Linearization Techniques for CMOS Low Noise Amplifiers: A Tutorial

Heng Zhang, *Student Member, IEEE*, and Edgar Sánchez-Sinencio, *Life Fellow, IEEE*

(Invited Paper)

***Abstract*—This tutorial catalogues and analyzes previously reported CMOS low noise amplifier (LNA) linearization techniques. These techniques comprise eight categories: a) feedback; b) harmonic termination; c) optimum biasing; d) feedforward; e) derivative superposition (DS); f) IM2 injection; g) noise/distortion cancellation; and h) post-distortion. This paper also addresses broadband-LNA-linearization issues for emerging reconfigurable multiband/multistandard and wideband transceivers. Furthermore, we highlight the impact of CMOS technology scaling on linearity and outline how to design a linear LNA in a deep submicrometer process. Finally, general design guidelines for high-linearity LNAs are provided.**

***Index Terms*—Broadband, CMOS technology scaling, derivative superposition, design guidelines, distortion canceling, feedback, feedforward, harmonic termination, IIP2, IIP3, IM2 injection, linearity, linearization, low noise amplifier (LNA), , multiband, multistandard, noise canceling, 1 dB compression point, optimum biasing, post-distortion, tutorial.**

I. Introduction

THE plethora of wireless communication standards employed in a single geographic region and moreover occupying narrow frequency bands tightly constrains RF-system linearity. Furthermore, the trend in radio research is to simplify/eliminate the expensive front-end module (FEM), which demands a highly linear receiver. In particular, since the low noise amplifier (LNA) is the first block in the receiver chain, it must be sufficiently linear to suppress interference and maintain high sensitivity.

LNA linearization methods should be simple, should consume minimum power, and should preserve noise figure (NF), gain, and input matching. Many traditional linearization techniques are not feasible for LNAs. For example, resistive source degeneration and floating-gate input attenuation reduce the gain and worsen NF or input matching. Hence, LNA linearization proves significantly more challenging than that of baseband circuits [1], often requiring innovative techniques.

Growing research on reconfigurable multiband/multistandard and broadband transceivers such as ultrawideband (UWB) and digital TV tuners has propelled interest in broadband LNA design. Radios in the same platform interfere with each other, and multiple channels applied simultaneously to an LNA without filtering act as in-band interferences. Consequently, broadband LNAs must maintain sufficient linearity over a wide frequency range. Emphasis on highly linear transceivers has sparked recent interest in linearizing LNAs [2]. Even though most previously reported techniques target narrowband applications and principally improve only the third-order intercept point (IIP3), we demonstrate why broadband systems require high second-order intercept point (IIP2) and 1 dB compression point $(P_{1\,\mathrm{dB}})$ as well. Because a broadband LNA is exposed to a wide frequency range, we investigate the dependence of IIP2/IIP3 on two-tone (center) frequency and frequency spacing.

Since LNAs typically have low-amplitude, high frequency inputs, the amplifier operates as a weakly nonlinear system with few relevant harmonics (typically only second and third). Thus, Volterra-series analysis [3] can capture the frequency-dependent distortion of LNAs and provide insight into how to compensate that distortion.

CMOS is the most promising technology for systems on a chip. Although MOSFETs are intrinsically more linear than bipolar transistors [4], they require higher dc current to achieve the necessary transconductance and linearity, thus linearization techniques must be employed to reduce the dc power. Deep submicrometer (DSM) technology challenges include nonlinear output conductance, mobility degradation, velocity saturation, and poly-gate depletion; which complicate CMOS LNA linearization, especially in the face of low supply voltages. We present multidimensional Taylor analysis to evaluate the effects of these nonidealities.

This paper is organized as follows. Section II analyzes previously reported CMOS LNA linearization techniques. Section III discusses new broadband-LNA-linearization issues arising in multiband/multistandard/wideband transceivers. Section IV investigates the impact of CMOS technology scaling on linearity, and provides insights into the design of linear LNAs in DSM processes. Remarks for high linearity LNA design are provided in Section V, and Section VI concludes the paper.

II. Linearization Techniques

A weakly nonlinear amplifier with input X and output Y can be approximated by the first three power series terms

$$\mathrm{Y} = \mathrm{g}_1\mathrm{X} + \mathrm{g}_2\mathrm{X}^2 + \mathrm{g}_3\mathrm{X}^3 \tag{1}$$

where $\mathrm{g}_{1,2,3}$ are the linear gain and the second/third-order nonlinearity coefficients of the amplifier, respectively. The goal of linearization is to make $\mathrm{g}_{2,3}$ small enough to be negligible, keeping only the linear term g_1, hence $\mathrm{Y} \approx \mathrm{g}_1\mathrm{X}$. The purpose

Manuscript received December 10, 2009; revised April 20, 2010; accepted June 02, 2010. Date of current version December 30, 2010. This paper was recommended by Editor-in-Chief W. A. Serdijn.

H. Zhang was with the Analog and Mixed-Signal Center, Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843-3128 USA. She is now with Broadcom Corporation, Irvine, CA 92617 USA (e-mail: heng@broadcom.com).

E. Sánchez-Sinencio with the Analog and Mixed-Signal Center, Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843-3128 USA (e-mail: sanchez@mail.ece.tamu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSI.2010.2055353

TABLE I
OVERVIEW OF DISTORTION SOURCES AND LINEARIZATION TECHNIQUES

Distortion Sources / Linearization Methods	g_m				g_{ds}
	Intrinsic 2nd-order	Intrinsic 3rd-order	2nd-order interaction	Higher order	
Feedback	√	√		√	
Harmonic termination		√	√		
Optimal biasing		√			
Feedforward	√	√		√	
Derivative superposition(DS)		√			
Complementary DS	√	√			
Differential DS	√	√			
Modified DS		√	√		
IM2 injection		√	√		
Noise/distortion cancellation	√	√			√
Post-distortion	√	√			

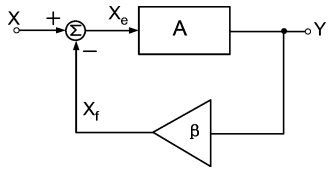


Fig. 1. Nonlinear amplifier with negative feedback.

of this paper is to discuss the main linearization techniques for LNAs.

LNA nonlinearity originates from two major sources:

1) nonlinear transconductance g_m, which converts linear input voltage to nonlinear output drain current; this effect is also termed "input limited";
2) nonlinear output conductance g_{ds}, whose effect becomes apparent under large output voltage swing and small drain-source voltage V_{ds} (i.e., when the device operates near linear region); also referred to as "output limited."

The MOSFET capacitances (the gate-source capacitance C_{gs}, the gate-drain capacitance C_{gd}, the drain-bulk capacitanceC_{db}) are roughly linear when the transistor operates in the saturation region, and when the frequency is less than $f_T/10$ [44]. Thus, for the most part, the capacitors contribute less distortion than g_m/g_{ds} [43]; however, C_{gd} influences the linearity indirectly through feedback, which will be discussed later.

The IIP3 is degraded by both the intrinsic third-order distortion and the "second-order interaction" (caused by intrinsic second-order distortion combined with feedback), while IIP2 originates from intrinsic second-order distortion.

We categorize previously reported CMOS LNA linearization techniques into eight clusters: a) feedback; b) harmonic termination; c) optimum biasing; d) feedforward; e) derivative superposition (DS); f) IM2 injection; g) noise/distortion cancellation; and h) post-distortion. Note that DS, IM2 injection, and noise/distortion cancellation are special cases of the feedforward technique.

Table I illustrates the distortion sources and the corresponding linearization methods. Most of the reported linearization techniques focus on suppressing second- and third-order distortion of transconductance. Therefore, linearization of higher order terms (beyond third order) and output conductance still remains an open problem.

A. Feedback

Fig. 1 shows the negative feedback scheme with a nonlinear amplifier A and a linear feedback factor β, where X and Y are the input and output signals, respectively. X_f is the feedback signal, and X_e is the difference between X and X_f.

Assuming the nonlinear amplifier A can be modeled by (1), we obtain the third-order closed-loop power series for Y

$$Y = b_1 X + b_2 X^2 + b_3 X^3 \tag{2}$$

$$b_1 = \frac{g_1}{1+T_0}$$

$$b_2 = \frac{g_2}{(1+T_0)^3}$$

$$b_3 = \frac{1}{(1+T_0)^4}\left(g_3 - \frac{2g_2^2}{g_1}\frac{T_0}{1+T_0}\right). \tag{3}$$

where $b_{1,2,3}$ are the closed-loop linear gain and second/third-order nonlinearity coefficients, respectively, and $T_0 = g_1\beta$ is the linear open-loop gain. The IIP2 and IIP3 [5] of the amplifier A and the closed loop system are

$$A_{IIP2,amplifier} = \sqrt{\frac{g_1}{g_2}} \tag{4a}$$

$$A_{IIP2,closeloop} = \sqrt{\left|\frac{b_1}{b_2}\right|} = \sqrt{\frac{g_1}{g_2}(1+T_o)^2} \tag{4b}$$

$$A_{IIP3,amplifier} = \sqrt{\frac{4}{3}\left|\frac{g_1}{g_3}\right|} \tag{4c}$$

$$A_{IIP3,closeloop} = \sqrt{\frac{4}{3}\left|\frac{b_1}{b_3}\right|} = \sqrt{\frac{4}{3}\frac{g_1}{g_3}\frac{(1+T_o)^3}{\left(1-\frac{2g_2^2}{g_1 g_3}\frac{T_o}{1+T_o}\right)}}. \tag{4d}$$

Hence, negative feedback improves A_{IIP2} by a factor of $(1+T_0)$; it also improves A_{IIP3} by a factor of $(1+T_0)^{3/2}$ when $g_2 \approx 0$. As shown by (4d), nonzero g_2 degrades IIP3 when g_1 and g_3 have opposite signs (this is the case for typical MOSFET biases). This phenomenon is called "second-order interaction" [6]. In other words, whenever the amplifier is in feedback, the third-order nonlinearity originates from two sources:

1) intrinsic amplifier third-order nonlinearity;
2) "second-order interaction" (originated from intrinsic second-order nonlinearity of the amplifier combined with feedback).

However, feedback linearity improvement is not as effective for LNAs as for baseband circuits because:

1) the open loop gain T_0 cannot be large due to stringent LNA gain, noise, and power requirement;
2) the second-order nonlinearity contributes to the IM3 indirectly through "second-order interaction."

To illustrate the "second-order interaction," we use the inductively source degenerated LNA [5] as an example. Fig. 2(a) presents the circuit, and Fig. 2(b) shows its small-signal model using the notation from Fig. 1. The inductor L_s acts as a frequency-dependent feedback element with $\beta = \omega L_s$, creating a feedback path between the output current i_d and the gate-source voltage v_{in}. For simplicity, we analyze these effects with a Taylor series—for a more accurate, frequency-dependent analysis refer to the results obtained using Volterra series in [11], [14], [23].

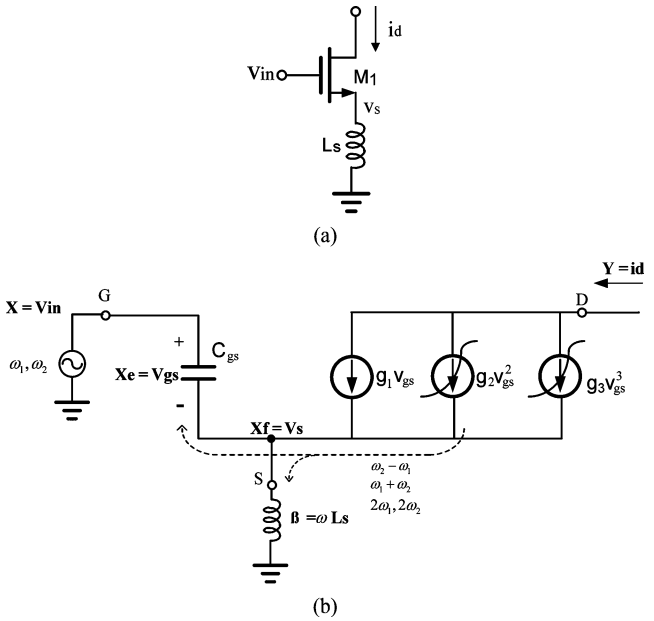


Fig. 2. (a) Inductively source-degenerated LNA. (b) Small-signal model.

First, $\mathrm{i_d}$ can be expressed as

$$\mathrm{i_d} = \mathrm{g_1}\left(\mathrm{v_{in}} - \mathrm{v_s}\right) + \mathrm{g_2}\left(\mathrm{v_{in}} - \mathrm{v_s}\right)^2 + \mathrm{g_3}\left(\mathrm{v_{in}} - \mathrm{v_s}\right)^3. \quad (5)$$

where $\mathrm{g_i}$ is the i^{th}-order coefficient of $\mathrm{M_1}$ obtained by taking the derivative of the drain-source dc current $\mathrm{I_{DS}}$ with respect to the gate-to-source voltage $\mathrm{V_{GS}}$ at the dc bias point

$$\mathrm{g_1} = \frac{\partial \mathrm{I_{DS}}}{\partial \mathrm{V_{GS}}}, \quad \mathrm{g_2} = \frac{1}{2!}\frac{\partial^2 \mathrm{I_{DS}}}{\partial \mathrm{V_{GS}^2}}, \quad \mathrm{g_3} = \frac{1}{3!}\frac{\partial^3 \mathrm{I_{DS}}}{\partial \mathrm{V_{GS}^3}}. \quad (6)$$

Assume input $\mathrm{v_{in}}$ has two frequency components ω_1 and ω_2; thus $\mathrm{v_s}$ contains components $2\omega_1$, $2\omega_2$, and $\omega_1 \pm \omega_2$ due to the second-order distortion. The product term $-2\mathrm{g_2 v_{in} v_s}$ from $\mathrm{g_2}(\mathrm{v_{in}} - \mathrm{v_s})^2$ generates IM3 terms $2\omega_1 \pm \omega_2$ and $2\omega_2 \pm \omega_1$. Therefore, the intrinsic second-order nonlinearity contributes to third-order intermodulation, IM3, when a feedback mechanism is employed. Note that this "second-order interaction" problem exists even if the LNA topology is differential because the term $-2\mathrm{g_2 v_{in} v_s}$ is an odd term and cannot be rejected by differential operation.

Though source degeneration mostly improves linearity, inductive source degeneration actually has two opposing effects on linearity:

1) increases $\mathrm{A_{IIP3}}$ by $\approx (1 + \mathrm{g_1}\omega \mathrm{Ls})^{3/2}$;
2) degrades $\mathrm{A_{IIP3}}$ due to "second-order interaction."

Fig. 3 shows $\mathrm{A_{IIP3}}$ versus source-degeneration inductor Ls for two cases: input tones at 2.4 GHz, 2.41 GHz, and at 5 GHz, 5.01 GHz. Note that this simulation only includes the distortion from input transconductance, while the loading and input-matching resonant network will also contribute to distortion in practice.

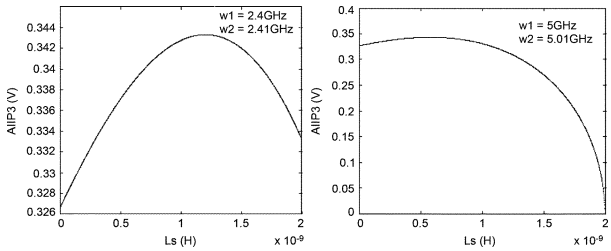


Fig. 3. $\mathrm{A_{IIP3}}$ versus source-degeneration inductance.

Reducing the degeneration inductance or adding a termination network such that $\mathrm{v_s} = 0$ at the IM2 frequency can mitigate "second-order interaction;" the latter approach is called "harmonic termination."

B. Harmonic Termination

"Harmonic termination" adds a termination network to accomplish one of two effects:

1) **BJT case**: sets $\mathrm{b_3} = 0$ in (3) with the "second-order-interaction" term;
2) **CMOS case**: forces a certain node voltage to zero at the IM2 frequency.

Equation (3) was obtained assuming a frequency-independent feedback factor β, which is only valid for pure resistive networks. For frequency-dependent networks as the case in Fig. 2, Volterra series [3] should be used to capture the memory effects. To obtain the third-order coefficient in the Volterra series, defined as $\mathrm{b_3}(\omega_\mathrm{x}, \omega_\mathrm{y}, \omega_\mathrm{z})$, a three-dimensional Fourier transform is performed on the third-order impulse response $\mathrm{h_3}(\tau_\mathrm{x}, \tau_\mathrm{y}, \tau_\mathrm{z})$ of the system. Thus, (3) becomes (7), as shown at the bottom of the page. $\mathrm{T}(\omega) = \mathrm{g_1}\beta(\omega)$ is the frequency-dependent linear loop gain, which also depends on the feedback components and termination impedances $\mathrm{Z_1}$, $\mathrm{Z_2}$, and $\mathrm{Z_3}$ shown in Fig. 4. The expressions $\left|\mathrm{b_3}(\omega_\mathrm{x}, \omega_\mathrm{y}, \omega_\mathrm{z})\right|$ and $\angle \mathrm{b_3}\ (\omega_\mathrm{x}, \omega_\mathrm{y}, \omega_\mathrm{z})$ give the magnitude and phase of a tone at frequency $\omega_\mathrm{x} + \omega_\mathrm{y} + \omega_\mathrm{z}$ generated by third-order nonlinearity. For example, given two input tones at $\omega_1 \& \omega_2$, to get the IM3 products at $2\omega_1 - \omega_2$, choose $\omega_\mathrm{x} = \omega_\mathrm{y} = \omega_1$ and $\omega_\mathrm{z} = -\omega_2$. Assuming two closely spaced tones, i.e., $\omega_1 = \omega, \omega_2 = -\omega - \Delta\omega$, and $\Delta\omega \approx 0$, we obtained (8), as shown at the bottom of the next page. From (8), the contribution of second-order distortion to IM3 is defined by the loop gain at subharmonic frequency $\Delta\omega$ and second-harmonic frequency 2ω, i.e., $\mathrm{T}(\Delta\omega)$ and $\mathrm{T}(2\omega)$. Therefore, by tuning the termination impedances at $\Delta\omega$ and/or 2ω, the amplitude and phase of the second-order interaction terms $\mathrm{A_2}$ can be adjusted to cancel the intrinsic third-order distortion term $\mathrm{g_3}$, so that $\mathrm{b_3} \approx 0$. For narrowband applications, $\Delta\omega$ and 2ω are usually out-of-band, keeping the in-band operation unaffected, hence the "harmonic termination" technique is also called "out-of-band tuning/termination" [7], [8].

$$\mathrm{b_3}\ (\omega_\mathrm{x}, \omega_\mathrm{y}, \omega_\mathrm{z}) = \frac{1}{\left(1 + \mathrm{T}\left(\omega_\mathrm{x} + \omega_\mathrm{y} + \omega_\mathrm{z}\right)\right)\left(1 + \mathrm{T}\left(\omega_\mathrm{x}\right)\right)\left(1 + \mathrm{T}\left(\omega_\mathrm{y}\right)\right)\left(1 + \mathrm{T}\left(\omega_\mathrm{z}\right)\right)} \times \left[\mathrm{g_3} - \frac{2\mathrm{g_2^2}}{3\mathrm{g_1}}\left(\frac{\mathrm{T}\left(\omega_\mathrm{y} + \omega_\mathrm{z}\right)}{1 + \mathrm{T}\left(\omega_\mathrm{y} + \omega_\mathrm{z}\right)} + \frac{\mathrm{T}\left(\omega_\mathrm{x} + \omega_\mathrm{z}\right)}{1 + \mathrm{T}\left(\omega_\mathrm{x} + \omega_\mathrm{z}\right)} + \frac{\mathrm{T}\left(\omega_\mathrm{x} + \omega_\mathrm{y}\right)}{1 + \mathrm{T}\left(\omega_\mathrm{x} + \omega_\mathrm{y}\right)}\right)\right] \quad (7)$$

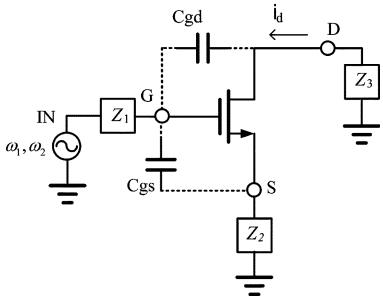


Fig. 4. Common-source LNA with termination impedances.

TABLE II
THREE INTRINSIC FEEDBACK PATHS.

Feedback Path	Path Components	
	CS-LNA	CG-LNA
Source-to-Gate	C_{gs} + source degeneration inductor Z_2	C_{gs} + input driving impedance [10]
Drain-to-Gate	C_{gd} + output load Z_3	
Input-to-Gate	Input matching network Z_1	

The second-order nonlinear current can mix with the input through three intrinsic feedback paths, as listed in Table II for the common source LNA(CS-LNA) and common gate LNA(CG-LNA).

The CG-LNA inherently has less drain-to-gate feedback than the CS-LNA since its gate is ac grounded, therefore the CG-LNA usually has better linearity.

Resonant tanks can be added to optimally tune $Z_i(\Delta\omega)$ and/or $Z_i\ (2\omega)$ such that the second-order remixing term cancels the IM3 term. Techniques have been reported to tune the input terminal $Z_1(\Delta\omega)$ for bipolar LNAs [7]–[9]. The terminations are commonly implemented with dedicated LC networks, which provide high impedance at ω but small impedance paths to ground at $\Delta\omega$ or 2ω. However, the required inductance value is usually quite large. The low Q factors of the on-chip passive inductors limit their distortion-cancellation effectiveness and also affect noise and input matching. Furthermore, on-chip active inductors add noise and nonlinearity, so in practice off-chip inductors are employed.

Though popular in BJT LNAs, harmonic termination is less effective for CMOS LNAs [8], [12]. For a stable design, the A_2 term in (8) has a positive real part. Thus, $|\varepsilon(\Delta\omega, 2\omega)|$ can be reduced below $|g_3|$ only if g_3 is positive, which is true for a BJT, but not for a MOSFET in saturation. Therefore, both g_3 and A_2 must be reduced to improve a CMOS LNA's IIP3.

One way to reduce A_2 is to reduce both $Z_i\ (\Delta\omega)$ and $Z_i\ (2\omega)$ [13]. A cascode configuration can reduce Z_3 to $1/g_1$ [13], and

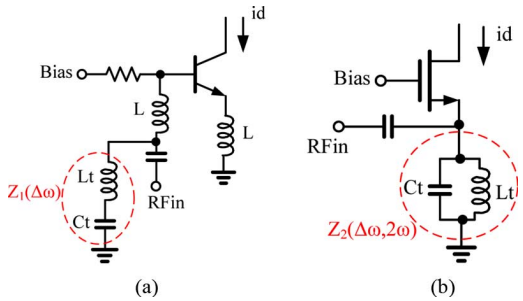


Fig. 5. Harmonic termination: (a) Common-emitter stage with low-frequency-trap network (L is the package inductance) [9]. (b) Common-gate stage with "RF current source"[13].

capacitive cross coupling in the cascode stage further reduces Z_3 to $1/(2g_1)$ [14]. Although their IIP3 improvement is not as great as that attainable from large passive LC components, it is more feasible. In [10] and [13] an LC-resonant RF current source reduces Z_2. Fig. 5 shows some termination examples, in which Lt and Ct form low-frequency/second-harmonic trap networks $Z_1(\Delta\omega)$ [Fig. 5(a)] and $Z_2(\Delta\omega, 2\omega)$ [Fig. 5(b)].

Harmonic termination only works well in narrowband systems because the tuning network is optimized at $\Delta\omega$ and 2ω, and only works for a narrow range of two-tone spacing/center frequencies [7]. For wideband applications, $\Delta\omega$ and 2ω vary considerably, so it is difficult to tune out the termination impedance. Furthermore, $\Delta\omega$ and 2ω may fall in-band, affecting the normal operations.

In summary, to improve CMOS LNA linearity, we should ensure a small intrinsic third-order coefficient g_3 of the transistor, and relax the "second-order interaction." Adding a harmonic termination network alleviates the latter. Next, we will discuss a few techniques to reduce the third-order coefficient g_3.

C. Optimal Biasing

Assume the main nonlinearity of a MOS transistor arises from transconductance nonlinearity, as modeled in (5). To characterize this single-transistor nonlinearity, we fixed its drain-source voltage V_{ds}, swept the gate-source voltage V_{gs}, and then took the first three derivatives of the drain-source dc current I_{ds} with respect to V_{gs} [as defined in (6)] at every dc bias point to obtain the plots in Fig. 6. While g_2 is always positive, g_3 has a sign inversion.

— Small V_{gs}: $g_3 > 0$ because the transistor operates in weak inversion, where the I_{ds} vs V_{gs} relation is exponential.
— Large V_{gs}: $g_3 < 0$ because mobility degradation/velocity saturation cause gain compression. The key idea of "optimum biasing" is to bias the transistor at the "sweet spot" $g_3 = 0$ [44].

$$b_3\left(\omega, \omega, -\omega-\Delta\omega\right) \cong \frac{1}{\left(1+\mathrm{T}\left(\omega\right)\right)^3\left(1+\mathrm{T}\left(-\omega\right)\right)} \times \left[\underbrace{g_3 - \underbrace{\frac{2g_2^2}{3g_1}\left(\frac{2\mathrm{T}\left(\Delta\omega\right)}{1+\mathrm{T}\left(\Delta\omega\right)} + \frac{\mathrm{T}\left(2\omega\right)}{1+\mathrm{T}\left(2\omega\right)}\right)}_{A_2}}_{\varepsilon(\Delta\omega,2\omega)}\right]. \tag{8}$$

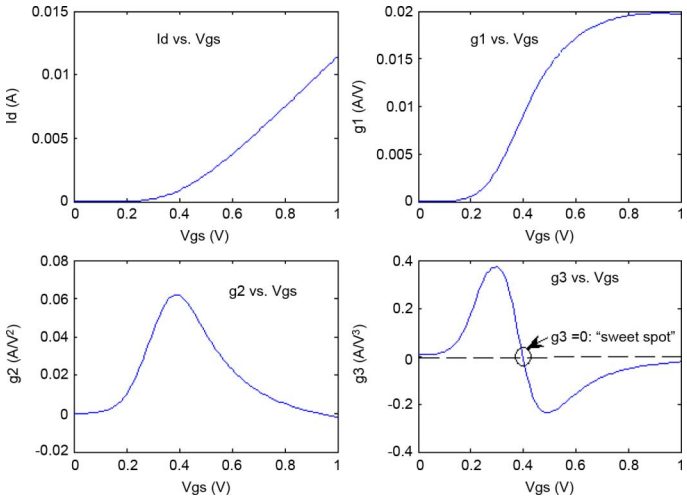


Fig. 6. NMOS transconductance characteristics (UMC 90 nm CMOS process, $\mathrm{W/L} = 20/0.08\ \mu\mathrm{m}$, $\mathrm{V_{ds}} = 1\ \mathrm{V}$).

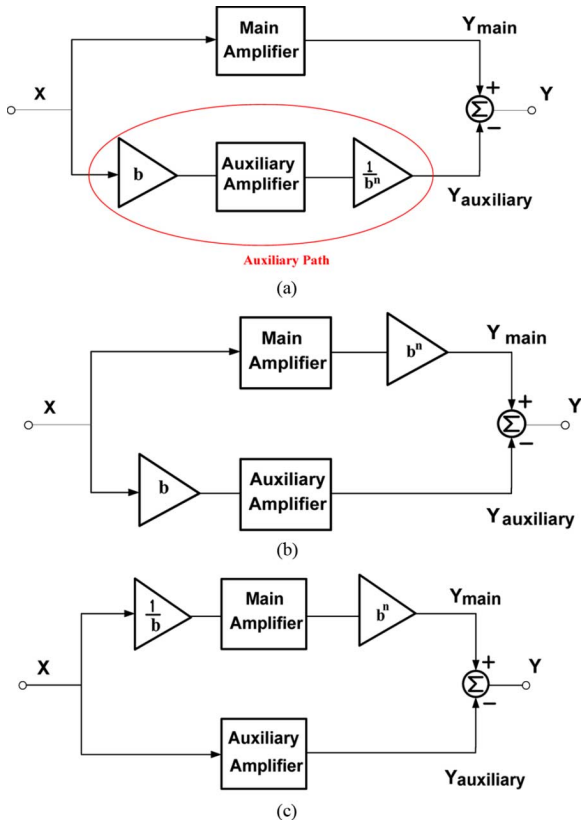


Fig. 7. Three representations of the feedforward linearization technique.

Though simple in principle, the optimal biasing technique has the following limitations.

1) The cancellation is sensitive to process variations (e.g., $\mathrm{V_{th}}$), so we recommend constant-current or constant-gm biasing over constant-voltage biasing.
2) The technique is sensitive to operating point, resulting in a limited input-signal amplitude range for effective distortion cancellation.
3) The sweet spot shifts to a lower bias current level as the gain increases, since the output swing increases and nonlinear output conductance starts to play a role.

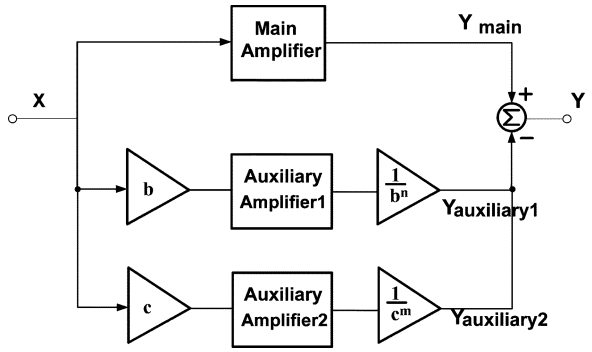


Fig. 8. Proposed dual-auxiliary-path feedforward linearization technique.

4) Due to the "second-order interaction," the IIP3 peak at the "sweet spot" decreases and will finally disappear as source degeneration inductance increases.
5) The sweet spot is frequency-dependent, and the IIP3 peak decreases due to parasitic effects [44].
6) Biasing the transistor at $\mathrm{g_3} = 0$ restricts the input-stage transconductance, lowering gain and increasing NF.

An automatic bias circuit could mitigate some of these effects [15]; however, this "automatic" bias circuit itself is prone to process variations and requires manual tuning in practice. The bias point for optimum IIP3 is shifted from the bias for zero $\mathrm{g_3}$ due to "second-order interaction."

In summary, the "sweet spot" is a single transistor characteristic and only signifies optimum intrinsic third-order transconductance nonlinearity. Many other factors will weaken the IIP3 improvement at the "sweet spot." Furthermore, some claim that no "sweet spot" exists in practical LNAs because of input/output networks and parasitics [44].

D. Feedforward

From (5), note that simultaneous cancellation of $\mathrm{g_2}$ and $\mathrm{g_3}$ with minimum effects on $\mathrm{g_1}$ requires more degrees of freedom. Generating additional nonlinear currents/voltages, and subsequently summing (subtracting) them accomplishes such simultaneous cancellation. These actions constitute feedforward, as illustrated in Fig. 7(a) [16]. An auxiliary path includes a replica amplifier and signal-scaling factors b and $1/\mathrm{b^n}$ at its input/output, respectively, to replicate the distortion in the main path. We use $\mathrm{n} = 2$ or 3 depending on whether IM2 or IM3 is to be canceled. Note that if the amplifiers are differential, the second-order distortion is ideally zero, and $\mathrm{n} = 3$ yields a linear output. Without loss of generality, the following discusses the single-ended case. To obtain the total output Y, we subtract the output of the auxiliary amplifier $(\mathrm{Y_{auxiliary}})$ from that of the main amplifier $(\mathrm{Y_{main}})$. Assuming $|\mathrm{b}| > 1$, by changing the location and value of scaling factor, we propose two alternate implementations, shown in Fig. 7(b) and 7(c), respectively.

Assuming the main and auxiliary amplifiers have the same nonlinearity coefficients $\mathrm{g_i}$, we have

a) See equation (9)–(11) at the bottom of the next page.
b) See equation (12)–(14) at the bottom of the next page.
c) See equation (15)–(17) at the bottom of the next page.

In these equations, $\mathrm{g_2} = 0$ for differential amplifiers. Comparing (11), (14), and (17), the implementation in a) has a gain-attenuation factor of $(1 - 1/\mathrm{b^{n-1}})$, thus gain is reduced by 2.5 dB with $\mathrm{b} = 2$ and $\mathrm{n} = 3$ as in [16]. On the other hand, the proposed implementations in b) and c) increase the gain. Note

that c)'s input attenuator 1/b worsens its NF. The implementations in Fig. 7 can only cancel one type of harmonic at a time; to reduce both second- and third-order distortion simultaneously, we need an additional degree of freedom, which we could attain with two auxiliary paths as shown in Fig. 8.

Assuming the main and auxiliary amplifiers have the same nonlinearity coefficients g_i, we have (18)–(20), at the bottom of the page. In (20), we have two equations (second and third term equals zero) and four variables (b, c, n, m), resulting in multiple solutions. A possible additional constraint is to bound the reduction in linear gain to be less than, say, 20%, and one reasonable solution set is: $b = -2$, $c = -3$, $n = 0$, $m = 1$. This choice causes the linear gain to double.

Note that if the amplifiers are differential, all even order harmonics are ideally zero, and the implementation in Fig. 8 can cancel both third- and fifth-order distortion.

This general feedforward technique improves linearity without relying on the amplifier's linearity characteristics; however, it has several disadvantages.

1) Accurate, noiseless, and highly linear scaling factors (b, c) are often not feasible. For instance, the off-chip coaxial assembly used in [16] is expensive and cannot be integrated.
2) The added active components introduce more noise.
3) Highly sensitive to mismatch between the main and auxiliary gain stages.
4) Power is doubled or tripled since the auxiliary amplifier is an exact copy of the main amplifier.

[17] reports an improved feedforward technique, where the auxiliary path only passes the IM3 products. Hence, its dynamic range is relaxed, resulting in only 21% power overhead. Next, we will discuss three special cases of the feedforward technique: derivative superposition, IM2 injection, and noise/distortion cancellation.

E. Derivative Superposition (DS)

The derivative superposition (DS) method [10], [11], [13], [18]–[20] is a special case of the feedforward technique. Notice that the DS method is obtained when $b = 1$ in Fig. 7 and when the main/auxiliary amplifiers are implemented with transistors operating in different regions. Fig. 9(a) depicts a dual-NMOS implementation of the DS method. $M_{A/B}$ denotes the main/auxiliary transistor, respectively, and the input matching network is omitted for simplicity.

This method is called "derivative superposition" because it adds the third derivatives (g_3) of drain current from the main and auxiliary transistors to cancel distortion. As discussed in

$$Y_{main} = g_1 X + g_2 X^2 + g_3 X^3 \tag{9}$$

$$Y_{auxiliary} = \left[g_1 (bX) + g_2 (bX)^2 + g_3 (bX)^3\right] \frac{1}{b^n} \tag{10}$$

$$\begin{aligned} Y &= Y_{main} - Y_{auxiliary} \\ &= g_1 \left(1 - \frac{1}{b^{n-1}}\right) X + \underbrace{g_2 \left(1 - \frac{1}{b^{n-2}}\right) X^2 + g_3 \left(1 - \frac{1}{b^{n-3}}\right) X^3}_{\text{Residue Distortion}}. \end{aligned} \tag{11}$$

$$Y_{main} = \left(g_1 X + g_2 X^2 + g_3 X^3\right) b^n \tag{12}$$

$$Y_{auxiliary} = g_1 (bX) + g_2 (bX)^2 + g_3 (bX)^3 \tag{13}$$

$$Y = g_1 b \left(b^{n-1} - 1\right) X + \underbrace{g_2 b^2 \left(b^{n-2} - 1\right) X^2 + g_3 b^3 \left(b^{n-3} - 1\right) X^3}_{\text{Residue Distortion}} \tag{14}$$

$$Y_{main} = \left[g_1 \frac{X}{b} + g_2 \left(\frac{X}{b}\right)^2 + g_3 \left(\frac{X}{b}\right)^3\right] b^n \tag{15}$$

$$Y_{auxiliary} = g_1 X + g_2 X^2 + g_3 X^3 \tag{16}$$

$$Y = g_1 \left(b^{n-1} - 1\right) X + \underbrace{g_2 \left(b^{n-2} - 1\right) X^2 + g_3 \left(b^{n-3} - 1\right) X^3}_{\text{Residue Distortion}} \tag{17}$$

$$Y_{auxiliary1} = \left[g_1 (bX) + g_2 (bX)^2 + g_3 (bX)^3\right] \frac{1}{b^n} \tag{18}$$

$$Y_{auxiliary2} = \left[g_1 (cX) + g_2 (cX)^2 + g_3 (cX)^3\right] \frac{1}{c^m} \tag{19}$$

$$\begin{aligned} Y &= Y_{main} - Y_{auxiliary1} - Y_{auxiliary2} \\ &= \underbrace{g_1 \left(1 - \frac{1}{b^{n-1}} - \frac{1}{c^{m-1}}\right) X}_{\text{Linear Gain}} + \underbrace{g_2 \left(1 - \frac{1}{b^{n-2}} - \frac{1}{c^{m-2}}\right) X^2 + g_3 \left(1 - \frac{1}{b^{n-3}} - \frac{1}{c^{m-3}}\right) X^3}_{\text{Residue Distortion}}. \end{aligned} \tag{20}$$

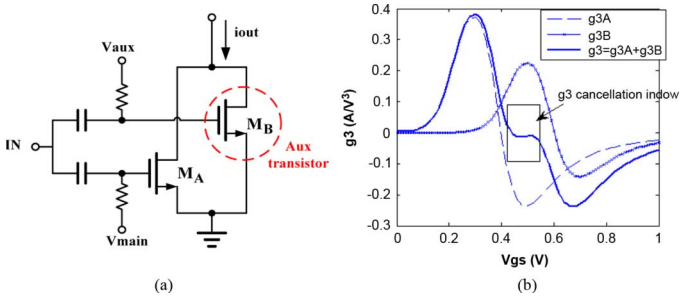


Fig. 9. (a) DS method with dual-NMOSs. (b) Third-order distortion terms of the main transistor (g_{3A}), auxiliary transistor (g_{3B}), and total output (g_3) (UMC 90 nm CMOS process, $(W/L)_{MA} = 20/0.08\ \mu m$, $(W/L)_{MB} = 12/0.08\ \mu m$, $V_{ds} = 1\ V$).

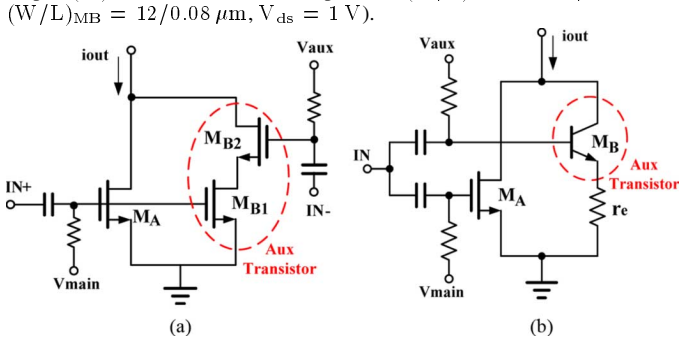


Fig. 10. DS method. (a) Additional transistor works in triode region [18]. (b) Use of a bipolar transistor [19].

Section II-C, g_3's sign changes at the boundary of moderate and strong inversion region. Thus, proper biasing creates net zero g_3, as shown in Fig. 9(b). Linearity is improved within a finite bias-voltage *range* instead of just a *point*. Hence the DS method is less sensitive to process variations than the optimum biasing technique. Moreover, the auxiliary path contains only one weak-inversion transistor, resulting in much smaller power consumption than the general feedforward technique. Since the DS method employs multiple transistors in parallel with their gates connected together, it is also called the "multiple gated transistor technique" (MGTR) [10], [11], [13]. Note that since positive and negative characteristic of g_3 are not symmetric, the g_3-cancellation window is fairly narrow with only one auxiliary transistor, but the window widens with more auxiliary transistors at the cost of degraded input matching, NF, and gain [11], [20].

Fig. 10(a) and 10(b) show alternate implementations of the DS method that use a triode region [18] or bipolar [19] transistor as the auxiliary device. In Fig. 10(a), M_{B1} and M_{B2} are driven by differential input signals. M_{B1} is biased in deep triode region, and M_{B2} helps to boost the positive g_3 peak of M_{B1} to be sufficiently large to cancel the negative peak in g_3 of input transistor M_A. In Fig. 10(b), a bipolar transistor M_B provides the positive g_3, and emitter degeneration resistor r_e reduces g_3 to match that of M_A for optimum distortion cancellation.

1) Complementary DS: Fig. 6 shows that the second-order term (g_2) has a positive sign for transistors working in either moderate or strong inversion region. Therefore, techniques, such as conventional DS, that improve third-order distortion usually worsen second-order distortion. The "complementary DS method" employs an NMOS/PMOS pair to improve IIP3 without hurting IIP2 [21], [28].

Fig. 11 shows the common-source and common-gate implementations, respectively. The ac current combiner in Fig. 11(b) could be either a large coupling capacitor with negligible impedance within signal bandwidth [28], or a current mirror

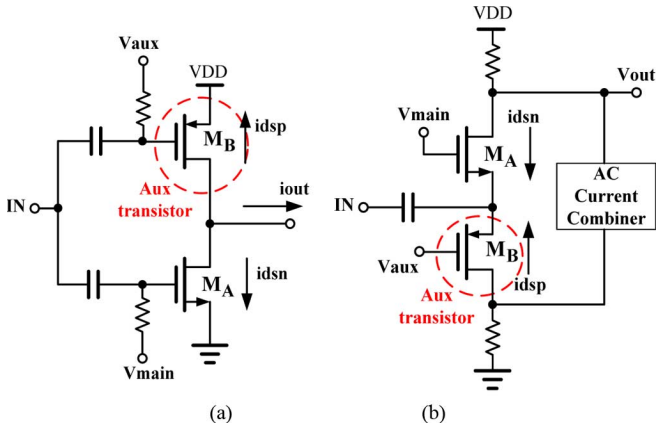


Fig. 11. (a) Complementary DS with common-source configuration [21]. (b) Complementary DS with common-gate configuration [28].

[21]. Since the ac input signal for NMOS/PMOS are out of phase, the output current is expressed as

$$i_{dsn} = g_{1A}v_{gs} + g_{2A}v_{gs}^2 + g_{3A}v_{gs}^3 \quad (21)$$

$$i_{dsp} = -g_{1B}v_{gs} + g_{2B}v_{gs}^2 - g_{3B}v_{gs}^3 \quad (22)$$

$$i_{out} = i_{dsn} - i_{dsp} = (g_{1A} + g_{1B})\,v_{gs} + (g_{2A} - g_{2B})\,v_{gs}^2 + (g_{3A} + g_{3B})\,v_{gs}^3 \quad (23)$$

The total transconductance increases, the IM2 term decreases because g_{2A} and g_{2B} have the same sign, and the IM3 term decreases because g_{3A} and g_{3B} have different signs. Fig. 12 compares the conventional DS and complementary DS in terms of second-order (g_2) and third-order (g_3) distortion of the output current. A cancellation window for g_3 exists in both cases at V_{gs} around 500 mV, but g_2 is maximized for conventional DS and minimized for complementary DS. Note that the g_3 cancellation window is narrower and less flat for complementary DS since PMOS and NMOS devices have different linearity characteristics, so the IIP3 improvement is not as good as that in a dual-NMOS implementation. Furthermore, as shown in (23), we can either match g_{3A} and g_{3B} for a good IIP3 while slightly canceling g_2, or we can match g_{2A} and g_{2B} for optimum IIP2, because IIP2 and IIP3 do not share the same optimum bias. The differential DS method is essentially the same as complementary DS, which also alleviates IIP2 problem [19], [22].

"Second-order interaction" ultimately limits the IIP3 improvement at higher frequencies after the intrinsic g_3-induced third-order distortion is canceled by the DS method. The "modified DS method" alleviates this issue [23], [24].

2) Modified DS: As discussed in Section II-B, three feedback paths exist for "second-order interaction": source-to-gate, drain-to-gate, and input-to-gate. The modified DS methods [23], [24] provide an on-chip solution to minimize the source-to-gate feedback. The vector diagram in Fig. 13 graphically explains the modified-DS concept. In conventional DS, the anti-parallel g_{3A} and g_{3B} result in a zero total g_3, but residual IM3 exists due to g_{2A} contributions (Note: here we neglect g_{2B}). In the modified DS method, g_{3B} is rotated properly such that the composite vector of g_{3A} and g_{3B} contribution is 180° out of phase with the g_{2A} contribution, yielding zero net IM3. Fig. 14(a) shows the circuit implementation of the modified DS method from [23]. Note that choice of L_2 determines the angle of g_{3B}.

Although the channel noise of weak inversion transistor M_B is negligible, its gate-induced noise is inversely proportional to drain current, and is added directly to the main transistor's (M_A) gate noise because their gates are connected together. M_B also

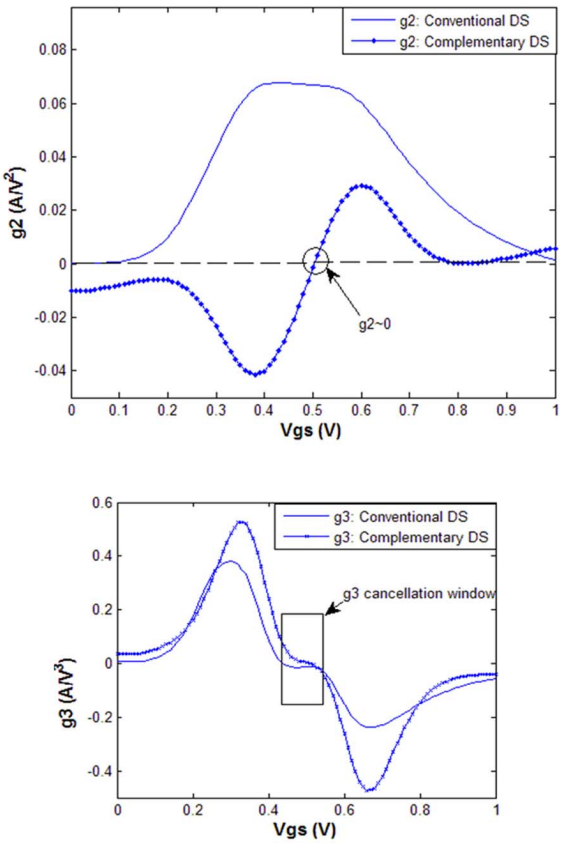


Fig. 12. Comparison of g_2, g_3 as a function of gate bias for conventional (dual-NMOS) DS and complementary (PMOS/NMOS) DS (UMC 90 nm CMOS process, $V_{ds} = 1\,V$).

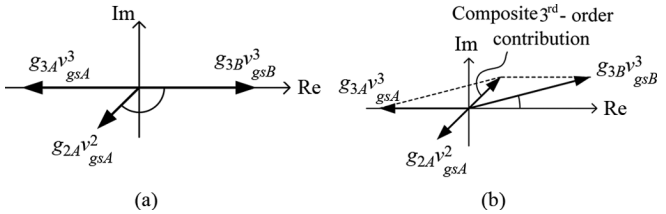


Fig. 13. Vector diagram for the distortion components of: (a) conventional DS method; (b) modified DS method [23].

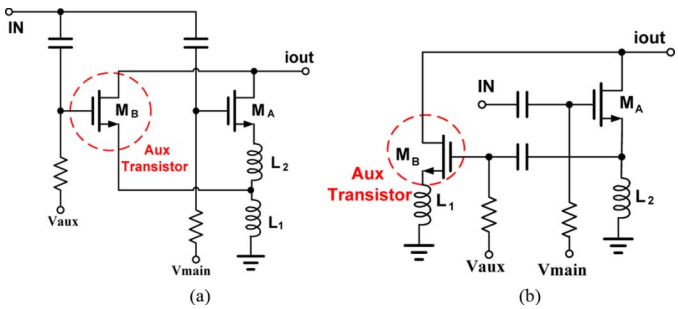


Fig. 14. Circuit implementation of modified DS method [23], [24].

affects the input impedance matching. An alternate implementation of the modified DS method reported in [24] [Fig. 14(b)] moves M_B to the source of M_A instead of directly connecting it to the input, thus minimizing the degradation in NF and input matching.

Limitations of the DS methods include the following.

1) The weak-inversion transistor may not operate at sufficiently high frequency.

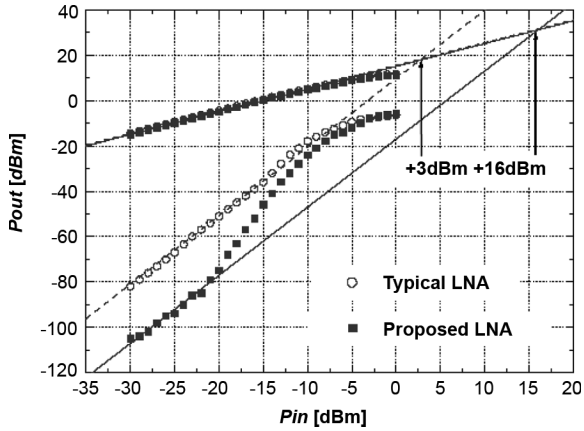


Fig. 15. Measured IIP3 of LNAs with/without DS method [18] (© 2003 IEEE).

2) The weak-inversion transistor cannot handle large signals or it will be turned off, resulting in a very limited distortion-cancellation range.
3) Weak-inversion transistor models are generally not accurate, resulting in considerable discrepancy between simulation and measurement.
4) Triode-region transistors' positive g_3 peaks decrease as technology scales down, thus complicating the task of matching their amplitudes with the negative peaks of g_3 in main transistors.
5) Matching transistors working in different regions or matching bipolar with MOS transistors is difficult if not impossible, resulting in a linearity improvement sensitive to PVT variations. Current bias with digital control bits [13] or manual adjustment is required for good results in practice.

Fig. 15 shows an example IIP3 measurement plot [18]. Although it is from a conventional DS method, similar characteristics can be observed with complementary, differential, and modified DS methods. We observe the following.

1) The DS method works well within the g_3-cancellation window annotated in Figs. 9(b) and 12 ($P_{in} < -20\,dBm$).
2) Even for inputs outside the g_3-cancellation window, the DS method can still reduce the third-order tone below that of the conventional LNA having a main transistor with negative g_3 as long as g_3 of the auxiliary transistor is positive.
3) The third-order curve shows a greater-than-three slope at much smaller input amplitudes after applying the DS method, because the fifth- and higher odd-order-distortion terms contribute more appreciably after g_3 is canceled.
4) The DS method does not improve the compression point.

F. IM2 Injection

The IM2 injection method eliminates the explicit auxiliary path entirely by merging it with the main path to reuse the active devices and the dc current [25]. To understand the concept, we first recall (8). To reduce IM3, we should minimize the $\varepsilon(\Delta\omega, 2\omega)$ term. As previously discussed in Section II-B, making A_2 cancel g_3 is difficult for CMOS LNAs because these two terms are out-of-phase in a typical design. Furthermore, the self-generated IM2 term has too small an amplitude to suppress g_3 sufficiently.

The IM2 injection technique externally generates and injects a low-frequency IM2 component into the circuit. The injected IM2 phase is inverted with boosted amplitude for IM3 cancellation. Hence, IM2 injection could also be viewed as a smart

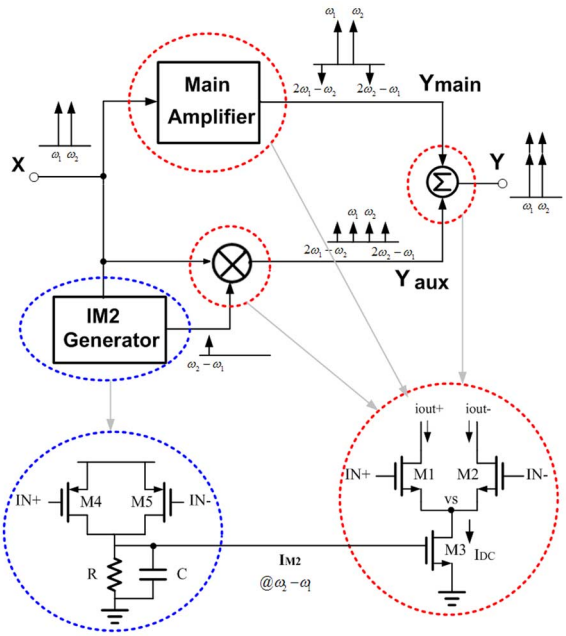


Fig. 16. Block diagram and basic cell implementation of "IM2 injection" [25].

implementation of harmonic termination. Fig. 16 illustrates the concept and basic cells. M1 and M2 are the input transistors of the LNA, and M4, M5, R, and C compose a squaring circuit to generate a low-frequency IM2 current at $\omega_2 - \omega_1$, which is then injected through M3 into the common source node v_s of the LNA. This technique utilizes second-order interaction to generate tones at $2\omega_2 - \omega_1$ and $2\omega_1 - \omega_2$ to cancel the IM3 tones arising from intrinsic third-order distortion. The design equation is

$$\underbrace{-\frac{2g_{1,M1}g_{3,M1}}{4g_{2,M1}}+\frac{3}{2}g_{2,M1}}_{\text{Main Circuit}} = \underbrace{-g_{1,M3}\times 2g_{2,M4}R}_{\text{Squaring Circuit}} \qquad (24)$$

where $g_{i,Mi}$ is the i_{th} transconductance coefficient of M_i. The injected IM2 tone should be in phase with the envelope of the RF input signal. Frequency component $\omega_2 - \omega_1$ is chosen over other IM2 components $(\omega_2+\omega_1, 2\omega_2, 2\omega_1)$ because it is easier to match the phase at low frequency. Since the linear gain is added in phase, and the noise injected from the IM2 generator appears as common mode noise (suppressed by differential operation), IM2 injection circumvents gain and NF penalties.

Limitations of IM2 injection include the following.

1) NMOS/PMOS transistors and resistors have independent PVT variations—hence more difficult to satisfy the IM3 cancellation criteria in (24) robustly.
2) Since R and C in the IM2 generator introduce extra phase shift, two tone spacing must be smaller than the RC-filter cutoff frequency for negligible phase mismatch. Cancellation performance degrades as tone spacing increases.
3) Frequency components at $\omega_2 \pm \omega_1$ and $2\omega_{1,2}$ injected by the IM2 generator may fall into signal band and degrade the IIP2.
4) Noise from the IM2 generator is negligible only for differential LNAs, but would result in appreciable NF degradation for single-ended LNAs.

In short, IM2 injection applies chiefly to narrowband, differential systems with small two-tone spacing.

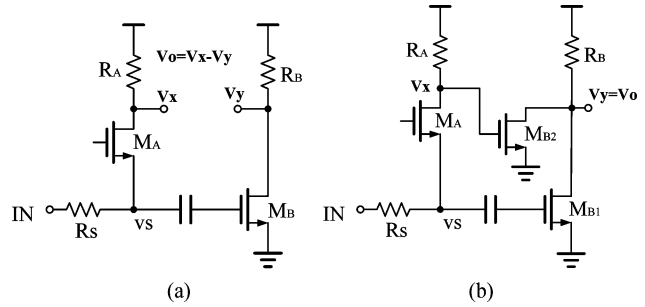


Fig. 17. Noise/distortion cancellation. (a) Differential output [27], [29]. (b) Single ended output [28].

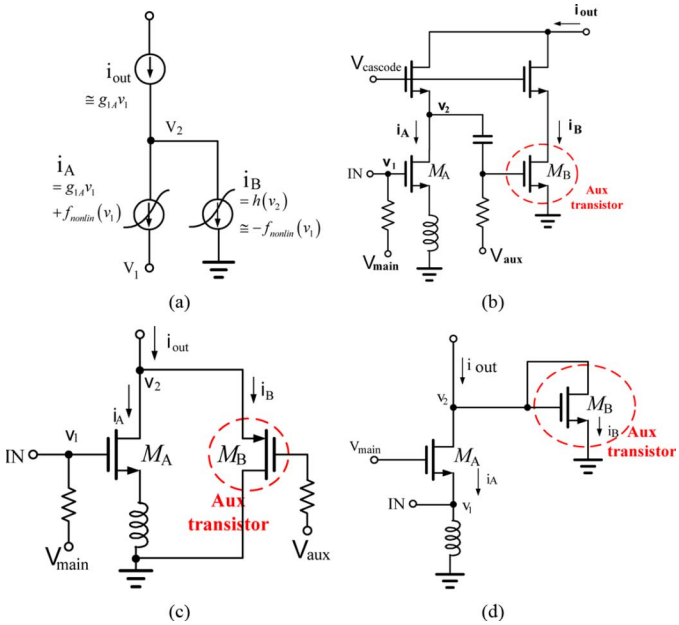


Fig. 18. Post-distortion. (a) Conceptual view. (b) Circuit implementation in [30]. (c) Circuit implementation in [31]. (d) Circuit implementation in [32].

G. Noise/Distortion Cancellation

Noise/distortion cancellation parallels CG (M_A) and CS (M_B) stages, as shown in Fig. 17 [26]–[29]. The circuit is driven by a voltage at node "IN." The nonlinearity of M_A can be modeled as a current source between its drain and source controlled by both V_{gs} and V_{ds}. Hence, both the channel thermal noise and distortion of M_A flowing through the CG and CS paths are subtracted at the output, whereas the signal is added. Noise/distortion cancellation requires that $V_x = V_y$, i.e.,

$$g_{1,M_A}R_A = g_{1,M_B}R_B \text{ (differential output)} \qquad (25a)$$
$$g_{1,M_{B1}}R_s = g_{1,M_{B2}}R_A \text{ (single-ended output)}. \qquad (25b)$$

Note this technique can cancel all intrinsic distortion generated by M_A, including both g_m and g_{ds} nonlinearity, while previous techniques could only compensate g_m nonlinearity.

After canceling the distortion from M_A, M_B's distortion dominates the residual nonlinearity, which comprises two terms: 1) M_B's intrinsic third-order distortion and 2) second-order interaction originating from the CG-CS cascade. Optimal biasing of M_B[27], [29] or employing complementary DS [28] could further improve the linearity.

TABLE III
PERFORMANCE COMPARISON OF SILICON-VERIFIED LINEARIZATION TECHNIQUES FOR CMOS LNAS.

Linearization Technique	Harmonic Termination [14]	Optimum biasing [15]	Feedforward [16]	Derivative Superposition [18]	Modified DS [23]	Complementary DS ***[21]	IM2 Injection **[25]	Noise/Distortion Cancellation ***[29]	Post Distortion [32]
*IIP3/ Δ IIP3	-4.4dBm/ +2.5dB	+10.5dBm	5dBm/ +13dB	2.7dBm/ +13.4dB	2dBm/ +20dB	3dBm	-10.4dBm/ +10.6dB	>0dBm	5dBm/ +9dB
*IIP2/ Δ IIP2	N/A	N/A	N/A	N/A	N/A	+44dBm	N/A	>+20dBm	+10dBm/ +10dB
*Gain/ Δ Gain	20.4dB/ +2dB	14.6dB/ 0dB	18dB/ -2.5dB	15.3dB/ -0.4dB	16dB/ -0.5dB	14dB	22dB/0dB	13-15.6dB	14.3dB/ -1.7dB
*NF/ Δ NF	1.92dB/ 0dB	1.8dB/ 0dB	2.6dB/ +0.2dB	2.9dB/ +0.1dB	1.4dB/ +0.25dB	3dB	5.3dB/ 0dB	<3.5dB	2.7dB/ +0.6dB
Power/ Δ Power	16.2mW/ 0%	5.4mW/ 0%	22.5mW/ +100%	20mW/ +17.5%	23.4mW/ +3.4%	34.8mW	19.6mW/ +0.7%	14mW	2.6mW/ +1%
Supply Voltage	1.8V	2.7V	3.0V	2.5V	2.6V	2.2V	1.5V	1.2V	1.3V
Frequency	2.2GHz	880MHz	900MHz	2.2GHz	900MHz	48-1200MHz	900MHz	0.2-5.2GHz	2.5-10GHz
Process	0.35μm	0.25μm	0.35μm	0.25μm	0.25μm	0.18μm	0.18μm	65nm	0.13μm
Robustness over PVT	moderate	poor	good	moderate	moderate	moderate	moderate	good	good
Wideband?	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes

H. Post-Distortion

Similar to the DS method, the post-distortion (PD) technique also utilizes an auxiliary transistor's nonlinearity to cancel that of the main device, but it is more advanced in two aspects.

1) The auxiliary transistor is connected to the output of main device instead of directly to the input, minimizing the impact on input matching.
2) All transistors operate in saturation, resulting in more robust distortion cancellation.

Fig. 18 displays a conceptual diagram of PD as well as three implementations [30]–[32]. The auxiliary transistor $\mathrm{M_B}$ taps voltage $\mathrm{v_2}$ and replicates the nonlinear drain current of the main transistor $\mathrm{M_A}$, partially canceling both second- and third-order distortion terms. The nonlinear drain currents of $\mathrm{M_A}$ and $\mathrm{M_B}$ can be modeled as

$$\mathrm{i_A} = \mathrm{g_{1A}v_1} + \underbrace{\mathrm{g_{2A}v_1^2} + \mathrm{g_{3A}v_1^3}}_{\mathrm{f_{nonlin}(v_1)}} \tag{26}$$

$$\mathrm{i_B} = \mathrm{g_{1B}v_2} + \mathrm{g_{2B}v_2^2} + \mathrm{g_{3B}v_2^3}. \tag{27}$$

Next, suppose $\mathrm{v_2}$ is related to $\mathrm{v_1}$ by

$$\mathrm{v_2} = -\mathrm{b_1v_1} - \mathrm{b_2v_1^2} - \mathrm{b_3v_1^3} \tag{28}$$

where $\mathrm{b_1} - \mathrm{b_3}$ are generally frequency dependent and can be extracted from simulation. In Fig. 18(a), the cascode devices were assumed to be ideal current buffers [30]. The two nonlinear currents $\mathrm{i_A}$ and $\mathrm{i_B}$ sum at node $\mathrm{v_2}$, yielding $\mathrm{i_{out}}$:

$$\begin{aligned}\mathrm{i_{out}} = \mathrm{i}_A + \mathrm{i}_B &= \left(\mathrm{g_{1A}} - \mathrm{b_1g_{1B}}\right)\mathrm{v_1} \\ &+ \underbrace{\left(\mathrm{g_{2A}} - \mathrm{b_1^2g_{2B}} - \mathrm{b_2g_{1B}}\right)\mathrm{v_1^2}}_{\mathrm{2nd-order\ distortion}} \\ &+ \underbrace{\left(\mathrm{g_{3A}} - \mathrm{b_1^3g_{3B}} - \mathrm{g_{1B}b_3} - 2\mathrm{g_{2B}b_1b_2}\right)\mathrm{v_1^3}}_{\mathrm{3rd-order\ distortion}}.\end{aligned} \tag{29}$$

Note that in the PD method, both the main and auxiliary transistors operate in saturation with the same $\mathrm{g_{1,2,3}}$ polarity. Hence, $\mathrm{M_B}$ partially cancels the linear term as well; however, it does not substantially degrade the gain/NF because $\mathrm{M_B}$ is designed to be more nonlinear than $\mathrm{M_A}$ (i.e., $\mathrm{g_{2,3B}/g_{1B}} \gg \mathrm{g_{2,3A}/g_{1A}}$). Finally, note that among the three implementations, Fig. 18(b) and 18(d) probably have better performance in practical implementations, since both $\mathrm{M_A}$ and $\mathrm{M_B}$ are NMOS, which can be matched very well in layout. In Fig. 18(c), NMOS/PMOS transistors must have commensurate nonlinearity, but are hard to match across PVT.

I. Summary

Table III compares the IIP2/IIP3 improvement and gain/NF/power penalties of the previously discussed, state-of-the-art linearization techniques. We chose only one representative reference for each technique for brevity. The best performance per row has been marked with gray color. The modified DS method achieves the best IIP3(> 20 dBm); the IM2 injection method yields minimum degradation in NF, gain, and power; and the PD method renders robust linearity improvement.

Note that transconductance linearization methods are inherently broadband, however, to apply it on wideband LNAs, we should match the delays and phases from the main and auxiliary paths, including input matching/loading network, at the desired frequency band, so that the distortion cancellation is carried out with sufficient accuracy. Most reported techniques (e.g., IM2 injection, modified DS, and harmonic termination) dealing with second-order interaction are only limited to narrowband applications.

IIP2 calibration is another linearization technique that has been extensively reported for mixers, but still remains an open problem for LNAs. The concept of IIP2 calibration is to sense and correct the dc offset with an analog or digital feedback loop [33]–[36]. Some correction approaches for mixers include adjusting the LO bias [33], the load resistor/capacitor banks [34], the current source load [35], or injecting current at the mixer output [36]. It might be possible to apply some of the methods currently employed in mixers to differential LNAs.

III. NEW ISSUES FOR WIDEBAND APPLICATIONS

Growing research on reconfigurable multiband/multistandard and broadband transceivers has increased interest in broadband LNA design. In these transceivers, hundreds of channels could enter the LNA without any prefiltering, acting as in-band interferers. Moreover, nearby radios and on-chip transmitter leakage cause increased adjacent blockers, creating severe cross modulation, intermodulation, and desensitization. Therefore, a big design challenge for broadband LNAs is to achieve high linearity over a wide frequency range, lest the SNDR at the LNA output be dominated by distortion instead of noise. Furthermore, the old textbook argument that the LNA receives a small input signal amplitude is not valid for wideband LNAs. We consider

three main concerns: IIP2, $P_{1\,dB}$, and IIP2/IIP3 vs. two-tone frequency and spacing.

A. IIP2

Most linearization methods target narrowband applications and only cancel the third-order distortion, since the second-order nonlinearity is generally out of band in narrowband system. However, for wideband receivers, many channels are present concurrently and act as in-band interferences. Thus, the second-order intermodulation products generated by certain combination of interferences are highly likely to fall into the signal band. Hence, broadband LNAs should have a good IIP2 as well as IIP3. Often, in applications like digital TV, the required IIP2/3 must be derived from a multitone test such as complex second-order distortion (CSO) and composite triple beat distortion (CTB) [21].

A fully differential LNA will improve IIP2, but requires a transformer, which is expensive for wideband systems. Other IIP2 improvement techniques include the complementary/differential DS method [21], [22], [28] and post-distortion [30]–[32]. Moreover, in DSM processes, biasing a CS-stage at the maximum gain yields a high IIP2 [29].

B. 1 dB Compression Point

The "large-signal" 1 dB compression point $(P_{1\,dB})$ has traditionally not been a major concern for LNA designers because the LNA typically has a small input signal. However, in wideband receivers, LNAs receive the accumulated power from multiple channels, which could range from -10 to 0 dBm. For example, in the A/74 standard developed by the Advanced Television Systems Committee (ATSC), many transmitters are in close spectral proximity, so the receiver is exposed to more multicarrier adjacent energy. The maximum input power (the average of multiple tones) could even exceed 0 dBm [41]. Furthermore, severe transmitter leakage, poor isolation between antennas, and single-tone blockers with large peak-to-average ratio all require a high signal-handling capability, i.e., high $P_{1\,dB}$, for the LNA to prevent desensitization, gain compression, and clipping.

IIP2/IIP3-improvement techniques typically only work over small signal ranges, and do not improve $P_{1\,dB}$ because it is a large-signal parameter. At higher input amplitudes clipping occurs, and the $P_{1\,dB}$ worsens due to limited supply voltage/dc-bias current.

$P_{1\,dB}$-improvement techniques include

1) Increasing V_{dd} above nominal values to maximize the voltage headroom and performing substantial PVT simulation to guarantee breakdown/overstress will not occur.
2) Using low$-f_T$, thick-oxide transistors to handle larger voltage swings to allow even larger V_{dd}. Using such transistors degrades NF and high-frequency performance and raises cost.

Achieving high $P_{1\,dB}$ with thin-oxide devices and low supply voltages remains an open problem. Some possible approaches include

1) Cancel higher-order distortion, e.g., IM5 & IM7, since these become prominent at larger inputs and contribute to $P_{1\,dB}$.
2) Extend the effective input range of IM2/IM3 cancellation. One solution is to employ more auxiliary transistors in parallel in the DS method [11], [20]. Note that weak-inversion transistors being turned on and off at large voltage swing will add more high-order harmonic components to the circuit. A more robust solution is to combine triode and weak inversion transistors as auxiliary transistors [20].

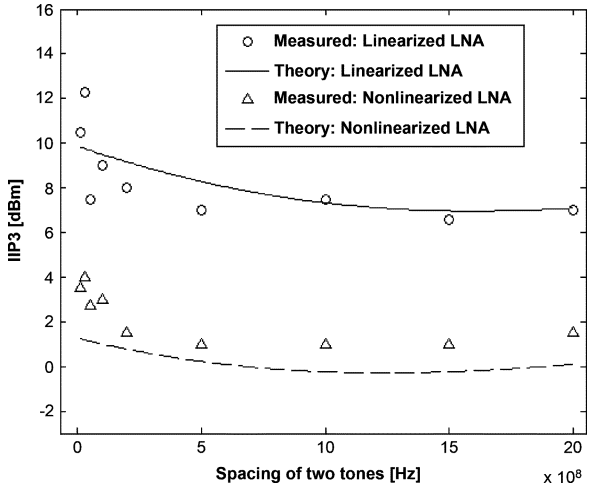


Fig. 19. Experimental and theoretical results of LNA IIP3 as a function of frequency spacing [32] (© 2009 IEEE).

3) Add source degeneration at the cost of extra noise.
4) Dynamic bias/dynamic supply [42].
5) Reduce the output voltage swing to relax the limitation from nonlinear output conductance. One option is to use a low-impedance load for the LNA, for example by choosing a passive mixer over an active mixer as the following stage. However, this choice requires a larger g_m stage and hence greater difficulty to linearize the transconductance.

C. IIP2/IIP3 Vs. Two Tone Frequency and Spacing

Broadband LNAs have flat gain/NF over the whole bandwidth. Likewise, IIP2/IIP3 should also be relatively flat over the signal band. Therefore, while narrowband systems typically use a specific interference frequency and a small tone spacing for the two-tone test, broadband systems require IIP2/IIP3 to be examined at various two-tone-spacing and center frequencies [32]. Fig. 19 shows an example plot.

Reactive components, such as those in the matching network, cause the frequency-dependence of IIP2/IIP3. Note that typically, this frequency-dependence is mild for operating frequencies below 1 GHz, so it is more of a concern for UWB systems (3.1–10.6 GHz) than for digital TV (54–880 MHz), for example.

IIP2 depends on two-tone-spacing. For two input-signal tones at $\omega_1 \& \omega_2$, the upper-frequency IM2 component is at $\omega_1 + \omega_2$, while the lower-frequency component is at $\omega_1 - \omega_2$. The IIP2 dependence on two-tone spacing is subtle when $\omega_1 - \omega_2$ is very small. There are two situations in which this dependence becomes more significant.

1) Large two-tone spacing, where larger frequency spacing yields correspondingly larger reactive effects.
2) Narrowband IM2 cancellation scheme. For example, in the complementary DS method with CG configuration shown in Fig. 11(b), the impedance from coupling capacitors increases with smaller two-tone spacing. Thus the ac-short condition worsens and degrades the IM2-cancellation effectiveness [21], [28].

The IIP3 dependence on two-tone spacing is mainly attributed to the "second-order interaction" as shown in (8). Therefore, the variations of $\Delta\omega$ cause the optimum point of the second-order interaction cancellation to change, resulting in worse linearity. For example, in the IM2-injection method [25] (Fig. 16), the squaring circuit experiences more phase shift at larger two-tone spacing, which degrades IIP3. In the harmonic-termination method [7], IIP3 degrades noticeably at larger $\Delta\omega$.

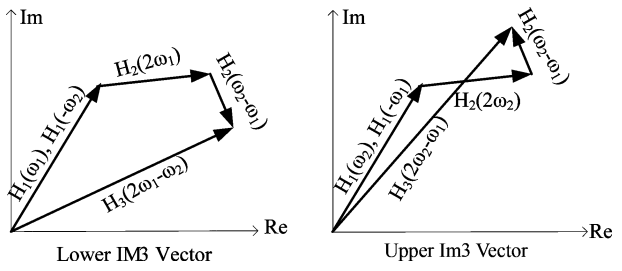


Fig. 20. Vector diagram showing the 180° out-of-phase contribution of $\omega_2-\omega_1$ term on the upper and lower IM3 components [38].

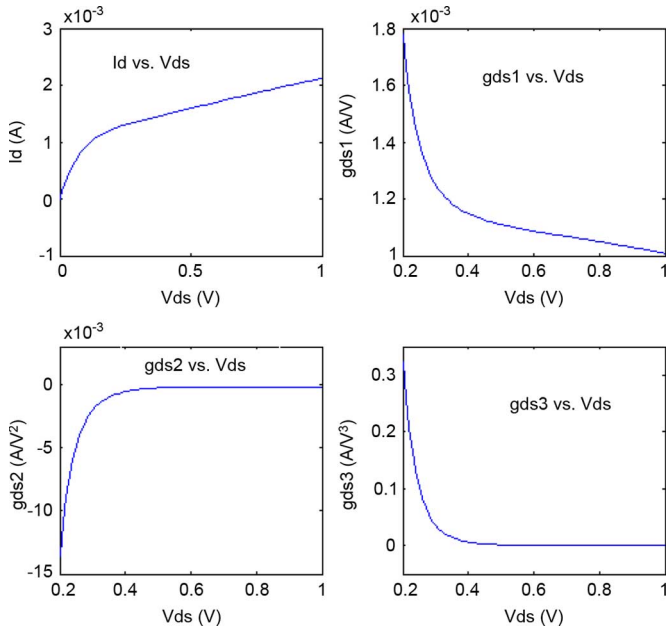


Fig. 21. NMOS output conductance nonlinearity characteristics (UMC 90 nm CMOS process, $\mathrm{W/L} = 20/0.08\ \mu\mathrm{m}$, $\mathrm{V_{gs}} = 0.5\ \mathrm{V}$, $\mathrm{V_{th}} = 0.26\ \mathrm{V}$).

Another major contributor to this $\mathrm{IIP_3}$ dependence is the IM3 asymmetry, also called "sideband asymmetry." IM3 asymmetry is attributed to various types of memory effects [37]–[40], but for CMOS LNAs specifically, it is because the reactive part of the circuit impedance (e.g., termination impedance) at $\omega_2-\omega_1$ has a 180°-out-of-phase contribution to the IM3 components at $(2\omega_1-\omega_2)$ and $(2\omega_2-\omega_1)$. This concept is qualitatively illustrated by the vector diagram in Fig. 20 [38], where the $H_{1,2,3}$ refers to the first-, second-, and third-order Volterra-Series coefficients. The IM3 components at $(2\omega_2-\omega_1)$ and $(2\omega_1-\omega_2)$ have different imaginary parts (i.e., reactance), resulting in IM3 asymmetry.

Note that this IM3 asymmetry depends on bias and frequency. For very small two-tone spacing, it is hard to see any IM3 asymmetry since the reactive-impedance effect at $\Delta\omega$ is negligible; but for larger $\Delta\omega$, the reactive impedances at the second-harmonic frequency also contribute differently to the lower/upper IM3 components, which worsens the IM3 asymmetry [7] and also indicates a more obvious IIP3 dependence on two-tone-spacing. However, proper bias can reduce this IM3 asymmetry [40]. Note that in the multitone case, Adjacent channel power ratio (ACPR) asymmetry is defined correspondingly.

IV. LNA Linearization in DSM Technology

A. Nonlinearity From Output Conductance $\mathrm{g_{ds}}$

Distortion of MOS transistors is mainly caused by the nonlinear transconductance $(\mathrm{g_m})$ and output conductance $(\mathrm{g_{ds}})$. Previously published linearization techniques mainly focus on linearizing $\mathrm{g_m}$, assuming that (1) drain current $\mathrm{i_{ds}}$ is controlled only by the gate-source voltage $\mathrm{V_{gs}}$, and (2) $\mathrm{g_{ds}}$ nonlinearity is negligible. These assumptions are valid for small load resistance, small voltage gain, small input signal, and a drain-source voltage $(\mathrm{V_{ds}})$ sufficiently large that the small-signal variation of $\mathrm{V_{ds}}$ does not appreciably perturb the bias point.

However, as technology scales down, the $\mathrm{g_{ds}}$ nonlinearity becomes more prominent. Current $\mathrm{i_{ds}}$ is controlled not only by $\mathrm{V_{gs}}$ but also the $\mathrm{V_{ds}}$, which can be approximated by the two-dimensional Taylor series [29], [44]

$$\begin{aligned}\mathrm{i_{ds}}(\mathrm{V_{gs}},\mathrm{V_{ds}}) = {} & \mathrm{g_1V_{gs}}+\mathrm{g_2V_{gs}^2}+\mathrm{g_3V_{gs}^3} \\ & +\mathrm{g_{ds1}V_{ds}}+\mathrm{g_{ds2}V_{ds}^2}+\mathrm{g_{ds3}V_{ds}^3} \\ & +c_{(1,1)}\mathrm{V_{gs}V_{ds}}+\mathrm{c_{(2,1)}V_{gs}^2V_{ds}} \\ & +\mathrm{c_{(1,2)}V_{gs}V_{ds}^2} \end{aligned} \tag{30}$$

where $\mathrm{g_i}$ is the $\mathrm{i^{th}}$-order transconductance as defined in (6); $\mathrm{g_{dsi}}$ represents the nonlinear output conductance effect which is proportional to the $\mathrm{i_{ds}}$ derivatives with respect to $\mathrm{V_{ds}}$; $\mathrm{c_{(m,n)}}$ is the cross modulation term describing the dependence of $\mathrm{g_i}$ on $\mathrm{V_{ds}}$ or $\mathrm{g_{dsi}}$ on $\mathrm{V_{gs}}$, as formulated in (31)

$$\begin{aligned}\mathrm{g_{dsi}} &= \frac{1}{\mathrm{i!}}\frac{\partial^{\mathrm{i}}\mathrm{I_{DS}}}{\partial\mathrm{V_{DS}^i}} \\ \mathrm{c_{(m,n)}} &= \frac{1}{\mathrm{m!n!}}\frac{\partial^{\mathrm{m+n}}\mathrm{I_{DS}}}{\partial\mathrm{V_{GS}^m}\partial\mathrm{V_{DS}^n}}\end{aligned} \tag{31}$$

To characterize the $\mathrm{g_{ds}}$ nonlinearity for a single transistor, we fix its $\mathrm{V_{gs}}$, and sweep the $\mathrm{V_{ds}}$, by taking the first three derivatives of the drain-source dc current $\mathrm{i_{ds}}$ with respect to $\mathrm{V_{ds}}$ [as defined in (31)] at every dc bias point, we can obtain Fig. 21. It is observed that the drain current is modulated a lot by $\mathrm{V_{ds}}$. g_{ds3} is large when the transistor operates at small $\mathrm{V_{ds}}$; while it decreases for large $\mathrm{V_{ds}}$ values.

Here we assume a negligible nonlinearity contribution from g_{mb}, otherwise three-dimensional Taylor series should be used instead.

From (30), the distortion is contributed by four parts.

1) $\mathrm{g_m}$ nonlinearity due to nonlinear $\mathrm{i_{ds}}-\mathrm{V_{gs}}$ relation.
2) $\mathrm{g_{ds}}$ nonlinearity from channel length modulation effect. Note that $\mathrm{g_{ds}}$ contributes less nonlinearity when device operates deeper into saturation region.
3) the dependence of $\mathrm{g_m}$ on $\mathrm{V_{ds}}$, (partially due to the drain induced barrier lowering (DIBL) effect [29].
4) the dependence of $\mathrm{g_{ds}}$ on $\mathrm{V_{gs}}$, especially in saturation region [44].

The cross modulation effect remains fairly constant for a broad range of $\mathrm{V_{gs}}$, while $\mathrm{g_m}$ is more linear and $\mathrm{g_{ds}}$ becomes more nonlinear as $\mathrm{V_{gs}}$ increases, $\mathrm{V_{ds}}$ decreases, and transistors operate close to the linear region. In [29] the $\mathrm{V_{gs}}$. $\mathrm{V_{ds}}$ cross term $c_{1,2}$ cancels the intrinsic second-order distortion $(\mathrm{g_2})$ to obtain an amplifying stage with high IIP2. Note that when $\mathrm{g_{ds}}$ nonlinearity dominates (i.e., output limited), the tradeoff between gain and linearity becomes more severe.

TABLE IV
DOMINANT CONTRIBUTOR TO DISTORTION UNDER VARIOUS CONDITIONS

	g_m	g_{ds}
Small load resistance	√	
Large load resistance		√
Small voltage gain Av	√	
Large voltage gain Av		√
High frequency	√	
Low frequency		√
Saturation region	√	

B. MOSFET Capacitance

For the most part, the capacitances of a saturation-region transistor are linear at an operating frequency less than $f_T/10$ [44]. Therefore they do not directly contribute to distortion [43]. However, if a strong blocker is present (e.g., in the order of 0dBm), C_{gs} varies significantly around the threshold voltage, and its nonlinearity becomes significant. Also, as previously mentioned, C_{gd} provides a feedback path for the "second-order interaction," and this C_{gd} effect becomes more visible as the load impedance increases. At high frequency, C_{gd} and C_{db} reduce total output impedance and hence the output voltage swing, helping to mitigate the nonlinear g_{ds} effect. Therefore, g_m nonlinearity dominates at high frequency, while g_{ds} nonlinearity dominates at low frequency [43]. However, in those circuits where capacitive components are tuned out for a matched load, g_{ds} nonlinearity is still prominent at high frequencies. The substrate affects linearity through C_{db} with higher operating frequency, and this effect varies with different substrate-doping profiles [43]. Generally, IIP3 improves as substrate doping increases [47]. The effect from substrate leakage current can typically be neglected [45].

Table IV provides a summary: the mark "√" denotes whether g_m or g_{ds} dominates the distortion under the given conditions.

C. Impact of Technology Scaling On Linearity

As channel length decreases, the velocity saturation effect becomes prominent, i.e., the drain current saturates at smaller V_{ds}. Thus, the long-channel equation for drain current in saturation region needs to be modified as [5]

$$\begin{aligned} I_{ds} &= \frac{\mu C_{ox}}{2}\frac{W}{L}(V_{gs}-V_t)\left[(V_{gs}-V_t) \parallel (LE_{sat})\right] \\ &\approx \frac{\mu C_{ox}}{2} W (V_{gs}-V_t) E_{sat} (\text{for small } L) \end{aligned} \tag{32}$$

where E_{sat} is the field strength at which the carrier velocity drops to half the value extrapolated from low-field mobility. g_m becomes more linear

$$g_m = \frac{\partial I_{ds}}{\partial V_{gs}} = \frac{\mu C_{ox}}{2} W E_{sat}. \tag{33}$$

The vertical-field mobility degradation effect also helps to linearize g_m in DSM process. The long-channel equation for drain current can be modified as

$$I_{ds} = \frac{\mu C_{ox}}{2}\frac{W}{L}\frac{(V_{gs}-V_t)^2}{1+\theta(V_{gs}-V_t)} \tag{34}$$

where $\theta \propto 1/t_{ox}$ models vertical-field mobility degradation. Equation (34) reduces to $I_{ds} \propto (V_{gs}-V_t)$ as $V_{gs}-V_t$ increases, resulting in a linear I-V curve, and g_m becomes constant with respect to bias voltage:

$$g_m \approx \frac{\mu C_{ox}}{2}\frac{W}{L}\frac{1}{\theta}. \tag{35}$$

On the other hand, g_{ds} is more nonlinear for shorter channel length, as proven by the experimental data in [43]. Furthermore, the reduced supply/V_{dsat} values result in the device being biased closer to the triode-saturation boundary, which worsens the g_{ds} nonlinearity. Consequently, the maximum OIP3 occurs with smaller load impedance (which mitigates the distortion contribution from nonlinear g_{ds}) and the peak IIP3 shifts to lower V_{gs}[47], since a smaller overdrive voltage allows the device to stay far away from the triode-saturation boundary while still keeping g_{ds} nonlinearity small.

The "sweet spot" in the optimal biasing technique(discussed in Section II-C) will systematically shift to higher bias-current density I_{ds}/W (i.e., larger overdrive voltage) as technology scales down [44], which means larger power is required to preserve linearity.

As oxide thickness decreases, poly-gate depletion increases, and the nonlinear gate capacitance develops strong second-order derivatives (C_{g2}) with respect to V_{gs}, which contribute to significant third-order distortion (g_3) in drain current, as shown in (36) and (37) [46]

$$I_{ds} = Qv \approx Qv_{sat} \tag{36}$$

$$\begin{aligned} g_3 &= \frac{\partial^3 I_{ds}}{\partial V_{gs}^3} \approx \frac{\partial^3 Q}{\partial V_{gs}^3} v_{sat} \approx \frac{\partial^2 C(V_{gs})}{\partial V_{gs}^2} v_{sat} \\ &= C_{g2} v_{sat} \end{aligned} \tag{37}$$

where Q is the channel charge density along the current direction, v is the carriers' velocity, and v_{sat} is the saturated velocity for sufficiently high field. Thus, distortion increases with thinner oxides.

DIBL becomes more severe in DSM process, besides a V_{ds}-dependent g_m, DIBL also affects the linearity by changing the effective V_{th} [48]. Measured results in [48] shows that the distortion is more sensitive to DIBL effect when the drain voltage increases and the MOSFET operates in moderate region (i.e., V_{gs} is slightly higher than V_{th}).

Finally, each process has a "low frequency limit" (LFL), below which the MOSFET exhibits fairly frequency-independent linearity. LFL is closely related to the device speed and can be approximated as $f_T/5$ [48]. Therefore, it is easier to achieve IIP2/IIP3 flatness over the signal band in smaller-size technology.

In summary, as technology scales down, lower supply voltages reduce the headroom and can lead to greater nonlinearity from g_{ds}, necessitating multidimensional Taylor analysis to model the nonlinear I_{ds}. Higher-order effects such as DIBL, velocity saturation, and poly-gate depletion all affect linearity. A key challenge resides in delivering high linearity with core transistors and with a low supply voltage in the DSM processes.

V. REMARKS FOR HIGH LINEARITY LNA DESIGN

Besides applying explicit linearization techniques to the circuit, some general guidelines are helpful for designing a high-linearity LNA.

A. To Reduce g_m-Induced Distortion

From (34), the low-frequency expressions for second- and third-order intercept points A_{IIP2} and A_{IIP3} are [3], [5]:

$$A_{IIP2}^2 = \left|\frac{g_1}{g_2}\right| = \left|V_{dsat}(2+\theta V_{dsat})(1+\theta V_{dsat})\right| \tag{38}$$

$$\begin{aligned} A_{IIP3}^2 &= \frac{4}{3}\left|\frac{g_1}{g_3}\right| \\ &= \left|\frac{4}{3}\frac{V_{dsat}}{\theta}(2+\theta V_{dsat})(1+\theta V_{dsat})^2\right| \end{aligned} \tag{39}$$

where $V_{dsat} = V_{gs} - V_{th}$. Equations (38) and (39) indicate that increasing V_{dsat} improves both IIP2 and IIP3. Therefore, given sufficient voltage headroom, maximizing V_{dsat} and minimizing transistor sizes helps to minimize parasitics and to linearize the circuit at the cost of increased dc current.

B. To Reduce g_{ds}-Induced Distortion

Increasing supply voltage mitigates the g_{ds} effect and also improves $P_{1\,dB}$, but the voltage drop across core transistors must be ensured not to exceed the safe operation value.

Provided sufficient voltage headroom, adding cascode device allows the output impedance from transistors to be much larger than load resistor, yielding a more linear output load.

With cascode transistor, most of the output swing will show as V_{ds} variation at the cascode transistor, while the V_{ds} of input transistor remains relatively constant. Therefore, the nonlinear output conductance of the cascode transistor has more contribution to the overall distortion. It is helpful to bias the cascode transistor at smaller V_{gs} (i.e., lower overdrive voltage) to tolerate a larger swing at the drain.

If supply voltage cannot be increased, we can:

1) use longer channel length to reduce the channel length modulation effect (assuming speed is not an issue);
2) reduce the load resistance of the LNA, which may affect the design of other building blocks in the receiver.

C. To Reduce Second-Order Distortion

1) Biasing a CS-stage at the maximum gain yields a high IIP2 in DSM process [29].
2) Biasing the device for maximum f_T yields minima in the second-harmonic [48]; this intrinsic distortion cancellation results from opposite contributions of gate capacitance and g_m, as the device enters the linear region from saturation.

D. Other Tips

For inductively degenerated CS-LNAs, we can reduce Q to mitigate the "Q boosting" effect [5], provided that there is enough margin in NF and gain. Since C_{gs} has negligible effect on linearity [45], an external capacitor can be added in parallel with C_{gs} to allow more freedom for input transistor sizing. On the other hand, CG-LNAs generally provide better linearity than CS-LNAs [32] because it does not have this "Q boosting."

Use cascode transistors whenever possible because they:

1) reduce second-order interaction through C_{gd};
2) reduce the voltage swing across each active device, improving reliability for DSM devices.

VI. Conclusions

We have reviewed eight categories of CMOS LNA-linearization techniques and discussed the tradeoffs among linearity, power, and PVT variations. We subsequently discussed wideband-LNA-linearization issues for the emerging broadband transceivers, noting that IIP2 is becoming just as important as IIP3, and that improving $P_{1\,dB}$ is also necessary for wideband applications to improve high-signal-handling capability. Issues in DSM processes, such as nonlinear output conductance were examined. A key challenge resides in delivering high linearity with core transistors and low supply voltage in the DSM processes. Linearization techniques for canceling higher-order distortion terms (beyond third order), linearizing output conductance, and improving LNA $P_{1\,dB}$ still remain open problems. Finally, we presented general design guidelines for high-linearity LNAs.

Acknowledgment

The authors wish to thank W. He, E. Pankratz, X. Chen, M. Onabajo, M. Mobarak, Dr. C. Xin, Dr. S. Krishnamurthy, and A. Amer for proofreading the manuscript. The authors also wish to thank Dr. S. Lou, J. Wardlaw, and J. Kim for helpful discussions. The authors also want to acknowledge support from the TI Wireless Group.

References

[1] E. Sánchez-Sinencio and J. Silva-Martinez, "CMOS transconductance amplifiers, architectures and active filters: A tutorial," *IEE Proc.–Circuits Devices Syst.*, vol. 147, no. 1, pp. 3–12, Feb. 2000.

[2] V. Aparin, "State-of-the-art techniques for high linearity CMOS low noise amplifiers," presented at the IEEE RFIC Symp. Workshop WSC, Honolulu, HI, 2007.

[3] B. H. Leung, *VLSI for Wireless Communication*. Englewood Cliffs, NJ: Prentice-Hall, 2002.

[4] W. Sansen, "Distortion in elementary transistor circuits," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 46, no. 3, pp. 315–325, Mar. 1999.

[5] T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[6] S. Narayanan, "Application of Volterra series to intermodulation distortion analysis of transistor feedback amplifies," *IEEE Trans. Circuit Theory*, vol. CT-17, no. 4, pp. 518–527, Nov. 1970.

Harmonic termination

[7] V. Aparin and C. Persico, "Effect of out-of-band terminations on intermodulation distortion in common-emitter circuits," in *IEEE MTT-S Int. Microwave Symp, Dig.*, Jun. 1999, vol. 3, pp. 977–980.

[8] V. Aparin and L. E. Larson, "Linearization of monolithic LNAs using low-frequency low-impedance input termination," in *Proc. Eur. Solid-State Circuits Conf.*, Sep. 2003, pp. 137–140.

[9] K. L. Fong, "High-frequency analysis of linearity improvement technique of common-emitter trans-conductance stage using a low-frequency trap network," *IEEE J. Solid-State Circuits*, vol. 35, no. 8, pp. 1249–1252, Aug. 2000.

[10] T. W. Kim, "A common-gate amplifier with transconductance nonlinearity cancellation and its high-frequency analysis using the Volterra series," *IEEE Trans. Microw. Theory Tech.*, vol. 57, no. 6, pp. 1461–1469, Jun. 2009.

[11] B. Kim, J.-S. Ko, and K. Lee, "Highly linear CMOS RF MMIC amplifier using multiple gated transistors and its volterra series analysis," in *IEEE MTT-S Int. Microwave Symp. Dig.*, May 2001, vol. 1, pp. 515–518.

[12] J. S. Fairbanks and L. E. Larson, "Analysis of optimized input and output harmonic termination on the linearity of 5 GHz CMOS radio frequency amplifiers," in *Radio Wireless Conf.*, Aug. 2003, pp. 293–296.

[13] T. W. Kim, B. Kim, and K. Lee, "Highly linear receiver front-end adopting MOSFET transconductance linearization by multiple gated transistors," *IEEE J. Solid-State Circuits*, vol. 39, no. 1, pp. 223–229, Jan. 2004.

[14] X. Fan, H. Zhang, and E. Sánchez-Sinencio, "A noise reduction and linearity improvement technique for a differential cascode LNA," *IEEE J. Solid-State Circuits*, vol. 43, no. 3, pp. 588–599, Mar. 2008.

Optimum biasing

[15] V. Aparin, G. Brown, and L. E. Larson, "Linearization of CMOS LNAs via optimum gate biasing," in *Proc. IEEE Int. Circuits Syst. Symp.*, Vancouver, BC, Canada, May 2004, vol. 4, pp. 748–751.

Feedforward

[16] Y. Ding and R. Harjani, "$A + 18$ dBm IIP3 LNA in 0.35 μm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2001, pp. 162–163.

[17] E. Keehr and A. Hajimiri, "Equalization of IM3 products in wideband direct-conversion receivers," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 204–205.

Derivative superposition (DS)

[18] Y. S. Youn, J. H. Chang, K. J. Koh, Y. J. Lee, and H. K. Yu, "A 2 GHz 16 dBm IIP3 low noise amplifier in 0.25 μm CMOS technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2003, pp. 452–453.

[19] C. Xin and E. Sánchez-Sinencio, "A linearization technique for RF low noise amplifier," in *Proc. IEEE Int. Circuits Syst. Symp.*, Vancouver, BC, Canada, May 2004, vol. IV, pp. 313–316.

[20] H. M. Geddada, J. W. Park, and J. Silva-Martinez, "Robust derivative superposition method for linearizing broadband LNAs," *IEE Electron. Lett.*, vol. 45, no. 9, pp. 435–436, Apr. 2009.

Complementary DS

[21] D. Im, I. Nam, H. Kim, and K. Lee, "A wideband CMOS low noise amplifier employing noise and IM2 distortion cancellation for a digital TV tuner," *IEEE J. Solid-State Circuits*, vol. 44, no. 3, pp. 686–698, Mar. 2009.

Differential DS

[22] T. W. Kim and B. Kim, "A 13-dB IIP3 improved low-power CMOS RF programmable gain amplifier using differential circuit transconductance linearization for various terrestrial mobile D-TV applications," *IEEE J. Solid-State Circuits*, vol. 41, no. 4, pp. 945–953, Apr. 2006.

Modified DS

[23] V. Aparin and L. E. Larson, "Modified derivative superposition method for linearizing FET low-noise amplifiers," *IEEE Trans. Microw. Theory Tech.*, vol. 53, no. 2, pp. 571–581, Feb. 2005.

[24] S. Ganesan, E. Sánchez-Sinencio, and J. Silva-Martinez, "A highly linear low noise amplifier," *IEEE Trans. Microw. Theory Tech.*, vol. 54, no. 12, pp. 4079–4085, Dec. 2006.

IM2 injection

[25] S. Lou and H. C. Luong, "A linearization technique for RF receiver front-end using second-order-intermodulation injection," *IEEE J. Solid-State Circuits*, vol. 43, no. 11, pp. 2404–2412, Nov. 2008.

Noise/distortion cancellation

[26] F. Bruccoleri, E. A. M. Klumperink, and B. Nauta, "Wide-band CMOS low-noise amplifier exploiting thermal noise canceling," *IEEE J. Solid-State Circuits*, vol. 39, no. 2, pp. 275–282, Feb. 2004.

[27] J. Jussila and P. Sivonen, "A 1.2-V highly linear balanced noise-cancelling LNA in 0.13-μm CMOS," *IEEE J. Solid-State Circuits*, vol. 43, no. 3, pp. 579–587, Mar. 2008.

[28] W. Chen, G. Liu, B. Zdravko, and A. M. Niknejad, "A highly linear broadband CMOS LNA employing noise and distortion cancellation," *IEEE J. Solid-State Circuits*, vol. 43, no. 5, pp. 1164–1176, May 2008.

[29] S. C. Blaakmeer, E. A. M. Klumperink, D. M. W. Leenaerts, and B. Nauta, "Wideband balun-LNA with simultaneous output balancing, noise-canceling and distortion-canceling," *IEEE J. Solid-State Circuits*, vol. 43, no. 6, pp. 1341–1350, Jun. 2008.

Post-distortion

[30] N. Kim, V. Aparin, K. Barnett, and C. Persico, "A cellular-band CDMA $0.25\ \mu$m CMOS LNA linearized using active post-distortion," *IEEE J. Solid-State Circuits*, vol. 41, no. 7, pp. 1530–1534, Jul. 2006.

[31] T.-S. Kim and B.-S. Kim, "Post-linearization of cascode CMOS LNA using folded PMOS IMD sinker," *IEEE Microw. Wireless Comp. Lett.*, vol. 16, no. 4, pp. 182–184, Apr. 2006.

[32] H. Zhang, X. Fan, and E. Sánchez-Sinencio, "A low-power, linearized, ultra-wideband LNA design technique," *IEEE J. Solid-State Circuits*, vol. 44, no. 2, pp. 320–330, Feb. 2009.

IIP2 calibration

[33] D. Kaczman, M. Shah, M. Alam, M. Rachedine, D. Cashen, L. Han, and A. Raghavan, "A single-chip 10-Band WCDMA/HSDPA 4-Band GSM/EDGE SAW-less CMOS receiver with DigRF 3 G interface and $+90$ dBm IIP2," *IEEE J. Solid-State Circuits*, vol. 44, no. 3, pp. 718–739, Mar. 2009.

[34] M. Hotti, J. Ryynanen, K. Kivekas, and K. Halonen, "An IIP2 calibration technique for direct conversion receivers," in *Proc. IEEE Int. Circuits Syst. Symp.*, Vancouver, BC, Canada, May 2004, vol. 4, pp. 257–260.

[35] W. Kim, S. Yang, Y. Moon, J. Yu, H. Shin, W. Choo, and B. Park, "IP2 calibrator using common mode feedback circuitry," in *Proc. Eur. Solid-State Circuits Conf.*, 2005, pp. 231–234.

[36] H. Darabi, H. Kim, J. Chiu, B. Ibrahim, and L. Serrano, "An IP2 improvement technique for zero-if down-converters," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2006, pp. 464–465.

IIP2/IIP3 dependence on two-tone spacing

[37] W. Bosch and G. Gatti, "Measurement and simulation of memory effects in predistortion linearizers," *IEEE Trans. Microw. Theory Tech.*, vol. 37, no. 12, pp. 1885–1890, Dec. 1989.

[38] J. F. Sevic, K. L. Burguer, and M. B. Steer, "A novel envelope-termination load-pull methods for ACPR optimization of RF/microwave power amplifiers," in *IEEE MTT-S Int. Microwave Symp. Dig.*, Baltimore, MD, 1998, pp. 601–605.

[39] N. B. de Carvalho and J. C. Pedro, "Two-tone IMD asymmetry in microwave power amplifiers," in *IEEE MTT-S Int. Microwave Symp. Dig.*, Boston, MA, 2000, pp. 445–448.

[40] N. Carvalho and J. Pedro, "A comprehensive explanation of distortion sideband asymmetries," *IEEE Trans. Microw. Theory Tech.*, vol. 50, no. 9, pp. 2090–2101, Sep. 2002.

Others

[41] N. Cowley and R. Hanrahan, "ATSC compliance and tuner design implications, video/imaging design line," *EE Times* Nov. 10, 2005 [Online]. Available: http://www.videsignline.com/showArticle.jhtml?articleID=173601582

[42] S. S. Taylor and J. S. Duster, "High-linearity low noise amplifier and method," U.S. Patent 0 278 220, Nov. 13, 2008.

LNA linearization in deep-submicrometer technology

[43] S. Kang, B. Choi, and B. Kim, "Linearity analysis of CMOS for RF application," *IEEE Trans. Microw. Theory Tech.*, vol. 51, no. 3, pp. 972–977, Mar. 2003.

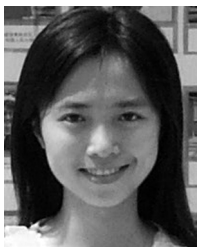
[44] B. Toole, C. Plett, and M. Cloutier, "RF circuit implications of moderate inversion enhanced linear region in MOSFETS," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 51, no. 2, pp. 319–328, Feb. 2004.

[45] R. A. Baki, T. K. K. Tsang, and M. N. El-Gamal, "Distortion in RF CMOS short-channel low-noise amplifiers," *IEEE Trans. Microw. Theory Tech.*, vol. 54, no. 1, pp. 46–56, Jan. 2006.

[46] C. H. Choi, Z. Yu, and R. W. Dutton, "Impact of poly-gate depletion on MOS RF linearity," *IEEE Electron Device Lett.*, vol. 24, no. 5, pp. 330–332, May 2003.

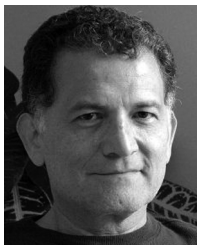
[47] R. van Langevelde, L. F. Tiemeijer, R. J. Havens, M. J. Knitel, R. F. M. Ores, P. H. Woerlee, and D. B. M. Klaassen, "RF-distortion in deep-submicron CMOS technologies," in *IEEE IEDM Tech. Dig.*, 2000, pp. 807–810.

[48] T. Lee and Y. Cheng, "High-frequency characterization and modeling of distortion behavior of MOSFETS for RF IC design," *IEEE J. Solid-State Circuits*, vol. 39, no. 9, pp. 1407–1414, Sep. 2004



Heng Zhang (S'07) was born in Guangzhou, China. She received the B.S. degree in electrical engineering from Peking University, Beijing, China, in 2004, and the Ph.D. degree in electrical engineering from Texas A&M University, College Station, in 2010.

In the summer and fall of 2006, she was an Analog IC Design Engineer (Co-op) with Texas Instrument, Dallas, TX, where she designed a low power ADC for hard disk applications. In summer 2007, she was with the RF and Analog Technologies Department, UMC, Sunnyvale, CA, where she researched digital calibration techniques for ADCs. Since August 2010, she has been with the Analog and Mixed Signal Group, Broadcom Corporation, Irvine, CA, working on high-speed transceivers for optical and backplane/cable applications. Her research interests include data converters and high-speed/RF circuits. Her Web site can be found at http://www.ece.tamu.edu/~hzhang.



Edgar Sánchez-Sinencio (S'72–M'74–SM'83–F'92–LF'10) was born in Mexico City, Mexico. He received the degree in communications and electronic engineering (Professional degree) from the National Polytechnic Institute of Mexico, Mexico City, the M.S.E.E. degree from Stanford University, Stanford, CA, and the Ph.D. degree from the University of Illinois at Champaign-Urbana, in 1966, 1970, and 1973, respectively.

He is currently the TI J. Kilby Chair Professor and Director of the Analog and Mixed-Signal Center at Texas A&M University. His research work has more than 2636 citations according to the Thomson Reuters Scientific Citation Index. He has graduated 42 M.Sc. and 34 Ph.D. students. He is a coauthor of six books on different topics, such as RF circuits, low-voltage low-power analog circuits, and neural networks. His present interests are in the area of power management, RF-communication circuits, analog and medical electronics circuit design.

Dr. Sánchez-Sinencio is a former Editor-in-Chief of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II. In November 1995 he was awarded a Honoris Causa Doctorate by the National Institute for Astrophysics, Optics and Electronics, Mexico. This degree was the first honorary degree awarded for microelectronic circuit design contributions. He is a corecipient of the 1995 Guillemin-Cauer Award for his work on cellular networks. He was also the corecipient of the 1997 Darlington Award for his work on high-frequency filters. He received the IEEE Circuits and Systems Society Golden Jubilee Medal in 1999. He also received the prestigious IEEE Circuits and Systems Society 2008 Technical Achievement Award. His Web site can be found at http://amesp02.tamu.edu/~sanchez/.