Complexity Reduced Models for Radio Frequency Power Amplifiers' Modeling and Linearization

by

Marie-Claude Fares

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Applied Science

in

Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2009

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Marie-Claude Fares

Abstract

Radio frequency (RF) communications are limited to a number of frequency bands scattered over the radio spectrum. Applications over such bands increasingly require more versatile, data extensive wireless communications that leads to the necessity of high bandwidth efficient interfaces, operating over wideband frequency ranges. Whether for a base station or mobile device, the regulations and adequate transmission of such schemes place stringent requirements on the design of transmitter front-ends. Increasingly strenuous and challenging hardware design criteria are to be met, especially so in the design of power amplifiers (PA), the bottle neck of the transmitter's design tradeoff between linearity and power efficiency. The power amplifier exhibits a nonideal behavior, characterized by both nonlinearity and memory effects, heavily affecting that tradeoff, and therefore requiring an effective linearization technique, namely Digital Predistortion (DPD). The effectiveness of the DPD is highly dependent on the modeling scheme used to compensate for the PA's nonideal behavior. In fact, its viability is determined by the scheme's accuracy and implementation complexity. Generic behavioral models for nonlinear systems with memory have been used, considering the PA as a black box, and requiring RF designers to perform extensive testing to determine the minimal complexity structure that achieves satisfactory results. This thesis first proposes a direct systematic approach based on the parallel Hammerstein structure to determine the exact number of coefficients needed in a DPD. Then a physical explanation of memory effects is detailed, which leads to a close-form expression for the characteristic behavior of the PA entirely based on circuit properties. The physical expression is implemented and tested as a modeling scheme. Moreover, a link between this formulation and the proven behavioral models is explored, namely the Volterra series and Memory Polynomial. The formulation shows the correlation between parameters of generic behavioral modeling schemes when applied to RF PAs and demonstrates redundancy based on the physical existence or absence of modeling terms, detailed for the proven Memory polynomial modeling and linearization scheme.

Acknowledgements

This document is the last milestone of a long journey that has been anything but ordinary. I would like to thank you as a reader, and hope you see and appreciate not just the work presented herein but everyone's effort behind it.

My sincere thoughts go to my supervisor, Dr. Slim Boumaiza, whose guidance and patience were invaluable for the completion of this work. His support seemed unlimited, always professional yet with real concern to his students. Most of all, he taught me by example that "when you work you are a flute through whose heart the whispering of the hours turns to music, and if you cannot work with love but only with distaste, it is better that you should leave your work and sit at the gate of the temple and take alms of those who work with joy." K.Gibran.

My family; Alphonso my Dad, for he taught me to have enough reason to persevere, yet enough passion to keep moving ahead. To keep a clear sight of what is essential and most valuable, so as not to fear challenges and bold steps. Mona my Mom, working her magic and being an infinite source of blessings and happiness. My brother, Fares, my dearest friend whose endeavors and visions in his career are most inspirational, Paul and Amin, sharing laughter and sarcasm every now and then.

My friends; my coworkers, members of the EMRG lab, friends who reminded me of work and helped me out, and those who made me get away at its expense. All was much appreciated and needed.

This work was supported by the University of Waterloo and the National Sciences and Engineering Research Council (NSERC).

Table of Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

3G, 4G	Third and Fourth Generation of Telecommunication Standards
CMOS	Complementary Metal-Oxide Semiconductor
DPD	Digital Predistortion
DUT	Device Under Test
DSP	Digital Signal Processing
FIR	Finite Impulse Response
IMD	Intermodulation Distortion
LNA	Low-Noise Amplifier
M-Polynomial	Memory Polynomial
ME	Memory Effects
OFDM	Orthogonal Frequency-Division Multiplexing
PA	Power Amplifier
P-Hammerstein	Parallel Hammerstein
RF	Radio Frequency
WCDMA	Wideband Code Division Multiple Access
WiMAX	Worldwide Interoperability for Microwave Access

Chapter 1
Introduction

1.1 Introduction

The last decade has seen an exponential growth in mobile and telecommunication services. Consumer requirements increased the demand for low-cost, low-power and reduced size and weight equipments. Adding to the complexity of designs to meet consumer demands are the stringent operational and technical requirements. While wireless usage became more complex demanding wider bandwidths, the frequency spectrum allocations available are constant and limited. Therefore an increasingly higher level of design and integration is needed to meet these requirements. Recent advances in semiconductor technologies (i.e. complementary metal-oxide semiconductor CMOS) allow highly efficient low-frequency digital signal processing. However the most power consuming and nonideal part in the wireless transmission remains the RF front end whose most critical block is the RF Power Amplifier (PA). Therefore, improving the PA performance greatly affects the performance of the overall wireless system on many levels including linearity and power efficiency. These two mentioned criteria are essential in the design of an implementable and operable efficient wireless communication system, in fact it is the trade-off every communication system, specifically front-end part, is trying to optimize.

Much effort has been put into improving the physical PA designs in order to achieve higher power efficiency in a linear region of operation of the PA. Significant advances in power efficiency have been achieved with newer technologies and innovative topologies such as Doherty PA. However, those alone were not sufficient to meet the strict requirements of new wideband data extensive communication schemes, urging designers to recur to combine physical advancements with adequate PA linearization techniques.

For that end, many linearization techniques have been devised, each presenting its own advantages and challenges. Current communication systems operate high frequency, amplitude varying wideband signals which, along with advanced digital signal processing (DSP) techniques, sets the Digital Predistortion (DPD) linearization method as a viable and applicable linearization solution. Ideally, DPD would distort the input signal of the PA in a way to compensate for the system's nonlinear behavior.

One tends to assume that synthesizing a digital predistortion scheme is merely a function fitting problem, which should be fairly straight forward with the wide availability of fitting and optimization algorithms. However, in practice, because of the complexity of the PA behavior, crunching high order numbers is extremely computationally expensive and can be unstable. Moreover a less than accurate inverse PA model would not only fail to compensate for the spurious PA signal components, but also introduce new unnecessary distortions to the system being linearized.

Most of the models for transistors/amplifiers can be divided into two classes: empirical, black-box models and physical, circuit level models. Empirical, black-box models are extracted from measurement data, without any information on internal device operation, whereas physical, circuit-level models provide information on device operation [1].

The Volterra model is recognized as one of the most comprehensive and accurate empirical PA modeling schemes. However, that scheme requires a large number of coefficients when modeling high nonlinearities making it a non viable solution because of its computational and implementation costs. Many attempts have been made to reduce redundant or unnecessary terms from the Volterra scheme, leading to the most popular DPD scheme, the Memory Polynomial (M-Polynomial) as introduced in [2]. Although the M-Polynomial is a less complex, proven modeling and linearization scheme, authors in [3] found redundant parameters in its structure, and were able to reduce the number of parameters compared to the full M-Polynomial scheme, as well as improved the scheme's stability.

Up to this point, all DPD linearization schemes were based on behavioral modeling of the Device Under Test (DUT) which is considered a black box, finding a mathematical model fit that is as accurate and efficient as possible to the function presented by the DUT, disregarding any real physical operational information behind the behavior observed. However, there's a physical correlation between the coefficients of the behavioral functions or schemes used until now, even the reduced ones where the elimination of coefficients has been experimental. Also, an ME formulation can be found based entirely on the physical circuit properties of the DUT, setting the link between high level behavioral modeling and physical models of the DUT.

In fact, as presented in this thesis a new close-form ME formulation was found based on circuit and PA topology properties. To test the validity of this new formulation, it is first related back to the well established Volterra behavioral model; find the similarities between their respective terms. And second, the proven reduction from Volterra to M-Polynomial is explained through the terms of the expression. The third and last step was to implement a model based on this formulation and test the accuracy of simulation of the PA behavior.

The thesis is divided into five chapters. The first chapter is an introduction laying the background information necessary to understand the purpose, the novelty and the contributions of the new work presented in the following 4 chapters. A brief summary of the contents of each chapter is presented in the following list.

Chapter 1 – this chapter serves as an introduction to first situate the scope of the thesis in the wide technological advancement, and where the work presented in the following chapters can be considered useful. An overview of wireless communications and the operation of Power amplifiers are presented, explaining the need for linearization techniques.

Chapter 2 – the second chapter delves into the various linearization techniques used for RF PAs. The techniques are compared against various criteria to demonstrate why digital predistortion was applied for the rest of the thesis. Digital predistortion calls for accurate modeling, hence the foundation for behavioral modeling is presented with proven models, explaining the reasoning behind the polynomial base that will be used in later developments.

Chapter 3 – as a first step in finding an efficient linearization scheme, this chapter presents deterministic approach to define the optimal number of coefficients to be used based on the parallel Hammerstein structure. The proposed approach is based on observing the filter impulse responses. The resulting model is tested experimentally for both modeling and linearization.

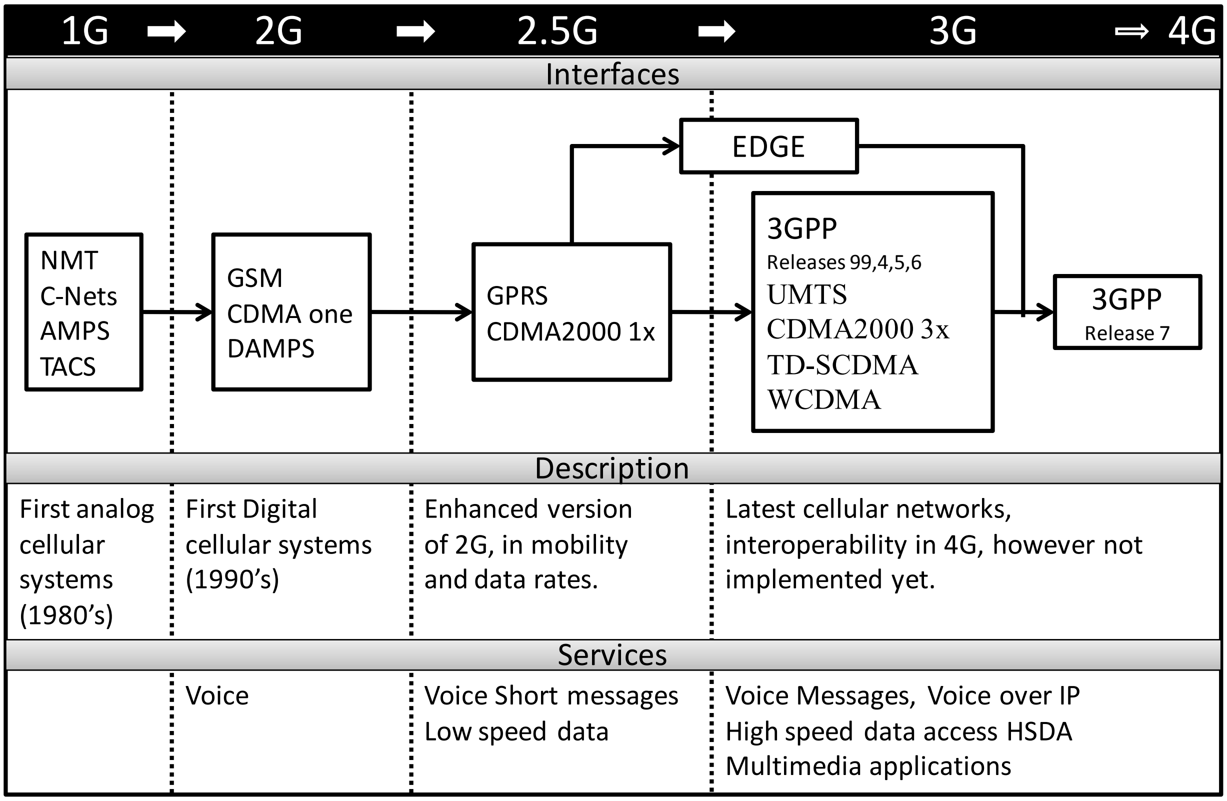
Chapter 4 – in this chapter, a detailed explanation of memory effects in RF PAs and their physical sources leads to a close form formulation of memory effects. The formulation is the basis of the novelty presented in this thesis, as a link is established between the behavioral modeling discussed and physical circuit properties. A model is proposed based on that formulation, and is compared against the Volterra series formulation.

Chapter 5 – the formulation presented in the previous chapter is explored further to understand the applicability of the Memory Polynomial to RF PA modeling. Some underlying assumptions for the memory polynomial are detailed, as well as justifying the necessity, or redundancy, of each coefficient in that scheme in terms of physical circuit properties.

1.2 Wireless Communication Evolution

1.2.1 Evolution of Standards

In an attempt to situate the scope of the work presented in this thesis, and where it fits in the wide technological developments, let's start with an overview definition of the different generations of wireless networks since their early deployment, summarized in Figure 1.1.



1G	2G	2.5G	3G
First analog cellular systems (1980's)	First Digital cellular systems (1990's)	Enhanced version of 2G, in mobility and data rates.	Latest cellular networks, interoperability in 4G, however not implemented yet.
	Voice	Voice Short messages Low speed data	Voice Messages, Voice over IP High speed data access HSDA Multimedia applications

Figure 1.1 - Wireless network evolution and definition

First generation (1G) wireless networks were targeted primarily at voice and data communications, requiring low data rates and hence narrow frequency bandwidth operation, Air interfaces operated in narrowband including the Advanced Mobile Telephone System (AMPS) for which the Federal Communications Commission (FCC) allocated a total of 40MHz of spectrum from the 800MHz band. Looking for mobility and speed, second generation (2G) and third generation (3G) wireless systems were developed, operating broadband signals with many added application features. These include mobility to networks, multimedia applications, quality of service (QoS) assurance, wideband wireless usage and high data transfer rates. New spectrally efficient transmission protocols were developed to achieve these parameters as shown in Figure 1.2, raising the complexity level of the schemes and their information density.

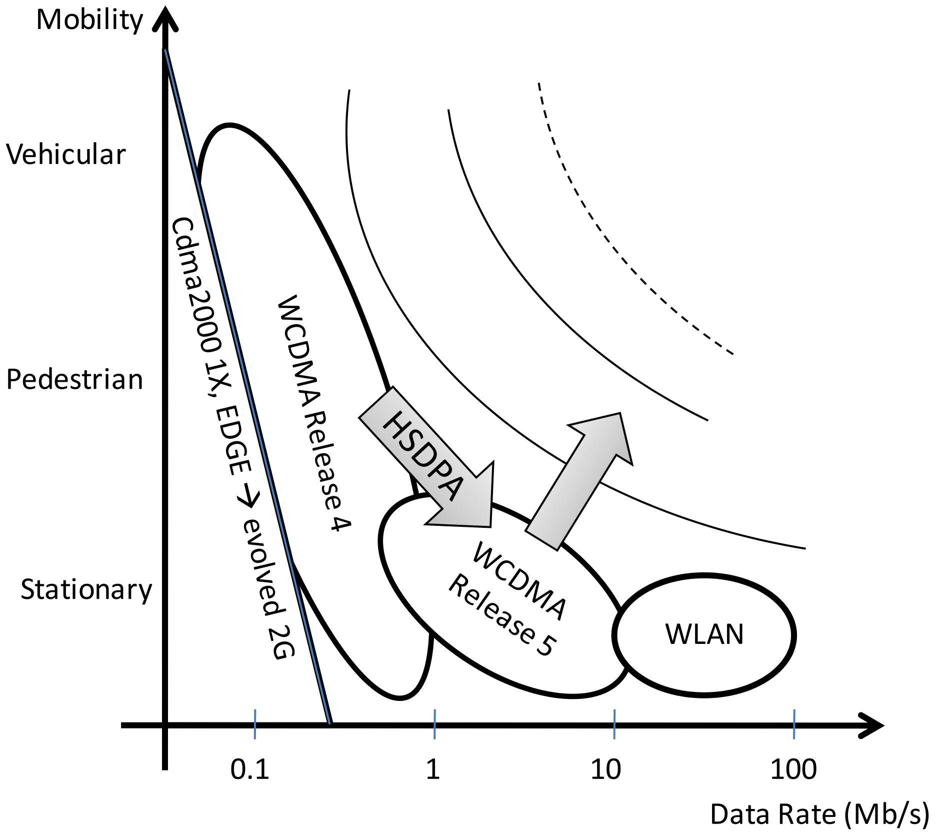


Figure 1.2 - Wireless interfaces evolution for mobility and speed. [4]

Third-generation (3G) and beyond (3.5 and 4G) mobile radio networks strive to provide multiple multimedia services that require high data transfer rates. User applications require that the transmission schemes support services with different data rates and bandwidths. Many wireless transmission schemes are evaluated for implementation, namely Wideband Code Division Multiple Access (WCDMA) and Worldwide Inter-operability for Microwave Access (WiMAX). Wide band code division multiple access (WCDMA) was first developed in Japan and Europe, becoming the leading 3G wireless standard in the world today. It is the technology used in the Universal Mobile Telecommunications System (UMTS). It can handle high bandwidth applications such as video, data, and image transmission necessary for mobile internet. WCDMA can also handle multiple users with variable application bandwidths and data rate requirements. It was introduced as framework of standards by the International Telecommunication Union (ITU) under the name IMT-2000 which brought high-speed access, support for broadband multimedia services, and universal mobility.

Different air interface are necessary for the emerging (4G) networks, to meet the need for interoperability. This has been increasingly recognized by the research community, leading to new developments such as recently multicarrier WCDMA (MC-WCDMA), combining principles of Orthogonal Frequency-Division Multiplexing (OFDM) and CDMA transmission.

In summary, limited resources of frequency spectrum have to accommodate an increasing demand for wider bandwidth, high speed data transfer wireless applications. This leads to the development of highly complex, wideband schemes, with varying signal envelope levels. Therefore, every new generation transmitter has to have the capability of properly and accurately operating instantaneous power variations of signals with wide bandwidths, while simultaneously not affecting adjacent channel communications. Those are the main requirements behind signal figures of merit, and measurement criteria set forth to characterize either the transmitter operating itself or the regulatory requirements of a wireless interface.

1.2.2 Peak-to-Average Power Ratio

With high varying signal envelopes of new modulation schemes such as WCDMA and OFDM, Peak to Average Power Ratio (PAPR) is a signal parameter that has become increasingly crucial. Schemes operating wide modulation bandwidths have fast varying signal envelope amplitude, therefore considerably changing the instantaneous power fed to the PA from an average level to high peaks. Even though the occurrence of such peaks is low in WCDMA signals', clipping them is not a solution as it reduces information transmitted contained in those signal peaks. The PAPR is a very important issue for any OFDM system as well. Since an OFDM transmitter makes use of multiple subcarriers to transmit data. The combined signal has a highly varying envelope. An RF power amplifier usually works in a saturated status to achieve relatively high power efficiency, and thus it will behave like a hard-limiter, which will cut off all useful data information if the dynamic range of the input signal exceeds a certain level. As a simple case scenario, if a 10W amplifier is passing a signal with a 10dB PAPR, that PA should be able to handle an input power up to a 100W.

The main concern is that, consequently, the back-off level at which the PA has to be operated at is a function of the input signal's PAPR. The higher the PAPR the higher back off needed for safe PA operation, leading to low power-efficiency designs.

Moreover, distortions from the PA, observed at these high power levels (at and around peaks) should be minimal as the effect of these distortions is amplified on performance measures of the transceiver, as detailed in the following sections. Therefore higher PAPR values impose higher linearity requirements on the PA. As an example, in the case of WCDMA, guidelines determine a maximum of 4.5 dB PAPR for handsets and 12 dB for basestation transmission.

1.2.3 Transmission Quality Measures

Every communication front end design is evaluated based on a few key measures set for signal quality. Depending on the wireless transmission scheme used, various sets of criteria are set by regulatory bodies to evaluate signal linearity and cleanliness to minimize cross talk. The various measures and criteria relevant to WCDMA are summarized in the following sections as an example of illustration.

1.2.3.1 Adjacent Channel Power Ratio

The adjacent channel power ratio (ACPR) can also be called adjacent channel leakage ratio (ACLR). This measure is used by many transmission standards as an important test parameter for the distortion of a transmission system, measuring the interference with a neighboring frequency band. Hence, ACPR is one of the main criteria set for standards by regulatory bodies.

The ACPR measures the ratio of power in an adjacent frequency channel (offset) away from the main signal, to the inband signal power. Therefore, ACPR is an indication of nonlinear distortions in the transmitter hardware especially when modulated signals such as WCDMA are to be transmitted. The wider bandwidth, the more spurious out-of-band frequency components are observed otherwise known as spectral regrowth. As an example, Table 1.1 details the ACPR requirements for the transmission of WCDMA signals.

Table 1.1 – Minimum ACPR levels required for transmitters using WCDMA signals

Frequency Offset from Carrier Δ*f*	**Minimum ACPR**	**Measurement Bandwidth**
2.5 - 3.5 MHz	-35-15*(Δ*f*-2.5) dBc	30 kHz (Note 1)
3.5 – 7.5 MHz	-35 - 1*(Δ*f*-3.5) dBc	1 MHz (Notes 2 and 3)
7.5 – 8.5 MHz	-39 - 10*(Δ*f*-7.5) dBc	1 MHz (Notes 2 and 3)
8.5 – 12.5 MHz	-49 dBc	1 MHz (Notes 2 and 3)

Notes: [5]

1. The first and last measurement position with a 30 KHz filter is 2.515 MHz and 3.485 MHz.
2. The first and last measurement position with a 1 MHZ filter is 2.515 MHz and 3.485 MHz.
3. The lower limit shall be -50 dBm /3.84 MHz or whichever is higher.

1.2.3.2 Error Vector Magnitude (EVM)

Error Vector Magnitude (EVM) or alternatively known as Signal Vector Error (SVE) is a measure for the performance of both transmitter and receiver. EVM requirements can place a more stringent requirement on amplifier linearity than does adjacent channel performance.[5] Vector error can be graphically interpreted as shown in the following figure:

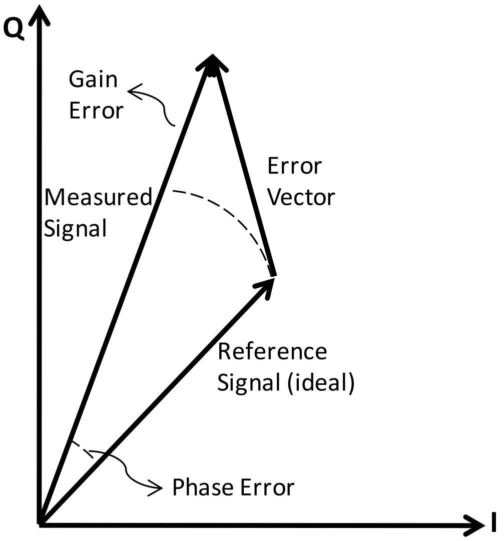


Figure 1.3 - Graphical representation of the Error Vector Magnitude (EVM)

In addition to the power amplifier in transmission, SVE is affected by many factors depending on the transmitter architecture. In the case of direct conversion transmitter, two dominant factors have non-negligible contributions. The first factor is the gain and phase imbalance which results in a distortion in the I and Q plane as shown in Figure 1.3. The second factor causing greater EVM values is phase noise of a local oscillator which might introduce random distortions on the I-Q plane of the original signal itself.

Calculating the EVM is given by:

$$\mathbf{EVM = 100\sqrt{[1+M^2]-\ 2Mcos(\varphi)}} \qquad \mathbf{(1.1)}$$

where M is the magnitude of the actual or measured vector, and φ is the phase error between measured and ideal vectors. [4][6] Similarly, percentage EVM can be defined as:

$$\mathbf{EVM(\%) = \sqrt{\frac{P_{error}}{P_{reference}}}*100\%} \qquad (1.2)$$

where P_{error} is the RMS power of the error vector, and $P_{reference}$ is the RMS power of ideal transmitted signal. As an example, for WCDMA transmission, the maximum allowable EVM in its specification is 7%.Specifying EVM is decibels is also common and is defined as follows:

$$\mathbf{EVM(dB) = 10log_{10}\left(\frac{P_{error}}{P_{reference}}\right)} \qquad \mathbf{(1.3)}$$

1.3 Wideband Power Amplifiers in RF Communications

A radio frequency and microwave frequency (RF/MW) transmitter is designed to operate at one or multiple carrier frequencies within those bands, with signals of up to a few megahertz of bandwidth. As signals evolved from relatively narrowband to wider modulation bandwidth, and highly varying signal envelopes especially in 3G, a wideband transmitter and thus power amplifier design is required for the transmission of such signals, but also linear to meet the previously mentioned signal quality criteria. A typical direct conversion transmitter topology is shown in Figure 1.4. The PA is the most power consuming, nonlinear device in a wideband transmitter, thus it is mainly the PA in a transmitter that should be linear, over the range of frequencies and amplitudes of signals to be transmitted.

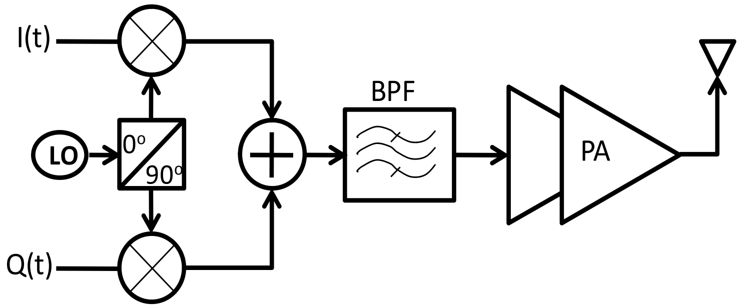


Figure 1.4 – Typical direct conversion RF Transmitter Topology

RF wireless communications are allocated operation bandwidths in the bands from 800 MHz to 3GHz frequency range, and the 10 to 18 GHz for microwave applications. Power amplifiers used for RF communications applications are typically differentiated into *Classes* of operation. Each class has a determined mode of operation, efficiency and linearity characteristics. Different approaches have been developed to characterize the in terms of linearity and efficiency, the two major design criteria. Both criteria can be determined by output power levels achieved, whether it is in terms of device input power, or frequency of operation.

1.3.1 Linearity of RF PAs

Linearity in power amplifiers across classes can be specified in several ways. In some cases, the classical intercept points are the most meaningful characterization. However, when wider bandwidths and higher PAPR signal operation is required, adjacent channel power ratio as defined previously, or equivalently, two-tone intermodulation levels at full output power are more meaningful.[7]

1.3.1.1 Intercept Points

Intercept points can be determined on a simple input output characteristic power curve of a PA, as shown in Figure 1.5.

1. *1dBcompression Point - 1dB* – compression is measure of the level of linearity of a device. As the amplitude of the input RF signal increases, the output level should follow linearly. However, for a nonlinear device, the output starts to deviate at one point from the linear case, at which the device starts to exhibit nonlinear behavior. The 1dB compression point is defined as the point where that deviation from the ideal linear output reaches the 1dB level.
2. *Second Order Intercept Point - IP2* – it is the theoretical point on the curve as shown in Figure 1.5 where the desired output signal and second order products become equal in amplitude.
3. *Third Order Intercept Point - IP3* – the theoretical point on the curve where the desired input signal and third-order products become equal in amplitude s the RF input is raised. IIP3 is the input referred IP3 which, multiplied by the small signal gain, yields the OIP3, or output referred IP3.

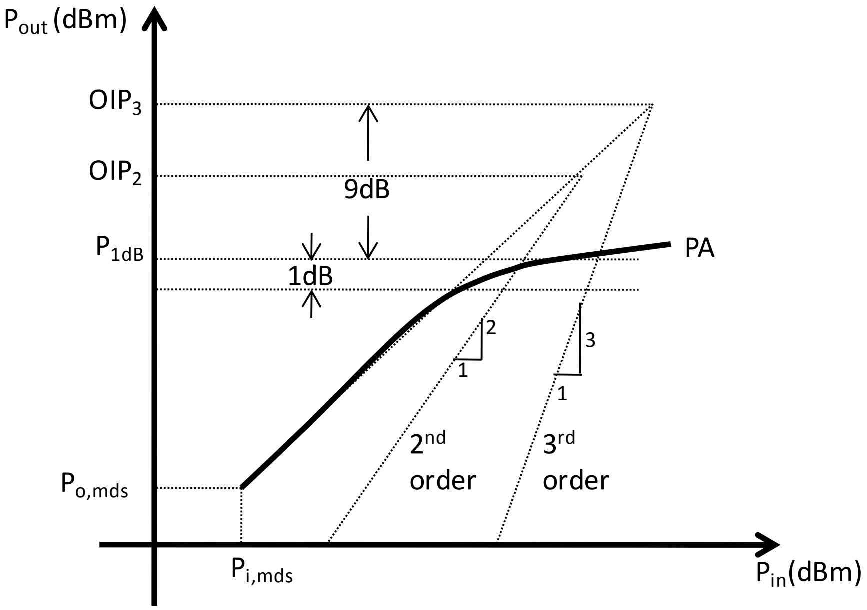


Figure 1.5 – The 1 dB Compression and Intercept Points defining the dynamic range of an RF PA

1.3.1.2 Intermodulation Distortion

When two sinusoidal frequencies are applied to a nonlinear amplifier, the amplifier nonlinearity generates new frequency components called intermodulation products (IMD), located at specific frequencies as shown in Figure 1.6. Those spread from baseband to higher harmonics of the input signal.

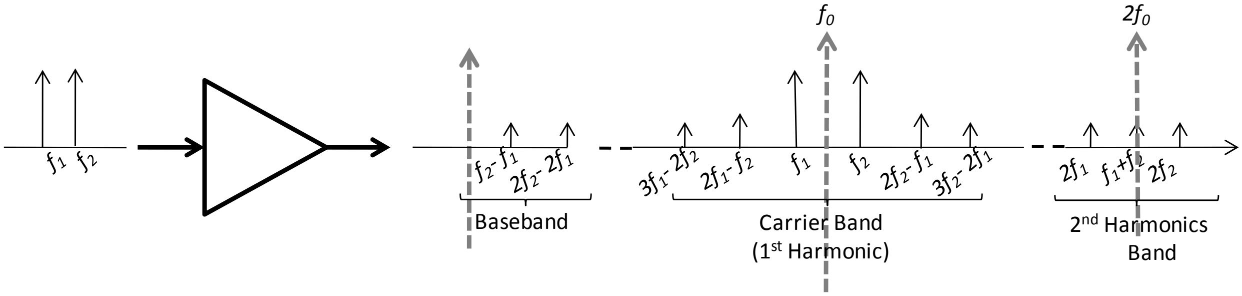


Figure 1.6 - Typical RF PA spectral output under two-tone input signal excitation

The spurious frequency components that are out of band do not necessarily affect transmission and reception of the signal itself but rather the users of the corresponding bands. However, when evaluating the linearity of the PA, it is necessary to define the levels of the spurious components that are inband, which potentially interfere with the signal itself. These spurious frequencies are the mainly components that fall at the 3rd and 5th order intermodulation frequencies (IMD3 and IMD5 respectively). As an illustration, Table 1.2 summarizes of the location of these IMD components under a two-tone test condition.

Table 1.2 - Intermodulation product frequency elements from a two-tone test that fall within, or in the vicinity of the signal bandwidth.

Order	Intermodulation Products	
1st Order	f_1	f_2
3rd Order – IMD3	$2f_1$ - f_2	$2f_2$ -f_1
5th Order – IMD5	$3f_1$ - $2f_2$	$3f_2$-$2f_1$

The IMD is thus defined as the difference in power levels between the resulting fundamental frequencies and the 3rd and 5th order frequencies respectively for IMD3 and IMD5, as shown in Figure 1.7.

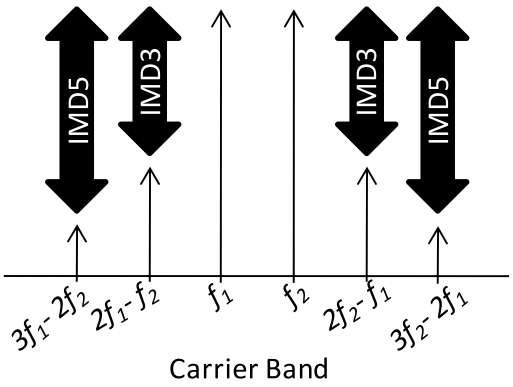


Figure 1.7 – IMD3 and IMD5 definition

Distortion in Power amplifiers can arise from two different phenomena. At low levels, distortion is caused by the same nonlinearities that affect small-signal amplifiers. As the amplifier is driven into saturation, however, distortion caused by clipping the amplitude peaks of the modulated carrier waveform becomes the dominant phenomenon, and the distortion generated in this manner is much greater than the small-signal distortion. As a result distortion increases significantly as the amplifier is driven into saturation. For all classes of operation, minimizing clipping distortion requires optimizing the load impedance. However, in most cases, the optimum output power and efficiency are achieved at significantly different load impedance than the impedance that minimizes distortion. At full output power, strong distortions are observed. One solution is to operate the amplifier at a lower power level or *back-off* input power which is several decibels below sinusoidal output power, decreasing the Efficiency of the PA, as discussed in the following section.

1.3.2 Efficiency of RF PAs

The following paragraph defines the main figure of merit for RF PA efficiency, followed by details on the efficiency of the various classes of operation of PAs, demonstrating how efficiency and linearity are the main tradeoff in the PA's operation.

1.3.2.1 Efficiency Figure of merit: Power Added Efficiency

Power added efficiency (PAE) is a key figure of merit for power amplifiers. It is expressed as the ratio of the additional RF power provided by the amplifier to the DC power

$$PAE = \frac{Pout_{RF} - Pin_{RF}}{P_{DC}} \qquad \textbf{(1.4)}$$

where Pin_{RF} is the power of the RF input signal fed to the PA, $Pout_{RF}$ is the RF power measured at the output of the PA, and P_{DC} is the DC power required for the operation of the PA.

The DC power required for the operation of the PA cannot be overlooked. DC power is usually dissipated in the form of heat. PAs with high DC power consumption require large heat sinks, which are bulky and costly to design and realize. Therefore, a low PAE implies the need for large DC power modules, bulky thermal dissipaters on fixed terminals such as base stations, or consequently short battery life for mobile terminals.

1.3.2.2 The Linear (Continuously Driven) Amplifiers

Classes A, AB and B are the three main classes of continuously driven amplifiers. Those linear amplifiers cannot achieve high efficiencies. In fact, class A is the most linear and least efficient of the three, dissipating a great amount of power even under quiescent conditions (not excited). The maximum efficiency of a class A amplifier is 50%. Figure 1.8 shows the improvement in efficiency in Class B, whose maximum efficiency is 78%, however inherently has lower gain than Class A. Another disadvantage of Class B is that it generates a high level of harmonics in the drain current by switching the FET on and off during each excitation cycle.

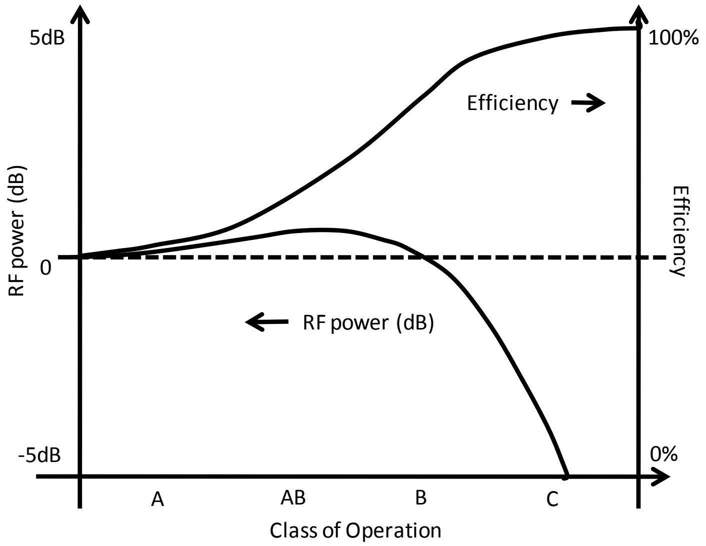


Figure 1.8 - RF output Power as inversely proportional to efficiency of an RF PA. [8]

RF power amplifiers are rarely operated in purely class A or B, but somewhere in between, creating the Class AB which is a compromise between the two classes. This mode of operation is attained either by saturating a class A PA or reducing the bias of a Class B, achieving better efficiency than Class A and better gain than Class B. Moreover, biasing an amplifier in a particular class depends on the performance required by the application. While the efficiency of the PA increases moving form Class A (50% efficiency), towards Class B (87% efficiency) or C (100% efficiency), maximum RF output power decreases as shown in Figure 1.8. Moreover, ideally for Class A and B operation there is a linear relation between Pout and Pin, while for the intermediate Class AB bias, the linear behavior is obtainable in an unsaturated operating condition. In other words, the PA should be operated at a higher output back-off power (OBO) to maintain its linear characteristics, thus reducing efficiency of the PA. Hence why, linearization schemes are sought to extend the linear region of the PA, and therefore alleviating the major efficiency-linearity trade-off, allowing a more efficient and linear operation of the RF PA. Several linearization schemes were developed, detailed in section (2.1), all aiming at distorting the input or output signals in a manner that compensates for the nonlinear behavior of the amplifier.

1.3.2.3 The Nonlinear Amplifiers

These are classes C, D, E, F, G, H, J and S. They are seldom used for RF and microwave applications. The efficiency of such amplifiers is high but the magnitude of the fundamental component of a signal is greatly reduced, therefore for RF applications, these classes are practically operable only at low power where device gain is high. Classes D and E are strongly nonlinear and are operable only where high levels of distortion are acceptable, or when the input signal has constant envelope (i.e. in frequency or phase modulated signals FM/PM). In classes E, D and S, transistors are used as RF switches, while in classes C, G and F, the transistor is still a voltage controlled source.

1.3.2.4 Load Modulated (Doherty) PAs

Another solution to alleviate the linearity-efficiency tradeoff of RF PAs was the Doherty amplifier. The Doherty power amplifier (DPA) technique proved to be a more efficient alternative to the classical PA topology. As shown in Figure 1.9, the DPA consists of two power amplifiers. The “main” amplifier is a Class AB which saturates at high input power levels, while the “auxiliary” or second PA is biased in Class C. In fact, the auxiliary PA turns ON and starts contributing in the output power at a predetermined input power threshold. Thus, the load impedance seen by the transistor varies depending on the input power level, which if designed properly, significantly improves the overall performance. In an ideal impedance matching case, when both amplifiers are delivering maximum power, they contribute equally to the output power.

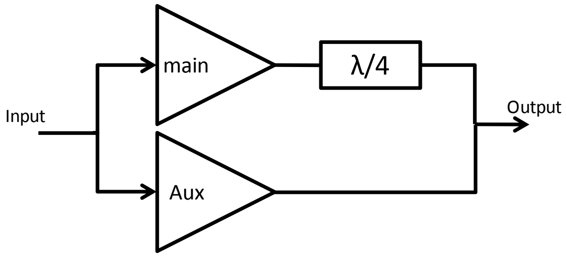


Figure 1.9 – Generic Load Modulated PA – Doherty Amplifier

The following figure shows the typical improvement in efficiency that the Doherty PA brings about compared to a regular class AB PA.

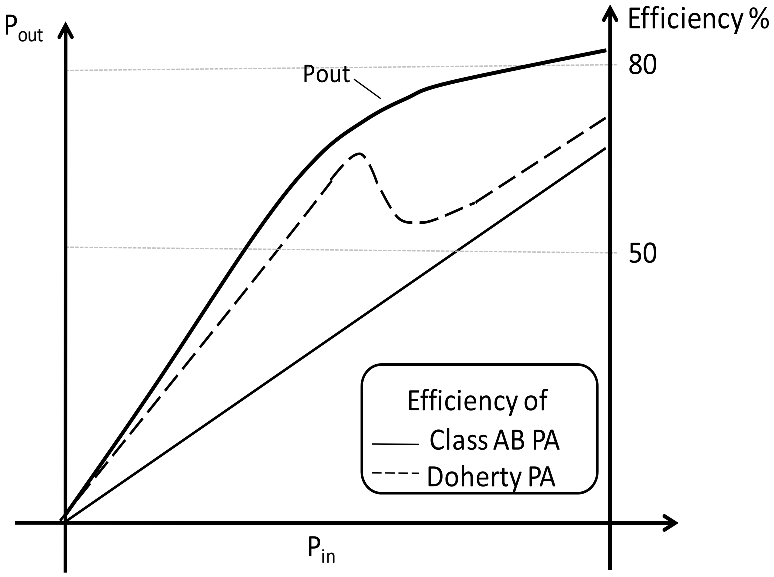


Figure 1.10 - Doherty's Improved efficiency over class AB, while both suffer from low efficiency in the linear region.

With highly modulated signals, the PA has to be operated at a certain back off power to accommodate for the high PAPR levels. As outlined in Figure 1.10, the efficiency of the Doherty PA is maintained high for a certain range of back off from saturation power, making the DPA a highly desirable topology for operating high modulation signals. For this reason, the experimental validation of the work presented in this thesis will be carried out based Doherty PA operation.

In summary, linearity and power efficiency are the two main criteria any PA designer is trying to optimize. Improving the linearity of a PA by only changing its mode of operation comes at the expense of its efficiency. The Doherty power amplifier brought an attractive improvement in terms of efficiency compared to the class AB. However, the linearity of such PAs is not at the level to achieve the transmission quality required by the wideband, high varying envelope signals discussed previously. In order to the minimum requirements of such schemes with acceptable efficiency levels, additional PA linearization methods are sought as described in the next sections.

Chapter 2
Linearization and Modeling of RF PAs

2.1 Linearization Techniques of Power Amplifiers

Linearization techniques for RF PAs are increasingly applied to enhance the main tradeoff of linearity versus efficiency and reduce the required power backoff level. Linearization can be performed using three basic techniques: feedback, feedforward and predistortion.

2.1.1 Feedback Technique

A basic compensation technique for the nonlinear behavior of an RF PA can be achieved through the implementation of a feedback loop. This technique, although fairly straight forward to implement, suffers from many drawbacks when implemented for an RF PA system. The main drawback of applying this technique to high frequency, wideband systems are the inherent gain-bandwidth product limitations and the sacrifice of gain for linearity. The block diagram of a generic feedback linearizing system is shown in Figure 2.1.

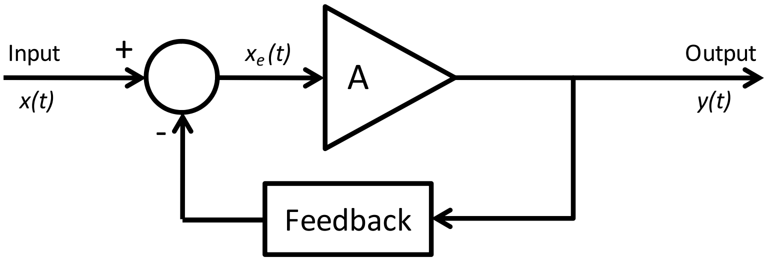


Figure 2.1- Feedback linearization block diagram. [5]

The Feedback Block determines the type of feedback structure and scheme of a linearizing system. These types include Passive RF feedback, Active RF feedback, frequency difference feedback, modulation feedback.

For a Passive RF feedback scheme, the Feedback block of Figure 2.1 reduces the gain of the amplifier at RF frequencies, i.e. its function can be simplified into (1/k) and it can be shown that:

$$\boldsymbol{y(t) = A\,x_e(t) = A\left(x(t) - \frac{1}{k}\,y(t)\right) \rightarrow y(t) = \frac{kA}{k+A}\,x(t)} \qquad \textbf{(2.1)}$$

In the case where *k* << *A*, the output of such system can be simplified as:

$$\boldsymbol{y(t) = k\,x(t)} \qquad \textbf{(2.2)}$$

which improves the stability of the system such as the output is independent of PA operational variations, i.e. temperature. However, realizing a gain (A) that is much higher at RF frequencies than k is expensive both in implementation and operation. Implementing an RF feedback using passive circuitry can be done through either shunt or series circuit configurations.

Alternatively, instead of using a passive circuit, the voltage divider or feedback block can be implemented with an amplifier stage. The resulting scheme is called Active RF feedback. The feedback amplifier can be adjusted to introduce distortions into the system that would cancel out the main amplifier that is being linearized. Another advantage of active RF feedback is that, since the PA introduces power to the stage, the power for the feedback path components is not all from the main power amplifier.

2.1.2 Feedforward Technique

Feedforward (FFW) linearization method aims to dynamically compensate for the gradual compression of the PA characteristic behavior as shown in Figure 2.2. The correction for distortions takes place at the output of the main PA, where an error PA (EPA) effectively adds power to the amplified signal from the main PA while also compensating for AM-PM distortions present as depicted in Figure 2.2.

Figure 2.2 - Basic Feedforward Linearization

The most basic feedforward amplifier linearization consists of the blocks shown in Figure 2.3. The signal in the top path is amplified by the Main PA, generating all the nonlinearities at its output. In order to singularize these nonlinearities, a delayed PA input is subtracted from a sample of its output. The resulting signal is essentially formed by the distortions only, and does not contain the original signal. That error signal is then linearly amplified by the Error PA (EPA) to a required level, in such a way that, coupled in antiphase with the delayed output of the main PA, it will cancel out the distortions produced by the main path. The delays in the loops are to account for the amplification time of the respective PAs. The result is an amplified version of the input signal, clean of distortions.

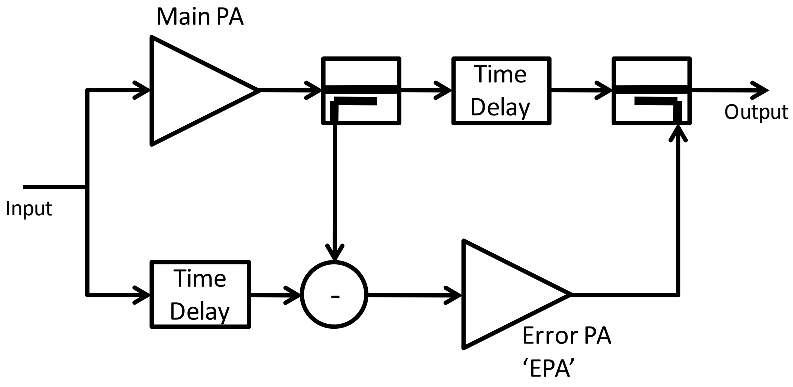


Figure 2.3 - Basic Feedforward Linearizer Block Diagram [5]

2.1.2.1 Implementation and Operation

Feedforward correction is a technology for cellular and base station applications to achieve the high linearity levels of -75 dBc and better. It can be employed for wideband applications; however, the practical realization of such a technique is quite complicated. Due to the many components required by such schemes, they are expensive to implement and operationally power consuming. One major concern in FFW loop design is the power addition which invariably involves power losses in power-combining devices shown in Figure 2.3.

Multiple feedforward loops can be implemented as a single operational block. That is, instead of having one feedforward loop, distortions are compensated for multiple times in sequence. Having multiple feedforward loops will improve the performance of the scheme on many levels; the ideal case assumes that the error amplifier is linear. However, in reality it will introduce some distortions of its own, whose effect will be reduced through additional loops. Also, if one loop fails to function, the remaining ones will still compensate for the errors perceived, therefore creating a fail-proof system. Moreover, in the amplifiers of the feedforward loops do not need to be high power amplifiers like the one amplifying the signal itself, and therefore can be designed to handle low power at a higher linearity level.

Advantages of the Feedforward Linearization:

1. Does not reduce amplifier's gain contrary to feedback systems.
2. Gain-Bandwidth is conserved within the bandwidth of interest, contrary to feedback systems which require high feedback bandwidths to provide the required levels of correction.

3. Correction is independent of Amplifier delays, unlike feedback where the system can potentially become unstable for high amplifier delays.
4. The correction is based on only current events, disregarding past ones which feedback is based upon.
5. The basic feedforward configuration is an unconditionally stable system.
6. Adding loops to the basic configuration incurs more costs but provides more reliability as in fault tolerance and accuracy, and a lower noise figure if needed for the application.

Disadvantages of the Feedforward Linearization:

1. Device characteristics are assumed to be constant over time, i.e. temperature effect is neglected, as is variation of device properties with time as it is an open-loop implementation.
2. Matching between the circuit elements must be maintained to high level of correction, over a wide bandwidth of operation.
3. Adding loops adds accuracy but also implementation size and cost.
4. A DSP control scheme can be added to monitor all loops of a feedforward system, keeping the reference and operational levels of the linearizer constant. However these control loops are fed with DC voltages that will fluctuate with time, changing yet again the reference voltages of a feedforward linearizer (long-term effect). This has been cited as objection to the practical use of feedforward by some system operators.[5]

2.1.3 Digital Predistortion (DPD)

Predistortion technique is the simplest form of power amplifier linearization. The concept of predistortion is similar to that of the feedforward linearization in that linearization is done by cancelling intermodulation products at the output of the amplifier. However, unlike the feedforward method, signals are predistorted before being amplified. There are three main types of predistortion, namely RF, intermediate frequency IF, and Baseband predistortion. The three techniques are discussed hereafter, RF and IF being in one section for they are highly similar.

The basic function of a predistorter is depicted in Figure 2.4. To achieve a power level A at the output of the amplifier, the input power level should be V_{in} of point B. Therefore V_p should be equal to V_{in} (B). Hence the concept of DPD linearization is to reverse the behavior of the amplifier by feeding it with the signal that would yield the ideal desired output given certain input conditions.

To synthesize the inverse function of the PA designated as *F(x)* in Figure 2.4, behavioral modeling schemes can be used as introduced in literature. With the growing demand in signal bandwidth, these modeling schemes evolved from simple memoryless [9][5][7], to more advanced schemes such as the Volterra series, which are more commonly used in the form of their simpler derivations: memory polynomials (M-Polynomial), Hammerstein [10] and Wiener models [11] as well as other non-polynomial based schemes such as Neural Networks [12]. These schemes used to model the PA, are now compensating for its non ideal behavior by simulating its inverse function. The performance evaluation of such models is not limited to their linearization capability, which is related to their modeling accuracy. The implementation complexity of the DPD is a major consideration for the practical implementation of the scheme, and is directly related to its number of coefficients and the complexity of their identification.

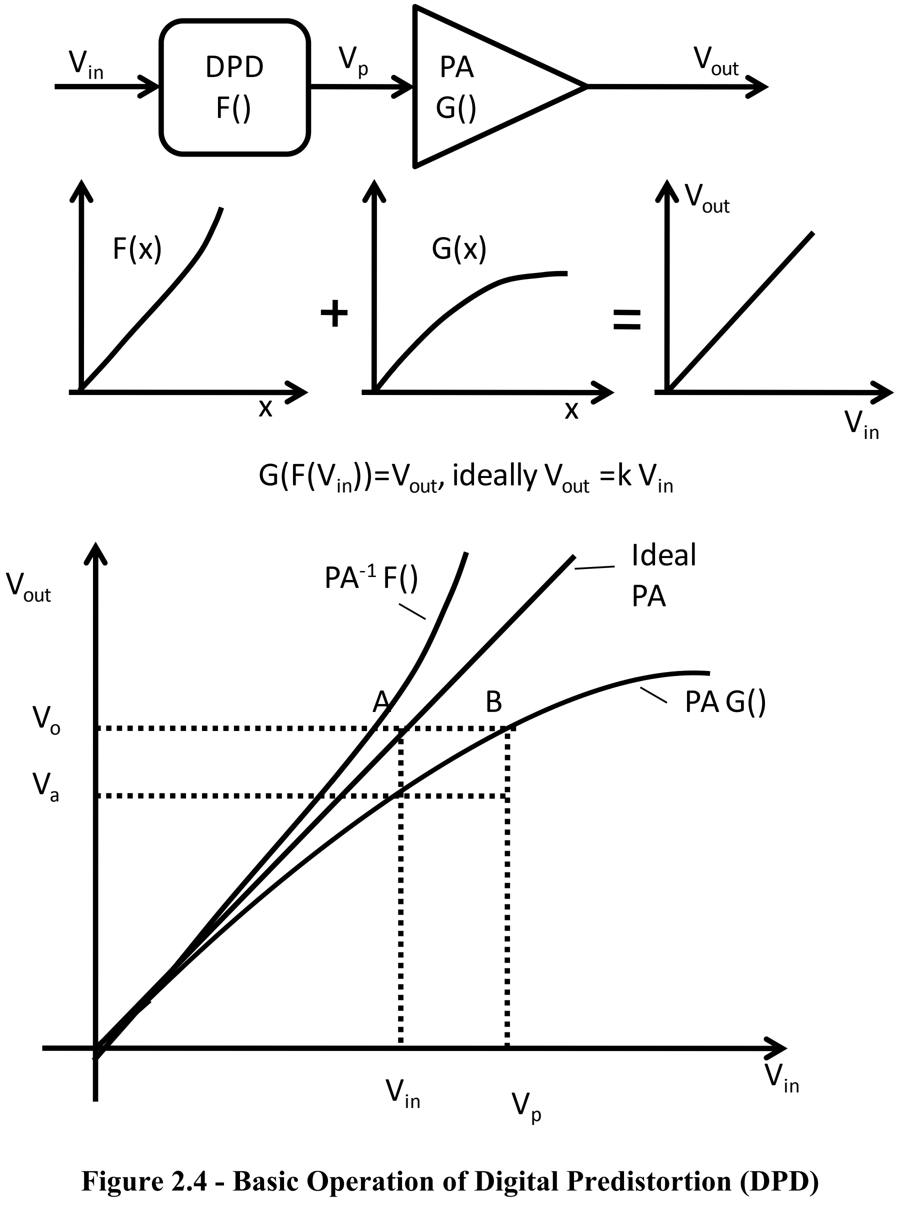


Figure 2.4 - Basic Operation of Digital Predistortion (DPD)

2.1.3.1 RF / IF Predistortion

The fundamental advantage of RF and IF predistortion is its ability to linearize the entire bandwidth of an amplifier or system simultaneously, making it ideal for wideband amplifier linearization such as base station or satellite amplifiers.

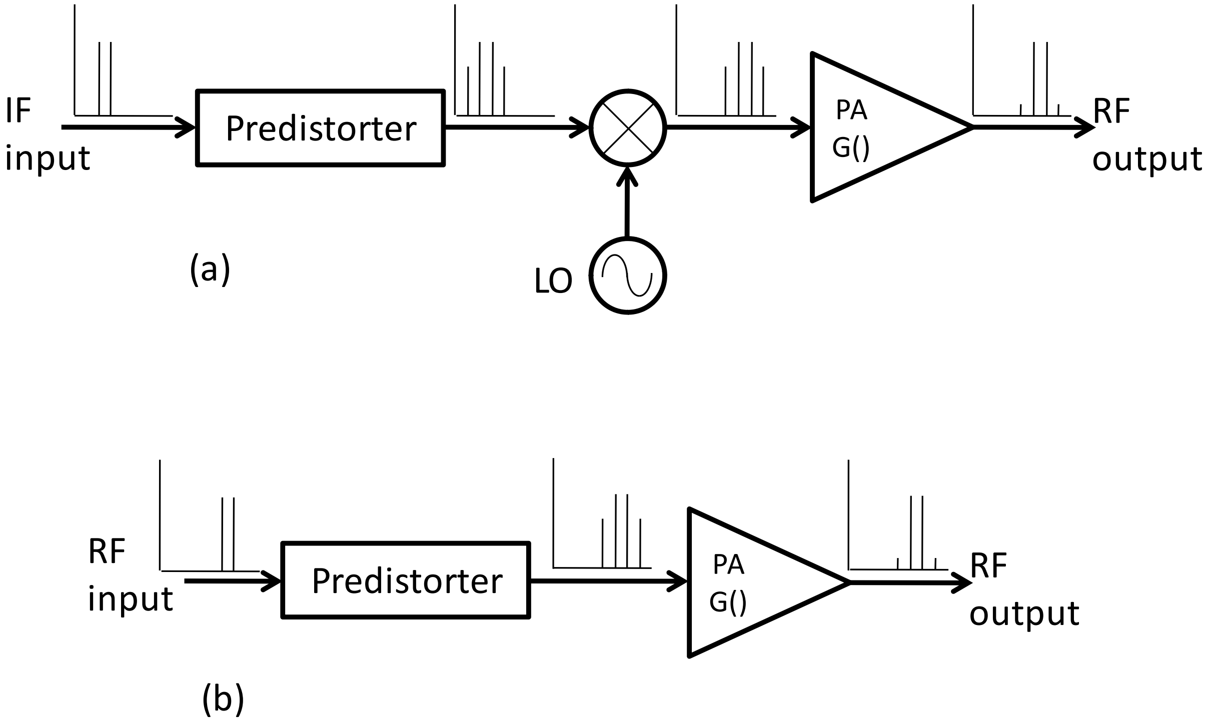


Figure 2.5 - Operation of: (a) IF predistorter requiring an LO, and (b) RF predistorter.

For these two types of predistortion, the predistorter operates at high frequencies (either IF or RF), the operation of such predistorters is summarized in the Table 2.1.

Table 2.1 - Operational Characteristics of IF/RF Predistortion

Implementation Complexity	Fairly simple implementation with a few number of components required.
Stability	Open loop scheme which makes it Unconditionally stable
Linearization Bandwidth	Very wide bandwidth, comparable to a feedforward system.
Linearization Capability	Modest, requires a high order of modeling which encumbers the ease of implementation and scheme viability.

2.1.3.2 Baseband Predistortion

Recent advances in DSP technologies have made it increasingly available, cheaper and more versatile. Operational power consumption was significantly reduced, and more importantly, higher computation rates allowed for wider bandwidth operation, making digital baseband predistortion a viable and competitive solution for PA compensation. In digital predistortion, the baseband signal is predistorted before it is converted to the analog domain, frequency translated to RF and amplified. The block diagram in Figure 2.6 shows a basic adaptive baseband predistortion implementation.

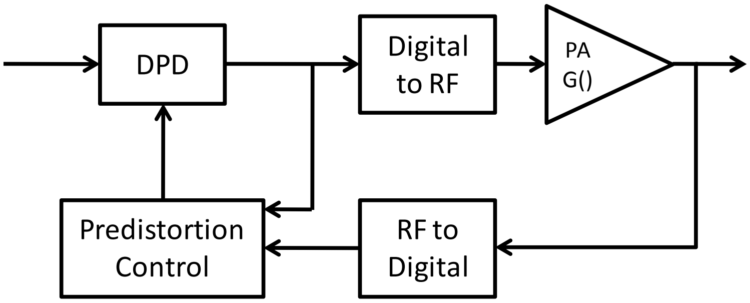


Figure 2.6 – Adaptive Baseband Digital Predistortion Block Diagram

The DPD-Digital to RF-PA path can be implemented on its own. However, an advantage of baseband predistortion is that it can be adaptive, that is to adapt the synthesizing function of the predistorter to the current amplifier behavior. The PA characteristic behavior can vary significantly with time where dc bias levels fluctuate, or with temperature changes, transistor degradation. Predistortion in such schemes takes place in the DSP part of the predistorter. A small fraction of the PA output is fed back and converted from RF to baseband. An adaptation algorithm compares this signal with the output of the predistorter. Different approaches exist: namely Direct- and Indirect Learning. However, in the scope of this thesis, adaptation has not been applied in the DPD process.

2.1.3.3 RF / IF versus Baseband

The choice of a proper high efficiency approach or linearity correction scheme depends on performance tradeoff as well as manufacturing capabilities [13]. Many factors are to be considered as shown in the comparison of Table 2.2, summarizing and comparing all main three techniques: feedback, feedforward and predistortion.

The most traditional linearization technique is the feedback one; however it is hard to implement it for RF frequencies. Conversion of the output RF spectrum down to the baseband frequency, whereupon the I and Q signals are picked out and fed back to the input of the amplifier as correction signals (Johansson et al. 1993). The method's cancellation performance is good, but the bandwidth is narrow, making the technique unsuitable for very wideband systems. Feedforward, on the other hand, can be employed for wideband linearization, but unfortunately the system is extremely complicated, resulting in great power waste and a large physical size. This technique entails comparison of the input and output spectra, and errors are corrected after amplification [5].

Table 2.2 - Performance comparison of the 3 main Linearization techniques.

Technique	Cancellation Performance	Bandwidth	PAE	Size
Feedback	Good	Narrow	Medium	Medium
Feedforward	Good	Wide	Low	Large
Predistortion	Medium	Medium	High	Small

When comparing a non-linearized to linearized system, the efficiency of the latter should be significantly higher than that of non-linearized system operated at a backoff level, and demonstrating the same linearity levels in terms of IMD or SVE [5]. Predistortion is therefore a useful efficiency enhancement technique where linearity is an issue in a system specification.

For the rest of the thesis, and based on the comparisons, digital baseband predistortion will be considered as the predistortion scheme used for Linearizing RF PAs. The thesis focuses on synthesizing the most accurate predistortion function, i.e. one that would best complement the PA behavior, with the least amount of complexity whether in synthesis or implementation and operation.

2.2 Behavioral Modeling of Nonlinear RF Systems

Linearization is necessary and crucial for adequate operation of a wideband RF PA. The efficiency of a DPD linearization depends largely on the accuracy of the modeling scheme used to compensate for the PA's nonideal behavior. The slightest inaccuracies in a PA model could fail to compensate for the spurious PA signal components, but also introduce new unnecessary distortions to the system being linearized.

2.2.1 Behavioral modeling

There are various levels of abstraction when modeling an RF PA. Great amount of research has been geared towards device modeling of PAs in order to understand the physical mechanisms behind the resultant complex behavior observed. On the other hand, behavioral modeling is at a much higher level of modeling, where the PA system is considered as a black box, from which the only information used in modeling are the input and output signals.

The most common method to capture that information (input and output signals) is through the envelope simulation approach, graphically summarized in Figure 2.7. It has in fact been the most commonly used technique for the last decade for RF PA characterization [8]. The resulting output signal from this simulation can be processed to obtain behavioral characteristic information such as EVM and ACPR mentioned previously, as well as used to model the RF PA system.

Characterizing an RF PA consists of measuring and evaluating its behavior under certain input conditions and constraints. As shown in Figure 2.7, the characteristics of the RF PA are obtained from the desired continuous wave input signal and its corresponding output [8]. From this information, all signal and PA characteristics can be obtained.

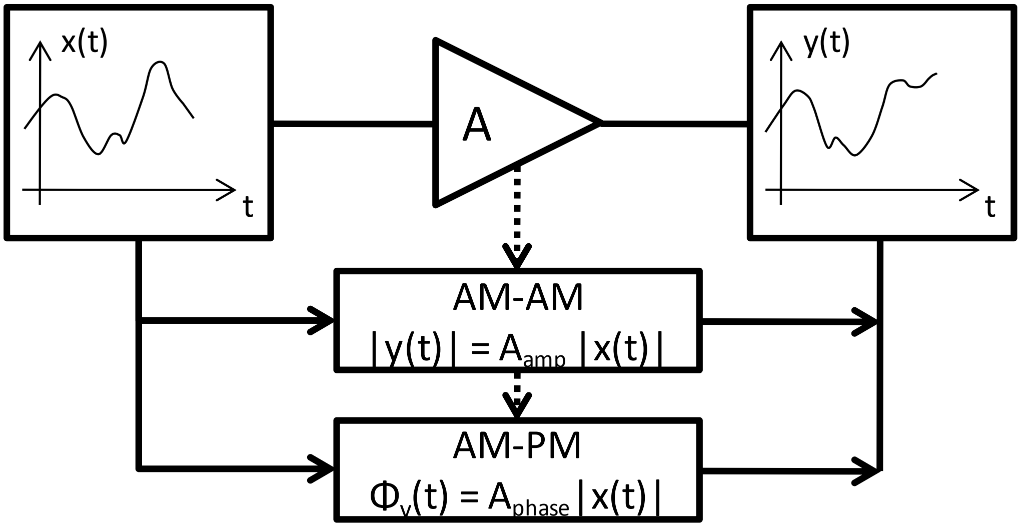


Figure 2.7 - Time domain characterization of a PA[8]

A useful representation of the characteristics of a PA is through AM/AM and AM/PM curves, which, respectively correspond to plotting the amplitude and phase of the output signal with respect to the amplitude of its input signal. As shown in Figure 1.5, linearity of the PA can be determined from the AM-AM plot.

Before discussing the models used for RF PAs, it is worth noting that the accuracy of any behavioral modeling scheme, since it uses only input/output data, is largely dependent on the characterization or measurement technique used. The envelope simulation described earlier is used for all experiments in the later sections of this thesis. This time domain characterization simplifies the task by making three major assumptions [8]:

1. The response of the device under test is assumed to be quasi-static.
2. Signals are assumed to have narrow bandwidths with respect to the RF spectral domain, The time-scale on which perceptible changes in the RF envelope occur is very slow in comparison to the RF time domain;

3. The measurement, hence simulation, bandwidth is restricted to a narrow frequency band by suitably filtering the immediate vicinity of the signal itself.

2.2.2 The Polynomial as a Basis for PA models

A narrowband PA is considered as a static, instantaneous or memoryless nonlinear input-output system, which can be modeled by a polynomial:

$$y(t) = \sum_{i=1}^{M} a_i x^i(t) \qquad \textbf{(2.3)}$$

where *y(t)* is the output of the amplifier in the time domain, *x(t)* its input, a_i is the complex coefficient of the i^{th} power, and *M* is the highest input power required to model the PA.

This simple expression has been used as the basis of many nonlinear models for amplifiers, In his book on “Advanced Techniques of Power Amplifiers”, Steve Cripps states that: “in radio frequency applications, there is fundamental justification for staying with a polynomial approach. The frequency domain, with its sinewave generators, bandpass filters, and spectrum analyzers gives integral polynomial powers and coefficients tangible and measurable reality.”[8]

Moreover, “the fact that some kinds of device may have characteristics that are more readily modeled by some other mathematical function is only of intermediate use if the final characteristic is to be transformed in the frequency domain, the FFT process itself infers a set of polynomial coefficients in the determination of harmonic frequency components.”

To illustrate the behavior described, let us consider a two-tone test applied to a PA modeled with a polynomial. The two-tone test is another characterization technique which uses an input signal consisting of two equal amplitude sinusoids at two different frequencies. The separation between the two tones is equivalent to the signal bandwidth when the DUT is excited by a real world modulated signal. Following the models presented in the above sections, the resulting output of the nonlinear network is expressed as follows:

$$\infty \qquad \textbf{(2.4),}$$

i.e. the output is composed of a very large number of mixing terms involving all possible combinations of . Assuming signal bandwidth is significantly lower than the RF center frequency, harmonic bands can be observed, each centered around multiples of the two tones, equivalently the carrier in real world RF signals as shown in the first row of Figure 2.8. All these newly generated components are called intermodulation distortions (IMD), or equivalently spectral regrowth. The amplitudes of these IMDs and their respective phase angles depend on the characteristic behavior of the amplifier under the specific conditions of operation, including the bandwidth and power levels of the input signal. The second row of Figure 2.8 shows the spurious spectral components that are created by a memoryless nonlinearity, i.e. the response of an RF PA modeled with simple polynomial when excited with a clean two tone signal. The AM/AM characteristic of the DUT shows nonlinearity with no dispersions. The Frequency domain equivalent of the output signal is composed of the fundamentals f_1 and f_2, as well as a set of spurious elements spread around the main signal bandwidth. The ones shown in the figure correspond to a maximum of 5 orders of nonlinearities, which spread over the following frequencies: $f_1,\ f_2,\ 2f_1-f_2,\ 2f_2-f_1,\ 3f_1-2f_2\ \textit{and}\ 3f_2-2f_1$. At this point it is worthwhile to note the following three concepts:

1. The spurious elements follow the principle of superposition. Every arrow at the various frequencies shown in Figure 2.8 is the resultant of spurious elements generated by many orders of nonlinearity.
2. The baseband and higher harmonics are eliminated when measuring the signal in the envelope domain, making it impossible to find the exact amplitude for each order of nonlinearity.
3. Only odd orders lie within the measurement bandwidth. Even orders do affect the amplitude and phase of those inband components, however for modeling purposes and when the characterization of the DUT is based on envelope time domain, it is sufficient to include odd orders in the model used and capture all nonlinearities within the band of interest, however those nonlinearities are formed or the individual contributions from even and odd orders to form the resultants observed.

2.2.3 Wideband PA behavior

In reality, a single complex polynomial can only capture the distortions brought about by the nonlinear component of the behavior of an RF PA. However a PA exhibits a more complex behavior as shown by the AM-AM distortions of the 3rd row of Figure 2.8. It is clear that for one input power level, many values of output power can be expected at the output of the DUT, implying that the instantaneous output signal is not only dependent on the instantaneous input. In fact, for wideband PAs, current output depends on current input but also on past occurrences of that input, defining that physical phenomenon as memory effects. The wider the bandwidth, the more pronounced the dispersions or memory effects are, and therefore should be taken into consideration by the modeling scheme.

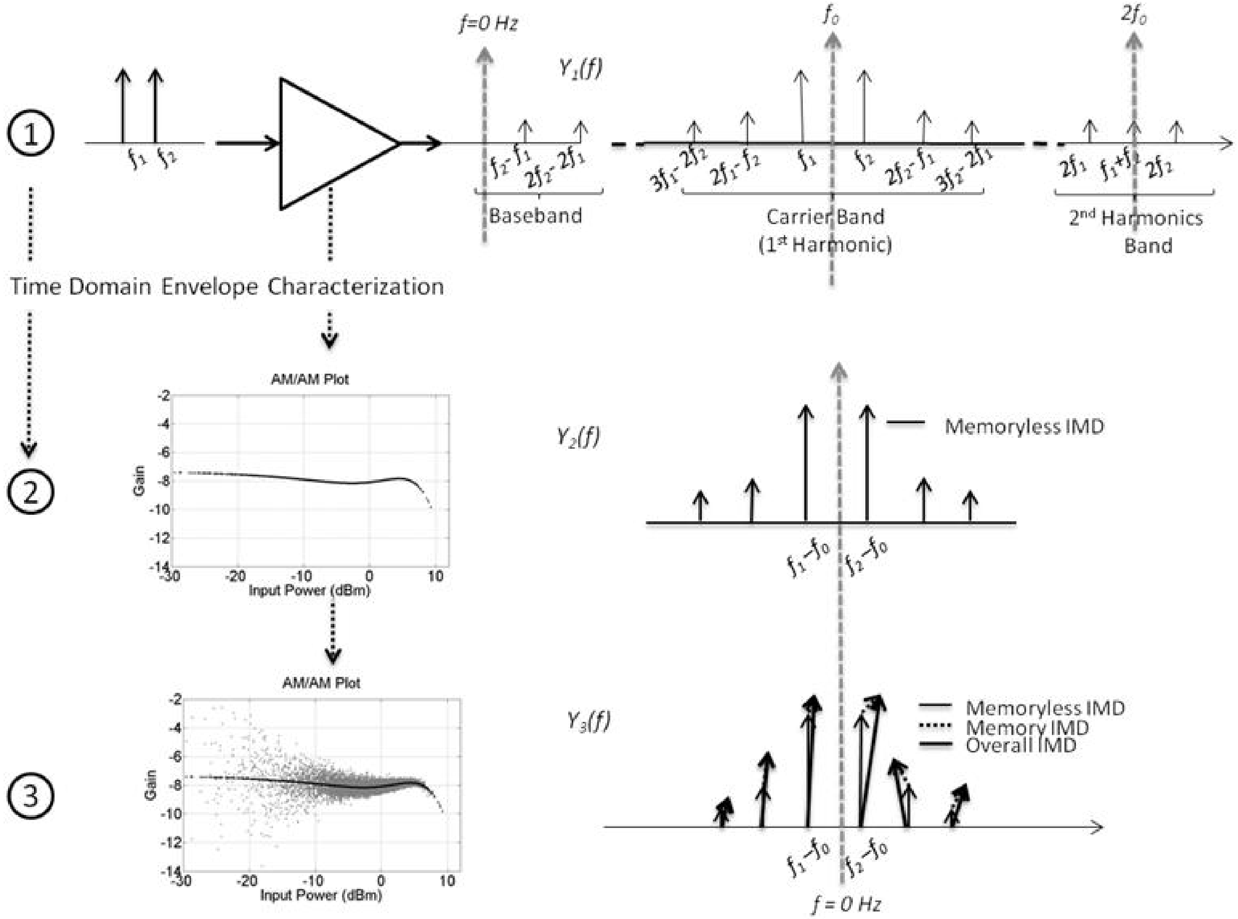


Figure 2.8 – Time and Frequency Domain Interpretation of a typical RF PA response in real time domain and RF envelope time domain – 1) Real time physical behavior of the PA with frequency components spanning from baseband to higher carrier harmonics – 2) Measured or simulated signal in the envelope time domain, with carrier band down converted to baseband, and 3) Baseband equivalent signal showing memory effects.

2.2.3.1 The Volterra Series

The Volterra series is one of the most comprehensive schemes for modeling nonlinear systems with memory. It is a generic modeling scheme that can be used to model an arbitrarily nonlinear system; it has been used for modeling biological (physiological) systems, nonlinear satellite links, multiple input devices such as mixers, and microwave circuits. Under suitable linearity conditions, the Volterra models with truncated nonlinearity order and memory can be used to represent nonlinearities of any order, to an arbitrary accuracy, over a given input amplitude range [14].

A Volterra series can be described as a 'Taylor series with Memory', defining the distorted output signal *y(t)* as an infinite series:

$$y(t) = \sum_{k=0}^{\infty} D_k(t) \tag{2.5}$$

where $D_k(t)$ is the k^{th} order response of the system, formed from an k-fold convolution of the input signal $x(t)$ with the n^{th} order nonlinear impulse response of the system $h_k(\tau_1,...,\tau_k)$, known as the Volterra Kernels [5]:

$$D_k(t) = \int_{-\infty}^{\infty}...\int_{-\infty}^{\infty} h_k(\tau_1,...,\tau_k)x(t-\tau_1)...x(t-\tau_k)d\tau_1...d\tau_k \tag{2.6}$$

The Volterra series model is expressed in a matrix form as

$$y[n] = f(D_k) \tag{2.7}$$

where f is a nonlinear system function with memory and the vector Dk is given by:

$$D_k = (D[n]\ , D[n-1]\ , D[n-2],...,D[n-M+1]) \tag{2.8}$$

M is the memory duration of the nonlinear system and $d[n]$ is the output of the Volterra model Predistorter defined as: $D[n] = h.X_T[n]$, where, the superscript T denotes the transpose of the matrix, h is the Volterra kernel vector, and $X[n]$ is the input vector.

The Volterra series presents the advantage of modeling each component separately, however, its main disadvantage is it has poor convergence properties. To characterize applications with a saturating behavior, it therefore requires a large number of terms, i.e. $k \geq 5$ for power amplifier modeling [15]. The extraction of higher order kernels (k>2) is practically and computationally ineffective, causing the Volterra series to be non-viable for modern communication circuit modeling.

2.2.3.1.1 Power Series

The Voltera series is a time-domain representation of the output of a nonlinear system. An equivalent frequency-domain power series representation can be found and the result is a much more efficient series, which can deal with severe nonlinearites. The input signal can be described as:

(2.9)

where are the magnitudes of the individual frequency components, . The output signal from the nonlinear system, $y(t)$, can then be expressed as a generalized power series:

$$y(t) = A\sum_{i=0}^{\infty} a_i \left\{ \sum_{n=1}^{N} b_n x_n (t - \tau_{n,i}) \right\}^i \qquad \textbf{(2.10)}$$

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

This technique has better modeling efficiency than the Volterra series because of its operation in the frequency-domain. However it is unclear how to generate the above coefficients from tabulated measured data. The inclusion of frequency-dependence of the time delay terms and the use of complex coefficients in the model allows a very wide range of nonlinearities to be characterized. This is therefore potentially a powerful modelling method, assuming that the relevant coefficients can be obtained.

2.2.3.1.2 Derivations from the Volterra series: The Memory Polynomial

The Memory polynomial as introduced by *Kim et. Al.* [2] is derived from the Volterra series formulation in an attempt to reduce the latter's complexity and take advantage of its modeling capability. The terms eliminated from the Volterra series are all the cross terms of samples of *x(n)*. Thus, considering the matrix formulation of the Volterra series, the remaining terms are the diagonal terms, corresponding to pure powers of the input signal samples. The memory polynomial formulation can be expressed as:.

$$\boldsymbol{y_{MP}(n) = \sum_{i=0}^{T} \sum_{m=1}^{M} c_{m,i} \; x(n-i)|x(n-i)|^{m-1}} \qquad \textbf{(2.11)}$$

In this formulation, *y_{MP}(n)* and *x(n)* are the output and input signals respectively, T is considered the memory length and M the highest order of nonlinearity of the polynomials. A Least square algorithm can be used to find the coefficients $c_{m,i}$ of that equation, with the following problem formulation:

$$\underbrace{\begin{bmatrix} x(1) & x^2(1) & \cdots & x^K(1) & x(2) & \cdots & x^K(2) & \cdots & x^K(M) \\ x(2) & x^2(2) & & & & & & & \\ \vdots & & \ddots & \vdots & & \ddots & & \ddots & \vdots \\ x(n) & x^2(n) & \cdots & x^K(n) & x(n+1) & & x^K(n+M) & \cdots & x^K(n+M) \end{bmatrix}}_{A} C = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(n) \end{bmatrix} \qquad \textbf{(2.12)}$$

The input data points are constructed as an nx$(M+K)$ matrix A, multiplied by the coefficient vector C, where . The conditioning of Matrix A is an important figure in the evaluation of the stability of the modeling schemes.

Whether one looks at the M-Polynomial as a derivation from the Volterra, or a simple expansion of the single polynomial described in the memoryless modeling in a way to include past input samples in current output, the M-Polynomial remains one of the most commonly used behavioral modeling scheme. Its proven modeling capability, along with its relatively lower complexity, makes it a gold standard model against which to evaluate the performance of other schemes.

2.2.3.2 Non-Polynomial based behavioral models

A few other behavioral models have been successful at capturing the PA's characteristics. These include Look up Tables (LUT), neural networks [16] and genetic algorithms. Other types of models are the two-box models known as Wiener or Hammerstein models, shown in Figure 2.9. These use a cascade of a nonlinear function and a linear filter to model dynamic nonlinear systems. The first box of the Hammerstein scheme accounts for the PA's static nonlinear behavior, while the second captures its memory effect. The static nonlinear part of the Hammerstein/Wiener model is implemented using polynomial functions, the LUT, and Neural Networks [16]. Parallel Hammerstein/Wiener models have also been suggested [17] to address the limited capability of traditional schemes to account for memory effects.

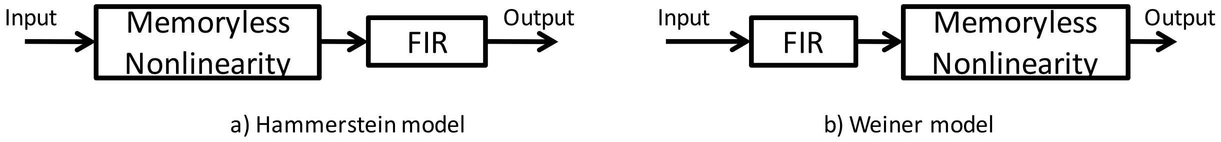


Figure 2.9 – Hammerstein and Weiner Model Diagrams.

2.3 Experimental Validation of Models

Models developed in this thesis, or procedures proposed are tested experimentally to verify their validity. The following paragraphs outline the experimental setup used, the device under test, and testing procedures for the accuracy and efficiency of the various PA modeling and linearization schemes presented.

2.3.1 Experimental Setup

The diagram of Figure 2.10 represents the setup used for measurement tests. Tests signals are first synthesized with Advanced Design System (ADS) software by Agilent, which controls the signal generator via GPIB (General Purpose Interface Bus). Signals generated are fed into the PA, which is described in the next paragraph as the device under test, whose output is captured by the Spectrum Analyzer. The latter communicates information back to ADS and the vector software analyzer.

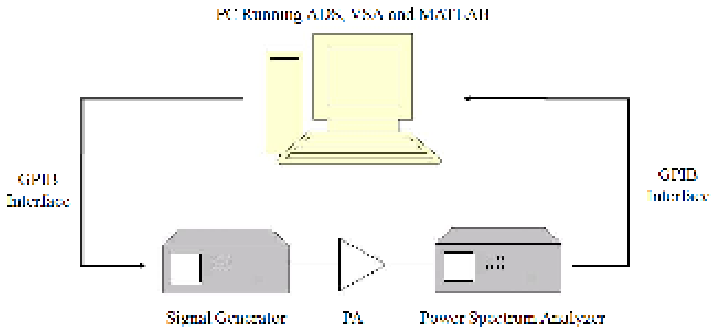


Figure 2.10 – Experimental Setup

2.3.2 Device Under Test (DUT)

Experimental validation of the results presented through the thesis is performed on a PA lineup that consists of three PAs all implemented using Laterally-Diffused Metal Oxide Semiconductor Field Effect Transistor (LDMOS-FET) technology:

1. 5 Watt IC driver (Freescale MHV5IC2215N)
2. 100 Watt class AB driver (Freescale MRF6S21100H)
3. 400 Watt Doherty PA (2x Freescale MRF7S21170H).

The output spectrum of the PA lineup is measured in both forward and reverse modeling scenarios. The construction of the dynamic AM/AM and AM/PM characteristics of the PA from this data, as well as that of the inverse function of these characteristics is then performed in MATLAB®.

A four-carrier WCDMA test signal, with a Peak to Average Power Ratio (PAPR) of 8.30dB, synthesized in Agilent Design System (ADS) was used as a test signal, as shown in Figure 2.11.

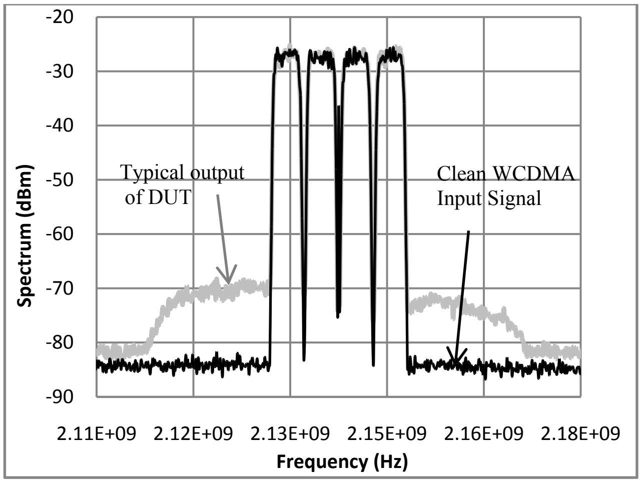


Figure 2.11 – Typical DUT Input and output spectra of a 4-carrier WCDMA signal.

2.3.3 Forward and Reverse Model Validation Tests

These tests go farther than just evaluating a model's accuracy. A model can be mathematically accurate, consisting of a function that theoretically 'fits' the input output response of the DUT. However, modeling this response consists of capturing the real physical, desired and spurious elements present in that response.

2.3.3.1 The Forward Validation Test

The Forward Validation test investigates the capability of the model to capture the memory effects observed in the behavior of the PA. In a typical PA, the nonlinear memory effects are perturbations, usually much smaller than the memoryless nonlinearities. Consequently, assessing the fidelity of the model by comparing the complete model and DUT signal spectra will emphasize the accuracy of the memoryless part of the model, and make it difficult to assess the accuracy of memory modeling because of these effects are relatively small perturbations. The test is based on the application of a memoryless predistorter as described in [17]. The objective of this test is to highlight the memory effects in the PA behavior, by eliminating the memoryless components of the nonlinear behavior. A memoryless predistorter is constructed and applied to the DUT and its simulated model, leaving only memory effects as distortions in the output. The two output signals, one from the linearized DUT and the other from the linearized model, are then compared, showing the extent of the model's capability in reproducing the small perturbations present in the DUT.

2.3.3.2 The Reverse, or DPD, Validation Test

The DPD implementation of a model requires the synthesis of the inverse function of that model. For several schemes, including the Volterra series, an analytical formulation of the inverse model can be obtained. However, a more direct method is to simply reverse the input and output data, since the reverse of a model that transforms A to B, is another model that transforms B to A. Therefore, a scheme that is considered accurate in forward mode, need not necessarily handle the reverse data accurately enough, thereby the necessity of testing a model in reverse, or DPD mode.

Chapter 3
A Deterministic Approach for Behavior Model Optimization

Many behavioral schemes are used to model RF PA systems, and are thereafter implemented as DPD schemes for front end linearization. Complexity of the behavioral models is an essential criterion in determining a scheme's viability. It determines the feasibility of the scheme's implementation as it greatly affects, among others, its synthesis (identification) complexity and numerical stability, and its execution burden (number of multiplications and additions).

3.1 The Memory Polynomial: Trial and error structure optimization

The M-Polynomial is the most popular model, combining the comprehensive nonlinear modeling properties of the Volterra series, from which it was derived, with lower complexity. As described in paragraph (2.2.3.1.2) it retains the diagonal terms of the Volterra series while ignoring its cross terms, and is therefore formulated as:

$$\boldsymbol{y_{MP}(n) = \sum_{i=0}^{T}\sum_{m=1}^{M} c_{m,i}\; x(n-i)|x(n-i)|^{m-1}} \qquad \textbf{(3.1)}$$

where *x(n)* and *$y_{MP}(n)$* are the complex input and output envelope signals respectively. T is the memory length of the system, equivalent to the number of polynomial branches. M represents the polynomial order of the branches and $c_{m,i}$ designates the *m*th polynomial coefficients of the *i*th branch. The structure of the M-Polynomial is shown in Figure 3.1, where

$$\boldsymbol{Poly\ i = \sum_{m=1}^{M} c_{m,i}\; x(n-i)|x(n-i)|^{m-1}} \qquad \textbf{(3.2)}$$

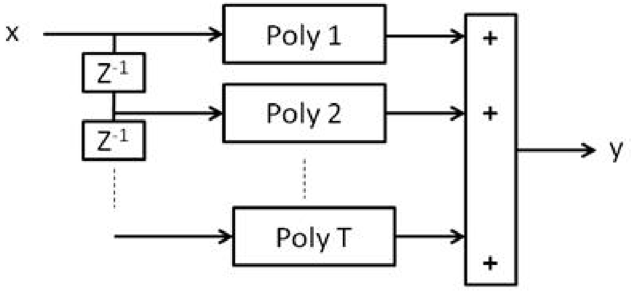


Figure 3.1 – Structure of the Memory Polynomial model

The design process aims at determining the optimal number of parameters in that structure, such that it has the least complexity while not compromising its modeling or linearization accuracy levels. Moreover, even the M-Polynomial has been found to have redundant parameters [3], and the real challenge presented to RF designers is to single those out. The current procedure to determine the optimal structure consists of performing a series of experimental measurements. Each measurement is done using a specific combination of parameters in the DPD. As implied by the term combination, this trial and error process is extremely time consuming, sometimes not even feasible if the designer does not have direct access to an experimental measurement setup.

3.2 The Parallel Hammerstein Model

To remedy to the inefficiency of that procedure, a systematic approach in determining the optimal modeling structure of RF PAs is proposed based on the parallel Hammerstein scheme described in [18]. The formulation of the parallel Hammerstein consists of the same terms as the M-Polynomial. In fact, the dispersion characteristics of the observed PA output, $y(n)$ in terms are defined in terms of the input $x(n)$ and its samples over time as follows:

$$\boldsymbol{y(n) = \sum_{m=0}^{M} \sum_{i=1}^{Tm} a_{i,m} \; x(n-i)|x(n-i)|^{m-1}} \tag{3.3}$$

where M is the maximum nonlinearity, T_m being the number of taps of the FIR filter corresponding to the mth nonlinear order.

The terms of that equation are nonlinear recombination of the input and its past samples, and it is clear that the terms constituting the M-Polynomial model are all found in the Parallel Hammerstein formulation. In fact, the M-Polynomial's *(m,i)* terms are the same as the *(i,m)* terms of the parallel Hammerstein. However, the repartition of these terms in a structure that is useful in modeling is shown in Figure 3.2 as the Parallel Hammerstein diagram.

However, the major difference between the two resides in their respective structures. The parallel Hammerstein's structure groups the nonlinear terms based on their order into a separate FIR filter for each nonlinear order, as shown in Figure 3.2.

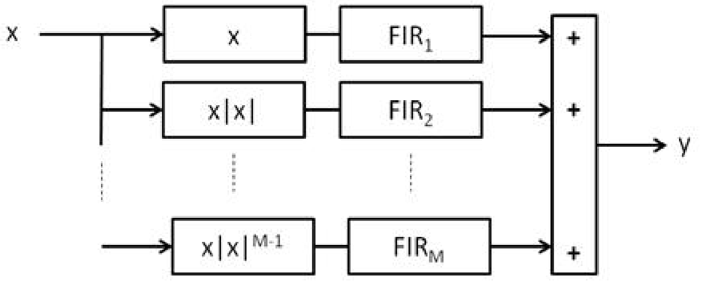


Figure 3.2 – Parallel Hammerstein Block Diagram

The advantage of the parallel Hammerstein resides in its structure: first, it allows for direct model extraction without the need for many iterations, or an optimization loop; second, a systematic determination of the optimal structure and number of coefficients is only made possible by visualizing the

FIRs of each nonlinear order as shown in the following sections; and third, identification of the sources of distortions in order to mitigate them from the power amplifier design stage when possible.

Extracting the model's coefficients consists of identifying the FIR filter taps from the PA input/output envelop records available from measurements. The measured output can be re-written in matrix form as

$$y = CA \tag{3.4}$$

where C a complex matrix is constructed based on $x(n-i)\left|x(n-i)\right|^{m-1}$ values and A is the set of filter coefficients to be extracted. Since the model is linear in a_{mt}, the coefficients identification algorithm uses the least square (LS) Error optimization method. This method first inverts matrix *C*, then extracts *A* using the pseudo inverse function (*pinv*) that involves the singular value decomposition (SVD), as

$$A = pinv(C)y \qquad \textbf{(3.5).}$$

The calculation complexity of the LS algorithm is proportional to the cube of the number of unknowns, which significantly improves with a lower number of model coefficients to be extracted. The accuracy and stability of the results are directly related to the numerical conditioning of the matrix *C* [19].

3.3 Initial Model Construction and Verification

The parallel Hammerstein model is constructed and validated experimentally using two approaches: forward and reverse (DPD) mode validations. The largest order of nonlinearity *M* and the memory order T_m were set to achieve a high accuracy prediction of the response of the circuit under test. The DUT used was detailed in section (2.3.2), and the validation tests in the following sections.

3.3.1 Forward Validation of the parallel Hammerstein

The accuracy of the model is first tested in forward mode, as described in section (2.3.3). The DUT is linearized using a memoryless predistorter; the output signal spectrum from the predistorted DUT will then contain only the nonlinear memory effects. This spectrum is shown as (A) in Fig. 2. Next, the DUT behavior is modeled using a Parallel Hammerstein model containing both nonlinear memoryless and memory components. This Parallel Hammerstein model is then linearized using the same memoryless predistorter that was used to linearize the actual DUT. The predistorted output spectrum from this model is shown as curve (B) in Fig. 2. Because we have removed the identical memoryless nonlinear effects from the DUT and model spectra, we can observe how well the model captures the DUT memory effects more clearly. As can be seen in Fig. 2, the spectra (A) and (B) are effectively superimposed, indicating that the Parallel Hammerstein model is very effective in recreating the slightest perturbations in the signal, seen here as the memory effects. Finally a 3rd test is implemented: a memoryless model of the

DUT is extracted, capturing only the memoryless nonlinearity, and linearized using the same memoryless predistorter as previously. The output spectrum from the predistorted memoryless model is shown as curve (C) in Fig. 2: it is a clean signal without any perturbations. This implies that the spectral perturbations observed in both spectra (A) and (B) are essentially the memory effects present in the DUT.

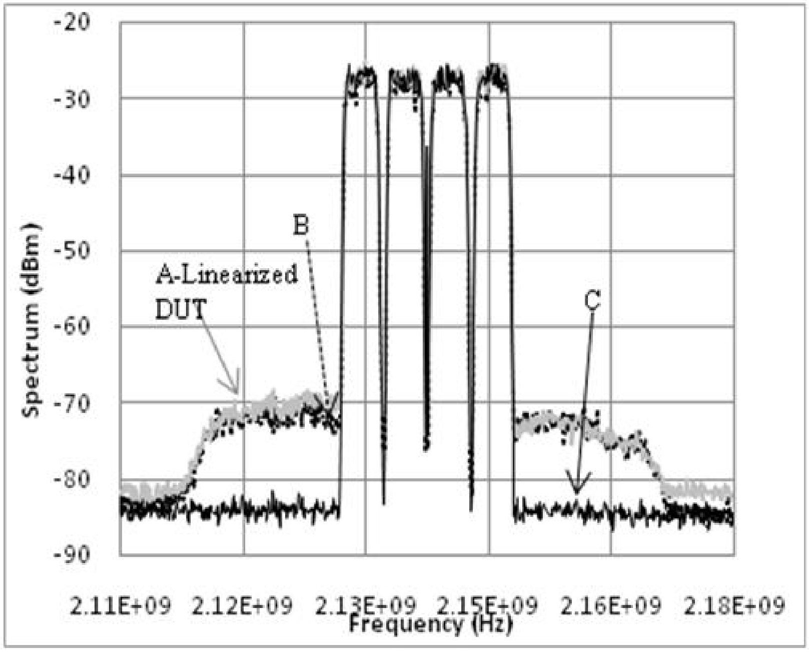


Figure 3.3 - Forward Parallel Hammerstein Validation Spectra: (A) Memoryless linearization of the actual DUT compared with (B) the memoryless linearization of simulated DUT Parallel Hammerstein, and (C) memoryless linearization of the memoryless Parallel Hammerstein extracted from the DUT.

3.3.2 Predistortion Validation of the parallel Hammerstein

The signal quality at the output of the test circuit is measured when a Parallel Hammerstein DPD is added in the signal chain. Figure shows the raw, non linearized output spectrum from the DUT, and the linearized output spectrum using a Parallel Hammerstein DPD scheme that employs a 10th order nonlinearity and memory length of 10. The DUT is driven with a 4-carrier WCDMA signal. The excellent linearization ability of the Parallel Hammerstein is clearly demonstrated through the reduction in the out-of-band spectrum emission of about 20dB, demonstrated in the output signal spectrum of Figure 3.4 and the adjacent channel power ratio, higher than 50dBc, shown in Table 3.1.

Table 3.1 – ACPR of Measurements for Linearization Validation of Parallel Hammerstein

DPD Configuration	ACPR (dBc)		
	1st Lower	2nd Lower	3rd Lower
Parallel Hammerstein DPD Nonlinear Order 10, FIRs length 10	-49.60	-52.06	-53.65

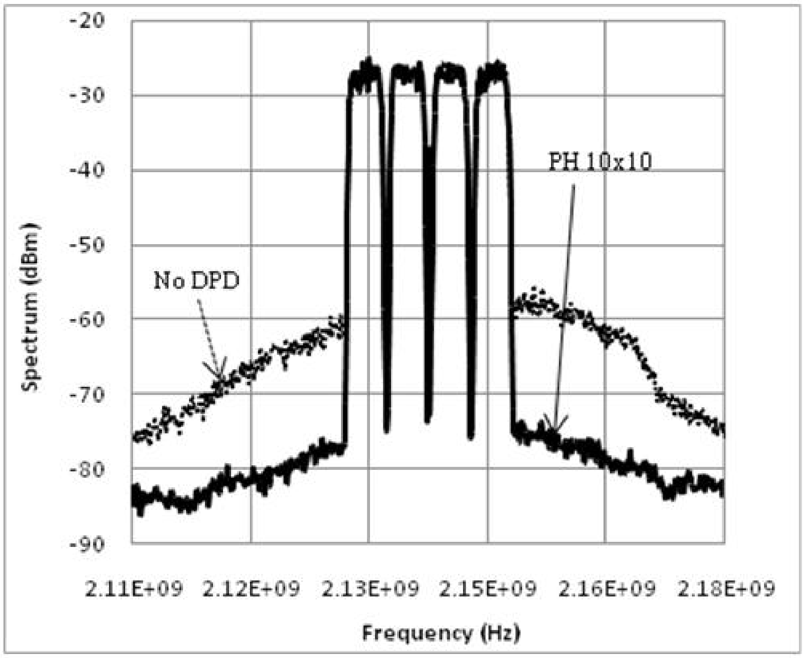


Figure 3.4 – Output spectrum of the linearized DUT using the Parallel Hammerstein based DPD (reverse validation)

3.4 Parallel Hammerstein Model Complexity Reduction:

3.4.1 Even Order Omission

As previously mentioned, the viability of models depends on their computational and implementation complexity. Advantageous and sound DPD design aims at lowering the required number of coefficients and operation (multiplication and addition). In this section, the omission of even order terms from the Parallel Hammerstein model is discussed for practical WCDMA signal linearization. While the omission of the even-order products from consideration is often done from considerations of the primary spectral components of the distortion, the following paragraphs present theoretical reasons for excluding these components, and then its effect on linearization based on the conditioning number and stability of the resulting DPD model.

3.4.1.1 Theoretical Explanation of Even Order Omission

In a simple two-tone test, even order distortions (i.e. $f_2 - f_1, 2f_1, f_2 + f_1, 2(f_2 + f_1)\ldots$) generated by a transistor are not physically present in the operation bandwidth of the amplifier. Moreover, unless the original signal input power is high, the recombined components have a minor effect on the resultant signal distortion, as their power is further reduced through the mixing process. The intermodulation distortions that fall in the vicinity of the carrier are the result of only odd order distortions of the input, i.e. components resulting from $x(t),\ x^3(t),\ x^5(t),\ldots$. Components produced from even order distortions such as $x^2(t),\ x^4(t),\ldots$ all fall either in the baseband region or higher harmonic signal bands, and it is only after they recombine with odd order components that their effect is observed within the carrier band. Therefore, although the even order products affect the resulting carrier band segment, none of them is physically present within this section.

Distortions around the carrier band consist of odd-only nonlinear terms. Therefore, for an efficient low pass equivalent modeling and predistortion purposes, it is necessary and sufficient to use only the odd orders to capture the "status" of the pass band, whether this information is originating from purely odd orders, or from the recombination of these with even orders. It is sufficient to capture the effects of such even nonlinearities on the output signal envelope. Accounting for only odd order nonlinearities would yield a satisfactory linearization and greatly simplify the model. Modeling RF PA with only odd order terms has proven to be an efficient complexity reduction method for the M-Polynomial model [3]. The following paragraphs present experimental results that show similar impact when the parallel Hammerstein model is reduced.

3.4.1.2 Impact on Signal Linearization

This complexity reduction strategy is first evaluated based on its effect on the DUT linearizability. The figure of merit used to measure the quality of signal is again the ACPR of the linearized signal, indicating the linearization capability of the odd-only parallel Hammerstein.

Table 3.2 - ACPR comparison for Even order omission strategy

Parallel Hammerstein DPD Configuration	ACPR (dBc)		
	1st Lower	2nd Lower	3rd Lower
Order 9 Even Odd FIRs: 10 All	-48.30	-49.80	-50.50
Order 9 Odd Only FIRs: 10 All	-49.15	-50.85	-52.38

The spectra of the test signals used are shown in Figure 3.5. It is clear that the linearization capability of the parallel Hammerstein is not reduced when even order terms are omitted from the DPD scheme.

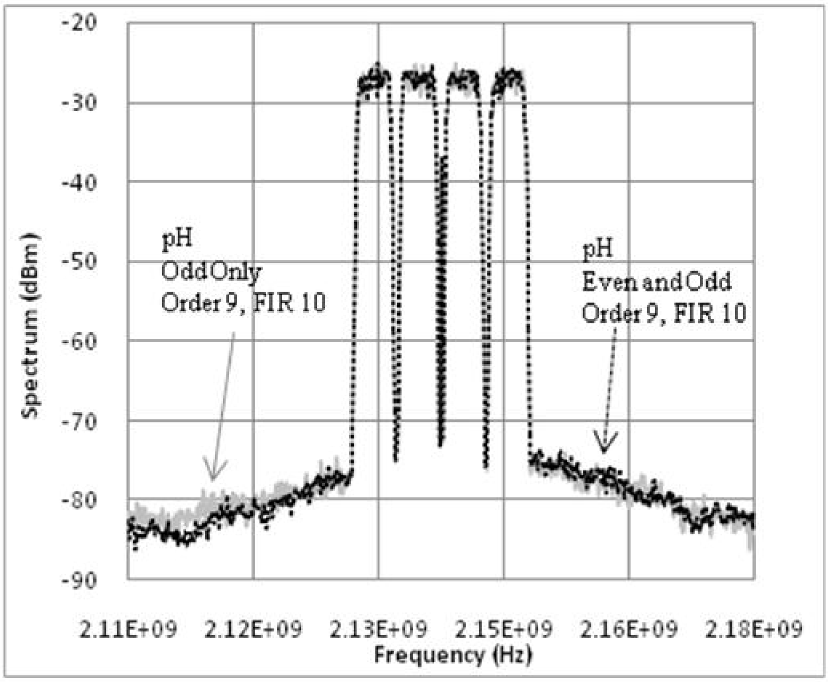


Figure 3.5 – Output spectrum of the linearized DUT using parallel Hammerstein (order 10) with and without even order parameters.

3.4.1.3 Impact on the numerical Stability of Model Extraction

The outcome of the even order omission is also evaluated based on its effect on model implementability. The main criteria to estimate ease of implementation are first, the numerical stability of the model extraction, second, the number of coefficients required by the model and third the number of operations needed.

The stability of the model extraction algorithm is a major issue that has been addressed for the memory polynomial models. The least squares algorithm is used to extract the coefficients of the filter taps. For high nonlinearity orders, the regressor matrix in the least squares coefficient estimation is ill-conditioned and causes the computations to be highly numerically unstable. Many attempts have been made to improve the Memory polynomial's conditioning number, one of which was to use orthonormal basis. In the case of the parallel Hammerstein, the conditioning number indicates a much better performance than the M-Polynomial. While it is observed that indeed the magnitude of this number is mainly affected by the highest nonlinearity order used in the model, it is not the only factor that comes into play. As shown in Table 3.3, for the same nonlinearity order, the conditioning number is reduced by almost 3 orders of magnitude when only odd-order nonlinearities are used. Also, a high conditioning number suggests a correlation between the data from the different nonlinearity orders, or that the different data samples are correlated. Since the same data is used in the comparison of Table 3.3, the main factor behind the reduction in the condition number is that the nonlinearity orders are not correlated when odd orders only are used. Therefore, omitting the even nonlinearities avoids the redundancy in the data used by the model, keeping only the pertinent information that can be modeled by the odd order nonlinearities only, as shown in the spectrum correction of Figure 3.5.

Table 3.3 - Conditioning Number as a measure of Stability

Parallel Hammerstein DPD Configuration	Conditioning Number
Order 9 FIR 10 Even + Odd	2.40E+07
Order 9 FIR 10 Odd Only	1.70E+04

From this point on, all the DPD schemes used employ only the odd nonlinearities.

3.4.2 Empirical Optimization of The Parallel Hammerstein's Structure

3.4.2.1 Determining the Optimal Model Structure

So far, the construction of the parallel Hammerstein scheme employed the same number of taps in the different FIR filters associated to the various order of nonlinearity. Furthermore, determining the memory length or equivalently the number of filter taps that is required for an efficient PA linearization is usually done by trial and error. The best DPD configuration is chosen based on figures of merit, such as Normalized Mean Square Error (NMSE) using an iterative and lengthy process. Therefore, the process of determining the memory length required is a sequential optimization process of simulation results followed by experimental checks. Conversely, in this section, the visualization of the impulse response of every digital filter in the parallel Hammerstein structure will be used to elaborate a deterministic way for finding the optimal FIR filter lengths of each nonlinearity order. This will also yield the smallest number of taps in each filter that maintains the modeling/linearization capability.

Starting from the coefficients of the parallel Hammerstein scheme, constructed in previous section with an order of nonlinearity of 9 and a memory order of 10, the different filters' coefficients are visualized in Figure 3.6. FIR filters visualized are FIR_1, $FIR_{3,}$… FIR_9 corresponding to the nonlinearity orders of 1, 3,…9. For a better visualization of the actual behavior of each FIR filter of Figure 3.6, the frequency domain responses of each was synthesized from its FIR taps and are shown in Figure 3.6.One can easily detect a generic trend in the amplitude of the filter taps. They all tend to converge to a certain range that is much lower than the first tap, while the rate of change between taps becomes minimal. This implies that the signal samples corresponding to those memory levels do not affect the output of the model, or equivalently the corresponding DPD. Therefore, eliminating those taps would reduce the complexity of the model without compromising its accuracy and linearization efficiency.

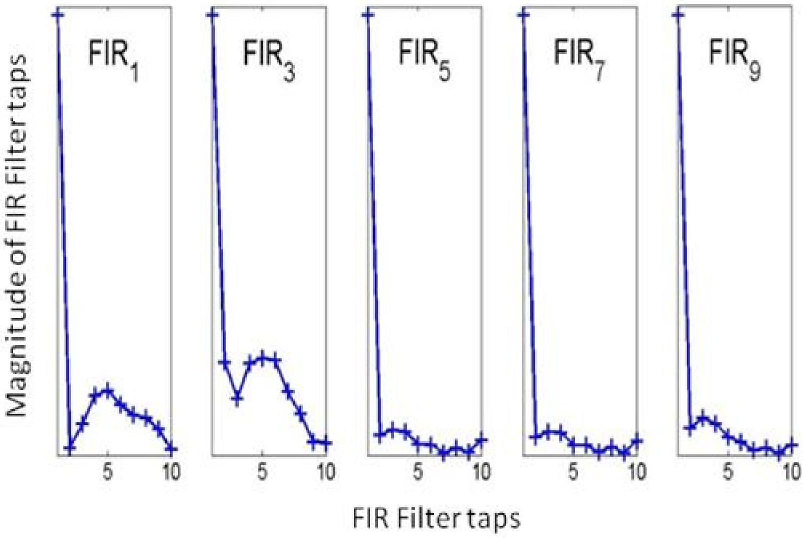


Figure 3.6 - Individual FIR Filter Coefficients Plots for a Parallel Hammerstein DPD employing Odd Only nonlinearities of maximum order 9 and FIR filter length of 10.

The FIR visualization of Figure 3.6 indicates that in this case, using a uniform filter length of 10 taps is over estimating the memory effects of the DUT. In fact, the comparison of the spectra of Figure 3.8 shows similar linearization capability of the Parallel Hammerstein when using filter lengths of 10 for all the FIR filters of the model, and the case where the lengths used were reduced for each order. The second measurement uses FIR filter lengths of 4, 6, 3, 3, 4 for filters FIR$_1$, FIR$_3$, FIR$_5$, FIR$_7$ and FIR$_9$ respectively. These lengths were determined by looking at the last significant filter tap on the visualization of Figure 3.6, and chopping off the taps after the change in amplitude trend, in other words, after the drop in tap amplitudes. Since the first tap of each filter corresponds to the purely nonlinear, memoryless response, its amplitude is much higher compared to the rest of the taps that correspond to the perturbations of smaller amplitude associated with memory effects. Hence, for a closer visualization of the FIRs in Figure 3.6, the first tap is left out, showing only the rest of the filter, i.e. taps 2 to 10, as shown in Figure 3.7. Common for all the filters responses, an increase in tap amplitude is observed followed by a relatively sharp decrease, most obvious in FIR3. Filters 5,7 and 9 clearly show that the taps following the sharp change in amplitude follow a mathematical fitting for the problem formulation of the model rather than a physical explanation for the existence taps.

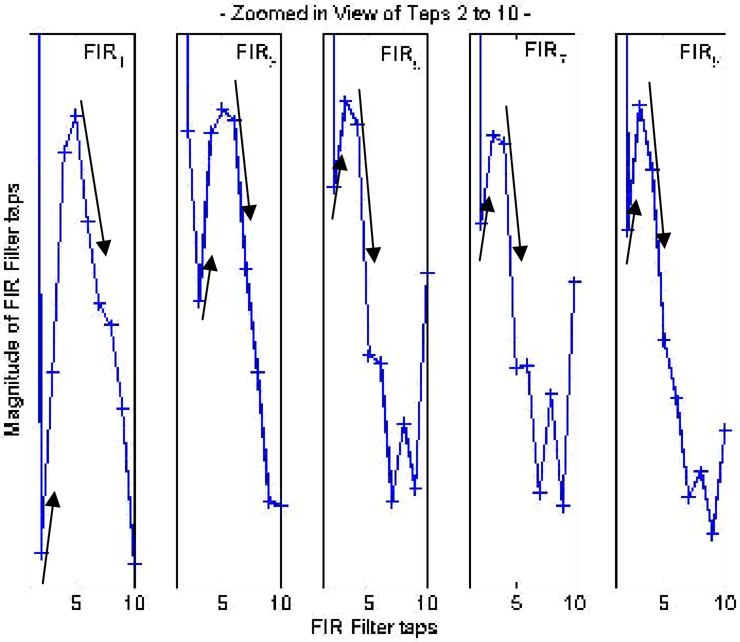


Figure 3.7 - Zoomed in View of the FIR response: Taps 2 to 10 -- the Arrows indicate where the amplitude suddenly decreases after a constant increase from tap 2.

Therefore, for every filter, cutting the response length to the last significant tap produced similar results as having an over fitted problem. Both full model and complexity reduced model achieve similar linearization results, demonstrated in Figure 3.8 as well as Table 3.4 which shows the ACPRs for the PA output signal when the PA is linearized using both DPD schemes. One can clearly observe similar linearization capability in both cases although the number of coefficients was substantially reduced.

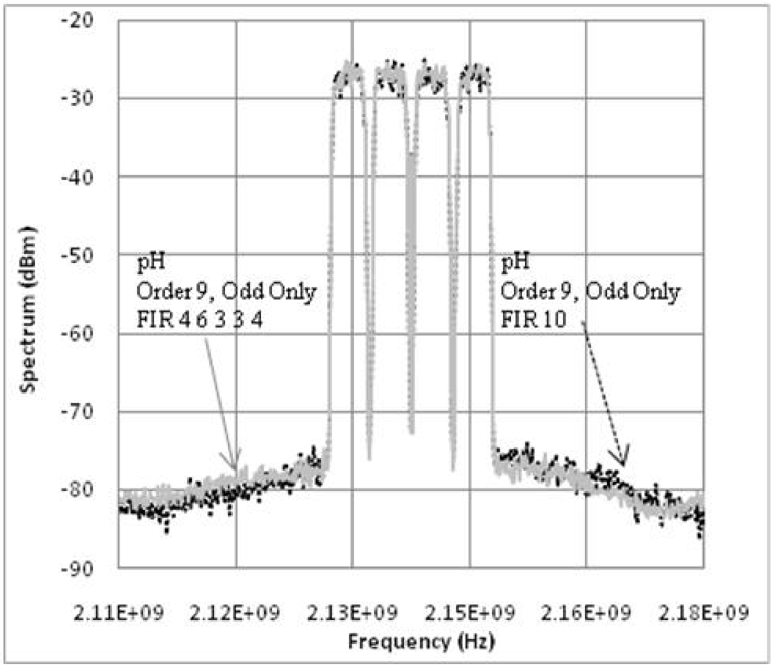


Figure 3.8 – Power Amplifier Output Spectra obtained using a 9th order parallel Hammerstein linearization with uniform and Optimized FIR Filter lengths

Table 3.4 - ACPR of Linearization schemes

Parallel Hammerstein DPD Configuration	ACPR (dBc)		
	1st Lower	2nd Lower	3rd Lower
Order 9 Even+Odd FIRs: 10 All	-48.30	-49.80	-50.50
Order 9 Odd Only FIRs: 4 6 3 3 4	-49.90	-51.20	-52.10

It is interesting to note that the NMSE and conditioning number of the two schemes are very similar as shown in Table 3.5, suggesting that it is impossible for a designer to determine an optimal DPD structure without having the advantage of the Parallel Hammerstein's filter visualization.

Table 3.5 - NMSE and Conditioning Number of parallel Hammerstein schemes

Parallel Hammerstein DPD Configuration	NMSE (dBm)	Conditioning Number
Order 9 Odd Only FIR 10	-36.08	1.70E+04
Order 9 Odd Only FIR: 4 6 3 3 4	-36.11	1.16E+04

The complexity of the model is directly dependent on the number of coefficients to be extracted for that model. The complexity reduction is not only essential in model implementability, but also in model extraction. Table 3.6 shows the reduction in the number of coefficients required for linearization. The first Parallel Hammerstein linearization configuration consists of the original Parallel Hammerstein scheme with no improvements, i.e. 9th order considering both even and odd orders with a common FIR filter length of 10. The 2nd scheme is the optimized Parallel Hammerstein consisting of only the odd orders coupled with optimal filter lengths of 4, 6, 3, 3 and 4 for the orders 1, 3, 5, 7 and 9 respectively.

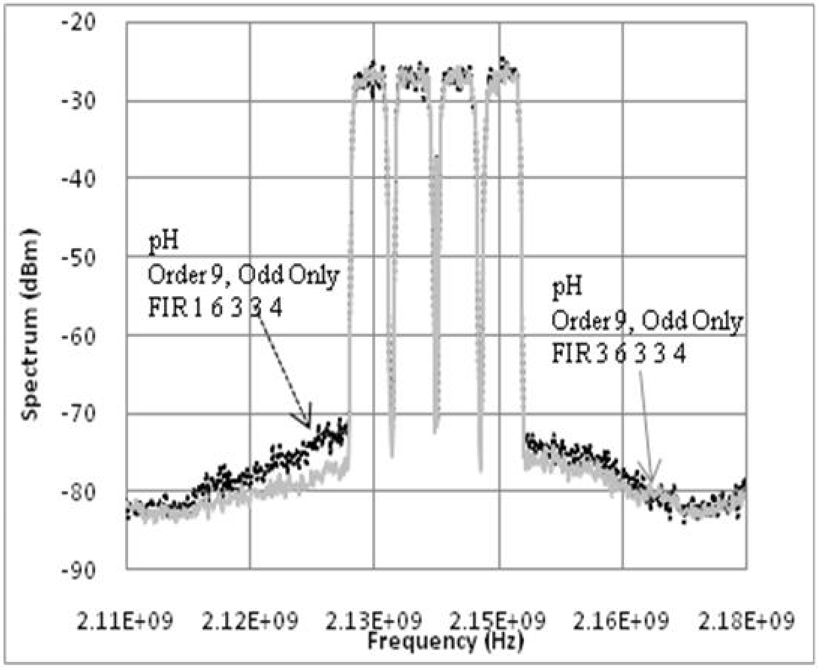
Table 3.6 – Total Number of Coefficients reduction achieved

DPD Configuration	Total Number Of Coefficients
Order 9 Even and Odd FIR 10	90
Order 9 Odd Only FIR 10	50
Order 9 Odd Only FIR: 4 6 3 3 4	20

With only one set of measurements, the number of coefficients required for implementation was reduced by a factor of 4.5. The simplification is even more significant with regards to model extraction whose complexity is proportional to the cube of the number of coefficients. The reduction is not only achieved when comparing Parallel Hammerstein to Parallel Hammerstein, but also in comparing the optimal scheme to M-Polynomial linearization which would use the same number of coefficients as the 1st scheme in Table 3.6.

3.4.2.2 Proving the non-Flatness at Baseband

Many models assume that the circuit under test has a flat gain across the signal baseband, implying a memoryless distortion of the inband frequency components. However, the frequency response of the first order filter as shown in Figure 3.6, shows indeed that the gain of the baseband FIR filter (FIR_1) is not flat for all frequencies. Refuting the assumption of flatness at baseband is strengthened by experimental measurements. The spectra of two DPD schemes devised for this test are compared in Figure 3.9, and their respective ACPRs in Table 3.7. The two DPDs have the same maximum nonlinearity order of 9, and use the same FIR filter lengths configurations except for the baseband filter, FIR_1 that was changed from 1 tap (memoryless) to 3 taps (with memory at baseband).



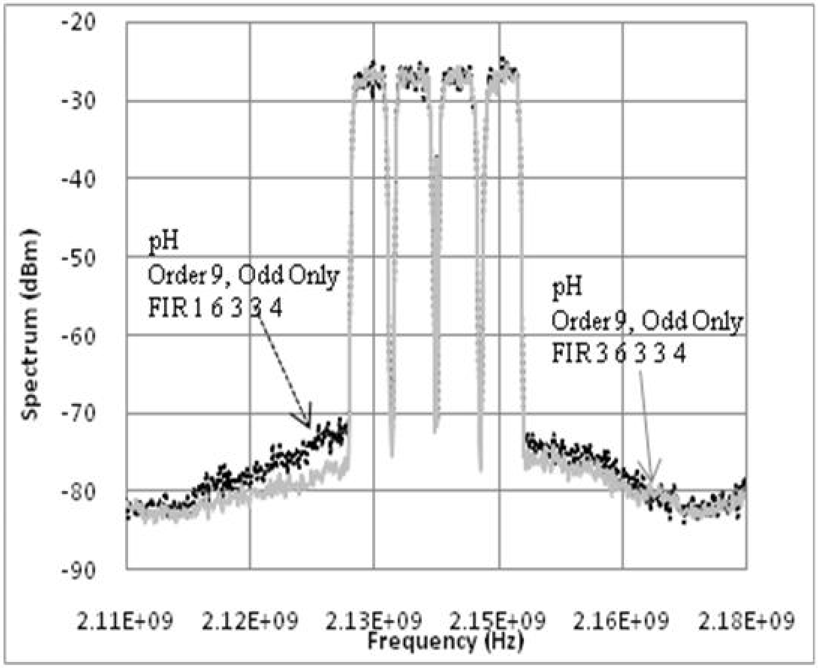


Figure 3.9 –Comparison of linearized spectra showing distortions arising from baseband distortions.

Table 3.7 – ACPR measurements showing the linearization losses incurred by Memoryless Baseband Linearization

Parallel Hammerstein DPD Configuration	ACPR (dBc)		
	1st Lower	2nd Lower	3rd Lower
Order 9 Odd Only FIRs: 4 6 3 3 4	-49.90	-51.20	-52.10
Order 9 Odd Only FIRs: 1 6 3 3 4	-45.40	-48.35	-50.12

Summarizing the chapter, the viability of linearization schemes is greatly dependent on their implementation complexity in real context. The main strength of the parallel Hammerstein scheme is explored through a proposed empirical approach that allows RF designers to systematically and reliably determine an optimal DPD/modeling structure. The minimal number of required coefficients was identified using a deterministic approach based on filters' impulse response visualization, testing the linearization on a 400Watt Doherty PA. The reduction of the number of coefficients was by a factor of about 4.5, implying substantial computation complexity reduction and numerical stability improvement, without compromising the linearization/modeling capability.

Furthermore, the examination of the Parallel Hammerstein's individual filter impulse response helped to identify certain memory effects contributors in the DUT circuit that can be mitigated at an early design stage. As an example, the non flat gain response of the PA under test, around the carrier, was detected by the simple visualization of the corresponding filter's impulse response.

Chapter 4
Interpretation of the Mechanisms of Memory Effects and their Close-Form Model Formulation

MEs are a major component in both the performance and linearization of wideband PAs, Understanding them and modeling them accurately will therefore allow designers to counteract their observed effects effectively. While behavioral models have proven to accurately model and effectively linearize PAs, they are mainly based on the input and output characteristics of the PA behavior without any linkage to the source of disturbances in that behavior. There is no closed form, solid and theoretical explanation of the rationale behind the applicability or non-applicability of these modeling schemes to model and linearize wideband PAs. To establish that link between theoretical modeling and physical behavior, it is essential to interpret electrical MEs not only as spurious frequency components at the output of the PA, but explaining their physical sources, then linking their characteristics in magnitude and phase to the physical properties of these sources. The latter sections suggest a model based on the physical characterization of the PA behavior, which will be shown to link back to the theoretical models previously introduced.

4.1 PA Building Blocks

Memory effects can be discriminated into two categories: Short term memory effects (STM) and Long term memory effects (LTM), depending on their frequency band location. Within the scope of this formulation, only STM are studied as they dominate the sources of ME when the PA is driven with wideband signals, i.e. the PA's response is considered quasi-static. The following paragraphs will explore the main sources of short term memory effects exhibited by a PA and discuss the key figures that determine the contribution of each of these physical sources on the overall characteristic of the circuit behavior.

Figure 4.1 is an illustration of a typical PA circuit and the connection of its main building blocks. The transistor on its own can be considered a quasi-memoryless nonlinear device providing that the signal bandwidth is relatively narrower than the carrier frequency. The transistor's output matching and biasing networks are subject to all the nonlinearities present at the output of the transistor. Depending on the nonlinearity order of the transistor, the spectra of such signals can span from baseband to high orders of the carrier harmonic frequencies.

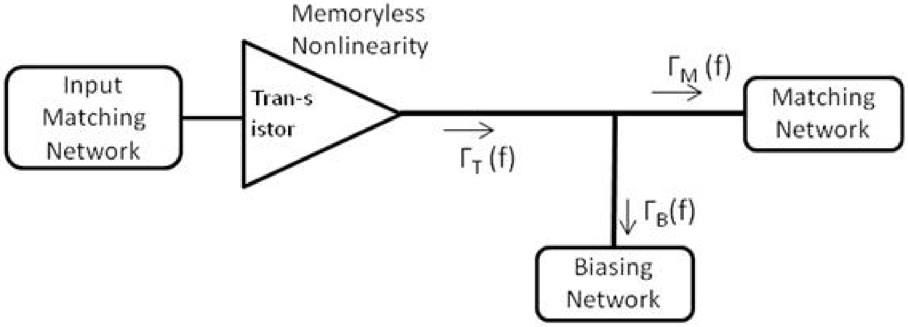


Figure 4.1 – Physical dominant sources of Short Term Memory Effects in a PA.

The output matching and biasing networks shown in Figure 4.1 are each characterized by their reflection coefficients, $\Gamma_M(f)$ and $\Gamma_B(f)$ respectively. These networks are typically designed to operate at a specific frequency or range of frequencies and their responses vary considerably when operating over a different or wider range of frequencies. Thus variations over the frequency range are observed in the amplitude and phase characteristic of their reflection coefficients.

A general approach to model inherent nonlinearities in the behavior of an RF transistor is through polynomial approximation of the transfer curves (AM/AM and AM/PM), where the time domain function $H_{tr}(t)$ defines the instantaneous output $y(t)$ as a function of the instantaneous input $x(t)$ as detailed in the following equation:

$$\begin{aligned} y(t) &= H_{tr}(x(t)) \\ &= p_1 x(t) + p_2 x^2(t) + \ldots + p_N x^N(t) \end{aligned} \tag{4.1}$$

The order of nonlinearity (*N*) is dependent on the physical characteristics of the transistor and its mode of operation, i.e. class, input power level, operation bandwidth. From this equation, the spectrum $Y(f)$ of the time domain output $y(t)$ can therefore be written as:

$$Y(f) = p_1 X(f) + p_2 X(f) \otimes X(f) + \ldots \tag{4.2}$$

where $X(f)$ is the frequency domain transform of the input signal $x(t)$. The spread of newly generated frequency components is determined by the order (*N*). Memoryless distortion of a signal does not alter the phase information of its components but only their amplitude. A new range of IMD frequencies is generated at the transistor output thus at the input ports of both the output matching network and biasing network as shown in Figure 4.1. These networks being non-ideal over all the frequency range reflect spurious components back into the transistor which initiates the STM effects as will be explained in the following sections.

4.2 Physical Explanation of Memory Effects

The two-tone test is used to illustrate the behaviors described in this section. Note that the spread of the frequency bands designated as baseband, carrier band, and harmonics bands, depends mainly on the order of nonlinearity considered (*N*) and the frequency separation of the two tones $\Delta f = f_2 - f_1$, i.e. signal bandwidth. The following paragraphs show in detail how the bandwidth determines the behavior of the amplifier circuit, and the effect of each source of dispersion on the overall memory effects observed at the output of the PA.

4.2.1 Sources of Memory Effects

4.2.1.1 Output Matching Network

The matching network at the output of the transistor is a linear device but contributes to the overall memory effects by reflecting several frequency components present at its input back into the transistor. This is governed by two key physical parameters of that network: its reflection and transmission coefficients. The reflection coefficient at the input port of the matching network ($\Gamma_M(f)$) is characterized by its amplitude and phase that vary with frequency. Although the design of such a network aims at a minimal and constant reflection factor, physical constraints of the design impose that this criterion cannot be maintained over the very wide range of frequencies observed at that node. Thus, the frequencies affected by the reflection coefficient of the output matching network are reflected back into the transistor thereafter recombining with the input signal's frequency components, with spurious elements falling back into baseband as shown in Figure 4.2.

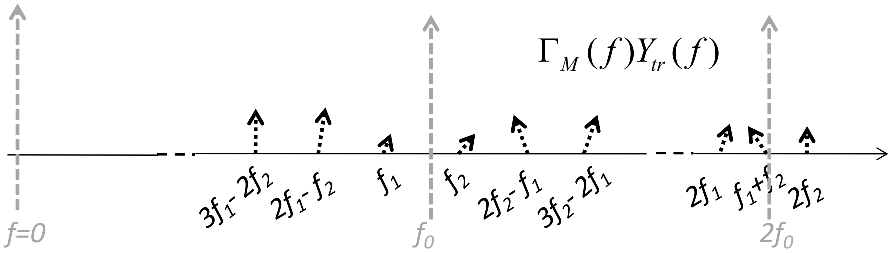


Figure 4.2 - Contribution of the Matching Network to the overall Memory Effects.

4.2.1.2 Biasing Network

$\Gamma_M(f)Y_{tr}(f)$ The biasing network affects the lower range of frequencies. Its reflection coefficient is designed to be minimal across the baseband section. Typically, baseband components range from DC to $(f_2 - f_1)$ and $2(f_2 - f_1)$. Both amplitude and phase of these components are altered by the reflection coefficient $\Gamma_B(f)$, as shown in Figure 4.3.

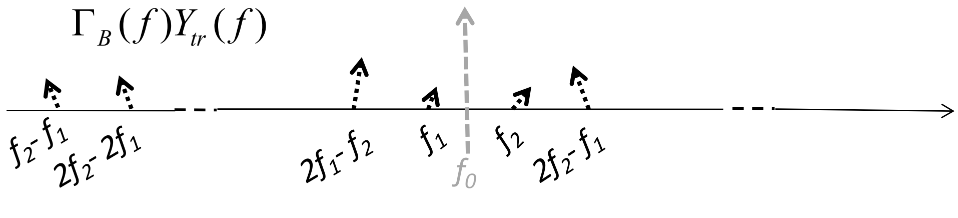


Figure 4.3 - Contribution of the Biasing Network to the overall Inband Memory Effects.

4.2.2 Overall Memory Effects

Although the carrier band is not directly affected by $\Gamma_M(f)$ which effect is observed at higher frequency bands, the remix of frequencies from those harmonic bands relates the high frequency response back into the carrier band. The distorted components are remixed with the carrier band frequencies to form components that fall within this carrier band, i.e. $(3f_2 - f_1) - f_2 = 2f_2 - f_1$.

Similarly, the biasing network $\Gamma_B(f)$ response is most critical at lower frequencies. However, those remix with other frequencies and fall back into the carrier band, i.e. $(f_2 - f_1) + f_2 = 2f_2 - f_1$.

Consequently, the effect of $\Gamma_M(f)$ and $\Gamma_B(f)$ over the carrier band can be seen as an equivalent filter which consists of not only $\Gamma_M(f)$ around f_0, but also $\Gamma_M(f)$ across the subsequent harmonic bands only now centered at f_0 instead of those higher harmonics (i.e. $2f_0$, $3f_0$). Therefore, when considering the RF signal, the contribution of the overall dispersions observed in the signal at the carrier frequency is effectively $\Gamma(f)$, which represents the resultant response of the networks. It is an equivalent filter that affects the carrier band through the recombination of baseband reflected signals and the carrier band frequencies. Considering both reflection effects from biasing and matching networks, their resultant $\Gamma(f)$ is such that

$$\Gamma(f) = \left(\Gamma_M(f), \Gamma_B(f)\right) \tag{4.3}$$

In fact, the biasing and matching networks, including the package effects, are seen by the transistor die as an analog filter which has $\Gamma(f)$ as reflection coefficient and $\left(1-\Gamma(f)\right)$ as a transmission coefficient. Once the overall reflection coefficient is defined, the spectrum of reflected frequencies can be formulated as:

$$\begin{aligned} Y_{\mathrm{R}}(f) &= \Gamma(f).Y_{tr}(f) \\ &= \Gamma(f_1).X(f_1)+\Gamma(f_2).X(f_2)+\Gamma(f_2-f_1).X(f_2-f_1)+\ldots \end{aligned} \tag{4.4}$$

where $Y_{tr}(f)$ is the spectrum of frequencies at the output of the transistor as defined in (4.2), part of which will be reflected as $Y_{\mathrm{R}}(f)$. Figure 4.4 shows the spectrum of the reflected frequencies as distorted in amplitude and phase by the resultant $\Gamma(f)$.

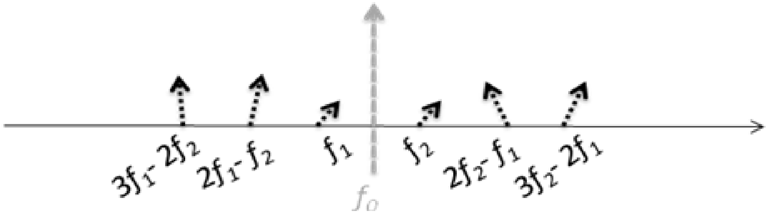


Figure 4.4 – Resulting Inband Reflected Spectrum (memory dispersions)

Defining $\gamma(t)$ as the equivalent time-domain impulse response of $\Gamma(f)$, then the time domain formulation of the reflected signal $Y_R(f)$ is:

$$y_R(t)=\gamma(t)\otimes y(t) \tag{4.5}$$

Hence, $y_R(t)$, which incorporates the frequency dependent response of the matching and biasing networks, can be considered as an extra input that will be remixed with the actual input signal as a result of the transistor's nonlinearity. Therefore, new frequency dependent inter-modulation distortion vectors (e.g. $2f_1-f_2$ as a result of the mix between the reflected f_1-f_2 component and f_1) will appear at the transistor output. These new vectors which originate not only from the input samples but also from its combination with the reflected signal $y_R(t)$ to result in frequency dependent overall inter-modulations distortion vectors, usually used to detect the memory effects. The spectrum of the resulting memory effects is shown in Figure 4.5. The distortions shown are the ones that are around the carrier only or in-band distortions.

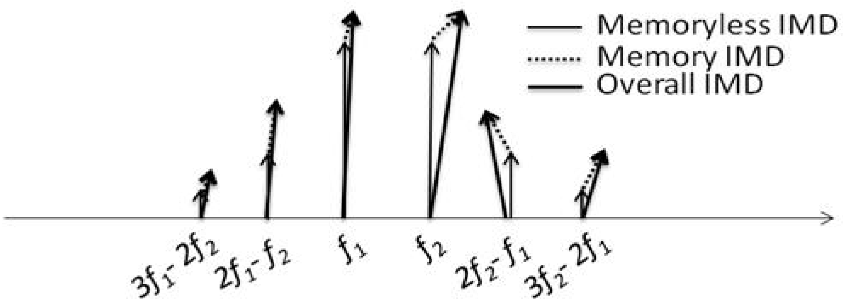


Figure 4.5 - Overall Carrier band Memory Effects

The overall IMDs are the addition of the memoryless distortion with the memory IMD. Although the transistor characteristic response does not involve phase distortions, the resulting signal components are altered both in amplitude and phase according to the resultant between Memoryless and Memory IMDs.

The following section explores the time domain equivalent of such memory effects in an attempt to construct a comprehensive scheme capable of accurately modeling these effects based on the physical aspects explained and the instantaneous samples of input and output signals in the time domain.

4.3 Physical Wideband PA Model Derivation

4.3.1 Time Domain Model Formulation

The continuous function, *γ(t),* which represents the time domain equivalent reflection coefficient of the biasing and matching networks can be transformed into discrete form as:

$$\gamma(n) = \left[\gamma_1 \ \ \gamma_2 \cdots \gamma_M\right]^T \tag{4.6}$$

where M designates the highest significant order of the discrete impulse response, i.e denotes the memory depth considered. Therefore, the time-domain discrete reflected signal can be expressed as:

$$y_R(n) = \sum_{i=1}^{M} \gamma_i \, y(n-i) \tag{4.7}$$

The output signal $y(n)$ formed by the recombination of nonlinearities ($y(n)$) and reflections ($y_R(n)$) can be derived, from the previous equation, as:

$$\begin{aligned} y(n) &= H_{tr}\left(x(n) + \sum_{i=1}^{M} \gamma_i y(n-i) \right) \\ &= H_{tr}\left(x(n) + \gamma_1 y(n-1) + \gamma_2 y(n-2) + \ldots + \gamma_M y(n-M) \right) \end{aligned} \tag{4.8}$$

A more general formulation of *y(n)* as a function of the past samples of *y* and the current input *x* would be:

$$y(n-k) = H_{tr}\left(x(n-k) + \sum_{i=1} \gamma_i y(n-k-i) \right) \qquad \textbf{(4.9)}$$

For simplification purposes, the highest nonlinear order considered in the equation expansions is 5. All the equations can be very easily generalized to higher orders. Let us now define *s(n)* within the argument of the previous equation as:

$$y(n) = H_{tr}\left(x(n) + s(n)\right) \qquad where \quad s(n) = \sum_{j=1}^{M} \gamma_j y(n-j) \qquad \textbf{(4.10)}$$

Thus having the *y(n)* expanded in the form of:

$$\begin{aligned} y(n) = {} & p_1 x(n) + p_2 x^2(n) + p_3 x^3(n) + p_4 x^4(n) + p_5 x^5(n) \\ & + s(n)\left[p_1 + 2p_2 x(n) + 3p_3 x^2(n) + 4p_4 x^3(n) + 5p_5 x^4(n)\right] \\ & + s^2(n)\left[p_2 + 3p_3 x(n) + 6p_4 x^2(n) + 10p_5 x^3(n)\right] \\ & + s^3(n)\left[p_3 + 4p_4 x(n) + 10p_5 x^2(n)\right] \\ & + s^4(n)\left[p_4 + 5p_5 x(n)\right] \\ & + s^5(n)\left[p_5\right] \end{aligned} \qquad \textbf{(4.11)}$$

Notice that the rearrangement the terms in equation (4.11) yields to the following implementable formulation of *y(n)*:

$$\begin{aligned} y(n) = {} & H_{tr}(x(n)) \\ & + \quad s(n)\left[H_{tr}^{(1)}(x(n))\right] \\ & + \frac{1}{2!} s^2(n)\left[H_{tr}^{(2)}(x(n))\right] \\ & + \frac{1}{3!} s^3(n)\left[H_{tr}^{(3)}(x(n))\right] \end{aligned}$$

(4.12)

Where the terms are the derivatives of the function H_{tr} defined as:

$$H_{tr}^{(i)}(x(n)) = {\partial^i H_{tr}(x(n))}\Big/{\partial(x(n))^i} \quad . \tag{4.13}$$

The model diagram that would implement expression (4.12) is outlined in Figure 4.6. The main memoryless nonlinearity of the transistor is represented as H_{tr}, whose output is fed to the overall model output, and to an FIR filter representing the reflections causing memory effects. In fact, since $s(n) = \sum_{j=1}^{M} \gamma_j y(n-j)$, then the FIR's response is essentially $\gamma(n) = [\gamma_1 \ \ \gamma_2 \cdots \gamma_M]^T$, in the time domain , equivalent to the overall reflections $\Gamma(f)$.

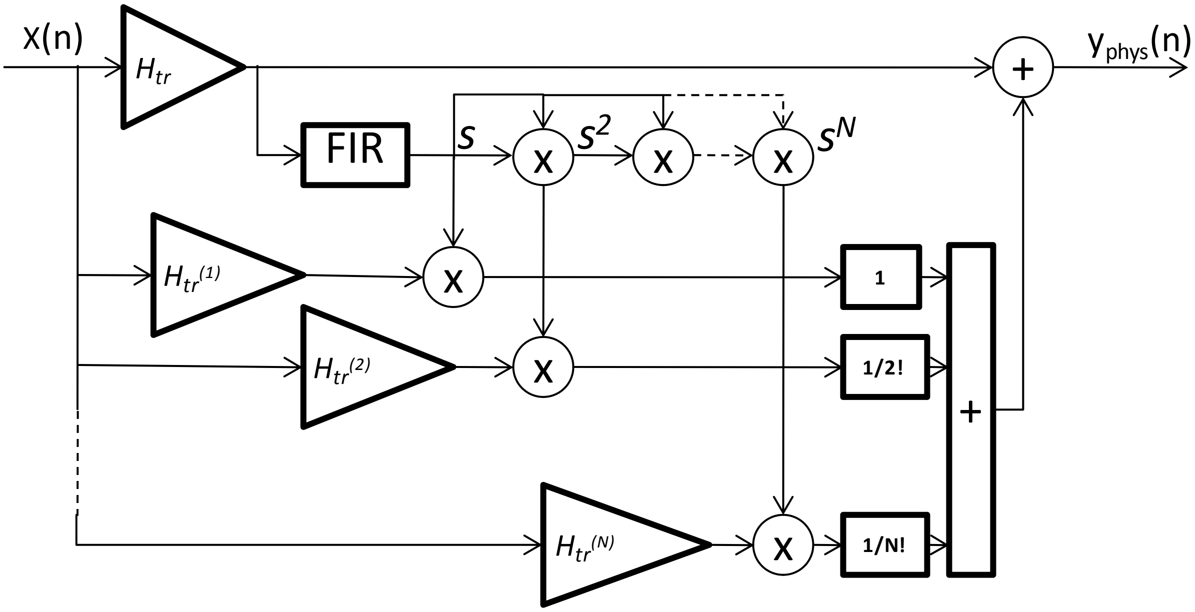


Figure 4.6 - Physical Model Block Diagram where the only unknown parameters are H_{tr} and FIR.

It is extremely interesting to note a few observations about this model, all explainable through physical effects:

1. If the FIR is nulled, in other words the memory effects were not taken into consideration in modeling, the model reduces to H_{tr}, which essentially capture the distortion arising from the memoryless nonlinearity in the system.
2. There is only one source of nonlinearity in the PA. Multiple completely separate polynomials add redundancy to the modeling scheme, as each polynomial represents a source of nonlinearity.
3. As shown clearly in the model block diagram of Figure 4.6, the model extraction process consists of identifying only two blocks: The unknown coefficients of that model are only those of H_{tr} and

FIR, for a total of *(N+M)* coefficients to be extracted. The subsequent derivative blocks use the same coefficients as H_{tr}.

4. Although the model accounts for the reflections and their recombination within the transistor with the current input, it is formulated as a forward model, where no feedback is necessary.
5. The model creates cross-terms at the level where the output of the FIR is mixed with the derivative blocks. As a matter of fact, in a physical circuit, the cross terms are the results of the multiple reflections reappearing at the input of the transistor and mixing , not only with the current signal, but with older reflections as well.

4.4 Validation of results

The previous section presented a physical model that accounts for the reflection mechanism, on which the memory effects. Although that mechanism intuitively translates into a feedback loop, equation analysis and circuit observations translated these complex mechanisms into simply a forward model. This reduces implementation costs and shortens synthesis procedures as the coefficients of such model can be extracted directly. The model is entirely based on the physical properties and mechanisms observed in typical RF PA circuits. *The analysis of these also led to linking the theoretical behavioral close form expression of the MP model to the new physical model's close form expression.* To test the validity of the physical model presented, it is imperative to perform appropriate measurements to test the performance of the model in its capability to capture the characteristic behavior of the RF PA circuit.

4.4.1 Experimental Validation

4.4.2 First Forward Validation: Capturing the PA's Overall Behavior

The first test aims to validate the model's ability to mimic the behavior of the PA circuit under test. This includes the memoryless nonlinearity coupled with memory effects. As shown in Figure 4.7, the PA introduces significant distortions to the clean WCDMA signal applied at its input. A significant imbalance is observed at the output of the DUT suggesting the presence of strong memory effects. For the theory presented in the previous sections to hold, and for the physical model's validity, the output of the physical model should match that PA Output observed. Indeed, when the physical model was synthesized and then driven by the same input signal as the actual PA, the output of that model matched *very closely* the PA's response as shown in the spectra of Figure 4.7.

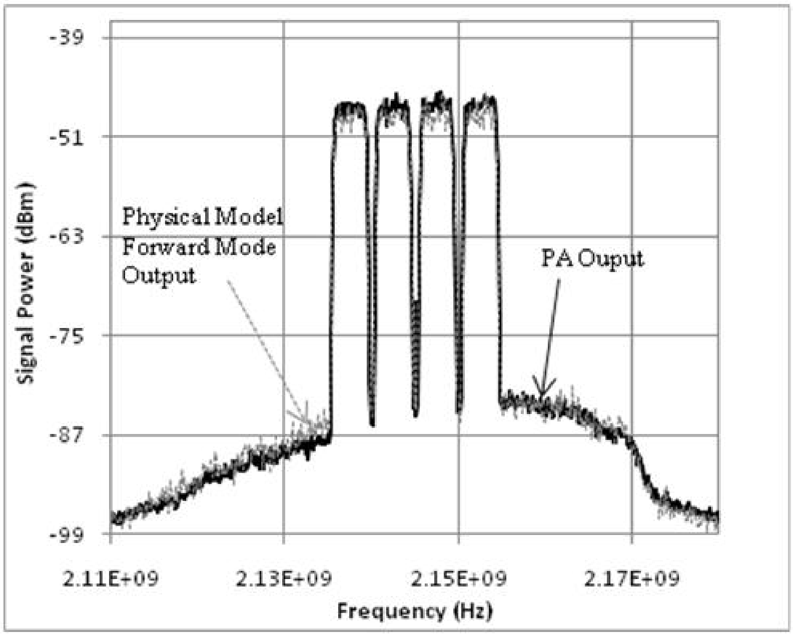


Figure 4.7 - Validation Test 1: Output of the PA when fed with a WCDMA signal, and that of the Physical Model.

The structure of the Physical Model implemented to attain the results of Figure 4.7 consists of a 5th order memoryless polynomial for the H_{tr} nonlinear block of and an FIR filter length of 8. The main advantage of that structure is that the total number of coefficients solved for was 13, (8+5) coefficients.

4.4.3 Second forward Validation: Capturing the PA's Memory Effects

Up until this point, the model's ability to capture the mechanisms of the PA circuit was tested on the combined nonlinearity and memory effects observed in the PA characteristic response. However, the distortions caused by the nonlinearity on its own are most significant in the overall response, and the memory effects represent small scale perturbations to that response. Therefore, a more rigorous test for the modeling accuracy can be achieved by linearizing the PA and the model with a memoryless DPD, and comparing their respective output spectra. First a memoryless DPD is constructed and applied at the input of both PA and model of PA. The remaining signal distortions are the minor ME perturbations, as shown in Figure 4.8.

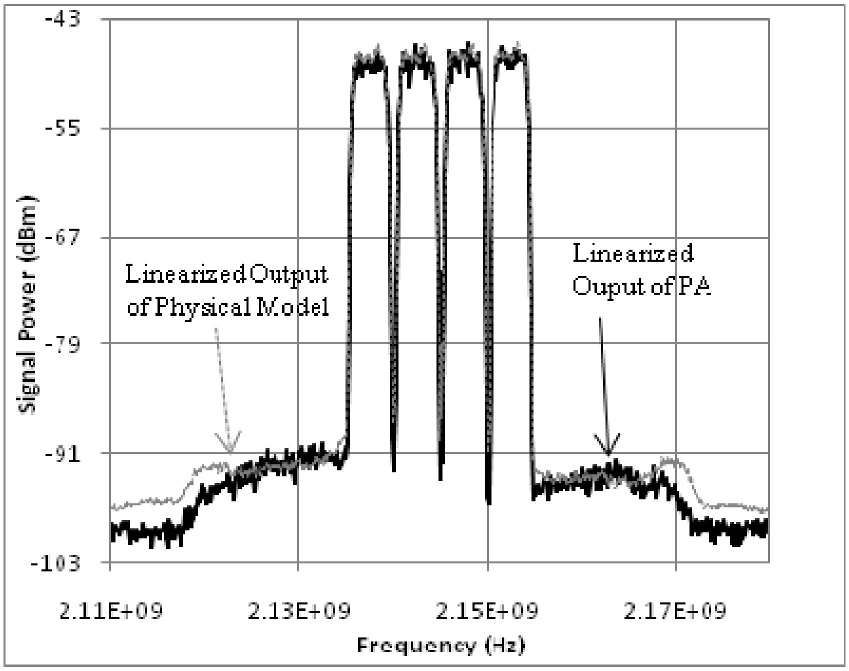


Figure 4.8 - Output of the Actual PA and that of the Model when linearized with a memoryless DPD.

The remaining perturbations are the memory effects of the DUT, which are very accurately captured by the physical model, implying the model's ability to closely reproduce the DUT's behavior.

4.5 Low Complexity of the Physical Model

The complexity of the model extraction algorithm depends mainly on the order of nonlinearity required and the number of coefficients to be extracted. The proposed physical model requires the extraction of only one nonlinear block (H_{tr}) of order N, and a set of FIR filter coefficients (*P* coefficients in total , where *P = (M+1) / 2*, where is the memoryless order of nonlinearity). Therefore, the order of nonlinearity required is N, and the number of coefficients is *(N+P)*. With the implementation scheme suggested, the model is demonstrating excellent modeling capabilities while requiring a minimal number of coefficients.

In fact, to obtain a comparable performance using a M-Polynomial scheme, one would need to identify a structure of *M* nonlinearities (*M* branches), each being a nonlinearity of order *N.* The total number of coefficients required to extract would be *(N*x*M)*, compared to *(N+P)* in the case of the proposed model.

In the reduced MP scheme, as recent papers present it [1], the number of coefficients was reduced by a factor of 2.6. However, the reductions achieved would not yield a requirement of lower than *N+P* coefficients.

Therefore, a significant reduction in the number of coefficients to be extracted is achieved with the physical model as only *N*+*P* instead of *N*x*M* coefficients are required. This reduction significantly reduces implementation costs most importantly in terms of model extraction complexity and stability. If the LSE algorithm is used to extract the model's coefficients, the complexity of that algorithm is proportional to the cube of the number of coefficients to be extracted. The physical model requiring only *N*+*P* instead of *N*x*M* coefficients is very advantageous in that aspect.

Table 4.1 - Coefficient Requirements of M-Polynomial versus Physical Model.

Scheme Structure	**Number of Coefficients**
Full M-Polynomial	NxM
Reduced M-Polynomial	NxM/2.6
Physical Model	N+P

4.6 Linking physical Model to Volterra Series

The Volterra series is a functional Taylor expansion expressed as:

$$\mathbf{y(t)} = \boldsymbol{h}\big(\boldsymbol{u}(\boldsymbol{t}-\boldsymbol{\sigma})\big) = \textstyle\sum_{\boldsymbol{i}=\mathbf{1}}^{\infty} \quad \int_{\mathbf{0}}^{\mathbf{t}} \cdots \int_{\mathbf{0}}^{\mathbf{t}} \boldsymbol{k_i}\,(\boldsymbol{\sigma}_{\mathbf{1}}, \cdots \boldsymbol{\sigma}_{\mathbf{i}})\boldsymbol{u}(\boldsymbol{t}-\boldsymbol{\sigma}_{\mathbf{i}})\boldsymbol{d\sigma}_{\mathbf{1}} \cdots \mathbf{d}\boldsymbol{\sigma}_{\mathbf{i}} \qquad \textbf{(4.14)}$$

where $k_i(\sigma_1 \cdots \sigma_i)$ is the i^{th} order kernel defined as:

$$\boldsymbol{k_i}(\boldsymbol{\sigma}_{\mathbf{1}}, \cdots \boldsymbol{\sigma}_{\mathbf{i}}) = {}^{\boldsymbol{\partial}^{\mathbf{t}}\,\mathbf{y(t)}}\Big/_{\boldsymbol{\partial}\mathbf{u}(\mathbf{t}-\boldsymbol{\sigma}_{\mathbf{1}}) \cdots \boldsymbol{\partial}\mathbf{u}(\mathbf{t}-\boldsymbol{\sigma}_{\mathbf{i}})} \qquad \textbf{(4.15)}$$

which is similar to the formulation of the model in equations (4.7) and (4.8).

Moreover, considering that the formulation of the physical model included both frequency and time domain considerations, let us examine a similar theoretical model, the power Series, as defined in section (Power Series2.2.3.1.1). The power series is a more powerful derivation of the Volterra series, defined as in equation (2.7) as:

(4.16)

The problem with using the power series is the identification of its coefficients. however, they are similar to those expressed in the early problem formulation of the physical model, which was stated in equation (4.5) as:

$$y(n) = H_{tr}\left(x\left(n\right) + s(n)\right) \qquad where \quad s(n) = \sum_{j=1}^{M} \gamma_j y\left(n - j\right) \tag{4.17}$$

which carries the same terms as the terms in the power series formulation. Therefore, the physical model formulation filled the gap between the generic theoretical formulation of nonideal systems with memory and the physical characteristics of that nonideal behavior. Coefficients of the generic schemes are correlated when modeling an RF PA. There is one source of nonlinearity in the system, and that is the transistor memoryless function. The rest of the nonideal behavior stems from network reflections. This implies a correlation between the coefficients of the theoretical formulations, which is shown by the physical model. Also, the terms of the theoretical formulation are given a physical explanation by interpreting the physical mechanisms behind their presence.

Therefore, the newly developed close-form expression of $y(n)$, to the authors' knowledge, for the first time creates a valuable bridge between the behavioral modeling and the physical properties of the circuit that was missing in the previously published behavioral models. It offers an extra dimension of analysis that complements the behavior modeling. Indeed, the numerous parameters of theoretical models can now be extracted more easily, and used to relate the overall behavior of the device under test with its topology and blocks' characteristics (matching and biasing networks).

Chapter 5

Systematic Complexity Reduction of the Memory Polynomial based on the Physical Formulation of Memory Effects

5.1 Memory effects analytical formulation

The previous chapter presented an explanation for the mechanism of memory effects, based on physical circuit properties of RF PAs. Subsequently, a close-form formulation of MEs was derived based on physical properties of a typical PA circuit. This chapter will build on that theory to first define the concept behind the formulation in a higher level of abstraction, where the networks and their reflections are considered as finite impulse response (FIR) filters. The formulation will then be applied and expanded into linking it with the very well established M-Polynomial behavioral model.

5.1.1 Theoretical Close-Form Physical Expression of Memory Effects

A concurrent model is that presented in [20] and shown in Figure 5.1. It is a model for the RF PA circuitry equivalent to what was described in the previous chapter. The close-form expression encompasses a static nonlinear function *G*, to account for the memoryless nonlinear behavior of the PA, along with FIR_{in}, FIR_{out},and FIR_{fb} filters to capture the frequency-dependent mechanism behind the memory effects. FIR_{in} and FIR_{out} have a negligible contribution to the PA memory effects as the signal bandwidth is kept relatively small enough compared to the carrier frequency. Hence, only the feedback filter, representing the total reflections from biasing and matching networks, is retained in the simplified model to be used as depicted in Figure 5.2.

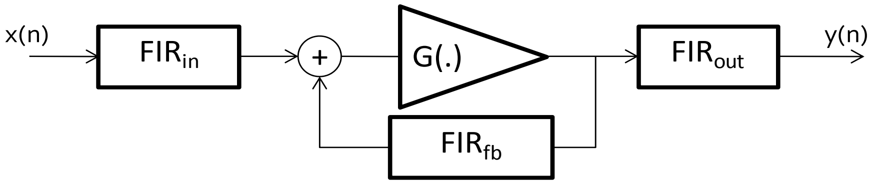


Figure 5.1 - Physical Model of the PA

Figure 5.2 - Simplified Physical Model of the PA

The static nonlinear behavior of the transistor on its own is a memoryless nonlinear function, which as discussed in earlier sections, is characterized by *G* defined as:

$$G(x(n)) = p_1 x(n) + p_2 x(n)^2 + ... + p_N x(n)^N \tag{5.1}$$

where *N* is the memoryless order of nonlinearity of the transistor. Consequently, as shown in Figure 5.2, the discrete output signal *y(n)* of the RF PA circuit is expressed in terms of the discrete input *x(n)* and its past samples as follows:

$$y(n) = G(x(n) + \sum_{j}^{M} \gamma_j y(n-j)) \tag{5.2}$$

where γ_j is the j^{th} instance of the discrete time domain representation of the feedback FIR filter, $\mathrm{FIR}_{\mathrm{fb}}$, that captures the dispersive behavior of the biasing and matching networks at the signal's envelope and harmonic frequencies.

Expanding the previous expression of *y(n)*, the following formulation is obtained:

$$\begin{aligned} y(n) &= G(x(n) + \sum_{j}^{M} \gamma_j y(n-j)) \\ &= p_1 \left[x(n) + \sum_{j}^{M} \gamma_j y(n-j) \right] + p_2 \left[\left((x(n) + \sum_{j}^{M} \gamma_j y(n-j)) \right)^2 \right] \\ &+ ... + p_N \left[\left((x(n) + \sum_{j}^{M} \gamma_j y(n-j)) \right)^{N-1} \right] \end{aligned} \tag{5.3}$$

Equation (5.3) suggests that if the memory effects are not taken into account, i.e. γ_j=0, the model reduces to memoryless nonlinear transistor expression *G*.

5.1.2 Elimination of Multiple Reflections Terms

Up to this point, the derivations are no different from the ones in the previous chapter. However, it is important at this stage to examine the summation of reflected signals in equation (5.3) which are the terms expressed as: . Expanding each term of that summation expression yields to the following expression in terms of the input signal *x(n)*:

$$\begin{aligned}\gamma_j y(n-j) = {} & p_1\gamma_j x(n-j) + p_1\sum_{k=1}\gamma_j\gamma_k y(n-j-k) \\ & + p_2\gamma_j x^2(n-j) + p_2\gamma_j\left(\sum_{k=1}\gamma_k y(n-j-k)\right)^2 \\ & +2p_2\gamma_j x(n-j).\sum_{k=1}\gamma_k y(n-j-k) \\ & +p_3\gamma_j x^3(n-j) + p_3\gamma_j\left(\sum_{j=1}\gamma_j y(n-j-k)\right)^3 \\ & +3p_3\gamma_j x^2(n-j)\sum_{j=1}\gamma_k y(n-j-k) \\ & +3p_3\gamma_j x(n-j)\left(\sum_{k=1}\gamma_k y(n-j-k)\right)^2\end{aligned} \qquad (5.4)$$

In the previous equation, terms similar to $\gamma_j\left(\sum_{k=1}\gamma_k y(n-j-k)\right)^2$ and $\gamma_j x(n-j)\sum_{k=1}\gamma_k y(n-j-k)$ can be rewritten as:

$$\begin{cases}\gamma_j\left(\sum_{k=1}\gamma_k y(n-j-k)\right)^2 = \left(\sum_{k=1}\sqrt{\gamma_j}\gamma_k y(n-j-k)\right)^2 \\ \gamma_j x(n-j).\sum_{k=1}\gamma_k y(n-j-k) = \sum_{k=1}\gamma_j\gamma_k x(n-j)y(n-j-k)\end{cases} \qquad \mathbf{(5.5)}$$

Those terms represent the 2nd , 3rd and higher stage reflections, which become minimal when taken to a total order that is higher than one, in other words when past samples of the input are multiplied by a factor of $\left(\gamma_j^a\gamma_k^b\right)$ where $a+b>1$. Such terms can consequently be omitted from the calculations and the previous expression of $\gamma_j y(n-j)$ is simplified to:

$$\begin{aligned}\gamma_j y(n-j) = {} & p_1\gamma_j x(n-j) + p_2\gamma_j x^2(n-j) \\ & + p_3\gamma_j x^3(n-j)\end{aligned} \qquad \mathbf{(5.6)}$$

Subsequently, expression (5.4) can be rewritten as:

$$
\begin{aligned}
y(n) &= G(x(n) + \sum_{j}^{M} \gamma_j \, y(n-j)) \\
&= G\left(x(n) + \gamma_j \sum_{j=1} G\left(x(n-j)\right) \right) \\
&= \left(p_1 x(n) + p_2 x(n)\left|x(n)\right| + \ldots + p_N \, x(n)\left|x(n)\right|^{N-1} \right) \\
&\quad + \left(p_1 + 2p_2 x(n) + \ldots + Np_N \, x(n)\left|x(n)\right|^{N-2} \right) \sum_{j}^{M} \gamma_j G\left(x(n-j)\right)
\end{aligned}
\tag{5.7}
$$

5.1.3 Elimination of Cross-terms

The previous expression of *y(n)* in equation (5.7) suggests the need for cross-terms to be present in the modeling scheme as *x(n)x(n-j)* elements are observed. However, in most RF PA linearization setups, the discrete input signal *x(n)* is relatively well oversampled compared to Nyquist theoretical minimum sampling limit. In other terms, the input signal is usually oversampled by a factor higher than 2. Therefore, for this type of discrete signals, it is safe to assume that:

$$
x(n-i)x(n-j) = x^2(n - floor(\frac{i+j}{2}))
\tag{5.8}
$$

This assumption is safe up to a certain limit of separation in the input signal samples, i.e. *(i-j)*. Applying this assumption within this limit to the close-form expression *y(n)* detailed in the previous section, all the cross-terms are assimilated into pure terms and the close form expression becomes:

$$
y(n) = \sum_{i=0}^{M} \sum_{j=1}^{P_k} a_{ij} x(n-i)\left|x(n-i)\right|^{j-1}
\tag{5.9}
$$

Where the coefficients a_{ij} are function of the memoryless polynomial coefficients (p_i) and the discrete coefficients of the feedback FIR filter coefficient (γ_j) of expression (1), and P_k is an integer that varies between 0 and 2N-1 as will be explained later. The expression of these coefficients in terms of p_i and γ_j differs depending on the order of memoryless nonlinearity (*N*) and memory depth (*M*) specified. Varying *N* and *M* led to observing two general key trends common to all values of a_{ij} as will be detailed in the next two sections. As an illustration of equation (5.9), Table I suggests the detailed expressions of the close-form expression coefficients a_{ij} and their repartition for the particular case where *N*=5 and *M*=8. It is worth mentioning that although the initial memoryless nonlinearity was equal to 5, Table I reveals an order of nonlinearity equal to 9 in the close form expression. Hence, hereafter, the nonlinear order will be designated with P and will be equal to the order of the nonlinearity obtained when memory effects were included in the model, i.e. P is equal to 9 when N is set to 5.

The examination of equation (5.9) and the expressions of a_{ij} coefficients will serve later in the development of a systematic approach for the complexity reduction of the M-Polynomial model especially when used to construct a DPD as will be detailed in the remainder of this paper.

5.2 Similarity with M-Polynomial Formulation

The M-Polynomial is a comprehensive modeling scheme, introduced by Kim *et al.* [2], which is derived from the Volterra model so that fewer coefficients are required while preserving its capability of capturing memory and nonlinearity effects. It is defined as:

$$y_{MP}(n) = \sum_{k=0}^{K} \sum_{p=1}^{P} c_{p,k} \left| x(n-k) \right|^{p-1} x(n-k) \qquad \textbf{(5.10)}$$

where *x(n)* and $y_{MP}(n)$ are the complex envelope signals at the input and output, respectively. K is the memory length of the system, equivalent to the number of polynomial branches. *P* represents the polynomial order of the branches and $c_{p,k}$ designates the p^{th} polynomial coefficients of the k^{th} branch. The M-Polynomial scheme has proven PA modeling and linearization performance although it ignores the cross-terms of the original Volterra series from which it was derived.

Comparing $y_{MP}(n)$ expression (5.10) with the expression (5.9) *y(n),* it is clear that the output in both is affected by similar combinations of input samples. Therefore, setting both equations equal will yield the following relationship between the coefficients of each:

$$a_{ij} = c_{p,k} \qquad \textbf{(5.11)}$$

Hence, a direct relationship is established between the coefficients of the physical formulation and the ones of the M-Polynomial scheme. The following section will further analyze the physical formulation to understand the implications of that direct relationship on the M-Polynomial scheme.

Table 5.1 - Expressions of the a_{ij} coefficients for N=5 and M=8 (P=9, M=8).

j \ i	11	2	3	4	5	6	7	8	9
poly1	p_1	$p_2(2p_1$ ${}^{\gamma_1}+1)$	$p_3+3p_1p_3({}^{\gamma_1}+{}^{\gamma_3}$ $+{}^{\gamma_2})+2p_2^{2\gamma_1}$	$p_4+4p_1p_4({}^{\gamma_3}+{}^{\gamma_2}$ $+{}^{\gamma_1})+5p_3p_2^{\gamma_1}$	$p_5+5p_5p_1({}^{\gamma_5}+{}^{\gamma_7}+$ ${}^{\gamma_6}+{}^{\gamma_4}+{}^{\gamma_3}+{}^{\gamma_2}+{}^{\gamma_1}$ $)+4p_2p_4^{\gamma_2}+4p_2$ $p_4^{\gamma_3}+3p_3^{2\gamma_1}$ $+6p_2p_4^{\gamma_1}$	$7p_2p_5^{\gamma_1}$ $+5p_2p_5({}^{\gamma_3}+{}^{\gamma_2}$ $)+7p_4p_3^{\gamma_1}$	$5p_3p_5^{\gamma_2}+5p_3p_5^{\gamma_3}$ $+4^{\gamma_1}p_4^2+8p_3p_5^{\gamma_1}$	$9p_4p_5^{\gamma_1}$	$5p_5^{2\gamma_1}$
poly2	p_1^2 γ_1	$p_2p_1(2$ ${}^{\gamma_3}+{}^{\gamma_1}+2$ ${}^{\gamma_2})$	$2p_2^{2\gamma_3}+3p_1p_3^{\gamma_5}$ $+3p_3p_1^{\gamma_4}+p_1p_3$ ${}^{\gamma_1}+2p_2^{2\gamma_2}$ $+3p_1p_3^{\gamma_6}$ $+3p_1p_3^{\gamma_7}$	$p_4p_1^{\gamma_1}+4p_1p_4^{\gamma_4}$ $+4p_1p_4^{\gamma_5}+5^{\gamma_2}$ $p_3\ p_2+3^{\gamma_3}$ $p_3p_2+4p_1p_4^{\gamma_7}$ $+4p_4p_1^{\gamma_6}$	$4^{\gamma_4}p_2p_4+4^{\gamma_5}$ $p_4p_2+3^{\gamma_3}$ $p_3^2+p_1p_5^{\gamma_1}$ $+2p_2p_4^{\gamma_2}+3^{\gamma_2}p_3^2$	$5\ p_2p_5^{\gamma_4}+5p_2p_5$ ${}^{\gamma_5}+7p_3p_4^{\gamma_2}+5$ $p_2p_5^{\gamma_7}+7^{\gamma_3}\ {}^{\gamma_2}$ $p_4p_3+2p_2p_5$ $+5p_2p_5^{\gamma_6}$	$4p_4^{2\gamma_3}+5p_5p_3^{\gamma_4}$ $+3p_3p_5^{\gamma_3}+5p_3p_5$ ${}^{\gamma_5}+3p_3p_5^{\gamma_2}+4p_4^2$ ${}^{\gamma_2}$	$9p_5p_4($ ${}^{\gamma_2}+{}^{\gamma_3})$	$5p_5^2($ ${}^{\gamma_2}+{}^{\gamma_3})$
poly3	p_1^2 γ_2	$p_2p_1({}^{\gamma_2}$ $+2^{\gamma_5}+2$ ${}^{\gamma_4})$	$p_1p_3^{\gamma_2}+2^{\gamma_5}$ $p_2^2+2^{\gamma_4}p_2^2$	$3^{\gamma_5}p_3p_2+p_4p_1^{\gamma_2}$ $+3^{\gamma_4}p_3p_2+2^{\gamma_3}$ p_3p_2	$p_5p_1^{\gamma_2}+3^{\gamma_5}p_3^2+3$ ${}^{\gamma_4}p_3^2+2p_2p_4^{\gamma_3}$ $+4p_2p_4^{\gamma_7}+4p_2p_4$ ${}^{\gamma_6}$	$7p_3p_4^{\gamma_5}+7p_3p_4$ ${}^{\gamma_4}+2\ p_2p_5^{\gamma_3}$	$4^{\gamma_5}p_4^2+4^{\gamma_4}$ $p_4^2+5p_3^{\gamma_7}p_5+5p_3$ ${}^{\gamma_6}p_5+3p_3p_5^{\gamma_5}$ $+3p_5p_3^{\gamma_4}$	$9p_5p_4($ ${}^{\gamma_4}+{}^{\gamma_5})$	$5p_5^2($ ${}^{\gamma_4}+{}^{\gamma_5})$
poly4	p_1^2 γ_3	$p_1p_2({}^{\gamma_3}$ $+2^{\gamma_7}+2$ ${}^{\gamma_6})$	$p_3p_1^{\gamma_3}+2^{\gamma_6}p_2^2$	$p_1p_4^{\gamma_3}+3p_2p_3^{\gamma_7}$ $+3p_2p_3^{\gamma_6}+2^{\gamma_5}$ $p_3p_2+2^{\gamma_4}p_3p_2$	$p_1p_5\ {}^{\gamma_3}+2\ {}^{\gamma_4}$ $p_2p_4+2^{\gamma_5}\ p_4p_{2+}3$ ${}^{\gamma_6}p_3^2+3^{\gamma_7}p_3^2$	$2p_2p_5^{\gamma_4}+4p_3p_4$ ${}^{\gamma_7}+2\ p_2p_5^{\gamma_5}+7*$ ${}^{\gamma_6}p_3p_4$	$3p_3^{\gamma_6}p_5+4p_4^{2\gamma_6}$ $+4p_4^{2\gamma_7}$	$p_4p_5(5$ ${}^{\gamma_7}+9^{\gamma_6})$	$5p_5^2($ ${}^{\gamma_6}+{}^{\gamma_7})$
poly5	p_1^2 γ_4	$p_1\ p_2^{\gamma_4}$	$p_3p_1^{\gamma_4}+2p_2^{2\gamma_7}$	$p_1p_4^{\gamma_4}+2p_2p_3^{\gamma_6}$	$2p_2p_4^{\gamma_6}+p_1p_5^{\gamma_4}$	$3p_3p_4^{\gamma_7}+2p_2p_5$ ${}^{\gamma_6}$	$3p_3^{\gamma_7}p_5$	$4p_4p_5^{\gamma_7}$	0
poly6	p_1^2 γ_5	$p_1p_2^{\gamma_5}$	$p_1p_3^{\gamma_5}$	$2p_2p_3^{\gamma_7}+p_1p_4^{\gamma_5}$	$p_1p_5^{\gamma_5}+2p_2p_4^{\gamma_7}$	$2p_2p_5^{\gamma_7}$	0	0	0
poly7	p_1^2 γ_6	$p_1p_2^{\gamma_6}$	$p_1p_3^{\gamma_6}$	$p_4p_1^{\gamma_6}$	$p_1p_5^{\gamma_6}$	0	0	0	0
poly8	p_1^2	p_1p_2	p_1p_3	p_1p_4	p_1p_5	0	0	0	0

5.3 Analysis of M-Polynomial Formulation

In this section, a systematic complexity reduction of the M-Polynomial scheme will be established based on the analytical formulation of the previously presented ME formulation. It will also be proven that the actual PA's memoryless nonlinearity order is not the order of the 1st polynomial in an M-Polynomial scheme, but a direct relationship between these two orders will be determined hereafter.

It is worth mentioning that the reformulation, analysis and complexity reductions will be tested and validated experimentally in reverse mode (DPD). Indeed, assessing the capability of the reduced complexity M-Polynomial scheme in constructing the accurate DPD for the DUT was preferred to conventional validation approach (forward mode) that relies on the capacity of predicting its output spectrum as it offers a rigorous modeling accuracy metric. From this point on, the spectra in figures will be designated by the '*P*x*M'* notation where *P* corresponds to the nonlinearity order and *M* to the memory depth of the linearizing scheme.

5.3.1 Systematic M-Polynomial Complexity Reduction

Based on Table 5.1, the first key trend that was observed in the reduction of the *y(n)* expression is that the coefficients of high order terms of the subsequent polynomials are equal to zero. In other words, modeling the effect of older input samples on *y(n)* requires a lower nonlinearity order than the 1st polynomial of the M-Polynomial structure. Therefore, a reduction in the order of the subsequent polynomials can be applied to the M-Polynomial scheme without any loss of its modeling or DPD capability. Each Cell of Table I represents a M-Polynomial coefficient, where each row designates a polynomial branch of the scheme. In other words, the cells of the 1st row in table are the coefficients of the 1st polynomial, Cell 1 being the coefficient of the term *x(n)*, cell 2 being the coefficient of the term $x(n)^2$ up to Cell 9, coefficient of $x(n)^{9}$. Same for the subsequent rows where for example cells of row 2 are the coefficients of *x(n-1)*, $x(n-1)^2$, and $x(n-1)^9$. The order of the 1st four polynomials remained unchanged; however, the order of the fifth to eighth was reduced from 9th order to successively 8th, 6th, 5th and 5th, by having null coefficients in the high orders of the subsequent branches of the M-Polynomial. The scheme will be denoted as 99998655, the numbers being in reference to the order of each branch.

It is essential to note that the reduction of the scheme eliminates only redundant terms, keeping the scheme's full modeling capabilities. In order to validate this fact, the first step was to find the M-Polynomial-DPD scheme that was most efficient in linearizing the lineup behavior. That is to maintain the best ACPR attained with the least number of coefficients. A high order, large number of branches M-Polynomial is used as a starting point. The complexity of this initial DPD is lowered by reducing the order of the M-Polynomial and the number of branches. It was found that a 9th order MP-DPD with memory depth of 8 was the most efficient linearization scheme when using all the M-Polynomial coefficients. A 7th order DPD did not achieve the same linearization performance, and orders higher than 9 do not improve the spectrum linearized with a 9th order and designated in Figure 5.3 as 'Full MP-DPD 9x8'. Similarly, it was found that a 8th as memory depth is a good configuration for the M-Polynomial as 9th order does not improve its linearization capability and 7th order somewhat compromises it.

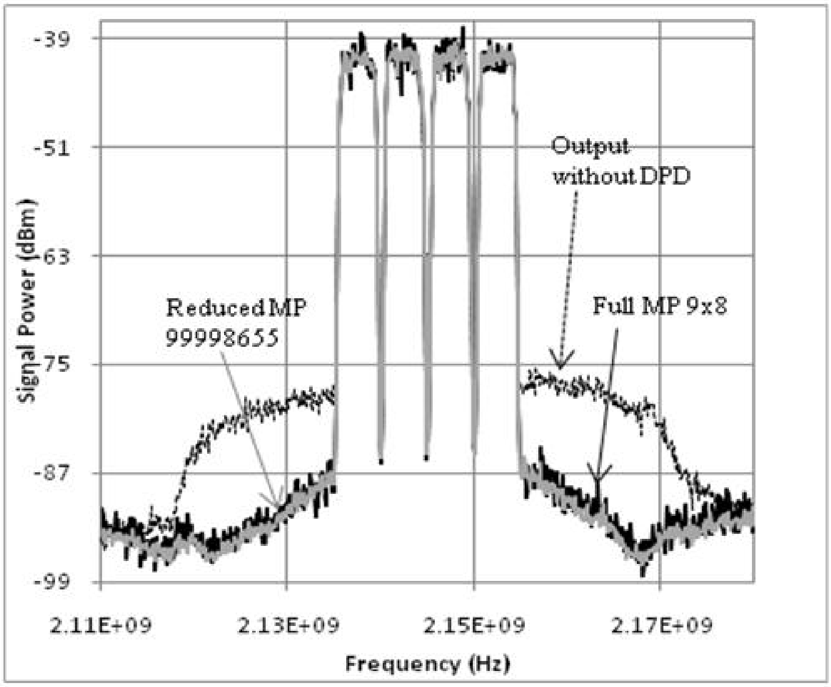


Figure 5.3 - Spectra of DUT response without any DPD applied (Output without DPD), and linearized with a full polynomial scheme of order 9 and memory length of 8 (Full DPD MP 9x8) and the reduced MP scheme 99998655.

As shown in the spectra of the Fig. 2 and the ACPR values of Table 5.2, the linearization performance is maintained when the higher orders of subsequent polynomial branches are replaced with zeroes according to the scheme suggested by the cross-term reduction of the physical equation. Hence, so far the number of coefficients was reduced by 12 without any degradation in linearization performance.

Table 5.2 – First Step of M-Polynomial DPD Order Reduction

DPD Used	ACPR (dBc)			Conditioning of *A*	Total # of Coeff
	1st OB	2ndOB	3rdOB		
Full MP 9x8	-47	-50.7	-57.7	1.405e10	72
Reduced MP 99998655	-47.6	-51.3	-58.3	1.186e10	60

5.3.1.1 Elimination of the 5th order Terms in M-Polynomial

To further reduce the complexity of the M-Polynomial scheme without trading any of its linearization capability, the contribution of the 5th order nonlinearities, p_5, of the transistor static nonlinearity to the memory effects is investigated here.

When considering the effect of the 5th order nonlinearities on the envelope of the signals being modeled, it can be said that although it might be present through recombination, the effect is minimal due to scaling factors and the multiple recombination needed for the 5th order to reappear in the fundamental band. Based on that, the coefficients of Table 5.1 were further reduced by nulling p_5 in all the subsequent branches of the MP while keeping it in the 1st branch as it is essential in capturing the memoryless part. This step led to a new DPD structure defined as: 97776544.

The resulting DPD is implemented with the indicated orders, and compared with the one from the full M-Polynomial scheme as shown in Figure 5.3. Table 5.3 gives the ACPR values for the 3 M-Polynomial configurations.

Table 5.3 - Second Step of M-Polynomial Order Reduction

DPD Used	ACPR (dBc)			Conditioning of *A*	Total # of Coeff
	1st OB	2ndOB	3rdOB		
Full MP 9x8	-47	-50.7	-57.7	1.405e10	72

Reduced MP 99998655	-47.6	-51.3	-58.3	1.186e10	60
Reduced MP 97776544	-47.6	-50.7	-58	1.175e10	49

5.3.1.2 Even Order Elimination

As previously discussed, when modeling the envelope of the RF signal, only odd orders are needed to capture the effects of both even and odd nonlinearity products that fall within the pass-band. As noted in Figure 5.4, the spectrum of the DPD does not deteriorate when the even orders are removed from the reduced scheme. Not only is ACPR maintained to the levels of a full MP-DPD scheme, but the number of coefficients is now reduced by a factor of 2.6, as shown in Table 5.4.

Table 5.4 - Third Step of M-Polynomial DPD Order Reduction.

DPD Used	**ACPR (dBc)**			**Conditioning of *A***	**Total # of Coeff**
	1st OB	**2ndOB**	**3rdOB**		
Full MP 9x8	-47	-50.7	-57.7	1.405e10	72
Reduced MP 97776544	-47.6	-50.7	-58	1.175e10	49
Reduced MP 97776544 Odd Only	-47.5	-50.6	-57.4	2.853e7	27

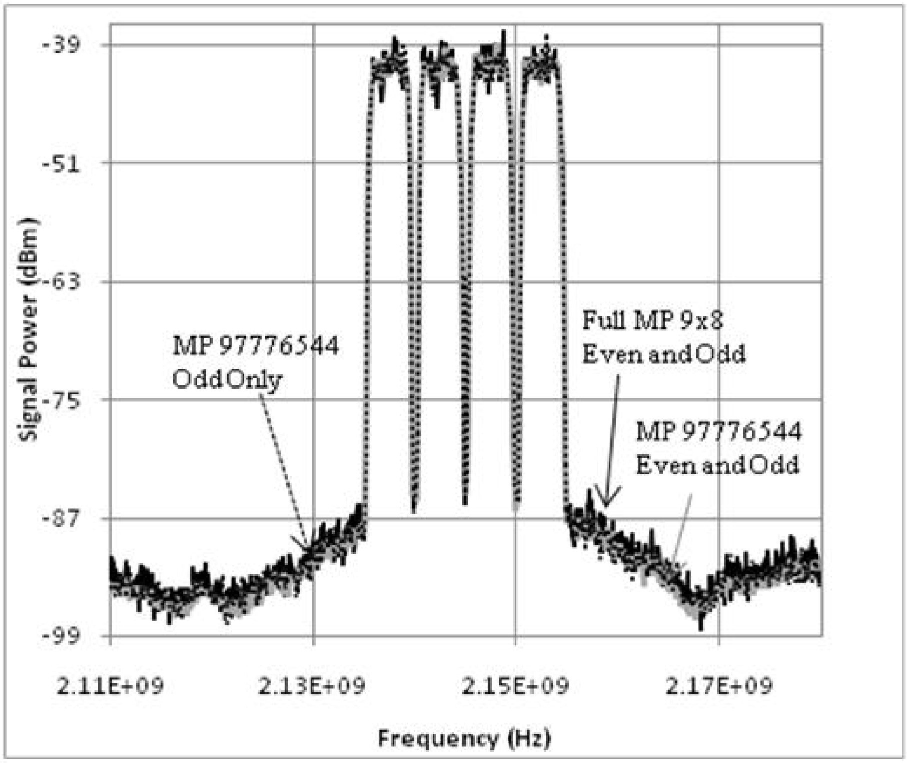


Figure 5.4 – Response of DUT linearized with full polynomial scheme: order 9 and memory length of 8 (Full MP 9x8), and the reduced MP schemes 97776544 after even order elimination

5.3.1.3 Overall Complexity Reduction of the MP scheme

The previous three subsections presented reduction and simplification strategies of the physical model which proved that in the full M-Polynomial scheme; some coefficients are redundant for the modeling of an RF PA behavior. In fact, as shown in Table 5.4, the complexity of the M-Polynomial was significantly reduced: the number of coefficients required by the scheme was reduced from 72 to 27 coefficients, or a factor of 2.6. Also, the stability of the model extraction is improved as the conditioning number of the scheme is reduced by 3 orders of magnitude.

These reductions, all based on circuit and device properties of the DUT, were achieved without any loss in the scheme's linearization performance, as shown by the spectra of Figure 5.4 and ACPRs of Table 5.4 presented in the previous sections.

To evaluate the effect of the DPD on the overall system's operational costs, the PAE was measured in the case where no DPD was applied to the signal and where the complexity reduced M-Polynomial

DPD was used to linearize the PA lineup. Applying an M-Polynomial DPD achieved a PAE of 32%, which is a significant improvement on the 15% PAE achieved where no DPD was applied.

Table 5.5 combines all the reduction steps, including the even order omission compared to the reduction presented in Table 5.1:

Table 5.5 – Final Reduction of the M-Polynomial for N=5, M=8 (equivalent to P=9, M=8).

i j	*1*	*3*	*5*	*7*	*9*
Poly1	p_1	$p_3+3p_1p_3(\gamma_1+\gamma_3+\gamma_2)+2p_2^2\gamma_1$	$p_5+5p_5p_1(\gamma_5+\gamma_7+\gamma_6+\gamma_4+\gamma_3+\gamma_2+\gamma_1)+4p_2p_4\gamma_2+4p_2p_4\gamma_3+3p_3^2\gamma_1+6p_2p_4\gamma_1$	$5p_3p_5\gamma_2+5p_3p_5\gamma_3+4\gamma_1 p_4^2+8p_3p_5\gamma_1$	$5p_5^2\gamma_1$
Poly2	$p_1^2\gamma_1$	$2p_2^2\gamma_3+3p_1p_3\gamma_5+3p_3p_1\gamma_4+p_1p_3\gamma_1+2p_2^2\gamma_2+3p_1p_3\gamma_6+3p_1p_3\gamma_7$	$4\gamma_4 p_2p_4+4\gamma_5 p_4p_2+3\gamma_3 p_3^2+p_1p_5\gamma_1+2p_2p_4\gamma_2+3\gamma_2 p_3^2$	$4p_4^2\gamma_3+5p_5p_3\gamma_4+3p_3p_5\gamma_3+5p_3p_5\gamma_5+3p_3p_5\gamma_2+4p_4^2\gamma_2$	
Poly3	$p_1^2\gamma_2$	$p_1p_3\gamma_2+2\gamma_5 p_2^2+2\gamma_4 p_2^2$	$p_5p_1\gamma_2+3\gamma_5 p_3^2+3\gamma_4 p_3^2+2p_2p_4\gamma_3+4p_2p_4\gamma_7+4p_2p_4\gamma_6$	$4\gamma_5 p_4^2+4\gamma_4 p_4^2+5p_3\gamma_7 p_5+5p_3\gamma_6 p_5+3p_3p_5\gamma_5+3p_5p_3\gamma_4$	
Poly4	$p_1^2\gamma_3$	$p_3p_1\gamma_3+2\gamma_6 p_2^2$	$p_1p_5\gamma_3+2\gamma_4 p_2p_4+2\gamma_5 p_4p_2+3\gamma_6 p_3^2+3\gamma_7 p_3^2$	$3p_3\gamma_6 p_5+4p_4^2\gamma_6+4p_4^2\gamma_7$	
Poly5	$p_1^2\gamma_4$	$p_3p_1\gamma_4+2p_2^2\gamma_7$	$2p_2p_4\gamma_6+p_1p_5\gamma_4$		
Poly6	$p_1^2\gamma_5$	$p_1p_3\gamma_5$	$p_1p_5\gamma_5+2p_2p_4\gamma_7$		
Poly7	$p_1^2\gamma_6$	$p_1p_3\gamma_6$			
Poly8	$p_1^2\gamma_7$	$p_1p_3\gamma_7$			

5.3.2 Memoryless Order vs. Order of MP scheme

It is widely believed that the memoryless nonlinear polynomial, that is necessary to model the memoryless nonlinear part of the behavior of the PA, is equivalent to the first polynomial of the M-Polynomial scheme. However, the expressions of coefficients obtained from the physical model do not reflect that fact.

For the purpose of illustrating the concept, and since a minimum order of 9 was required for the MP-DPD as shown in the previous subsections, let us consider a 9x8 M-Polynomial structure, and examine the expression of its coefficients shown in Table 5.1. In the previous subsections, it was found that a minimum nonlinearity order of 9 in the first branch of the M-Polynomial was necessary in order to capture both nonlinearity and memory effects of the PA. However, if the memory effects need not be accounted for, the reflection filter coefficients (γ_j) are nulled. Its implication on the M-Polynomial coefficients can be seen through the expression of the coefficients presented in Table 5.1. Each row of that table represents the coefficients of a branch of the M-Polynomial, where columns indicate the order of the polynomial elements multiplied by that cell. For example, the coefficient of row 1, column 1 is equivalent to c_{11} in M-Polynomial expression (5.10).

It is evident that once γ_j=0, the effect of all subsequent branches of the M-Polynomial (*Poly2* to *Poly8*) is nulled as all their coefficients become equal to zero too.

The observation of interest is the effect on the 1st branch polynomial (*Poly1*). The coefficients of orders higher than 5 in this case will be cancelled as well due to the fact that a γ_j factor multiplies all the elements of the expression. As expected, the coefficients of orders lower than 5 will be reduced to the transistor's characteristic expression *G* defined in the previous section. In other words, if the feedback FIR filter was not taken into consideration in the physical model, the memoryless nonlinear part would have been simply a 5th order polynomial.

The 9x8 structure was used only to clarify explanations, but these observations in the coefficients are general trends that apply to every combination of memory and nonlinearity orders. The nonlinearity order needed in the 1st branch of the M-Polynomial scheme is equal to $P_1=2N-1$, where N is the actual memoryless nonlinear order of the system being modeled.

The fact that only $(P_1+1)/2$ order of nonlinearity is needed to capture the memoryless behavior was tested in 2 ways. The first one is to assess the modeling capability of the 9th and 5th polynomial functions in capturing the static nonlinearity of the PA. As shown in Figure 5.5, both polynomials perform similarly in capturing the memoryless nonlinear part of the AM/AM and AM/PM characteristics.

Furthermore, the 5th order model was tested experimentally for linearization capability. As shown in Figure 5.6, the DPDs of both 9th and 5th order have the same linearization capabilities. This suggests that even though the 9th order was essential in linearizing the model with memory effects, it is over modeling the memoryless behavior and only a 5th order is needed to capture its memoryless characteristics

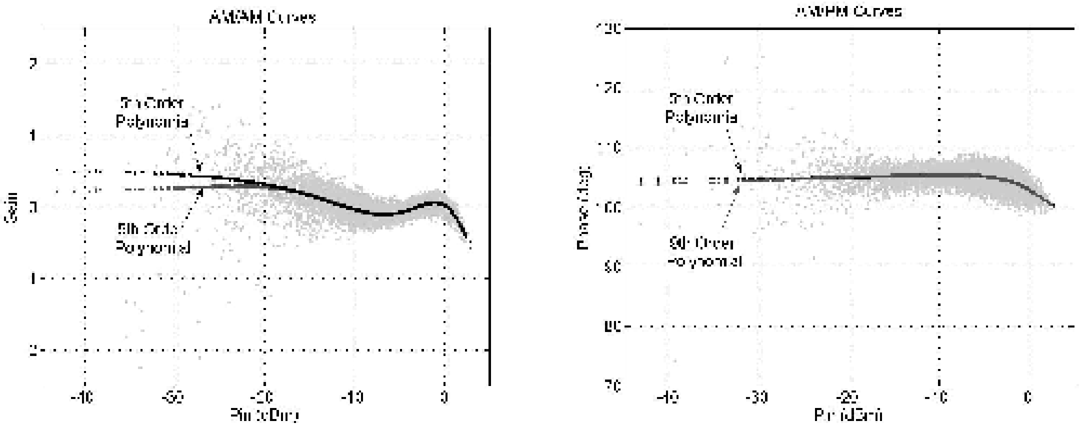


Figure 5.5 - AM/AM and AM/PM curves of the measured PA response, modeled with both 9th and 5th order memoryless polynomials.

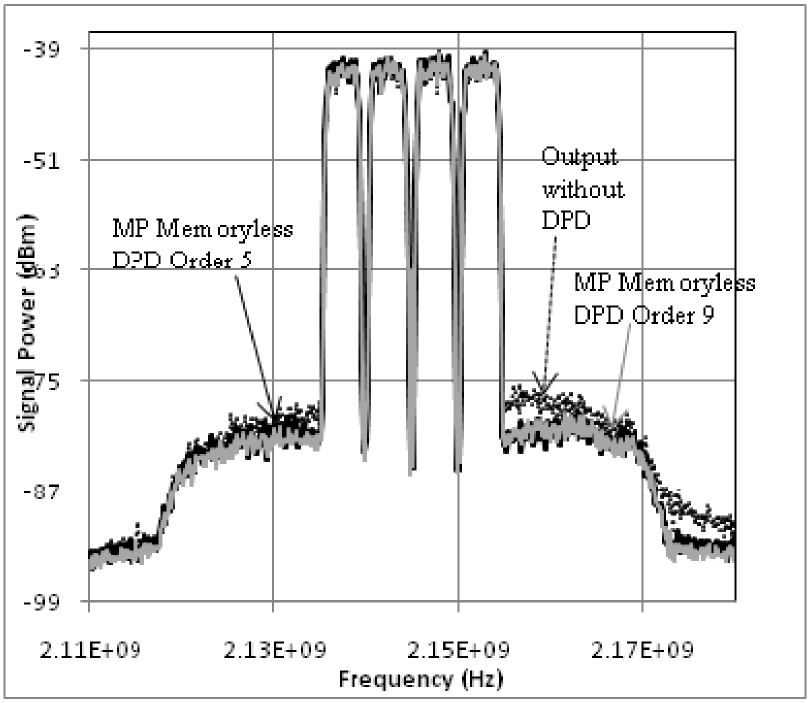


Figure 5.6 - Spectra of DUT response without any DPD applied (Output without DPD), and linearized with a memoryless DPD of order 9 and 5.

5.4 Approach Summary

A close form deterministic approach to M-Polynomial complexity reduction was presented. The reduction is based on the proposed close –form analytical expression of its coefficients, obtained from a physical PA model. Three main reduction steps were applied to the expression obtained from the physical model, namely: cross-term elimination, elimination of high order nonlinearities from memory effects, and even order elimination. The number of coefficients required by the M-Polynomial was reduced by a factor of 2.6, reducing complexity and costs of implementation of the f the scheme without losing any of its proven linearization capabilities as shown through experimental validations of results. Applying the reductions suggested by the analytical analysis of the coefficients guarantees that no losses will be observed in the linearization performance of the reduced scheme compared to that of the full M-Polynomial scheme.

Another concept that was proposed and verified, both in simulation and experimentally, was the fact that the order of the memoryless nonlinearity of the RF PA was almost half the order required by the 1st branch polynomial of the M-Polynomial scheme. Additionally, a PAE of 32% was achieved when employing the DPD scheme, which is an improvement on the PAE of the system with no DPD (15%).

Chapter 6
Conclusions and Future Work

6.1 Conclusion

The work presented in this thesis aimed mainly at understanding and improving the behavior of RF PAs, mainly measured by the linearity-efficiency tradeoff. The main contributions from this work can be summarized in three main points. First, a deterministic approach for parallel Hammerstein reduction was proposed. The proven parallel Hammerstein structure is used to demonstrate a deterministic approach where optimal filter lengths of the structure can be determined with one set of measurements. The reduced scheme using this method is shown to perform comparably to the full, unreduced schemes.

Another main point is the interpretation and close form formulation of memory effects. A link has been established between behavioral and physical models of RF PAs. This link is essentially a close form formulation for memory effects that is based entirely on physical PA circuit properties. The origin of memory effects is explained as reflections from networks back into the nonlinear transistor.

The last point is linking behavioral modeling to physical close form formulation. The physical formulation is linked back to behavioral models in two steps. The first step is linking the Volterra model to the physical model, showing the correlation between the theoretical Volterra coefficients when applied to model RF PAs. The second step is to explore assumptions and subtle reductions in the physical formulation presented that yields identical terms to those of the M-Polynomial. From that point, the necessity or redundancy of each coefficient in the memory polynomial is justified based on physical properties of the DUT.

6.2 Future Work

Many topics, presented in detail or mentioned briefly during this thesis, are worth further investigation. Here are a few directions of possible future research:

1. When exploring the FIR filter response for the Parallel Hammerstein reduction, it might be interesting to consider the omission of a few taps that are within the range of interest. Sampling rates used could be high compared to required rates for certain orders. Just as an

illustration, even if the FIR length has to be 6 to cover the physical time constant response of the PA, it is possible that the 2nd and 4th taps are not necessary in that filter.

2. The physical formulation gives a definition for the coefficients of the behavioral models in terms of physical circuit properties, namely reflection coefficients of networks in the PA circuit. These reflection coefficients are not measurable quantities, but would, if quantified, give great insight to PA designers as to where, and at what frequencies to improve their design. When finding numerical values for behavioral models, they can be linked back to identify the value of those reflection coefficients so that modeling is now quantifying non-measurable operation criteria of the PA components.

3. Sampling Rate / PA time constant

   The required sampling rate of signals has been the subject of extensive computational research in this field. The required sampling rate in this case is an essential factor to be determined as it determines the frequency at which DPD computations take place, therefore keeping the sampling rate to a minimum would save computational power for the DPD scheme operation. It is interesting to explore an optimal sampling rate which accurately covers the PA time constant. Then the term ``Memory depth`` of the PA would be comparable from experiment to experiment, since for now it is the number of sample points that cover the time constant, but only relative to a specific sampling rate.

4. Testing the M-Polynomial shortcomings based on the assumptions made in the transition from physical model to M-Polynomial model, as for example the extent to which it will still perform accurately with a low sampling rate.

Bibliography

[1] J. Vuolevi, “Analysis, measurement and cancellation of the bandwidth and amplitude dependence of intermodulation distortion in RF power amplifiers,” Department of Electrical Engineering, University of Oulu, 2001.

[2] J. Kim and K. Konstantinou, “Digital predistortion of wide-band signals based on power amplifier model with memory,” Electron. Lett., Vol.37, no.23, pp.1417-1418, Nov. 2001.

[3] N. Messaoudi, M-C. Fares, S. Boumaiza and J. Wood, “Complexity Reduced Odd Order Only Memory Polynomial Predistorter for Multicarrier Doherty Power Amplifier Linearization”, IMS2008.

[4] H.Honkasalo, K.Pehkonen, M.T.Niemi and A.T.Leino, “WCDMA and WLAN for 3G and Beyond”*, Nokia, 2002*.

[5] P. B. Kenington“High-linearity RF Amplifier Design”, Artech House Publishers, 2000.

[6] “WCDMA TX Theory and Measured Results from Maxim's WCDMA Reference Design v1.0”, Maxim, http://www.eetindia.co.in, April 2009.

[7] Stephen Mass, Nonlinear Microwave and RF circuits``, 2nd edition, Artech House, 2003.

[8] Cripps, “Advanced Techniques in RF Power Amplifier Design”, Artech House, 2002.

[9] Adel M. Saleh, “Frequency-Independent and Frequency-Dependent Nonlinear Models of TWT Amplifiers,” *IEEE Trans. on Communications*, vol. 29, issue 11, pp. 1715-1720, Nov., 1981.

[10] T. Liu, S. Boumaiza and F. M. Ghannouchi, "Augmented Hammerstein Predistorter for Linearization of Broadband Wireless Transmitters", IEEE Trans. Microwave Theory and Techniques, Vol. 54, Issue: 4, April 2006, pp. 1340-1349.

[11] M. Schetzen, “Nonlinear system modeling based on the wiener theory,” *Proc. IEEE*, vol. 69, no. 12, pp. 1557–1573, Dec. 1981.

[12] J. Xu, M. Yagoub, R. Ding and Q. Zhang, “Neural-Based Dynamic Modeling of Nonlinear Microwave Circuits”, IEEE Trans. Microwave Theory Tech., Vol. 50, no.12, pp. 2769-2780, December 2002.

[13] Andrei Grebennikov, “RF and Microwave Power Amplifier Design”, McGraw-Hill Professional Engineering, 2005.

[14] John Tsimbinos, “Identification and Compensation of Nonlinear Distortion”, University of South Australia, Doctor of Philosophy thesis, 1995.

[15] John Wood, David E. Root, “Fundamentals of Nonlinear Behavioral Modeling for RF and Microwave Design”, Artech House, 2005.

[16] Farouk Mkadem and Slim Boumaiza, “Extended Hammerstein Behavioral Model Using Artificial Neural Networks”, *IEEE Trans. Microwave Theory and Tech.*, vol. 57, no. 4, April 2009

[17] T. Liu, S. Boumaiza, F.M. Ghannouchi, “De-embedding Static Nonlinearities and Accurately Identifying and Modeling Memory Effects in Wide-Band RF Transmitters,” IEEE Trans. Microwave Theory Tech., Vol. 53, No.11, pp. 3578-3587, November 2005.

[18] M. Isaksson and D. Wisell, "Extension of the Hammerstein Model for Power Amplifier Applications," presented at ARFTG 63, Fort Worth, 2004.

[19] L. Ding, “Digital Predistortion of Power Amplifiers for Wireless Applications,” Elect. & Comp. Engineering faculty, Georgia Institute of technology, March 2004.

[20] T. R. Cunha, J. C. Pedro, E. G. Lima, “Low-Pass Equivalent Feedback Topology for Power Amplifier Modeling”, *IMS 2008.*