_t(E_{fn}).$$ (C.18)

The gate referred noise is not bias independent even if $N_t(E_{fn})$ is assumed uniformly distributed in energy. Bias dependence comes from $N = C_{ox}(V_{GS} - V_T)/q$ which is assumed uniform along the channel, α which is a function of N, and the mobility which decreases with increasing gate voltage.

At low gate voltage, N is small and the model becomes identical to this of the number fluctuation theory. At high gate voltage the model resembles this of the mobility fluctuation theory with a (bias dependent), Hooge's constant

$$\alpha_H = \frac{\lambda T}{q^2}(\alpha \mu)^2 N N_t(E_{fn})$$ (C.19)

while an intermediate situation results for moderate gate voltage.

The suggested expression for the dependence of $N_t^*(E_{fn})$ on bias is

$$N_t^*(E_{fn}) = A + BN + CN^2$$ (C.20)

where A,B and C are technology dependent process parameters. It is emphasized that the implementation of the new unified model in a circuit simulator does not differ from

the implementation of the classical number fluctuation theory with a bias dependent trap density, given by (C.20).

The model is extended in [33] to any gate bias using Reimbold's ratio R from (C.5), as well as to the nonlinear region and saturation. The bias dependent mobility is also taken into account.

A similar theory is developed in [64]. The derivation there starts with the assumption that the mobility is linearly dependent on the channel charge instead of the trapped charge. The dependence of the scattering coefficient on the distance from the interface is preserved throughout the derivation. This does not result in a simple model as in [35], but it leads to useful conclusions about the influence of the distance dependent scattering coefficient on the noise spectrum. Because the scattering coefficient decreases with distance, a $1/f^\gamma$ spectrum with $\gamma < 1$ results, even assuming uniform oxide trap distribution in space and energy and in addition γ decreases with increasing gate voltage. A technique is developed to extract the profile of the oxide traps with the distance from the interface and with energy.

At last [24], presents a re-derivation of the model developed in [35] using a the fluctuation of the flat-band voltage caused by the oxide traps.

C.3 Flicker Noise of Devices with Time-Varying Operating Point

There is a traditional belief that flicker noise is associated only with the DC current flowing through a device [61], and is not present unless a DC current is present. However,

experiments on resistors by Bull and Bozic [7] showed that flicker noise on resistors is generated by fluctuation in the value of the resistance itself, independent of the current flowing. When an AC current is applied, an $1/f$ spectrum appears on the spectrum of the voltage across the resistor, around all the frequency components of the excitation current, which is called $1/\Delta f$ noise. A DC current component would only sense the resistance fluctuation and generate an $1/f$ spectrum around DC. Since then, a few publications have followed [46, 43, 51, 81, 80] on this subject for resistors. Some researchers observed a small $1/f$ spectrum around DC when only AC excitation was present, but this has been disputed as a measurement error. Bull and Bozic also observed $1/\Delta f$ noise on diodes and bipolar transistors, but there existed some doubt if this noise was similar to the $1/\Delta f$ noise of the resistors, or simply the result of frequency distortion.

There exists a belief that since the flicker noise generating mechanisms in MOS-FETs have time constants much longer than the period of the operating point variation, their flicker noise depends on some effective bias value. It is not however clear what this effective value is. It could for example be the average gate-source voltage, the average drain current, or some other value. Theoretical and experimental work needs to be done to illuminate this aspect of the CMOS device model.

Bibliography

[1] A. A. Abidi. High-frequency noise measurements on fet's with small dimensions. *IEEE Transactions on Electron Devices*, ED-33(11), November 1986.

[2] R. Adler, H. A. Haus, R. S. Engelbrecht, M. T. Lebenbaum, S. W. Harrison, and W. W. Mumford. Description of the noise performance of amplifiers and receiving systems. *Proceedings of the IEEE*, pages 436–442, March 1963.

[3] V. Aparin and C. Persico. Effect of out-of-band terminations on intermodulation distortion in common-emitter circuits. In *IEEE MTT-S International Microwave Symposium Digest*, volume 3, pages 977–980, June 1999.

[4] Several Authors. Representation of noise in linear twoports. *Proceedings of the IRE*, pages 69–74, January 1960.

[5] F. Berz. Theory of low frequency noise in si mosts. *Solid State Electronics*, 13:631–647, 1970.

[6] B. Boukriss, H. Haddara, S. Cristoloneanu, and A. Chovet. Modeling of the 1/f noise overshoot in short-channel mosfet's locally degraded by hot-carrier injection. *IEEE Electron Devices Letters*, 10(10), October 1989.

[7] C. S. Bull and S. M. Bozic. Excess noise in semiconductor devices due to fluctuations in their characteristics when signals are applied. *Brittish Journal of Applied Physics*, 18:883–895, 1967.

[8] J. Chang, A. A. Abidi, and C. R. Viswanathan. Flicker noise in cmos transistors from subthershold to strong inversion at various temperatures. *IEEE Transactions on Electron Devices*, 41(11), November 1994.

[9] C. H. Cheng and C. Surya. The effect of hot-electron injection on the properties of flicker noise in n-channel mosfets. *Solid State Electronics*, 36(3):475–479, 1993.

[10] S. Christensson, I. Lundstrom, and C. Svensson. Low frequency noise in mos transistors - i theory. *Solid State Electronics*, 11:797–812, 1968.

[11] J. Crols and M. S. J. Steyaert. A single-chip 900 mhz cmos receiver front-end with a high performance low-if topology. *IEEE Journal of Solid-State Circuits*, 30:1483–1492, December 1995.

[12] J. Crols and M. S. J. Steyaert. Low-if topologies for high-performance analog front ends of fully integrated receivers. *IEEE Transactions on Circuits and Systems - II*, 45:269–282, March 1998.

[13] J. Crols and S. J. Steyaert. A 1.5-ghz highly linear cmos downconversion mixer. *IEEE Journal of Solid-State Circuits*, 30:736–742, July 1995.

[14] A. Demir. *Analysis and Simulation of Noise in Nonlinear Electronic Circuits and Systems*. PhD thesis, UC Berkeley, Berkeley, CA, 1997.

[15] Patrik Eriksson and Hannu Tenhunen. The noise figure of a sampling mixer: Theory and measurement. In *International Conference on Electronics, Circuits and Systems*, pages 899–902, 1999.

[16] D. M. Fleetwood, P. S. Winokur, Jr R. A. Reber, T. L. Meisenheimer, J. R. Schwank, M. R. Shaneyfelt, and L. C. Riewe. Effects of oxide traps, inteface traps and 'border traps' on metal-oxide-semiconductor devices. *Journal of Applied Physics*, 73(10), May 1993.

[17] K. L. Fong and R. G. Meyer. High-frequency nonlinearity analysis of common-emmiter and differential pair transconductance stages. *IEEE Journal of Solid-State Circuits*, 33:548–555, April 1998.

[18] H. T. Friss. Noise figures of radio receivers. *Proceedings of the IRE*, pages 419–422, July 1944.

[19] W. A. Gardner. Stationarizable random processes. *IEEE Transactions on Information Theory*, 24(1):8–22, January 1978.

[20] W. A. Gardner. Common pitfalls in the application of stationary process theory to time-sampled and modulated signals. *IEEE Transactions on Communications*, 35(5):529–534, May 1987.

[21] W. A. Gardner. *Introduction to Random Processes with Applications to Signals and Systems*. McGraw Hill, second edition, 1989.

[22] G. Ghibaudo. Calculation of surface charge noise at the $si - sio_2$ interface. *Phys. Stat. Sol. (a)*, 104:917, 1987.

[23] G. Ghibaudo. A simple derivation of reimbold's drain current spectrum formula for flicker noise in mosfets. *Solid State Electronics*, 30(10):1037–1038, 1987.

[24] G. Ghibaudo, O. Roux, Ch. Nguen-Duc, F. Balestra, and J. Brini. Improved analysis of low frequency noise in field-effect mos transistors. *Phys. Stat. Sol. (a)*, 124:571, 1991.

[25] B. Gilbert. A precise four quadrant multiplier with subnanosecond response. *IEEE Journal of Solid-State Circuits*, pages 365–373, December 1968.

[26] B. Gilbert. The micromixer: a highly linear variant of the gilbert mixer using a bisymmetric class-ab input stage. *IEEE Journal of Solid-State Circuits*, pages 1412–1423, September 1997.

[27] P. R. Gray and R. G. Meyer. Future directions in silicon ics for personal communications. In *Custom Integrated Circuits Conference*, 1995.

[28] S. A. Hayat and B. K. Jones. 1/f noise in mos inversion layers. *Noise in Physical Systems and 1/f Noise - 1985, A. D'Amico and P. Mazzetti Editors Amsterdam: North-Holland*, 1986.

[29] F. N. Hooge. 1/f noise. *Physica*, 83B:14, 1976.

[30] F. N. Hooge and L. K. J. Vandamme. Lattice scattering causes 1/f noise. *Physics Letters*, 66A:315, 1978.

[31] C. Hull. *Analysis and Optimization of Monolithic RF Downconversion Receivers*. PhD thesis, University of California, Berkeley, 1992.

[32] C. D. Hull and R. G. Meyer. A systematic approach to the analysis of noise in mixers. *IEEE Transactions on Circuits and Systems - I*, 40:909–919, December 1993.

[33] K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng. A physics-based mosfet noise model for circuit simulators. *IEEE Transcations on Electron Devices*, 37(5), May 1990.

[34] K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng. Random telegraph noise of deep-submicrometer mosfets. *IEEE Electron Devices Letters*, 11(2), February 1990.

[35] K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng. A unified model for flicker noise in metal-oxide-semiconductor field-effect-transistors. *IEEE Transcations on Electron Devices*, 37(3), March 1990.

[36] R. P. Jindal. High frequency noise in fine line nmos field effect transistors. In *IEDM*, pages 68–71, 1985.

[37] A. N. Karanicolas. A 2.7-v 900-mhz cmos lna and mixer. *IEEE Journal of Solid-State Circuits*, pages 1939–1944, December 1996.

[38] M. S. Keshner. 1/f noise. *Proceedings of the IEEE*, 70(3):212–218, March 1982.

[39] C. D. Keys. *Low Distortion Mixers for RF Communications*. PhD thesis, UC Berkeley, Berkeley, CA, 1994.

[40] P. R. Kinget and S. J. Steyaert. A 1-ghz cmos up-conversion mixer. *IEEE Journal of Solid-State Circuits*, 32:370–376, March 1997.

[41] F. M. Klaasen. High-frequency noise of the junction field-effect transistor. *IEEE Transactions on Electron Devices*, 14:368–373, 1967.

[42] T. G. M. Kleinpenning and L. K. J. Vandamme. Model for 1/f noise in metal-oxide-semiconductor transistors. *Journal of Applied Physics*, 52(3), March 1981.

[43] A. Kumar and W. I. Goldburg. Unsuccessful attempt to observe unrectified noise and nonlinear behavior in various resistances. *Applied Physics Letters*, 39(1), July 1981.

[44] K. Kundert and A. Sangiovanni-Vincentelli. *Sparse 1.3, A Sparse Linear Equation Solver.* available at www.netlib.org/sparse.

[45] Thomas H. Lee. *The Design of CMOS Radio-Frequency Integrated Circuits.* Cambridge University Press, 1998.

[46] J. H. Lorteije and A. M. H. Hoppenbrouwers. Amplitude modulation by 1/f noise in resistors results in $1/\delta f$ noise. *Philips Res. Repts*, 26:29–39, 1971.

[47] Z. J. Ma, H. Shin, P. K. Ko, and C. Hu. Effects of plasma charging damage on the noise performance of thin-oxide mosfets. *IEEE Electron Devices Letters*, 15(6), June 1994.

[48] S. Maas. Two-tone intermodulation in diode mixers. *Transactions on Microwave Theory and Techniques*, MTT-35, March 1987.

[49] S. Maas and D. Neilson. Modeling mesfet's for intermodulation analysis of mixers and amplifiers. *Transactions on Microwave Theory and Techniques*, MTT-38(12), December 1990.

[50] Stephen A. Maas. *Microwave Mixers.* Artech House, second edition, 1993.

[51] E. J. P. May and H. H. Mehdi S. H. Al-Charchafchi. 1/f and $1/\delta f$ noise in carbon-loaded polyurethane. *Noise in Physical Systems and 1/f Noise - 1985, A.D'Amico and P. Mazetti editors*, 1985.

[52] T. Melly, A.-S. Porret, C. C. Enz, and E. A. Vittoz. An analysis of flicker noise rejection in low-power and low-voltage cmos mixers. *IEEE Journal of Solid-State Circuits*, pages 102–109, January 2001.

[53] R. G. Meyer. Noise in transistor mixers at low frequencies. *Proceedings of the IRE*, 114:611–618, May 1967.

[54] R. G. Meyer. Signal processes in transistor mixer circuits at high frequencies. *Proceedings of the IRE*, 114:1605–1612, November 1967.

[55] R. G. Meyer. Noise in transistor mixers at high frequencies. *Proceedings of the IRE*, 115:487–495, April 1968.

[56] R. G. Meyer. Intermodulation in high-frequency bipolar transistor integrated-circuit mixers. *IEEE Journal of Solid-State Circuits*, pages 534–537, August 1986.

[57] R. G. Meyer. Ee242, advanced integrated circuits for communications. In *Class Notes, University of California, Berkeley*, 1996.

[58] R. G. Meyer and A. K. Wong. Blocking and desensitization in rf amplifiers. *IEEE Journal of Solid-State Circuits*, 30:944–946, August 1995.

[59] A. M. Niknejad. *Analysis, SImulation and Applications of Passive Devices on Conductive Substrates*. PhD thesis, UC Berkeley, Berkeley, CA, 2000.

[60] P. Orsatti, F. Piazza, and Q. Huang. A 20-ma-receive, 55-ma-transmit, single-chip gsm transceiver in 0.25-μm cmos. *IEEE Journal of Solid-State Circuits*, pages 1869–1880, December 1999.

[61] R. G. Meyer P. R. Gray. *Analysis and Design of Analog Integrated Circuits*. John Wiley, third edition, 1993.

[62] David M. Pozar. *Microwave Engineering*. John Wiley, second edition, 1998.

[63] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipies in C; The Art of Scientific Computing*. Cambridge University Press, second edition, 1992. pp. 650–655.

[64] C. G. Sodini R. Jayaraman. A 1/f noise technique to extract the oxide trap density near teh conducation band edge of silicon. *IEEE Transcations on Electron Devices*, 36(9), September 1989.

[65] K. S. Ralls and et. al W. J. Skocpol. Discrete resistance in submicrometer silicon inversion layers: Individual interface traps and low frequency (1/f?) noise. *Physical Review Letters*, 52(3):228, 1984.

[66] B. Razavi. A 2.4-ghz cmos receiver for ieee 802.11 wireless lan's. *IEEE Journal of Solid-State Circuits*, pages 1382–1385, October 1999.

[67] B. Razavi, R.-H. Yan, and K. F. Lee. Impact of distributed gate resistance on the performance of mos devices. *IEEE Transactions on Circuits and Systems - I*, 41(11), November 1994.

[68] Behzad Razavi. *RF Microelectronics*. Prentice Hall, 1998.

[69] G. Reimbold. Modified 1/f trapping noise theory and experiments in mos transistors biased from weak to strong inversion - influence of interface states. *IEEE Transcations on Electron Devices*, ED-31(9), September 1984.

[70] H. Rothe and W. Dahlke. Theory of noisy fourports. *Proceedings of the IRE*, pages 811–818, June 1956.

[71] J. Roychowdhury, D. Long, and P. Feldman. Cyclostationary noise analysis of large rf circuits with multi-tone excitations. *IEEE Journal of Solid-State Circuits*, 33(3):324–336, March 1998.

[72] J. C. Rudell, J.-J. Ou, T. B. Cho, G. Chien, F. Brianti, J. A. Weldon, and P. R. Gray. A 1.9 ghz wide-band if double conversion cmos receiver for cordless telephone applications. *IEEE Journal of Solid-State Circuits*, 10(9):2071–2088, December 1997.

[73] D. K. Shaeffer and T. H. Lee. A 1.5v, 1.5ghz cmos low noise amplifier. *IEEE Journal of Solid-State Circuits*, 32:745–759, May 1997.

[74] A. R. Shahani, D. K. Shaeffer, and T. H. Lee. A 12-mw wide dynamic range cmos front-end for a portable gps receiver. *IEEE Journal of Solid-State Circuits*, 32:2061–2069, December 1997.

[75] M. Stegherr. Flicker noise in hot electron degraded short channel mosfets. *Solid State Electronics*, 27(12):1055–1056, 1984.

[76] T. Strom and S. Signell. Analysis of periodically switched linear circuits. *IEEE Transactions on Circuits and Systems*, CAS-24(10), October 1977.

[77] P. J. Sullivan, W. H. Ku, and B. A. Xavier. Active doubly balanced mixers for cmos rfics. *Microwave Journal*, October 1997.

[78] P. J. Sullivan, B. A. Xavier, and W. H. Ku. Doubly balanced dual-gate cmos mixer. *IEEE Journal of Solid-State Circuits*, 34:878–881, June 1999.

[79] P. J. Sullivan, B.A. Xavier, and W. H. Ku. Low voltage performance of microwave cmos gilbert cell mixer. *IEEE Journal of Solid-State Circuits*, pages 1151–1155, July 1997.

[80] H. Sutcliffe and Y. Ulgen. 1/f and $1/\delta f$ noise produced by radio-frequency current in carbon resistor. *Electronics Letters*, 12(4), February 1976.

[81] H. Sutcliffe and Y. Ulgen. Spectra of ac-induced noise in resistors. *Electronics Letters*, 13(14), July 1977.

[82] H. L. Swain and R. M. Cox. Noise figure meter sets records for accuracy, repeatability and convenience. *Hewlett-Packard Journal*, April 1983.

[83] R. B. Swerdlow. Analysis of intermodulation noise in frequency converters by volterra series. *Transactions on Microwave Theory and Techniques*, MTT-26(4), April 1978.

[84] K. Takagi and A. Van der Ziel. Excess high frequency noise and flicker noise in mosfets. *Solid-State Electronics*, 22:289–292, 1979.

[85] S. Tedja, J. Van der Spiegel, and H. H. Williams. Analytical and experimental studies of thermal noise in mosfets. *IEEE Transactions on Electron Devices*, 41(11), November 1994.

[86] R. Telichevesky and K. Kundert. Efficient ac and noise analysis of two-tone rf circuits. In *Design Automation Conference*, pages 292–297, 1995.

[87] R. Telichevesky, K. Kundert, I. Elfadel, and J. White. Fast simulation algorithms for rf circuits. In *Custom Integrated Circuits Conference*, 1996.

[88] R. Telichevesky, K. Kundert, and J. White. Receiver charactrization using periodic small-signal analysis. In *Custom Integrated Circuits Conference*, 1996.

[89] M. T. Terrovitis and R. G. Meyer. Noise in current-commutating cmos mixers. *IEEE Journal of Solid-State Circuits*, pages 772–783, June 1999.

[90] M. T. Terrovitis and R. G. Meyer. Intermodulation distortion in current-commutating cmos mixers. *IEEE Journal of Solid-State Circuits*, pages 1461–1473, October 2000.

[91] D. P. Triandis, A. N. Birbas, and D. Kondis. Thermal noise modeling for short-channel mosfet's. *IEEE Transactions on Electron Devices*, 43(11), November 1996.

[92] M. H. Tsai and T. P. Ma. 1/f noise in hot-carrier damaged and mosfets: Effects of oxide charge and interface traps. *IEEE Electron Devices Letters*, 14(5), May 1993.

[93] Y. Tsividis, K. Suyama, and K.Vavelidis. Simple 'reconciliation' mosfet model valid in all regions. *Electronics Letters*, pages 506–508, March 1995.

[94] University of California, Berkeley. *BSIM3 Version 3 manual, Final Version*, 1995. available at http://www-device.EECS.Berkeley.EDU/~bsim3.

[95] M. J. Uren, D. J. Day, and M. J. Kirton. 1/f and random telegraph noise in silicon metal-oxide-semiconductor field-effect-transistors. *Applied Physics Letters*, 47(11), December 1985.

[96] L. K. J. Vandamme. 1/f noise model for most biased in nonohmic region. *Solid-State Electronics*, 23:325–329, 1980.

[97] L. K. J. Vandamme. Model for 1/f noise in mos transistors biased in the linear region. *Solid-State Electronics*, 23:317–323, 1980.

[98] G. D. Vendelin, A. M. Pavio, and U. L. Rohde. *Microwave Circuit Design, Using Linear and Nonlinear Techniques*. John Wiley & Sons, third edition, 1990.

[99] V. Volterra. *Theory of Functionals and of Integral and Integro-Differential Equations*. New York: Dover, 1959.

[100] D. K. Waever. A third method of generation and detection of single-sideband signals. *Proceedings of the IRE*, 44:1703–1705, December 1956.

[101] P. Wambacq and W. Sansen. *Distortion Analysis of Analog Integrated Circuits*. Kluer Academic Publishers, 1998.

[102] B. Wang, J. R. Hellums, and C. G. Sodini. Mosfet thermal noise modeling for analog integrated circuits. *IEEE Journal of Solid-State Circuits*, pages 833–835, July 1994.

[103] HongMo Wang. A 1-v multigigahertz rf mixer core in 0.5-μm cmos. *IEEE Journal of Solid-State Circuits*, 33:2265–2267, December 1998.

[104] D. D. Weiner and J. E. Spina. *Sinusoidal Analysis and Modeling of Weakly Nonlinear Circuits*. Van Nostrand Reinhold Company, 1980.

[105] S. Wu and B. Razavi. A 900-mhz/1.8-ghz cmos receiver for dual-band applications. *IEEE Journal of Solid-State Circuits*, pages 2178–2185, December 1998.

[106] Wei Yu, Subhajit Sen, and Bisco H. Leung. Distortion analysis of mos track-and-hold sampling mixers using time-varying volterra series. *IEEE Transactions on Circuits and Systems - II*, 46:101–113, February 1999.

[107] L. Zadeh. Frequency analysis of variable networks. *Proceedings of the IRE*, pages 291–299, March 1950.

[108] A. Van Der Ziel. Gate noise in field effect transistors at moderately high frequencies. *Proceedings of the IEEE*, March 1963.

[109] A. Van Der Ziel. Theory of flicker noise in mosfets. *University/Industry/Government Microelectronics Symposium*, pages 102–106, 1979.