Universidade de Aveiro
2017
Departamento de Electrónica, Telecomunicações e Informática

Filipe Miguel Esturrenho Barradas

Modelação Comportamental e Pré-distorção Digital de Transmissores de Rádio-Frequência

Modeling and Digital Predistortion of Wireless Radio Frequency Transmitters

Universidade de Aveiro
2017
Departamento de Electrónica, Telecomunicações e Informática

Filipe Miguel Esturrenho Barradas

Modelação Comportamental e Pré-distorção Digital de Transmissores de Rádio-Frequência

Modeling and Digital Predistortion of Wireless Radio Frequency Transmitters

Tese apresentada à Universidade de Aveiro para Cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Eletrotécnica, realizada sob a orientação científica do Prof. Dr. Pedro Miguel Lavrador, Professor auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro e sob a coorientação científica do Prof. Dr. Telmo Reis Cunha, Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro.

Apoio financeiro da Fundação para a Ciência e Tecnologia (FCT) no âmbito de uma bolsa de investigação, ref. SFRH/BD/90103/2012.

Dedico este trabalho à minha namorada e família.

O Júri / The Jury

Presidente / President	**Prof. Dr. Artur Manuel Soares da Silva** Professor Catedrático, Universidade de Aveiro
Vogais / Examiners Committee	**Prof. Dr. Nuno Miguel Gonçalves Borges de Carvalho** Professor Catedrático, Universidade de Aveiro
	Dr. John Wood Technical Consultant, Obsidian Microwave
	Dr. Francesc Purroy Martin Power Amplifiers Team Leader, Huawei Technologies Sweden AB
	Prof. Dr. Pere Lluis Gilabert Pinal Professor Associado, Castelldefels School of Telecommunications and Aerospace Engineering
	Prof. Dr. Pedro Miguel Lavrador Professor Auxiliar, Universidade de Aveiro (Orientador)

Agradecimentos / Acknowledgements

Em primeiro lugar, gostaria de agradecer à minha família e amigos próximos pelo apoio e encorajamento e por tomarem conta de mim todos estes anos. Á minha namorada por estar sempre presente e me trazer alegria todos os dias.

Gostaria de agradecer aos meus orientadores e a todo o grupo de Circuitos e Sistemas de Rádio do Instituto de Telecomunicações de Aveiro: Prof. José Carlos Pedro, Prof. Telmo Cunha, Prof. Pedro Lavrador, Prof. Pedro Cabral e Dr. Luís Cótimos Nunes pelas sugestões, bons-tempos e pelo desafio.

Agradeço à Huawei Technologies, em particular à Huawei Sweden PA Research and Development Team, pela oportunidade única de trabalhar para um projecto desafiante e que contribuíu também para o meu trabalho. Nomeadamente gostaria de agradecer ao Dr. Francesc Purroy e Mr. Richard Hellberg pelas discussões em prédistorção e modelação.

Agradeço à Universidade de Aveiro, ao Departamento de Electrónica, Telecomunicações e Informática e ao Instituto de Telecomunicações por me terem facultado todos os meios e ambiente de trabalho necessários à execução do meu trabalho.

À Fundação para a Ciência e Tecnologia pelo apoio financeiro concedido sob a forma de uma bolsa de Doutoramento.

Muito obrigado a todos, e bem haja.

Palavras-Chave

Amplificadores de Potência (PA), Pré-distorção Digital (DPD), Método dos Mínimos Quadrados, Teoria de Aproximação, Transmissores de Rádio-frequência, Modelação de sistemas

Resumo

Nos atuais sistemas de telecomunicações, os transmissores de rádio-frequência são desenvolvidos tendo maioritariamente em conta a eficiência da conversão da potência fornecida da fonte em potência de rádio-frequência. Este tipo de desenho resulta em amplificadores de potência com características de transmissão não-lineares, que distorcem severamente o envelope de informação no processo de amplificação, gerando distorção fora da banda. Para corrigir este problema utiliza-se um processo de compensação não linear, sendo que a pré-distorção digital se tem favorecido pela sua flexibilidade e precisão. Este método é tipicamente aplicado de uma forma cega, por força bruta até se obter a compensação desejada. No entanto, quando o método se mostra ineficaz, como se verificou em amplificadores de potência baseados em transístores de nitreto de gálio, é difícil saber o que modificar nos sistemas para os tornar de novo úteis. De forma a compreender e desenhar sistemas de pré-distorção digital robustos é necessário, por um lado, perceber o comportamento dos amplificadores de rádio-frequência, por outro, perceber as limitações e relações entre os modelos digitais e o comportamento real do amplificador. Nesse sentido, esta tese explora e descreve estas relações de forma a suportar a escolha de modelos de pré-distorção, desenvolve novos modelos baseados no comportamento dos transístores, e propõe métodos de caracterização para os amplificadores de RF.

Keywords Power Amplifiers (PA), Digital Predistortion (DPD), Least Square Approximations, Radio Transmitters, System Modeling

Abstract In current telecommunication systems, the main concern when developing the radio frequency transmitter is power efficiency. This type of design generally leads to a highly nonlinear transmission characteristic, mainly due to the radio frequency power amplifier. This nonlinear transmission severely distorts the information envelope, leading to spectral regrowth, out-of-band distortion. To correct this problem a nonlinear compensation process is employed. For this application, digital predistortion is generally favored for its flexibility and accuracy. Digital predistortion is mostly applied in a blind manner, using brute force until the desired compensation is achieved. Because of this, when the method fails, as it has in gallium nitride based power amplifiers, it is difficult to modify the system to achieve the desired results. To understand and design robust predistortion systems, it is both necessary to have knowledge of the power amplifiers' behavior, on one hand, and understand the limitations and relations between the digital models and these behaviors, on the other. To do this, this thesis explores and describes these relationships, granting support to the digital predistortion model choice, it further develops new predistortion models based on the physics of the transistors' behaviors, and it proposes methods for the characterization of radio frequency power amplifiers.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AM	Amplitude Modulation
AM/AM	Amplitude Modulation to Amplitude Modulation
AM/PM	Amplitude Modulation to Phase Modulation
BW	Bandwidth
CW	Continuous Wave
DDR	Dynamic Deviation Reduction
DDR1	First Order Dynamic Deviation Reduction
DHT	Doherty
DHTPA	Doherty Power Amplifier
DPD	Digital Predistortion/Predistorter
EDGE	Enhanced Data GSM Environment
EMP	Envelope Memory Polynomial
FIR	Finite Impulse Response
GaN	Gallium Nitride
GMP	Generalized Memory Polynomial
GSM	Global System for Mobile Communications
IF	Intermediate Frequency
IIR	Infinite Impulse Response
IMD	Intermodulation Distortion
IMR	Intermodulation Ratio
LDMOS	Laterally Diffused Metal Oxide Semiconductor
LMS	Least Mean Squares
LPE	Low-Pass Equivalent
LS	Least Squares
LTE	Long Term Evolution
LUT	Look-Up Table
MP	Memory Polynomial
MNOP	Mixed Nonlinear Order Polynomial
PA	Power Amplifier
PH	Parallel Hammerstein

PAPR	Peak to Average Power Ratio
PD	Predistortion/Predistorter
PM	Phase Modulation
PoD	Postdistorter
RLS	Recursive Least Squares
RF	Radio Frequency
RFPA	Radio Frequency Power Amplifier
RPV	Radially Pruned Volterra
SDDR2	Simplified Second Order Dynamic Deviation Reduction
Tx	Transmitter
UMTS	Universal Mobile Telecommunication System
VBW	Volterra Behavioral Model for Wideband PAs

1. Introduction

This work develops methods for analysis, modeling and compensation of nonlinear systems in the scope of telecommunication systems. The focus is on the behavior of wireless Radio Frequency (RF) Transmitter (Tx). The main nonlinear component of such systems is typically recognized as the Radio Frequency Power Amplifier (RFPA), which is responsible for bringing the output power to the desired levels.

The behavior of these devices (the RFPA) can typically be split into the behavior at RF and the behavior for the information (the envelope or modulation imposed in the RF carrier). The main interest in this work is the behavior for the information. In this sense, this work explores system level characterization, modeling and compensation techniques that can be applied to understand or correct the operation of the RFPA.

This initial chapter is focused on laying the background and motivation for this work, as well as going through the state-of-the-art in terms of RFPA models and nonlinear effects. This chapter ends with a description of the thesis organization and objectives, and the main contributions of this work. In this way, this chapter lays the initial foundations on which this work was developed as well as what developments it has achieved and the roadmap it has followed.

1.1. Background and Motivation

With the massive and widespread use of wireless RF communications, a large number of base stations to support these systems have been deployed. With this high number of wireless RF Tx, the energy consumption of the units quickly became an important factor in the overall cash expenditure to operate such a network. Furthermore, for lower efficiency transmitters, the higher heat generation leads to higher cooling requirements, further increasing the overall energy consumption and setup cost.

Because of the power and cooling requirements, the efficiency of wireless RF Tx rapidly became of major importance for the Telecom industry. Taking into account the typical architecture of these systems, shown in Fig. 1.1, for the current telecommunication generation, the highest power consumption (about 40%) is in the RFPA [1, 2].

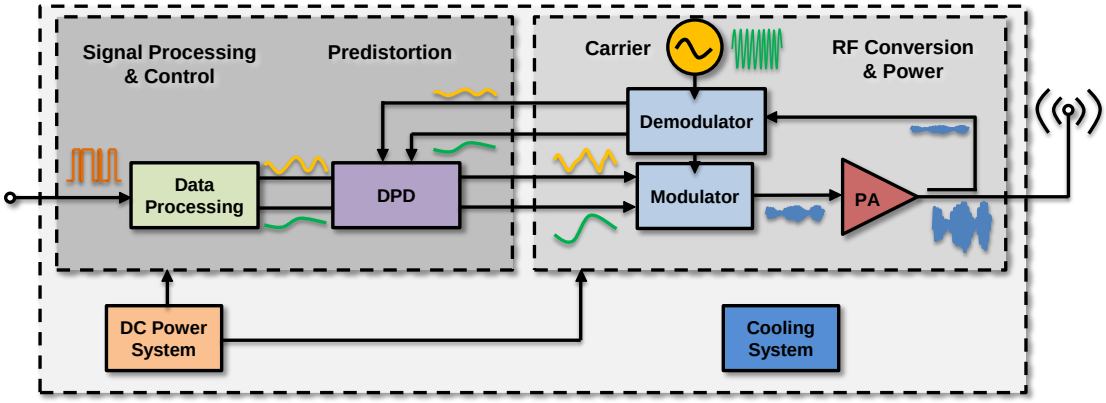


Fig. 1.1: Conceptual wireless RF Tx architecture block diagram using high level blocks.

Since the highest power consumption is in the RFPA it became important to increase the efficiency of this component, as this will directly translate into overall power savings. Furthermore, efficient RFPAs directly lead to savings in the cooling system's power and size, possibly even leading to their elimination [3].

Typical RFPAs reach higher efficiencies in saturated operation, however, PAs

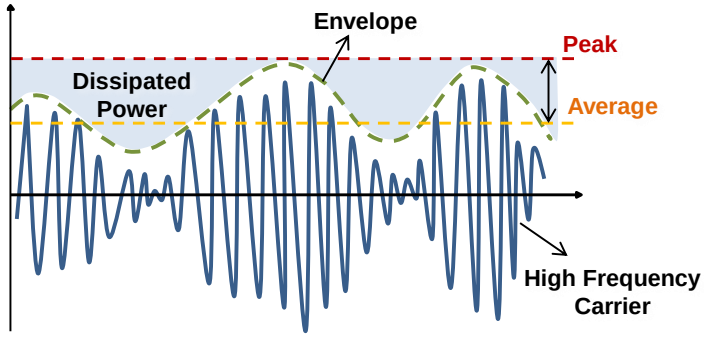


Fig. 1.2: Illustration of the PAPR problem. The PA must reach the required peak power but operates in back-off long periods of time.

cannot maintain a saturated operation when the input signal is backed-off. The average efficiency of the transmitter is then connected to the Peak to Average Power Ratio (PAPR) of the transmitted signals. Higher PAPR signals will lead to a decrease of the global efficiency due to longer periods of back-off operation. An illustration of this effect can be seen in Fig. 1.2. This effect has become particularly relevant for modern communication standards. In fact, the rapid growth of data rates while maintaining a tight spectral occupancy has led to the development of high PAPR signals for modern communications. Table 1-I shows the evolution of telecommunication signals (download path) in terms of BandWidth (BW) and PAPR.

Standard	Modulation	PAPR*	Data Rate*	BW
GSM (2G)	GMSK	0 dB	22.8 kbps	200 kHz
EDGE (2.5G)	8-PSK	3.2 dB	59.2 kbps	200 kHz
UMTS (3G)	QPSK	3.5 – 7 dB	2 Mbps	5 MHz
LTE (4G)	OFDM	9 – 12 dB	20 Mbps	1.25 – 20 MHz

*Typical values

Table 1-I: Communication standard's modulation, PAPR, Data Rate, Bandwidth

To further aggravate the issue, modern transmitters make use of the same RFPA to transmit closely spaced signal bands. This aggregation of multiple signals into the same RFPA increases the PAPR of the excitation. The multicarrier Global System for Mobile communications (GSM) signals are good examples of this effect, as shown in Fig. 1.3. Originally, GSM makes use of Gaussian Minimum Shift Keying (GMSK) modulation, which has an almost constant envelope (excluding ramp up and down),

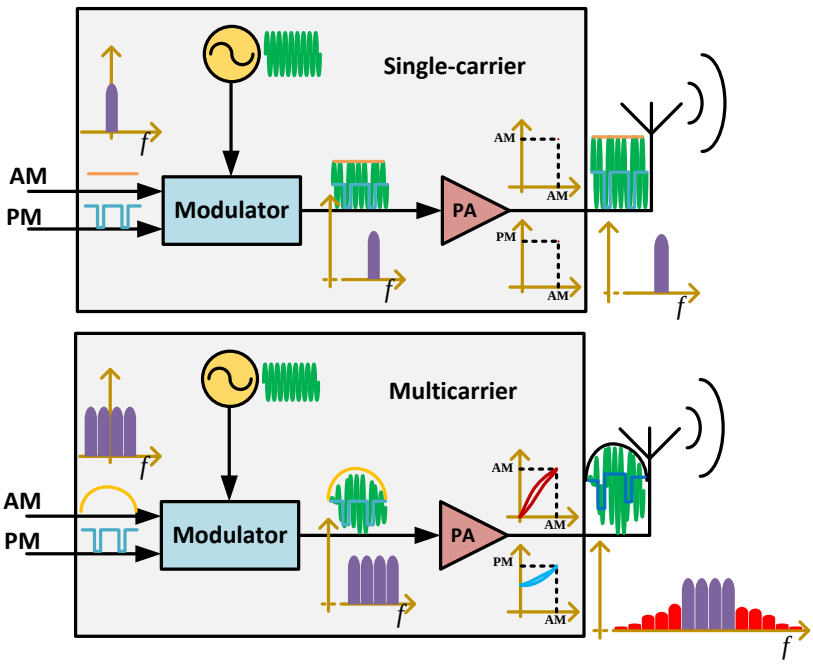


Fig. 1.3: Example of GSM amplification with several or one carrier. Besides the increase in bandwidth, the multicarrier signals have AM modulation which is not present in the single carrier case.

however, if multiple envelopes are combined, Amplitude Modulation (AM) is generated with a PAPR depending on the number of combined carriers.

To reach higher efficiencies in back-off operation modern transmitters make use of alternative RFPA topologies. The most widespread topology for base station power amplifiers is the Doherty (DHT) topology, illustrated in Fig. 1.4.

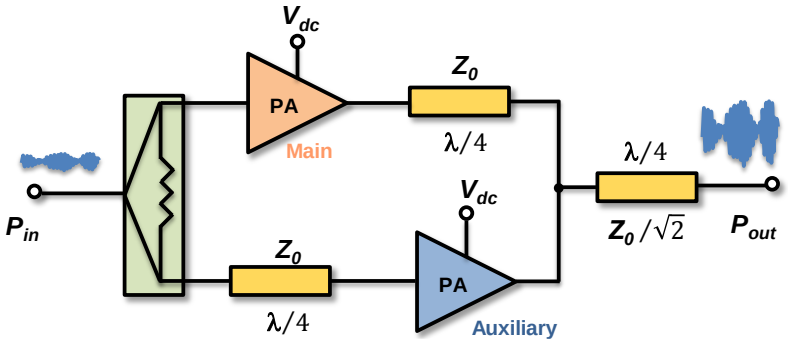


Fig. 1.4: Block diagram of the DHT architecture.

Even in these alternative, efficient topologies, designing RFPAs for maximum efficiency goals leads to nonlinear operation because the devices need to be pushed farther into saturated power regions [4]. Accommodating these nonlinearities into the transmit chain requires pre-compensation methods, due to the stringent spectral emission regulations. As seen in Fig. 1.3, nonlinear amplification produces spectral regrowth, which must be eliminated. However, the regrowth is too close to the signal for filtering and therefore requires pre-compensation methods (as it cannot be emitted and compensated in the receiver). It also becomes important to understand the distortion generation mechanisms, to characterize, and to model them. While characterization should allow a clear observation of the distortion effects and a direct relationship to the measurable quantities, the nonlinear modeling should allow faster system level simulations of the wireless Tx.

For nonlinear compensation, a number of techniques have been proposed, including digital and analog, feedforward and feedback techniques [5, 6, 7]. These techniques can be further divided depending on whether the compensation is generated at RF at an Intermediate Frequency (IF) or in the complex baseband, as well as such other characteristics as passivity.

In wireless RF Tx, Digital PreDistortion (DPD) in the digital complex baseband (also called Low-Pass Equivalent (LPE)) domain has become the most popular option [6], when implementation is possible – up to a few hundreds of megahertz of original signal bandwidth. The main advantage of DPD over other compensation techniques is flexibility and accuracy. DPD makes use of low to high complexity models applied in the digital complex envelope domain which generate distortion to exactly compensate the one generated by the subsequent analog stages [7]. DPD is then the attempt at

estimating the correct pre-inverse of the PA, involving a choice of model and its correct identification.

Trimmed dynamic polynomial models (Volterra, Wiener series) [8, 9] have been used with success. Several trimming methods have been proposed based on intuition, physical arguments, simplicity arguments and empirical evidence [10, 11, 12, 13, 14, 15, 16, 17, 18]. A number of other transformations can also be found in the literature to diminish the number, or include some particular set, of coefficients [19, 20, 21, 22, 23]. In general, one of the main advantages of such models is that the kernel estimation can be reduced to a linear leas squares (LS) estimation problem, [9] using the indirect learning technique [24] or some direct learning techniques [25].

In DPD, the main objective is to accurately linearize the device to fulfill a certain distortion metric. For some cases, traditional techniques have shown to be insufficient to reach these required linearity goals for four main reasons. First, the clock frequency limit imposes a severe bandwidth restriction which in turn limits the amount of distortion the digital compensator can inject in the signal [26, 27]. Second, the increased excitation richness reveals distinct nonlinear behavior from what was previously expected [28]. Third, accurate description and compensation of medium and long term memory is now taking its first steps [29, 30, 31]. Fourth, the used signal processing techniques may lead to unstable and biased solutions, due to ill-conditioning and estimator biases [32, 33, 34]. This thesis addresses some of these problems, initially focusing on relating the used DPD models to the phenomena observed on the RFPAs, then upgrading the model representation for improved extraction conditioning and orthogonality, and finally developing models for long-term compensation.

The trimmed dynamic polynomial models have also been widely used in RFPA modelling. In fact, in modeling, where the input and output of the systems are measurable, the problem is further reduced as it does not require iterative procedures. While, in DPD, the general objective is accurate compensation, modeling can have different objectives. For instance, it is oftentimes important to include in a model specific aspects of the RFPAs' behavior. Another example is relating the model coefficients to the measurable device response, while this has been achieved in the linear domain, using the S-parameters, it is far from being accomplished in the nonlinear domain, even with recent achievements [35]. Characterization and modeling in this way is important for both simulation and design. In this thesis, typical RFPA measurements (two-tone Intermodulation Distortion (IMD) measurements) are used for characterizing the different components of the devices' response. The focus is not on the device level but on the input-output relationship of the PA, so the techniques are limited to transmission measurements.

Concluding, DPD requires the development of new techniques and advances to deal with the rapid increase in signal complexity and the new device effects, where simplistic *ad hoc* model design methodologies have proven insufficient. Furthermore, the characterization of the RFPAs themselves is still insufficient from the nonlinear point of view. This thesis will focus on developing robust (well-conditioned and unbiased) signal processing techniques while maintaining a connection to the physical effects on the PA.

1.2. State-of-the-Art

This section presents the models found in the literature for DPD and behavioral modeling. The models are approached with digital approximation in the LPE in mind. In the signal processing community, the distinction between using a model for approximating the forward response of the PA or the inverse response of the PA is not very clear. In fact, most models found in the literature are used for both effects without changes. While this might seem strange, looking at the limit cases, a general polynomial can certainly approximate the inverse of another polynomial, a Finite Impulse Response (FIR) filter can also approximate the inverse response of another and the same is true for a complete Volterra series. Nonetheless, since the models themselves are instantiated without relationship to the physics of the devices, their use for one purpose or the other is blurred. Studying RFPA, Volterra based models, for DPD or for behavioral modeling will then lead to studying the same models.

Since a great deal of models are solidly rooted on the Volterra series, this section starts off by presenting the series itself. It then proceeds by going through the main models found throughout the literature. To help visualize the difference between the models here described, the third order kernel is plotted for each presented model. After the different structures for DPD and modeling have been presented, a brief overview of the state-of-the-art for nonlinear effects in RFPAs is presented as support for the second chapter of this thesis. Finally, an overview of DPD extraction techniques is given.

1.2.1. Predistortion and Behavioral Models

Volterra Series

The Volterra series is capable of describing a general, stable (with fading memory), time-invariant, continuous, nonlinear system [8]. It is the dynamic equivalent of a polynomial series and so can be seen from two different perspectives. The first use of the series is in the “Taylor sense”, where there is some operation point and each

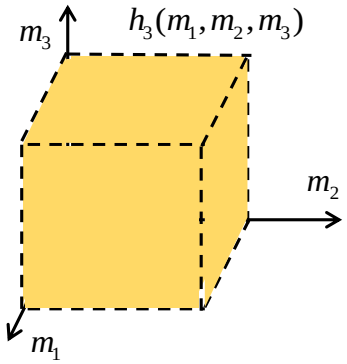


Fig. 1.5: Schematic representation of the 3rd order Volterra series kernel

kernel describes deviations from this steady-state. The other use of the series is in the "LS sense", where some input-output relationship is fit to minimize an error. In the DPD and behavioral modeling realm, the least squares sense is generally more appropriate.

The Volterra series has the formulation shown in (1.1), in the discrete-time, LPE domain.

$$\tilde{y}_{2p-1}(n) = \sum_{m_1=0}^{n} \dots \sum_{m_{2p-1}=0}^{n} \tilde{h}_{2p-1}(m_1, \dots, m_{2p-1}) \tilde{x}(n - m_{2p-1}) \times \prod_{r=1}^{p-1} \tilde{x}(n - m_r) \prod_{k=p}^{2p-2} \tilde{x}^*(n - m_k) \quad (1.1)$$

$$\tilde{y}(n) = \sum_{p=1}^{P} \tilde{y}_{2p-1}(n)$$

The series is composed by P kernels of increasing dimensional order. For modeling and DPD in the LPE domain, the kernels are limited to the odd orders. In a full Volterra series, the $(2p-1)^{th}$ kernel has dimensional order $2p-1$. So, each kernel is defined on hypercubes of increasing dimensions. As will be seen, other models defined in this series use kernel truncations, which limit the dimensionality of these kernels. The third order kernel of a Volterra series is defined in all of the three dimensional space as seen in Fig. 1.5.

Using the full Volterra series for modeling and DPD is generally ill-advised for several reasons. First, the signals typically do not present enough richness to fully excite the Volterra kernels. Second, the high number of coefficients difficults the extraction. Third, the problem rapidly becomes ill-conditioned. For these reasons, a number of other models have been developed by truncation of the original series.

Memory Polynomial/Parallel Hammerstein

The Memory Polynomial (MP), or Parallel Hammerstein (PH), is one of the most widely used models, due to its simplicity, easiness to control and simple implementation. The MP model was originally presented in [10] and has the mathematical formulation shown in (1.2).

$$\tilde{y}(n) = \sum_{p=0}^{P} \sum_{m=0}^{M} \tilde{h}_{2p+1}(m) x(n-m) |x(n-m)|^{2p} \quad (1.2)$$

Interestingly, the MP model can also be seen as a parallel Hammerstein model [36], where the functions have been expanded using polynomials.

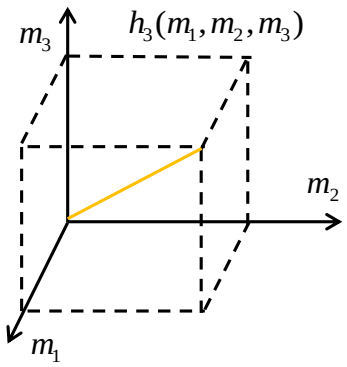


Fig. 1.6: Schematic representation of the 3rd order MP kernel

The MP model has a much more restricted third order kernel, as seen in Fig. 1.6. In fact, all kernels of the MP have been reduced to dimensionality one. Evidently, this reduces the number of coefficients significantly when related to the original series.

The MP model is one of the models in which predistortion using model inversion has been attempted [37, 38]. Even though this approach is not very popular, since direct DPD extraction is simpler, the MP model is sufficiently manageable to attempt it.

Envelope Memory Polynomial

The Envelope Memory Polynomial (EMP) was originally presented in [12]. One of the main arguments for its use being the reduction of hardware complexity, since most operations are performed on the absolute value which is a real value instead of a complex one. The EMP has the mathematical formulation shown in (1.3).

$$\tilde{y}(n) = \sum_{p=0}^{P} \sum_{m=0}^{M} \tilde{h}_{2p+1}(m) x(n) |x(n-m)|^{2p} \tag{1.3}$$

Again the kernels have been reduced to one dimension, as shown in Fig. 1.7. The main difference from the MP model is that the memory is only applied to the amplitude. One of the problems of this approach is that it fails to approximate the linear, frequency dependent response, of the PA.

One common upgrade over the EMP is to include a parallel linear filter to manage

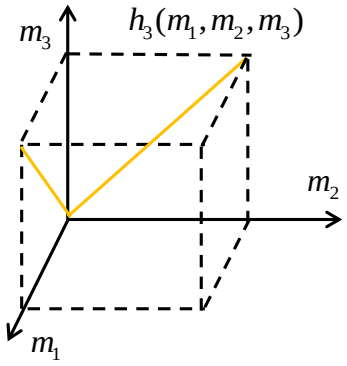


Fig. 1.7: Schematic representation of the 3rd order EMP kernel

the linear dynamics, resulting in the model in (1.4).

$$\tilde{y}(n) = \sum_{m=0}^{M} \tilde{h}_1(m)x(n-m) + \sum_{p=1}^{P}\sum_{m=0}^{M} \tilde{h}_{2p+1}(m)x(n)|x(n-m)|^{2p} \quad (1.4)$$

More complex parallel topologies can be used. For instance, using an MP and an EMP in parallel [39, 19].

Generalized Memory Polynomial

The Generalized Memory Polynomial (GMP) was first proposed in [11], it has the mathematical formulation shown in (1.5).

$$\begin{aligned}\tilde{y}(n) = &\sum_{p=0}^{P}\sum_{m=0}^{M} \tilde{h}_{2p+1}(m,0)\tilde{x}(n-m)|\tilde{x}(n-m)|^{2p} \\ &+ \sum_{p=1}^{P}\sum_{m=0}^{M}\sum_{l=1}^{L} \tilde{h}_{2p+1}(m,l)\tilde{x}(n-m)|\tilde{x}(n-m-l)|^{2p} \\ &+ \sum_{p=1}^{P}\sum_{m=0}^{M}\sum_{l=1}^{L} \tilde{h}_{2p+1}(m,-l)\tilde{x}(n-m)|\tilde{x}(n-m+l)|^{2p}\end{aligned} \quad (1.5)$$

Unlike the previously presented models, the GMP has bi-dimensional kernels, which will significantly increase the number of parameters when compared to the MP or EMP models. However, the GMP model has presented very good results in the literature, making it one of the default go to models for accurate linearization.

The third order kernel of the GMP model is defined by two planes intersecting each other on the MP line, as shown in Fig. 1.8.

The GMP has a great deal of flexibility compared to the previously presented models, at the expense of higher coefficient account. However, it still presents a much reduced set of parameters compared to the original Volterra series. The GMP generally presents a good tradeoff between complexity and accuracy.

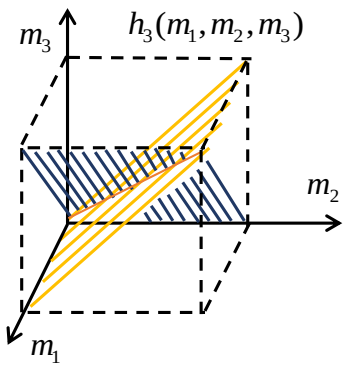


Fig. 1.8: Schematic representation of the 3rd order GMP kernel.

Dynamic Deviation Reduction (1st order)

The first order Dynamic Deviation Reduction (DDR1) was proposed in [14], it has the mathematical formulation shown in (1.6). The Dynamic Deviation Reductions (DDR) of several orders are derived within the framework of the deviation error presented in [40].

$$\tilde{y}(n) = \sum_{p=0}^{P} |\tilde{x}(n)|^{2p} \sum_{m=0}^{M} \tilde{h}_p^1(m)\tilde{x}(n-m) + \sum_{p=0}^{P-1} |\tilde{x}(n)|^{2p} \tilde{x}(n)^2 \sum_{m=1}^{M} \tilde{h}_p^2(m)\tilde{x}^*(n-m) \tag{1.6}$$

Again (similarly to the MP and EMP), it presents a one dimensional memory structure. The DDR1 also shares the EMP problem, it has trouble describing the linear dynamic response of the PA.

Unlike the previous models, the DDR1 is dependent on a conjugate term, which is an interesting addition. The existence of these terms is due to the restriction to the fundamental bandpass zone. In fact, signal components at other bands are converted to the fundamental bandpass by mixing down or up. In this case, the second harmonic zone is down converted (hence the conjugate) by mixing with the fundamental zone.

The third order kernel of the DDR1 model is shown in Fig. 1.9.

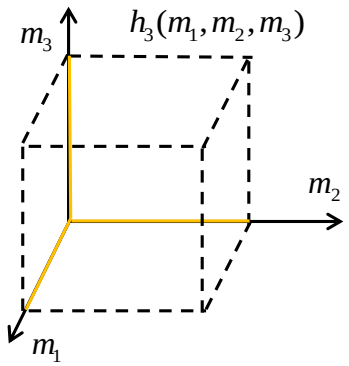


Fig. 1.9: Schematic representation of the 3rd order DDR1 kernel.

Simplified Dynamic Deviation Reduction (2nd order)

To improve the representation capabilities of the DDR1 model, the dynamic deviation order can be increased. However, this increases the number of parameters to unmanageable orders. A simplified second order model was proposed in [15] to avoid this parameter growth rate. The Second Order Simplified Dynamic Deviation Reduction (SDDR2) has the same mathematical structure as the DDR1 model with new

components introduced in parallel, as shown in (1.7).

$$\begin{aligned}\tilde{y}(n) = &\sum_{p=0}^{P}|\tilde{x}(n)|^{2p}\sum_{m=0}^{M}\tilde{h}_p^1(m)\tilde{x}(n-m)\\ &+\sum_{p=0}^{P-1}|\tilde{x}(n)|^{2p}\tilde{x}(n)^2\sum_{m=1}^{M}\tilde{h}_p^2(m)\tilde{x}^*(n-m)\\ &+\sum_{p=0}^{P-1}|\tilde{x}(n)|^{2p}\tilde{x}(n)^*\sum_{m=1}^{M}\tilde{h}_p^3(m)\tilde{x}^2(n-m)\\ &+\sum_{p=0}^{P-1}|\tilde{x}(n)|^{2p}x(n)\sum_{m=1}^{M}\tilde{h}_p^4(m)|\tilde{x}(n-m)|^2\end{aligned} \tag{1.7}$$

The third order kernel for the SDDR2 model is shown in Fig. 1.10. While the kernel components of the DDR1 model are preserved, new ones are added including those of the EMP for the third order.

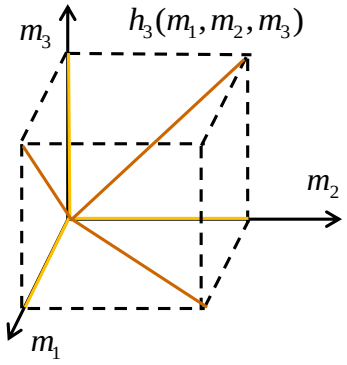


Fig. 1.10: Schematic representation of the 3rd order SDDR2 kernel.

Volterra Behavioral Model for Wideband PAs

The previous models were come upon by truncation of different series structures. The MP, EMP and GMP by truncating the original series representation, and the DDR1 and SDDR2 by truncating the series reformulated using the deviation error. The Volterra Behavioral Model for Wideband PAs (VBW) model, proposed in [16], is one of the few Volterra based models founded on the RF PA physics. To come upon the VBW, the memory in the PA is assumed to come mainly from the bias networks, this yields the mathematical formulation shown in (1.8).

$$\tilde{y}(n) = \tilde{h}_1\tilde{x}(n) + \sum_{p=1}^{P}\sum_{m_1=0}^{M_1}\dots\sum_{m_p=0}^{M_p}\tilde{x}(n)\tilde{h}_{2p+1}(m_1,\dots m_p)\prod_{k=1}^{p}|\tilde{x}(n-m_k)|^2 \tag{1.8}$$

The VBW model has memory in the amplitude terms only, similarly to the EMP

model. This model being rooted in the physics, and assuming memory from the PA bias networks, shows that these terms are connected to this memory type.

The VBW model also suffers from the lack of approximation capability for the linear memory. This is normal as the model is developed disregarding the RF matching networks, responsible for this memory.

The third order kernel of the VBW is equal to the EMP model third order kernel, as shown in Fig. 1.11. The models differ for higher orders, the VBW allowing more memory flexibility.

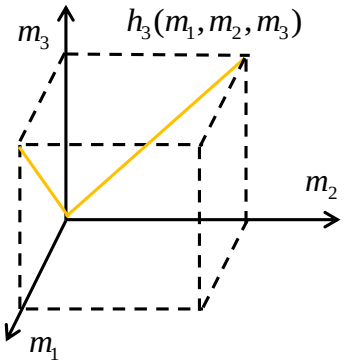


Fig. 1.11: Schematic representation of the 3rd order VBW kernel

The VBW model is not very popular because the construction of its terms is difficult and typically not required for a good approximation. The formulation is also prone to growing the number of terms very rapidly.

Radial Pruned Volterra Model

The Radial Pruned Volterra (RPV) model was proposed in [17]. The model proposes pruning all the Volterra kernels along the diagonals. Observing the previously presented third order kernels, shows that the models defined up to this point tend to preserve only the radial terms of the kernel. This is because these radial kernel directions happen when the memory lags are zero for some terms and equal for a number of terms.

Performing this radial pruning becomes more and more complex for increasing

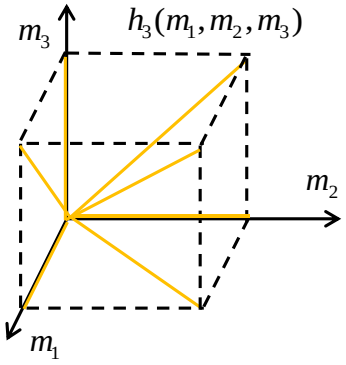


Fig. 1.12: Schematic representation of the 3rd order RPV kernel

model order and is usually not justified for application in modeling and DPD. For instance, the third order of this model has five different terms, the four from the SDDR2 model and one from the MP, as shown in (1.9).

$$\begin{aligned}\tilde{y}_3(n) = |\tilde{x}(n)|^2 \sum_{m=0}^{M} \tilde{h}_3^1(m)\tilde{x}(n-m) \\ +\tilde{x}(n)^2 \sum_{m=1}^{M} \tilde{h}_3^2(m)\tilde{x}^*(n-m) \\ +\tilde{x}(n)^* \sum_{m=1}^{M} \tilde{h}_3^3(m)\tilde{x}^2(n-m) \\ +x(n) \sum_{m=1}^{M} \tilde{h}_3^4(m)|\tilde{x}(n-m)|^2 \\ + \sum_{m=1}^{M} \tilde{h}_3^5(m)\tilde{x}(n-m)|\tilde{x}(n-m)|^2\end{aligned} \tag{1.9}$$

The third order kernel for this model is shown in Fig. 1.12.

Mixed Nonlinear Order Memory Polynomial

The Mixed Nonlinear Order Memory Polynomial (MNOMP) proposed in [21] is an expansion of the GMP model which is obtained by analysis of the memory processes in the RFPA, according to the feedback model proposed in [41]. The obtained terms are similar to the terms of the GMP, but expanded to have a nonlinear description of the signal mixing with the baseband terms. The mathematical formulation of the MNOMP model is shown in (1.10).

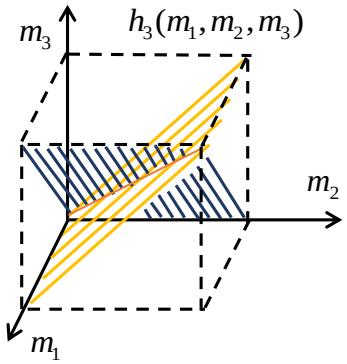


Fig. 1.13: Schematic representation of the 3rd order MNOMP kernel.

$$\tilde{y}(n) = \sum_{p=0}^{P} \sum_{r=0}^{R} \sum_{m_1=0}^{M_1} \sum_{m_2=0}^{M_2} \tilde{h}_{2(p+r)+1}(m_1, m_2) \tilde{x}(n-m_1)|\tilde{x}(n-m_1)|^{2p}|\tilde{x}(n-m_2)|^{2r} \tag{1.10}$$

The third order kernel of the MNOMP model is similar to the GMP model as shown in Fig. 1.13.

Physical Knowledge Pruning

Another physics based approach by exploring the model in [41] was presented in [42]. This model yields a similar formulation to the VBW model. However, since the memory is not limited to the bias but is limited to the mixing products, the model has more terms than the VBW including mixing products from higher orders to the fundamental. The fifth order of this model is shown in (1.11). For higher orders this model becomes more and more complex and difficult to manage.

$$\begin{aligned}\tilde{y}(n) = \tilde{h}_1\tilde{x}(n) &+ \sum_{m=0}^{M} h_3^1(m)x(n)|x(n-m)|^2 + \sum_{m=0}^{M} h_3^2(m)x^*(n)x^2(n-m) \\ &+ \sum_{m_1=0}^{M_1}\sum_{m_2=0}^{M_2} h_5^1(m_1,m_2)x(n)|x(n-m_1)|^2|x(n-m_2)|^2 \\ &+ \sum_{m_1=0}^{M_1}\sum_{m_2=0}^{M_2} h_5^2(m_1,m_2)x(n)x^2(n-m_1)x^{*^2}(n-m_2) \\ &+ \sum_{m_1=0}^{M_1}\sum_{m_2=0}^{M_2} h_5^3(m_1,m_2)x^*(n)x^2(n-m_1)\,|x(n-m_2)|^2\end{aligned} \tag{1.11}$$

Again, this model has problems approximating the linear response of the PA. The third order kernel for this model is similar to the EMP model and is shown in Fig. 1.14.

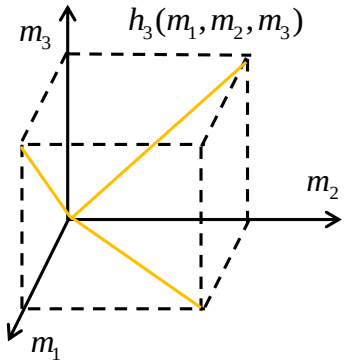


Fig. 1.14: Schematic representation of the 3rd order kernel of Zhu's physics based model.

Besides truncating the Volterra series, other transformations of this original model can also be useful to improve its description capabilities. Several changes to the original series can be found in the literature with different applications and showing different capabilities of the Volterra formulation.

Polar Volterra Series

One of the problems verified in the early LPE models was that the Volterra series for the LPE only possessed odd order terms. This comes from the LPE phase enforcement when the model is obtained from the series applied to the RF signal. In [43], it was shown that the Volterra series can be converted to the LPE domain using a polar formulation and that, under this formulation, the amplitude terms are unrestrained by the phase enforcement. This being the case, even and odd order polynomials of the amplitude can be used in the Volterra models. This can be applied to all the presented models. The polar Volterra series for the LPE has the formulation shown in (1.12).

$$\tilde{y}(n) = \sum_{p_1=0}^{P_1}\sum_{p_2=0}^{P_2}\sum_{m_1=0}^{M_1}\cdots\sum_{m_{p_1}=0}^{M_{p_1}}\sum_{l_1=0}^{L_1}\cdots\sum_{l_{p_2}=0}^{L_{p_2}}\sum_{l_{p_2+1}=0}^{L_{p_2+1}}\cdots\sum_{l_{2p_2+1}=0}^{L_{2p_2+1}} h_{p_1,2p_2+1}(m_1,\ldots,m_{p_1},l_1,\ldots,l_{p_2},l_{p_2+1},\ldots,l_{2p_2+1}) \quad (1.12)$$

$$a(n-m_1)\ldots a(n-m_{p_1})e^{j\phi(n-l_1)}\ldots e^{j\phi(n-l_{p_2})}e^{-j\phi(n-l_{p_2+1})}\ldots e^{-j\phi(n-l_{2p_2+1})}$$

The Polar Volterra series proved that the use of the initially non-orthodox terms in the Volterra based models was, in fact, correct.

Kautz/Laguerre Volterra

The Volterra series as originally proposed is based on unitary delays of the original signal. In fact, the Volterra formulation can be inserted into the Canonical Wiener architecture which has the structure shown in Fig. 1.15.

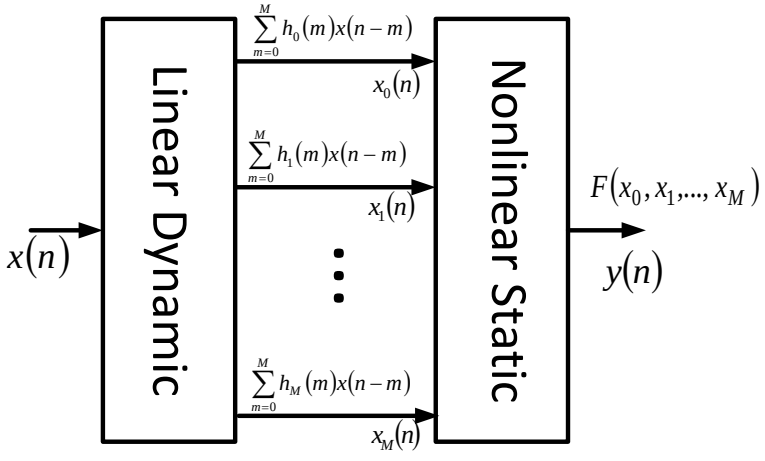


Fig. 1.15: Canonical Wiener structure.

The Canonical Wiener architecture shows that the Volterra series is a one to many linear system followed by a many to one nonlinear map. In fact, as long as the filters in the linear map are a basis of the filter space, the generality of the series is preserved. Using this information, in [44], the series was expanded using Kautz/Laguerre filters which are Infinite Impulse Response (IIR) filters. The use of IIR filters is not problematic, since the filter response can be easily calculated and studied before its application. This is in contrast with feeding back from the output of the nonlinear block. The change of filter basis functions can help in including memory effects with longer time-constants using fewer parameters.

Other Physics Based Power Amplifier Models

One useful formulation of a power amplifier was presented in [41, 45, 46]. In this formulation, the PA can be seen as a nonlinear bi-dimensional source working in a feedback loop, as shown in Fig. 1.16.

This system level model allows the analysis of the PA behavior and creating

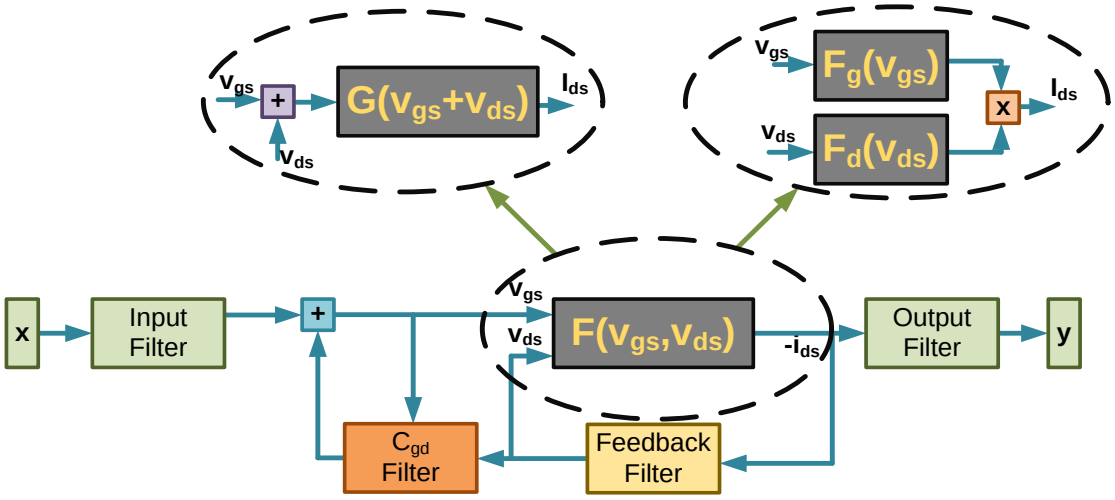


Fig. 1.16: PA behavioral model as a nonlinear source working under a feedback dynamic loop.

relationships between the digital models and the physics of the RFPA. Feedforward expansions of this feedback topology are also found in the literature, [47, 48, 49], that simplify the extraction and use of this structure. However, these models retain a high complexity that makes their popularity lower than Volterra based formulations.

1.2.2. Power Amplifier Distortion Mechanisms

RFPA distortion can be divided into static, or memoryless, and dynamic. The response of the amplifier is observed in terms of gain and phase-shift as a function of input, or output, power. These plots, generally called Amplitude Modulation to Amplitude Modulation (AM/AM) and Amplitude Modulation to Phase Modulation (AM/PM), were initially measured in Continuous Wave (CW) tests by varying the input power and measuring the output power and phase (versus the input) [50]. Eventually,

their use became generalized for dynamic envelopes where the instantaneous gain and phase-shift of the PA are used. These instantaneous quantities are defined using the complex envelopes of the carrier at the input and output of the PA. For memoryless PAs, or sufficiently narrowband signals, these traces are straightforward to interpret, otherwise it is often difficult to relate distortion effects and the AM/AM and AM/PM observations.

In Fig. 1.17 AM/AM and AM/PM traces of two-tone measurements for two closely spaced tones and two widely spaced tones in a Laterally Diffused Metal Oxide Semiconductor (LDMOS) RFPA are shown. As can be seen, for the narrowband case, the RFPA behaves as a nonlinear function, while in the wideband case, the RFPA has two possible gains for each input amplitude value, depending on the derivative of the signal. As the complexity of the signals increases more dynamic effects are excited in the system, making these plots difficult to interpret and inappropriate for characterization.

In order to properly understand the predistortion requirements of each PA it is important to understand the several effects that generate signal distortion. This section describes a number of effects presented in the literature, which are responsible for nonlinear distortion in a PA.

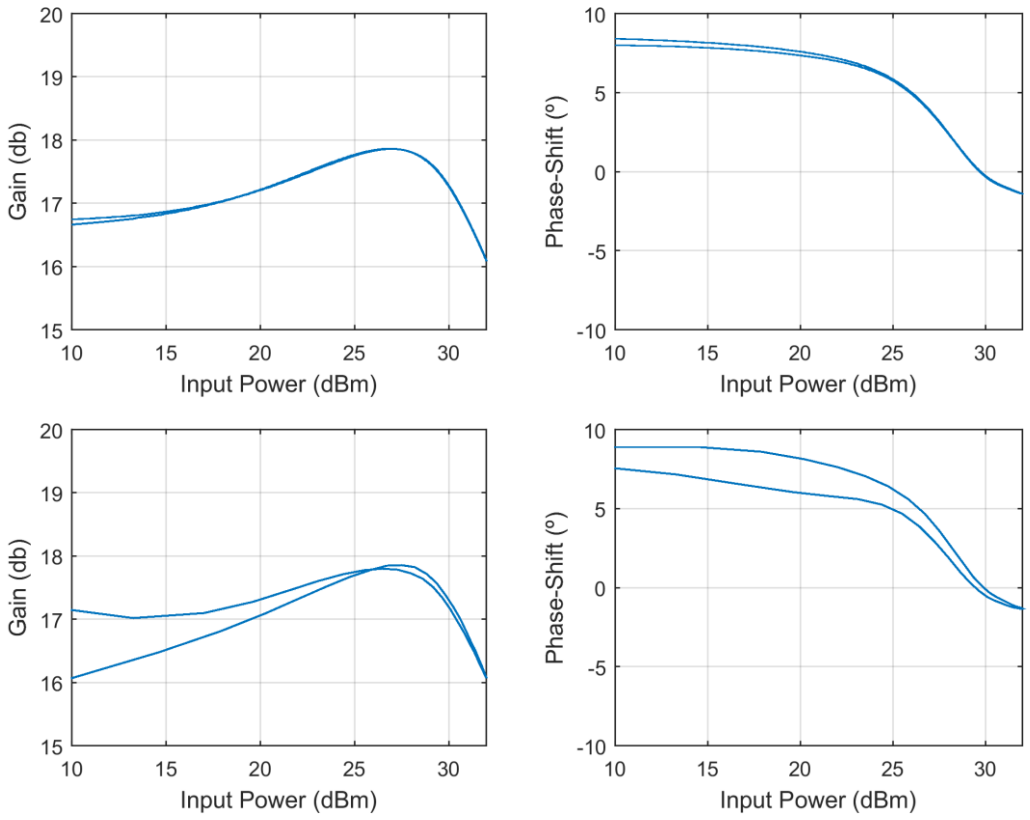


Fig. 1.17: Measured AM/AM, on the right, and AM/PM, on the left, of an MRF9045N LDMOS PA test board, using two narrow spaced tones, on top, (100kHz) and two widely spaced tones, on bottom, (5MHz).

Static and Short Term Effects

The power limitation of the transistors in an RFPA will force the eventual compression of its gain. This is typically the main cause for the nonlinear AM/AM conversion. The AM/PM, however, has more interesting root causes. In fact, it can be shown that a fully static system cannot have an AM/PM behavior. The description of AM/PM as a static effect comes from the observation of the transistor as a bandpass system, where the relationship is made between the modulation envelopes of the input and output carriers. The AM/PM behavior is due to the capacitances of the device. First, the nonlinear capacitances, C_{gs} and C_{ds}, change as the device is driven with different powers, changing the phase of the waves. Second the feedback capacitance, C_{gd}, will impact the AM/PM when the device is driven into compression due to the voltage gain variation at the transistor ports [51]. Effectively, the AM/PM is a short term behavior, which is so fast as to be static from the envelope's point of view.

In Doherty Power Amplifiers (DHTPAs) this effect is also felt when the peaking PA forces the main into voltage saturated operation. Another effect in this type of PAs is the combination of the waves in the combiner which generates AM/PM and AM/AM when the devices have different characteristics [52].

Other short term effects in the transistor are mainly due to the matching networks. The matching networks of typical PAs have a reasonably high coherence bandwidth. This being the case, these effects will manifest only for high-bandwidth signals, tens or hundreds of megahertz wide. This type of effects can be evaluated by obtaining the AM/AM and AM/PM behavior at several carrier frequencies since they are related to the variation of the networks around the central frequency and harmonics. As shown in Fig. 1.16 the transistor with these networks can be approximately interpreted as a Wiener-Hammerstein system, where the matching networks are the input and output filters. In this sense, the networks will shift the AM/AM and AM/PM characteristics of the device horizontally and vertically [53, 54].

Bias Effects

Bias effects are typically longer term than the effects introduced by the matching networks. A particular characteristic of these effects is that they are only excited by the bandwidth of the signal. In fact, the baseband current generated by the transistor for CW signals is at dc and does not excite the bias networks. Another particularity of these effects is their insensitivity to phase modulation, the phase modulation does not generate bias current variation and will not excite the bias networks.

For higher bandwidths, using amplitude modulated signals, the bias networks are excited and will cause the bias of the transistor to change according to some function

of the envelope amplitude. This will generate a dynamic bias of the transistor which will generate dynamic distortion at the fundamental due to changes in the saturation region [55, 56].

The fact that the bias effects are mainly sensitive to the envelope amplitude and the bandwidth only, is consistent with what was previously observed with the Volterra based models that mainly addressed bias memory, namely the VBW model.

Thermal and Trapping Effects

Thermal and Trapping effects introduce even longer time constants in RFPAs than bias effects [29, 31, 57]. Thermal effects are due to changes in the average dissipated power of the PA, which generate gain fluctuations. Thermal constants can go from millisecond to the microsecond range depending on the thermal resistance [58, 59, 60]. Trapping constants are in the same range for the new generation of the Gallium Nitride (GaN) devices, with the strong difference that trap charging is generally much faster than discharging, which generates unusual AM/AM and AM/PM observations [61].

These long term memory effects generate behavior fluctuations, dependent on the envelope. Thermal effects will typically be a lesser problem since, once thermal steady-state is achieved the temperature of the device tends to be relatively constant. However, this is dependent on the thermal resistances and capacitances of the PA. Trapping effects, on the other hand, work as a peak detector, due to the asymmetric time-constants, which make it harder to reach a steady-state. Overall, the high time-constants of these effects make them difficult to include in FIR based models, due to the number of coefficients required for this effect [62].

Trapping effects have been shown to have very slow time constants, but at the same time have a severe impact on the device's operation, as shown in Fig. 1.18, found in [61]. Thermal effects have also been identified as major role-players in the

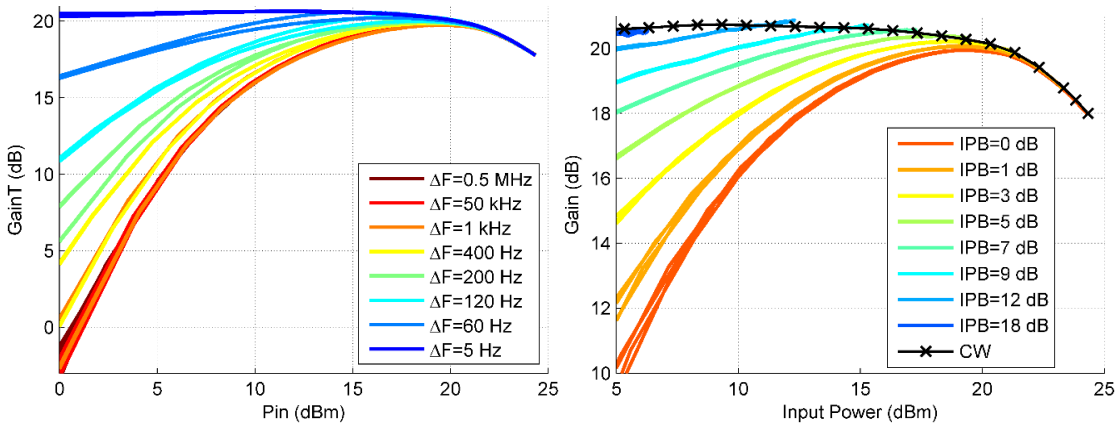


Fig. 1.18: GaN PA AM/AM responses for two-tone signals of different frequency spacing (left) and for the same frequency spacing with different peak powers (right), in [61].

response of high power GaN HEMTs, not only due to the static variation but also introducing important dynamics [63].

Similarly to the bias effects, CW testing cannot reveal the effects due to thermal and trapping behavior.

1.2.3. Digital Predistortion Methodology

Currently, the most popular DPD methods work in the digital LPE domain. This is achieved through an IQ demodulation and sampling in the feedback pass, as shown in Fig. 1.19. In this way, the required sampling frequency is related to the signal bandwidth around the carrier and is independent of the carrier frequency. For DPD, the sampling rate is traditionally three to five times the original signal bandwidth to accommodate the spectral regrowth. However, it has been noted that the sampling rate in the feedback path can be reduced when using direct learning for the DPD system [64, 65].

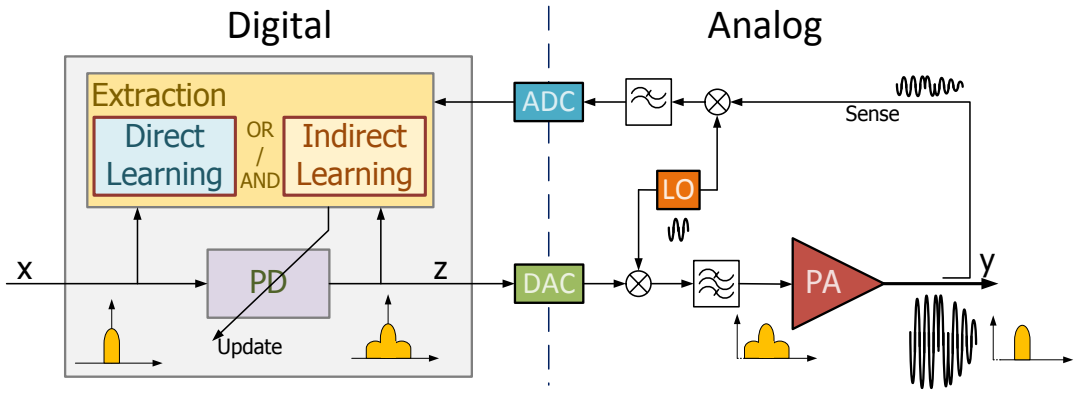


Fig. 1.19: General DPD setup, the predistorter works on the signal's complex envelope which is then translated to the proper carrier frequency, the feedback loop demodulates the signal and feeds it to the DPD system for updating.

The LPE domain represents the envelope of the carrier as an amplitude and phase signal at baseband. This means the observation is limited to the frequency zone around the fundamental carrier frequency, which is the zone of interest. The spectrum at other harmonics is assumed to be easily filtered before the signal is radiated.

It is important to understand correctly what the LPE domain represents since this imposes differences from traditional systems, in terms of model design and descriptions, for instance in terms of the nonlinear functions, [66, 43].

The DPD method works in the following way: some of the PA's output power is fed back but, instead of traditional feedback control, signal processing is used to generate, or update, an estimate of the PA's inverse. This technique achieves very accurate compensation up to significant bandwidths, in general, limited by the model's description capabilities. The severe cost is the vast increase of processing power it requires as well as the necessary high-speed for wide bandwidth compensation.

For each update, a significant length of signal must be captured to perform a correct identification and subsequent update of the predistortion (PD) model. The current PD estimate can be updated in one of two ways: direct or indirect learning, which are fundamentally different.

Direct learning attempts to estimate the PD directly, meaning, tries to tune the model coefficients so that it produces the signal required for the output of the PA to be the desired signal. This might be difficult to estimate because, even though there is an objective, the output of the PD is not known, simply the desired PD+PA output. A number of techniques that make use of different algorithms can be categorized as direct learning [10, 38, 65, 67, 68]. Direct learning can be reduced to a least squares estimation under some assumptions of the RFPA response [25].

Indirect learning approximates the PD by estimating the Postdistorter (PoD) [24], this technique is considerably simpler, since both the input and output of the PoD system are known. Generally, indirect learning estimation is performed using least squares or similar, but iterative, solvers: Least Mean Squares (LMS) or Recursive Least Squares (RLS).

Both techniques, direct and indirect learning, take several iterations to converge to the optimum PD. At each iteration, the input signal to the PA is modified, which will induce a variation of the PA's response. A possible cause for this is that the PA will stabilize around different "steady-state" conditions. Meaning, the average temperature, trapped charge, etc. is different at each iteration, before the PD converges. Even after the PD has stabilized in a particular solution, an update is performed every couple of minutes or tens of seconds to correct environment changes, PA aging and other very slow variations.

Generally, any of the previously described models can be used for DPD applications. To obtain higher linearity, solutions with several models working in parallel or cascade can also be found in the literature [69, 70, 39, 71].

From an implementation standpoint, polynomial models are avoided due to their inherent ill-conditioning [72, 73, 74, 75]. In fact, for DPD implementation, the use of LUTs is typically favored [76, 77]. Most Volterra based models can be converted to LUT representations using different function expansions [78, 32].

Since long-term memory effects have been singled out as problematic to include in FIR based topologies popular in DPD, the use of models with controlled coefficients has been put forward to manage their inclusion. This technique works with auxiliary models that generate control signals, which modify the original model coefficients. The main problem with this technique is developing the auxiliary models as well as determining how the coefficients are varied with the control signals [31, 79].

1.3. Main Objectives and Organization

As pointed out in the Background and Motivation, there currently exists a severe problem in the linearization of novel PA technology excited with modern communication signals. DPD is currently incapable of correctly compensating all of the nonlinear effects that are being excited.

The main linearization problem seems to be the new GaN transistors that have not only inherited previous limitations but also present new, long term, effects (trapping). In order to address the long term memory issue it is first necessary to understand what the current DPD models are capable of compensating. With this in mind, this thesis has been divided into several major objectives that will lead to the final goal of accurate linearization, including long term effects.

The objectives can be arranged in the following structure:

- Study of PA distortion mechanisms – the objective is to obtain a simple, conceptual representation of the PA's behavior that can be used to roughly understand the distortion sources in single-ended PAs.
- Study of current DPD models' compensation capability – taking the previous study into account, the objective is to understand if the used DPD models are adequate for PA distortion compensation. Selecting a DPD topology as a comparison standard for future work should be achieved here.
- Study, development and implementation of signal processing techniques for correct model extraction – the idea is to be certain that, when using each model, numerical pitfalls (ill-conditioning) are avoided. Moreover, the technical know-how for DPD implementation is also developed here.
- Study, development and implementation of models for accurate distortion compensation – this stage addresses the creation of new modelling techniques to accurately describe the long term behavior of the devices. The possibility to include the long term models as refinements of the typical models should be considered.

With these objectives in mind this thesis has been divided into multiple chapters as follows.

Chapter 1 provides the background for this work and its motivation, as well as the state-of-the-art on which it is based. In addition this chapter presents the objectives and main contributions of this thesis.

Chapter 2 presents an overview of PA distortion mechanisms and relationships between the models and these mechanisms. This chapter is intended as a support for the model selection in the subsequent chapters.

Chapter 3 focuses on the study of model robustness from a signal processing stand point, evaluating the polynomials as good model descriptors and suggesting different model descriptions to provide more reliable structures.

Chapter 4 focuses on using the previously acquired knowledge for PA characterization and modeling, providing methods to generate models from meaningful device measurements.

Chapter 5 focuses on RFPA nonlinear compensation, mainly long-term compensation using controllable models. A methodology for the development of the auxiliary models is suggested, as well as the extraction procedure for the newly developed models. Chapter 5 also proposes an alternative approach for PA compensation using a statistical approach with the objective of simplifying the measurements and modeling, and expediting the inversion.

Chapter 6 concludes this thesis summarizing the achievements and putting forth future research and development topics.

1.4. Main Contributions

This PhD work achieved contributions in:

1. Volterra model descriptions;
2. RFPA characterization and modeling;
3. RFPA nonlinear compensation.

On what Volterra model description is concerned, the work here developed presented methods to expand the nonlinear basis functions as spline interpolated LUTs [C1]; further developed this technique showing an expansion similar to LUTs using polynomial models [C2]; and showed that the interpolated LUTs and the polynomials can be joined under the same theory and provide similar model characteristics and approximation capabilities [J1]. The presented model modifications allowed achieving highly improved conditioning in the model extraction.

In terms of characterization and modeling of RFPAs, this thesis developed methods to transform large signal IMD measurements into models [C4, C5]. In one case, the information from the IMD measurements directly fits the model parameters making the model a useful characterization technique [C4], directly extracted from measurement data. In the other case, the model is expanded to include two memory types, direct (carrier sensitive) and cross (bandwidth sensitive) memory [C5], unfortunately the model loses orthogonality, making it less useful for characterization.

For RFPA nonlinear compensation this thesis has contributed with a technique for accurate compensation using cascaded models [M1], with the development of nonlinear models with long-term compensation [J2], and with a new approach to achieving predistortion using statistical information [C7].

Other works which this thesis has contributed to include a procedure for frequency selective predistortion [C3], the development of a new behavioral and predistortion model [C6], and a workshop on modeling and predistortion [W1].

Papers in Conferences

[C1] **F. M. Barradas**, P. M. Lavrador, T. R. Cunha and J. C. Pedro, "Using spline basis functions in Volterra series based models," in *2014 International Workshop on Integrated Nonlinear Microwave and Millimetre-wave Circuits (INMMiC)*, Leuven, 2014, pp. 1-3.

[C2] **F. M. Barradas**, T. R. Cunha, P. M. Lavrador and J. C. Pedro, "Higher locality non-linear basis functions of Volterra series based models to improve extraction conditioning," *2014 IEEE MTT-S International Microwave Symposium (IMS2014)*, Tampa, FL, 2014, pp. 1-4.

[C3] T. R. Cunha, **F. M. Barradas**, and J. C. Pedro, “DPD tuning with frequency selective distortion minimization,” *2015 IEEE MTT-S International Microwave Symposium*, Phoenix, AZ, 2015, pp. 1-3.

[C4] **F. M. E. Barradas**, P. M. Lavrador, T. R. Cunha and J. C. Pedro, "RF PA modeling with one chirp measurement," *2015 European Microwave Conference (EuMC)*, Paris, 2015, pp. 1200-1203.

[C5] **F. M. Barradas**, P. M. Lavrador, T. R. Cunha and J. C. Pedro, "Characterizing Direct and Cross Memory in RF Nonlinear Systems Using Simple Two Tone Measurements," *2016 European Microwave Conference (EuMC)*, London, 2016.

[C6] T. R. Cunha, **F. M. Barradas** and J. C. Pedro, "The Two-Tone Model for Power Amplifier Modeling," *2016 European Microwave Conference (EuMC)*, London, 2016.

[C7] **F. M. Barradas**, P. M. Lavrador, T. R. Cunha and J. C. Pedro, "Using Statistical Information for Fast Static DPD of RF PAs," *2017 IEEE Topical Conference on Power Amplifiers for Wireless and Radio Applications (PAWR),* (accepted).

Papers in Magazines

[M1] **F. M. Barradas**, L. C. Nunes, J. C. Pedro, T. R. Cunha, P. M. Lavrador and P. M. Cabral, "Accurate Linearization with Low-Complexity Models Using Cascaded Digital Predistortion Systems," in *IEEE Microwave Magazine*, vol. 16, no. 1, pp. 94-103, Feb. 2015.

Papers in Journals

[J1] **F. M. Barradas**, T. R. Cunha, P. M. Lavrador and J. C. Pedro, "Polynomials and LUTs in PA Behavioral Modeling: A Fair Theoretical Comparison," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 62, no. 12, pp. 3274-3285, Dec. 2014.

[J2] **F. M. Barradas**, L. C. Nunes, T. R. Cunha, P. M. Lavrador, P. M. Cabral and J. C. Pedro, "Compensation of GaN Long-Term Memory Effects on GaN HEMT Based Power Amplifiers," in *IEEE Transactions on Microwave Theory and Techniques*, (submitted)

Workshops

[W1] T. R. Cunha, P. M. Cabral, L. C. Nunes, **F. M. Barradas**, and J. C. Pedro, “Wideband power amplifier linearization: from active device distortion to DPD compensation,” workshop at the *IEEE Radio and*

Wireless Week, RWW2016, Austin, TX, January 2016.

2. Nonlinear Models for Radio Frequency Transmitters Behavioral Modeling and Digital Predistortion

In order to model and compensate the nonlinear operation observed in RF transmitters, the first step is to identify the main components responsible for this behavior. It is generally agreed that the PA is the main generator of nonlinear behavior, therefore, one of the initial concerns when developing this work was the study of PA distortion mechanisms, what could be observed due to these mechanisms and what the existing digital models could compensate and describe.

The main objective of this initial study is to provide support when choosing models to use in DPD and PA modeling. The focus of the distortion generation mechanisms is on single-ended PA architectures to simplify the analysis, avoiding active load or drain modulation effects.

This chapter is divided into two main parts. The first part is dedicated to the study of PA distortion mechanisms from a system level point of view, rarely going into circuitry details. These effects are divided into static and dynamic, with further partitioning of the dynamic effects into those caused by the bias and by the matching networks. The second part is dedicated to fitting the previously described effects into the PA behavioral models, attributing certain kernel formulations to particular effects.

2.1. Power Amplifier Distortion Generation Mechanisms

The study of the PA distortion generation mechanisms was divided into several parts. First, an analysis considering only the static behavior (for the signal envelope) of the transistor is performed. Since the transistor is considered to be the only nonlinear component of the PA, it is the sole responsible component for the static nonlinear behavior. The memoryless behavior is then expanded to include the changes in the PA due to the matching networks, bias networks and, finally, thermal and electron trapping behaviors. The division is performed in this way because the effects vary in terms of time-scale for each of these parts, starting with the faster responses to the slower responses of the PA.

To avoid studying the PA in terms of circuit theory, the study of the PA distortion mechanisms was mainly performed under the framework presented in [41, 46]. Initially, the PA is converted into an equivalent system-level model, divided into several blocks as shown in Fig. 2.1.

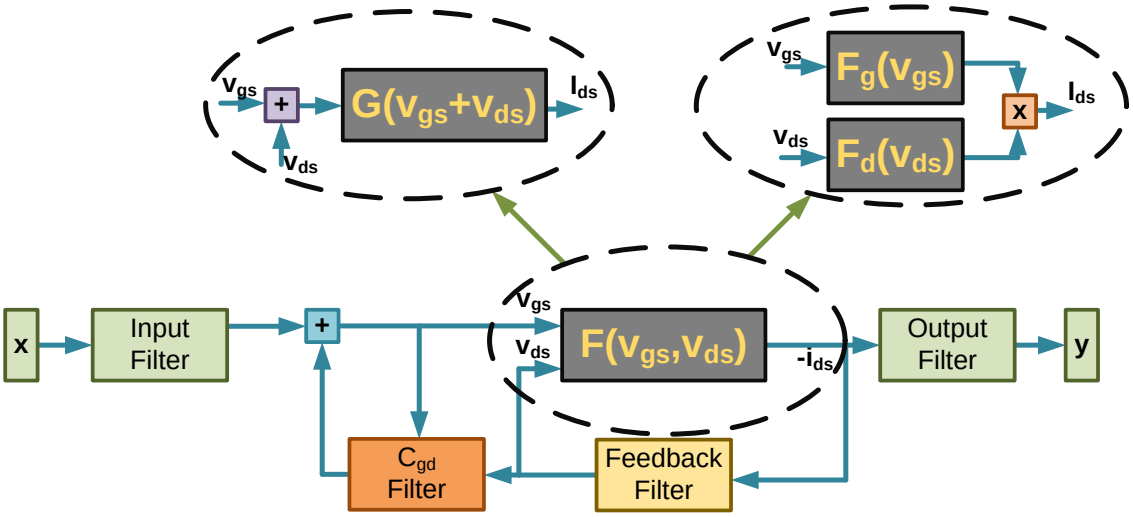


Fig. 2.1: System level behavioral model of a single-ended PA, typically either the sum, [41], or product, [46], are used to approximate the bi-dimensional function. The intrinsic and extrinsic elements of the transistor are incorporated into the filters, only the nonlinear current source is kept. Typically, the nonlinear bi-dimensional function is approximated by simpler expression to simplify the analysis.

In this model, it is assumed that the transistor is a quasi-static device, meaning that thermal and electron trapping behaviors are not formulated here. Furthermore, the intrinsic and extrinsic components of the transistor are grouped into the linear filters, this assumes that these intrinsic and extrinsic components are, themselves, linear.

In Fig. 2.1, the input filter reflects the change from the input voltage to the PA, and the voltage observed by the nonlinear current source. The output filter reflects the changes from the current produced by the nonlinear source, to the voltage output by the amplifier. The feedback and C_{gd} filters reflect the changes generated by the current produced from the nonlinear source, and the voltages sensed by the nonlinear current

source, which impact the current itself. Therefore, there are several processes included in this simplified amplifier model that should be able to account for many of the mechanisms of distortion observed in these devices.

2.1.1. Static Mechanisms

To analyze the static nonlinearity generation mechanisms, the input and output matching networks, as well as the bias networks, are assumed to have no memory for the envelope. The memory at the fundamental is preserved, since it is required to generate AM/PM behavior. In fact, fully static systems cannot generate AM/PM because the phase change requires the existence of capacitive or inductive loads. In envelope static systems, this memory is so fast that a sinusoidal steady-state is achieved instantaneously, from the envelope's perspective. Furthermore, it is assumed that the achieved steady-state is independent of the center frequency. This is a highly idealized scenario but can be approximately achieved within a given band of operation.

The steady-state the system settles into depends on the amplitude of the envelope. This amplitude will be responsible for shifting the phase and changing the gain of the overall system. The way this happens can be understood using the system-level model from Fig. 2.1. To remove the memory from the envelope it suffices that all the filters be constant in a sufficiently large bandwidth around the fundamental frequency and its harmonics. This bandwidth must be large enough to include the envelope bandwidth at the fundamental and at those harmonics. In this scenario, the filters can be represented by a gain and phase at each harmonic frequency, and the relation in (2.1) holds true for all filters.

$$\begin{aligned} x(t) &= \sum_{n=0}^{+\infty} a_n(t)\cos\big(n\omega t + \phi_n(t)\big) \\ y(t) &= \int_0^t h(\tau)x(t-\tau)dt = \sum_{n=0}^{+\infty} K_n a_n(t)\cos(n\omega t + \phi_n(t) + \theta_n) \end{aligned} \tag{2.1}$$

In (2.1), $x(t)$ is the input, $y(t)$ is the output and $h(t)$ is the impulse response of the filters, $a_n(t)$ and $\phi_n(t)$ define the time-varying envelope at each harmonic of the fundamental frequency.

Equation (2.1), shows that the filters can induce a gain and rotate the phase of each harmonic, but do not produce a change in the envelope of the harmonic. Consequently, the envelopes of the several signals in the system are synchronous and the system is memoryless for the envelope. The induced gains and phases for each harmonic are responsible for setting the class of operation of the PA, since they will

settle the time waveform of the currents and phases at the transistor ports.

Equation (2.1) shows that the output of the filters does not depend on the past history of the envelope. Therefore, this system can be analyzed for each envelope state independently. Assuming that the system is excited at the input only at the fundamental, if all the phases are referenced to the phase of the fundamental at the input port, the system can be analyzed only for varying amplitude of the input envelope signal. Ultimately, v_{gs}, v_{ds}, i_{ds} and y can be expanded into similar expressions, dependent only on the input envelope amplitude, as shown in (2.2), where $x(t)$ is the input to the system and $u(t)$ is the general expression into which the currents and voltages settle. For each harmonic, A_n is the AM/AM conversion and P_n is the AM/PM conversion. The main interest is in A_1 and P_1 of y.

$$x(t) = \mathrm{a}(t)\cos\big(\omega t + \phi(t)\big)$$
$$u(t) = \sum_{n=0}^{+\infty} A_n\big(\mathrm{a}(t)\big)\cos\Big(n\big(\omega t + \phi(t)\big) + P_n\big(a(t)\big)\Big) \tag{2.2}$$

Under small signal excitation, the nonlinear function provides conversion from v_{gs} to i_{ds}, and is insensitive to v_{ds}. This conversion may be nonlinear depending on the bias point of the PA but is dependent only on v_{gs} drive. As the v_{gs} signal drive increases, the i_{ds} current also increases, increasing also the magnitude of v_{ds}. For a sufficiently high v_{ds} signal, the current generation becomes sensitive to v_{ds} (due to reaching the triode region of the device), at this point the PA enters into saturation, reducing the current generation and, therefore, compressing the signal gain. This

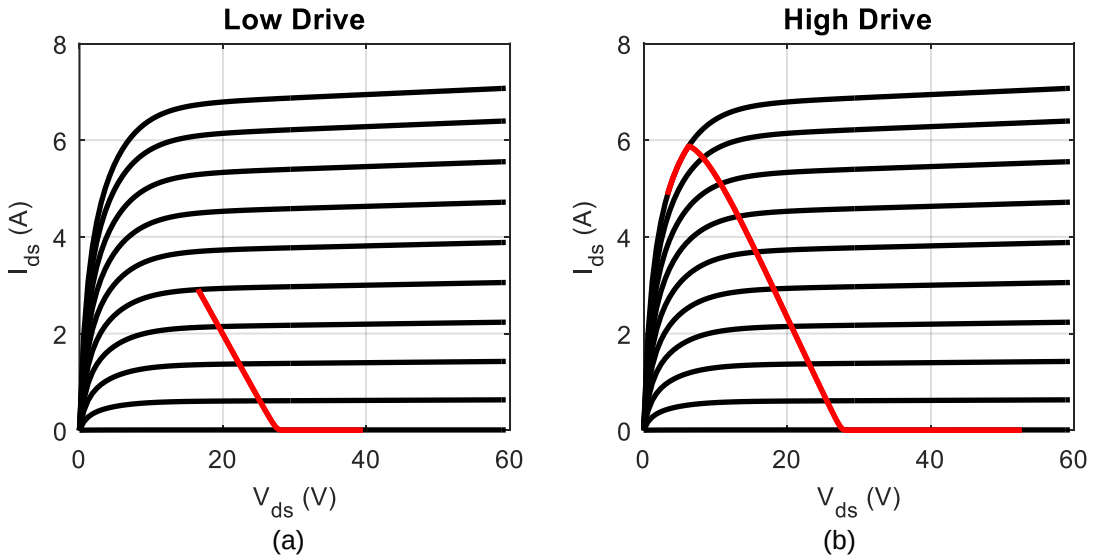


Fig. 2.2: Behavior of the current generation under low v_{gs} drive, (a), and high v_{gs} drive, (b), assuming the v_{ds} filter cuts-off all frequencies except the fundamental. The DC curves are plotted in black and the signal behavior in red.

behavior is shown in Fig. 2.2, assuming a typical current source equation (in annex a.1) for a typical transistor working in class B.

This compression of the fundamental current generation is the reason for AM/AM generation in the overall PA. The fundamental current output by the nonlinear source for varying input drive is shown in Fig. 2.3, under the same conditions from Fig. 2.2. As can be seen, for higher input drives the fundamental current enters into compression.

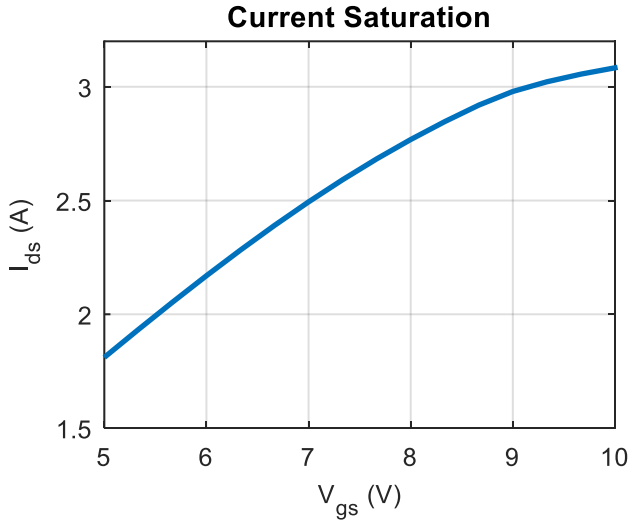


Fig. 2.3: Fundamental current from the nonlinear source for increasing gate fundamental voltage drive.

This current compression is also responsible for generating the AM/PM of the amplifier. In fact, under the system-level model from Fig. 2.1, the AM/AM and AM/PM are closely linked. The feedback process through the C_{gd} capacitor will generate a phase shift in the gate voltage, which is translated to the output of the PA [51]. Initially, when the amplifier operates in a linear mode, the increment in gate voltage due to the input and the increase in output current are proportional. However, as the transistor is driven into saturation the increase in input voltage exceeds the increase in current, leading to a phase-shift of the gate voltage towards the input phase, as shown in Fig. 2.4.

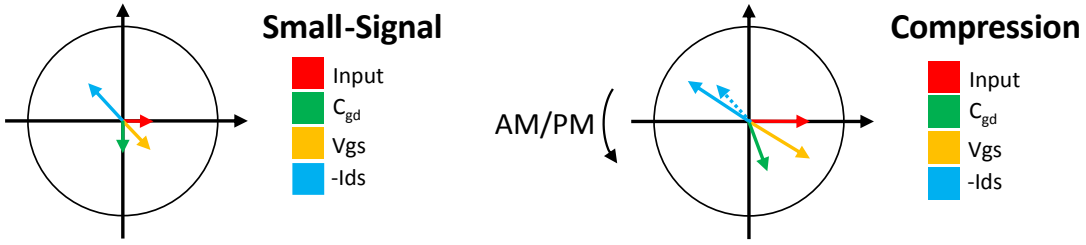


Fig. 2.4: AM/PM generation due to the feedback changing the phase of the driving voltage when in compression.

Other Static Mechanisms

In the physical devices the intrinsic capacitances of the transistor are generally not linear. In this case, the capacitances may contribute to changes in the gain and

phase-shift of the amplifier.

When the capacitances are nonlinear, as the signal drive increases, the equivalent linear capacitance value will change, changing the signal conversion from the input of the PA to the transistor drive, as well as the conversion from the transistor current into output voltage, in terms of phase and amplitude. Nonlinear capacitances are therefore also a source of AM/AM and AM/PM generation.

2.1.2. Matching Networks

In PAs, one of the memory inducing elements are the input and output matching networks. The matching networks vary in frequency, within the operation bandwidth, resulting in a change of the amplifier characteristics. The matching networks define the impedances at the fundamental and higher harmonics, the impedance around the DC component is defined by the bias network, which is explored in subsection 2.1.3. For the case of the matching networks two situations are worth noting: (1) the frequency variation of the networks is negligible in the bandwidth of the envelope, (2) the frequency variation of the networks is not negligible in the bandwidth of the envelope. The first situation can be seen as a narrowband case, and the second as a wideband case.

In the narrowband case, (1), the situation is very similar to the static case. Even though the networks are no longer invariant with the center frequency, they are still static for the envelope. In this case, (2.1) is no longer valid, but it is valid for a given operating center frequency, using this rationale (2.3) is obtained.

$$x(t) = \sum_{n=0}^{+\infty} a_n(t)\cos\big(n\omega t + \phi_n(t)\big)$$
$$y(t) = \int_0^t h(\tau)x(t-\tau)dt = \sum_{n=0}^{+\infty} K_n(\omega)a_n(t)\cos\big(n\omega t + \phi_n(t) + \theta_n(\omega)\big) \tag{2.3}$$

Again, according to (2.3), the output of the system is not dependent on the history of the envelope, so, according to the rationale developed for the static case, the system will also settle into a sinusoidal steady-state dependent on the instantaneous amplitude of the input envelope. The only difference is that the steady-state now also depends on the center frequency at which the PA is excited. For this situation, the currents and voltages of the amplifier will settle into expressions similar to (2.2) but also dependent on the center frequency, as shown in (2.4), which is a more reasonable approximation of the PA's behavior.

$$x(t) = \mathrm{a}(t)\cos\big(\omega t + \phi(t)\big)$$
$$u(t) = \sum_{n=0}^{+\infty} A_n(\mathrm{a}(t),\omega)\cos\Big(n\big(\omega t + \phi(t)\big) + P_n(a(t),\omega)\Big) \tag{2.4}$$

According to (2.4), and the rationale behind obtaining this expression, when measuring A_1 and P_1 of the output of the PA, a dependence on the amplitude and the center operating frequency should be expected (for a narrowband signal). This is what is observed when a PA is tested under CW operation for different carriers.

In the wideband case, (2), the matching networks generate memory in the bandwidth of the envelope. In this case, it can no longer be assumed that the filters only rotate and scale the harmonic frequencies produced by the nonlinear source. Because of the nonlinear feedback process and the fact that the envelope is stochastic (it transports information), it is difficult to evaluate how the output will behave. However, if the effect of the filters is limited, some expression can still be derived. To study the effects of the matching networks under the wideband stimulus case, it is assumed that the networks are sufficiently smooth in frequency so that (2.5) is valid (where v_x is representing either v_{gs} or v_{ds}.

$$i_{ds}(t) = F\Big(v_{gs}(t), v_{ds}(t)\Big) + \Delta i_{ds}(t)$$
$$\Delta i_{ds}(t) = \frac{\delta F}{\delta v_{gs}}\Big(v_{gs}(t), v_{ds}(t)\Big)\Delta v_{gs}(t) + \frac{\delta F}{\delta v_{ds}}\Big(v_{gs}(t), v_{ds}(t)\Big)\Delta v_{ds}(t) \tag{2.5}$$
$$\Delta v_x(t) = v_x(t) - \sum_{n=0}^{+\infty} A_n(\mathrm{a}(t),\omega)\cos\Big(n\big(\omega t + \phi(t)\big) + P_n(a(t),\omega)\Big)$$

For (2.5) to be valid the effect of the filters on the voltages at the transistor ports must be sufficiently small so that the behavior of the device can be interpreted as a linear deviation from the equilibrium point of the static case. Generally speaking, since the bandwidth of the signals is typically a small fraction of the center frequency and harmonics, this approximation may be valid.

In this scenario, the perturbations induced by the filter are transformed in the nonlinear current source along the differential relative to the static operation, for both voltages. These differentials vary with the instantaneous voltages of the static case and are themselves nonlinear functions. This means that even under small variations induced by the filters there will be conversion from higher harmonics to the fundamental.

The differential of the current function can be expanded into the formulation in

(2.4) since it is controlled by the static operation. Since the interest is in the relationship between the input fundamental envelope and the output fundamental envelope (a filtered version of the fundamental current envelope), the perturbation to i_{ds} can be calculated only at the fundamental. Expanding both the differentials and the voltage perturbations, as shown in (2.6), the mixture terms can be calculated. Note that, since the matching and bias networks have been differentiated, it is considered that the perturbations of the voltage do not exist around the DC component.

$$\frac{\delta F}{\delta v_x}\left(v_{gs}(t), v_{ds}(t)\right) = \sum_{n=0}^{+\infty} A_n(\mathrm{a}(t), \omega)\cos\left(n\left(\omega t + \phi(t)\right) + P_n(a(t), \omega)\right)$$

$$\Delta v_x(t) = \sum_{n=1}^{+\infty} b_n(t)\cos\left(n\left(\omega t + \phi(t)\right) + \beta_n(t)\right)$$

$$\begin{aligned}\frac{\delta F}{\delta v_x}\left(v_{gs}(t), v_{ds}(t)\right)\Delta v_x(t) &= \frac{1}{2}\sum_{n=0}^{+\infty}\sum_{k=1}^{+\infty} A_n(\mathrm{a}(t), \omega)b_k(t)\cos\Big((n+k)\left(\omega t + \phi(t)\right) \\ &+ P_n(a(t), \omega) + \beta_k(t)\Big) \\ &+ \frac{1}{2}\sum_{n=0}^{+\infty}\sum_{k=1}^{+\infty} A_n(\mathrm{a}(t), \omega)b_k(t)\cos\Big((n-k)\left(\omega t + \phi(t)\right) \\ &+ P_n(a(t), \omega) - \beta_k(t)\Big)\end{aligned} \tag{2.6}$$

From all the mixture terms in (2.6), only the ones at the fundamental are preserved in (2.7). These are the terms that will produce changes in the envelope of the fundamental, at the output. In (2.7) some dependencies have been omitted to avoid

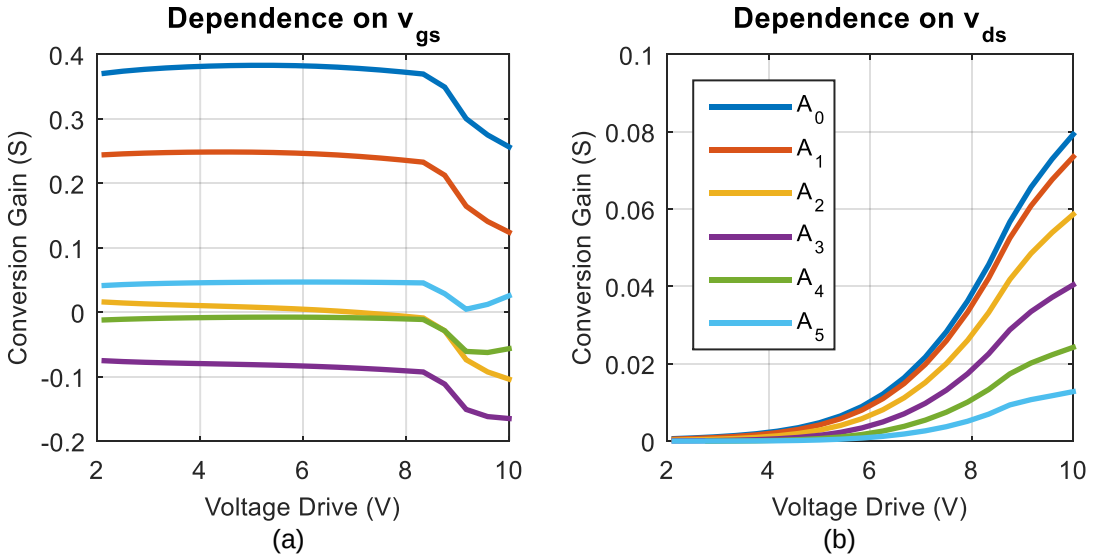


Fig. 2.5: Conversion gain from perturbations in v_{gs}, (a), and v_{ds}, (b), to the fundamental i_{ds} current as a function of the input drive for class B operation.

cluttering the equation.

$$\left.\frac{\delta F}{\delta v_x}\Delta v_x(t)\right|_{fund} = A_0 b_1(t)\cos\left(\omega t + \beta_1(t)\right)$$
$$+\frac{1}{2}\sum_{n=2}^{+\infty} A_n b_{n-1}(t)\cos\left(\omega t + \phi(t) + P_n - \beta_{n-1}(t)\right) \quad (2.7)$$
$$+\frac{1}{2}\sum_{n=1}^{+\infty} A_n b_{n+1}(t)\cos\left(\omega t + \phi(t) - P_n + \beta_{n+1}(t)\right)$$

Using the transistor functions in annex a.1 and assuming a class B operation the A_n coefficients from (2.7) can be calculated for both the v_{gs} and v_{ds} dependence, as shown in Fig. 2.5.

It is also important to understand if the source is generating harmonics of the current which are filtered to obtain the harmonics of the voltages. Using the same transistor functions the harmonic generation for class B operation can be calculated, this is shown in Fig. 2.6.

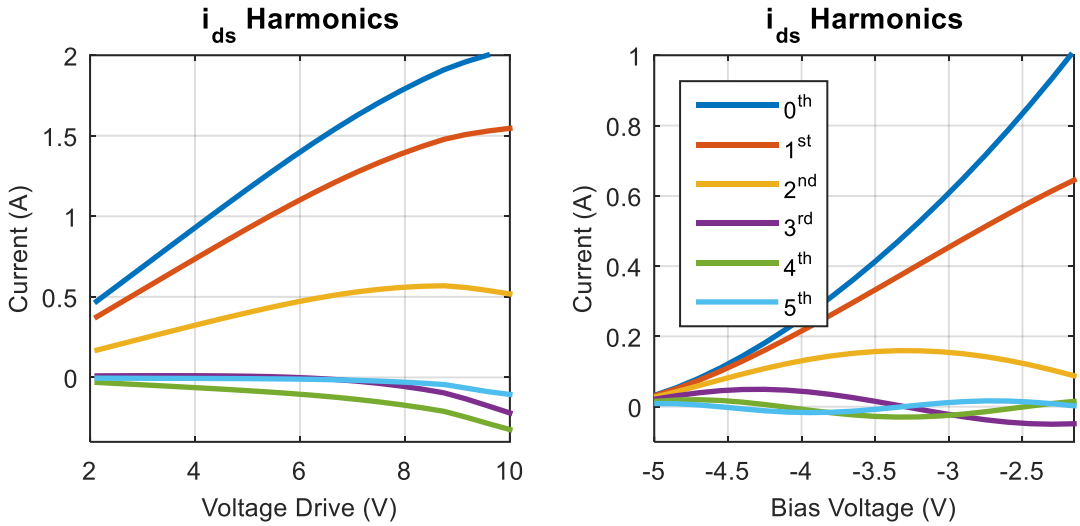


Fig. 2.6: Current generation at several harmonics depending on input drive, (a), when in class B operation (-3.42V gate bias); and depending on bias voltage, (b), when in small-signal (2V drive).

From Fig. 2.5 it can be seen that the main contributor from converting perturbations in the voltage to perturbations in the output current is the A_0 coefficient related to the gate voltage. The current is mostly insensitive to perturbations in the drain voltage, even when the PA is driven into saturation.

From Fig. 2.6 it can be understood that the main harmonic component generation is around DC and at the fundamental. The second harmonic may be important in the Class B case but harmonics above the second have a significantly smaller current.

Taking into account the results presented in Fig. 2.5, (2.5) can be shortened by

ignoring the contribution of v_{ds}, since it is negligible when compared to changes in the gate voltage. Furthermore, from Fig. 2.5, Fig. 2.6 and (2.7), the contribution of A_0 is much higher than the contribution of all the other coefficients even in saturation. To further support this claim, Fig. 2.7 shows the variation of the coefficients with the bias voltage in small signal, when moving into class A the dependence on A_1 is reduced, while the dependence on A_0 continues to increase.

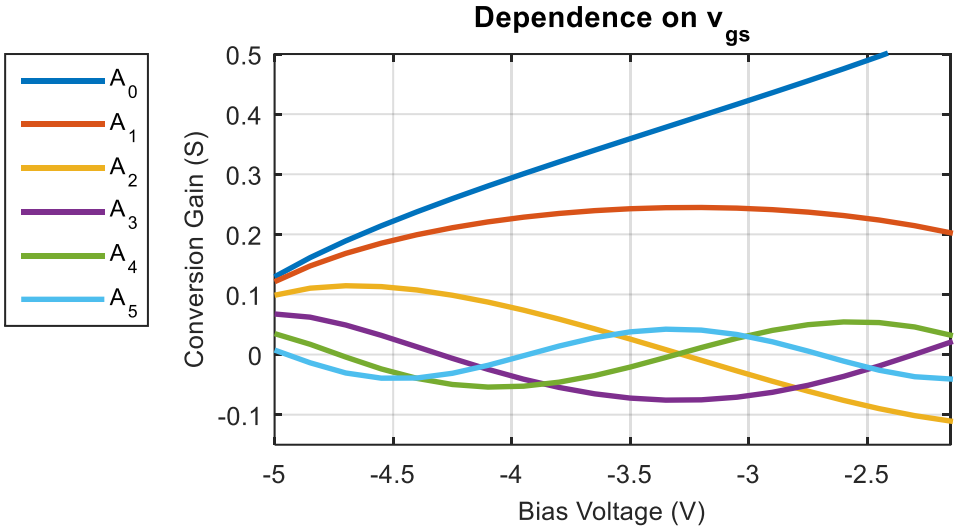


Fig. 2.7: Dependence of the conversion gain from perturbation on the gate voltage to perturbations of the drain current for varying bias voltage, when in small signal (2V drive).

From the observations in Fig. 2.5, Fig. 2.6 and Fig. 2.7, and taking into account (2.7) the main influence in the fundamental of the output current, due to the matching networks, is created by the perturbation of the fundamental of the gate voltage (unless the bias move deeply into class C where higher harmonics of the gate voltage start to be mixed down to the fundamental).

Now it is necessary to relate the gate voltage to the input excitation. The gate voltage is controlled by the input and the drain current due to the feedback loop generated by the drain to gate capacitance. While linearly filtering the input will not generate spectral regrowth, the drain current's envelope is already distorted. This means that, generally speaking, the envelope of the gate voltage at the fundamental is a distorted, filtered version of the envelope at the input, as shown in (2.8), where $\tilde{x}_n(t)$ is the complex envelope of $x(t)$ at the nth harmonic.

$$v_{gs}(t) = \int_0^t h_i(\tau)x(t-\tau)d\tau + \int_0^t h_{c_{gd}}(\tau)i_{ds}(t-\tau)d\tau$$

$$v_{gs}(t)\big|_{fund} = Re\left(\left(\int_0^t \tilde{h}_{i_1}(\tau)\tilde{x}_1(t-\tau)d\tau + \int_0^t \tilde{h}_{c_{gd_1}}(\tau)\tilde{\imath}_{ds_1}(t-\tau)d\tau\right)\exp(j\omega t)\right) \quad (2.8)$$

$$\tilde{\imath}_{ds_1}(t) = A_1\big(a(t)\big)\exp\Big(j\Big(\phi(t) + P_1\big(a(t)\big)\Big)\Big)$$

For small perturbations induced by the filters the perturbation of the fundamental drain current can therefore be reduced to (2.9), where $\Delta h(t) = h(t) - h(t)\delta(t)$ are the filters without the static component.

$$\widetilde{\Delta\imath}_{ds_1}(t) = A_0^{\frac{\delta F}{\delta v_{gs}}}\left(a(t)\right)\left(\int_0^t \widetilde{\Delta h}_{i_1}(\tau)\tilde{x}_1(t-\tau)d\tau + \int_0^t \widetilde{\Delta h}_{c_{gd_1}}(\tau)\tilde{\imath}_{ds_1}(t-\tau)d\tau\right) \quad (2.9)$$

To avoid falling into another nonlinear feedback loop due to the dependence of the drain current perturbation with itself, it is admitted that this process is dominated by the drain current imposed for the static case. Another approach is to approximate the nonlinear coefficient A_0 by a constant.

2.1.3. Bias Networks

The bias networks produce dynamic shifts of the bias of the transistor according to the envelope excitation. One of the special characteristics of these networks is that, since the memory is around DC, the filters are always frequency conjugate unlike the filters of the matching networks (from the envelope's perspective). Another characteristic is that the bias networks tend to produce longer time constants than matching networks because the energy storage components associated to these networks are substantially larger.

The source that excites the bias network is the generated DC component of the drain current. This component is controlled by the input envelope amplitude, however, it depends on the bias of the transistor, as shown in Fig. 2.8.

As seen in Fig. 2.8, the transistor will produce higher variations in DC current when it is moved towards Class C. In fact, when biased in Class A the variation of the DC current is zero before the transistor enters into the triode region. This indicates that

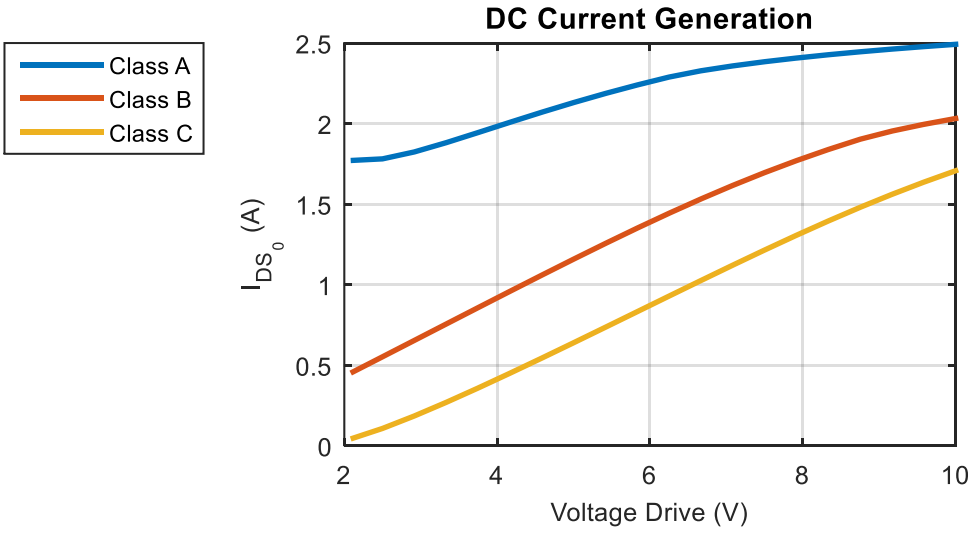


Fig. 2.8: Dependence of the DC current generation with the transistor gate voltage bias and gate voltage drive.

biasing a transistor closer to class A can reduce the bias memory of the transistor, simply because less variation of the DC current is generated even though the bias current is higher.

The feedback to v_{gs} from the DC current should be negligible since the transistor is isolated from drain to gate at DC and the feedback capacitance should be sufficiently small not to allow low frequencies through. The impact of the DC current perturbation is, therefore, limited to changing the DC drain voltage of the transistor.

As noted before, the transistor is mostly insensitive to perturbations of the v_{ds} voltage. This means that the impact of the bias networks should be limited to more demanding linearity goals, unless the induced variation is very high. Furthermore, it also means that the feedback of a variation of the drain current onto itself, through a variation of the drain voltage, should be very small. Fig. 2.9 shows the generation of DC drain current under static operation for several drain voltage bias for varying input drive, as can be seen in this figure, the drain current generation is mostly insensitive to the drain voltage, and therefore the feedback is confirmed to be very small. Despite all this, and mainly in the compression region, the transistor shows sufficient sensitiveness to the drain voltage to impact linearity. This impact is more evident as a demand for higher efficiency forces PA operation in the saturation region, as well as an increase of signal bandwidth excites these networks farther in frequency. Due to these facts, bias networks can be sources of troublesome nonlinear memory behaviors, typically observed as gain changes in the high power regions of the PA.

Taking into account Fig. 2.9, the induced perturbation of the drain voltage at DC can be directly calculated, to a good approximation, from the static $\tilde{\imath}_{ds_0}$, which is a function of the input fundamental envelope, as shown in (2.10). The envelope at the

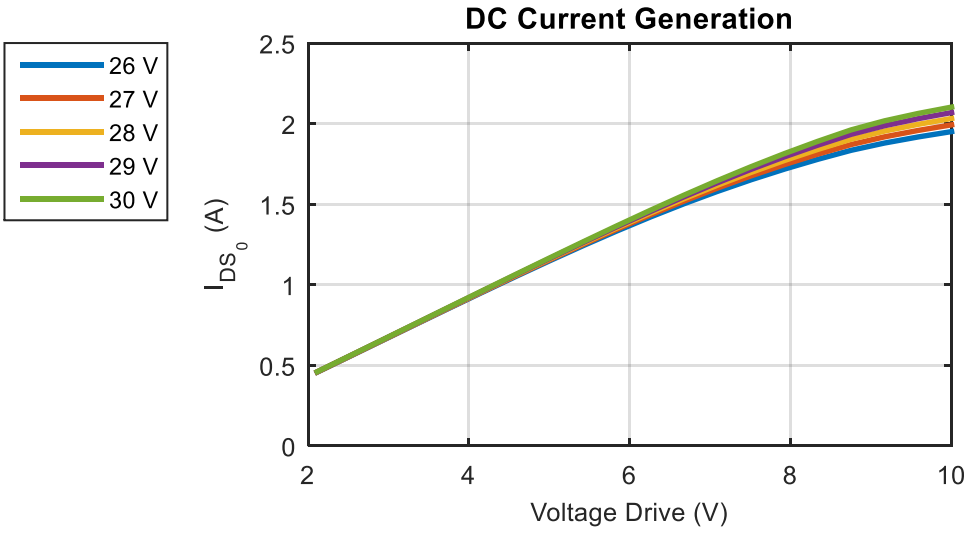


Fig. 2.9: Dependence of the DC current generation with the transistor drain voltage bias and gate voltage drive for Class B operation.

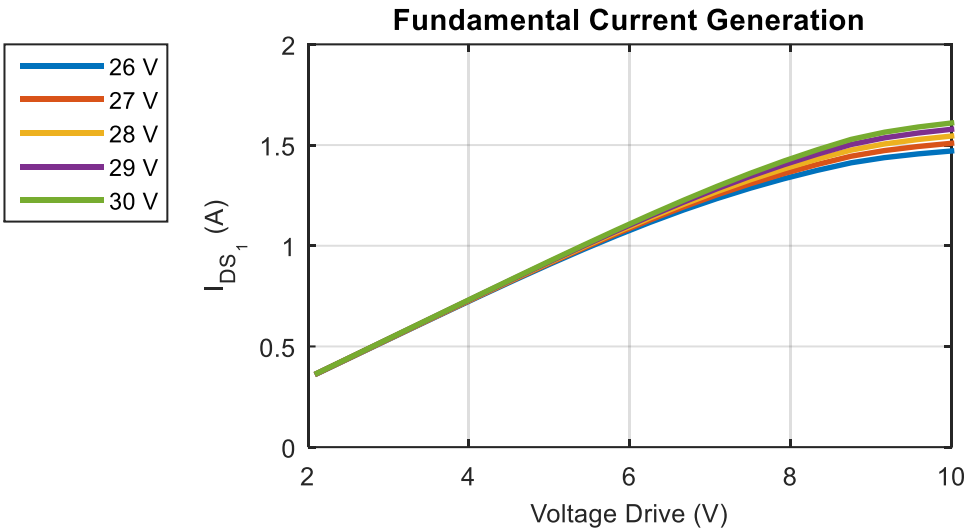


Fig. 2.10: Dependence of the fundamental current generation with the transistor drain voltage bias and gate voltage drive for Class B operation.

fundamental can then be calculated from this DC $\tilde{v}_{ds_0}$.

$$\tilde{v}_{ds_0}(t) = \int_0^t \tilde{h}_{f_0}(\tau)\tilde{\imath}_{ds_0}\big(a(t-\tau)\big)d\tau \tag{2.10}$$

To calculate the envelope of the current at the fundamental, first admit that the current generation function is the product of two functions, one dependent on the gate voltage and the other dependent on the drain voltage, as shown in (2.11).

$$i_{ds}(t) = F\left(v_{gs}(t), v_{ds}(t)\right) = F_g\left(v_{gs}(t)\right)F_d\big(v_{ds}(t)\big) \tag{2.11}$$

Under the approximation of (2.11) the changes around DC of the drain voltage are mixed up to the fundamental through the gate voltage function. In this scenario, the effect on the fundamental current of the bias networks can be represented as shown in (2.12).

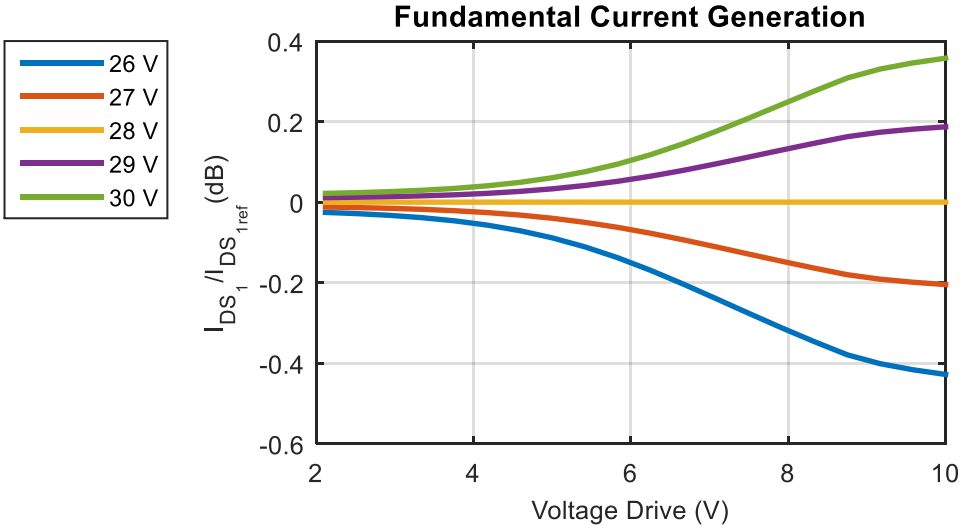


Fig. 2.11: Variation of the fundamental current versus the current for the center biasing condition with the transistor drain voltage bias and gate voltage drive for Class B operation.

$$\tilde{\imath}_{ds_1}(t) = F_{g_1}\left(v_{gs}(t)\right) F_{d_0}\left(\int_0^t \tilde{h}_{f_0}(\tau)\tilde{\imath}_{ds_0}\left(a(t-\tau)\right)d\tau\right) \quad (2.12)$$

One particularity of the described bias network effect is that, since it does not impact the gate side, the changes to the transistor current are more dominant for high input amplitudes. This is because, for low amplitudes, changing the drain voltage does not impact the current generation. In Fig. 2.10 the fundamental current is shown as a function of input drive for various drain voltages, showing this effect.

As seen in Fig. 2.9 and Fig. 2.10, the fundamental current change is slightly higher than the DC current change, to have a clearer idea of the expected variation, Fig. 2.11 shows the variation in dB versus the center biasing condition (28 V drain voltage). From Fig. 2.11 a variation of 2V in the DC drain voltage of the transistor creates 0.8dB of variation of the fundamental current under large signal operation.

Other Bias Networks Concerns

At the gate side, the bias network may also be excited due to DC current generation by the nonlinear transistor gate impedance. If the bias network at the gate presents a very high impedance, these small perturbations may produce an impact in the PA behavior. In the case of non-insolated gates the produced current may generate a bias shift towards Class C.

2.1.4. Thermal and Electron trapping

Thermal and electron trapping effects are not included in the system-level model from Fig. 2.1. However, these effects can impact the current generation of the transistor, therefore, it is interesting to gain some knowledge as to how they operate.

Thermal effects are dependent on the power dissipated in the transistor, which is

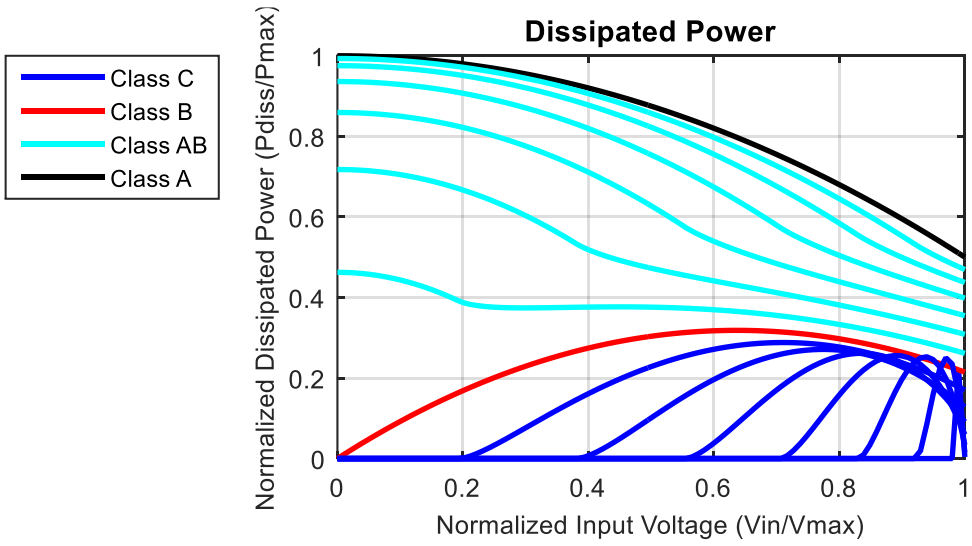


Fig. 2.12: Power dissipated in the transistor as a function of the input voltage drive.

heavily dependent on the class of operation. The normalized dissipated power for varying amplitude and class of operation, considering optimum power loads, is shown in Fig. 2.12, calculated according to (2.13).

$$
\begin{aligned}
P_{dc} &= \frac{1}{\pi}\left(v_{in}\sin\left(\frac{\xi}{2}\right) - \cos\left(\frac{\theta}{2}\right)\frac{\xi}{2}\right) \\
P_L &= \frac{R_L}{2\pi^2}\left(\frac{v_{in}}{2}(\xi + \sin(\xi)) - 2\cos\left(\frac{\theta}{2}\right)\sin\left(\frac{\xi}{2}\right)\right)^2 \\
P_{diss} &= P_{dc} - P_L \\
R_L &= \frac{2\pi}{\theta - \sin(\theta)} \\
\xi &= \begin{cases} 2\cos^{-1}\left(\frac{1}{v_{in}}\cos\left(\frac{\theta}{2}\right)\right), v_{in} \geq \left|\cos\left(\frac{\theta}{2}\right)\right| \\ \pi\left(1 + \text{sign}(\theta - \pi)\right), v_{in} < \left|\cos\left(\frac{\theta}{2}\right)\right| \end{cases}
\end{aligned}
\tag{2.13}
$$

As shown in Fig. 2.12 the dissipated power varies with the amplitude of the envelope of the carrier. For Class A, the dissipated power has a purely quadratic dependence on the amplitude and, for Class B, it has a linear and quadratic dependence, on the amplitude of the envelope.

The dissipated power conversion into temperature is a low pass process, as shown in Fig. 2.13. This process is dependent on how the temperature can flow from the transistor's internal junction to the environment. Looking at Fig. 2.13, the power dissipated in the transistor acts as a heat flow source that will dissipate through the junction, to the package and the sink, reaching the environment. Each interface will have its own associated thermal time constant and will reach a slightly different temperature.

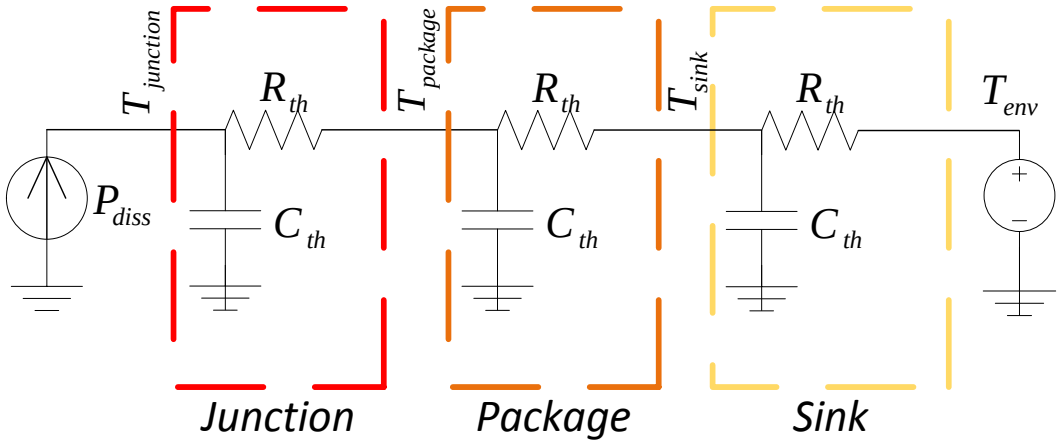


Fig. 2.13: Thermal circuit showing how the dissipated power impacts the temperature of the device, the temperature will then impact how the transistor generates current.

One difference between the generated temperature changes, for different transistor classes, is that Class B operation generates a power dissipation with lower frequency content than Class A, since in Class A the dissipated power is a purely quadratic function of the input excitation. This means that, the change in operation temperature in Class A should be smaller than in Class B. Once again, Class A operation shows more resilience to these memory effects.

Another effect that has shown impact in recent transistor technology, namely GaN transistors, is electron trapping. Electron trapping is a long-term memory process that creates changes in the transistor current generation. Depending on the signal this process may reveal itself as an apparent change of the transistor bias, with very slow

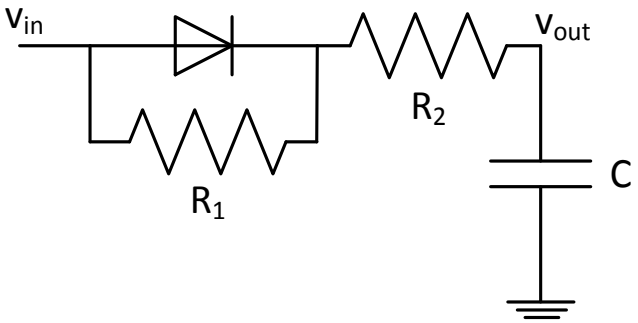


Fig. 2.14: Trapping circuit equivalent, the diode is responsible for generating a different charging and discharging behavior of the capacitor, the voltage in the capacitor impacts the current generation of the transistor. Typically the voltage is considered to cause a change of the threshold voltage.

recovery. One of the characteristics of this process is that the recovery (de-trapping) takes longer than the setting (trapping). A typical approach, in circuit models, to model this behavior is a circuit similar to the one shown in Fig. 2.14 [80].

Two trapping mechanisms have been identified in transistors, gate and drain trapping, depending on the controlling voltage. Both effects have an impact on the threshold voltage of the transistor or, equivalently, in the gate voltage. Because these effects are very slow, their impact is a slow variance of the class of operation over time depending on the excitation. In recent years, the gate related effects have mostly been eliminated and so, the remaining effects should mainly be controlled by the drain voltage [81].

2.1.5. Physics Effects Wrap Up

This section on nonlinear effects observed in PAs should be support for the developments in modeling and predistortion developed throughout this thesis. As so, a number of concepts should be retained from this initial exposition.

First, starting in a narrowband approximation, PAs are mostly static devices sensitive only to the carrier frequency. This means that PAs settle into a steady-state sinusoidal operation depending on the input instantaneous amplitude and carrier

frequency but are otherwise insensitive to the envelope. The change with the carrier frequency is due to the matching networks changing the system impedances with frequency.

As the bandwidth is increased, the matching networks will change sufficiently within the excitation bandwidth to react to the envelope. Furthermore, the bias networks are also excited by the variation of the current consumption. In this scenario, the PA should no longer behave as a static function. Since the matching networks are typically operated within a small fractional bandwidth, the change of behavior associated with these networks should be a small perturbation of the steady-state imposed in the narrowband case. For the bias networks, this claim may be false, the excitation of the bias networks will produce a variation of the transistor drain bias which will essentially impact the large signal gain. Observing the AM/AM of a transistor, changes that are restricted to the high power zone of the gain should be due to the bias networks while changes throughout the gain curve can be due to the matching networks.

Other effects that may impact PA behavior are temperature and electron trapping. Temperature effects have a low pass dependency on the dissipated power, which is itself determined by the amplitude of the envelope in the drain voltage. Similarly, the trapping effects show a low-pass dependency on this voltage, the difference being the existence of two time-constants, one for the charging and another for the discharging processes.

2.2. Volterra based behavioral model description mechanisms

Since the main interest in this work is in describing and correcting the mapping from the fundamental envelope at the input to the fundamental envelope at the output, the models are described in the LPE domain.

One of the most widely used model formulations in this domain is the Volterra series and its pruned versions. This was also the chosen formulation for this work because of its numerous advantages. First, the series is polynomial based which eases the analysis and the understanding of models. Second, the series is linear in the parameters, it is a linear combination of nonlinear basis functions, and this simplifies the extraction procedures and the transformations of the models. Third, the series has been proven to be a general descriptor of nonlinear models. Fourth, the several models developed from the series have been successfully used in modelling and DPD in the past. The original Volterra series is shown in (2.14).

$$y(t) = h_0 + \sum_{p=1}^{P} y_p(t)$$
$$y_p(t) = \int_0^t \dots \int_0^t h_p(\tau_1, \dots, \tau_p) \prod_{n=1}^{p} x(t-\tau_n)\, d\tau_n \tag{2.14}$$

This series can be converted into an LPE model to fit our description goals, taking the form of (2.15), where $\tilde{x}(t)$ is the envelope of the input at the fundamental, $\tilde{h}_p(\tau_1, \dots, \tau_p)$ are the envelope of the filters (nonlinear kernels) also at the fundamental, and $\tilde{y}(t)$ is the envelope of the output at the fundamental.

$$x(t) = \frac{\tilde{x}(t)e^{j\omega t} + \tilde{x}^*(t)e^{-j\omega t}}{2}$$
$$\tilde{y}_{2p-1}(t) = \int_0^t \dots \int_0^t \tilde{h}_{2p-1}(\tau_1, \dots, \tau_{2p-1}) \times \prod_{r=1}^{p} \tilde{x}(t-\tau_r) \prod_{k=p+1}^{2p-1} \tilde{x}^*(t-\tau_k)\, d\tau_1 \dots d\tau_{2p-1} \tag{2.15}$$
$$\tilde{y}(t) = \sum_{p=1}^{P} \tilde{y}_{2p-1}(t)$$

When analyzing the models in the envelope domain, the fundamental frequency is down converted to DC. In this sense, there is always the underlying assumption that the device is operating at a certain frequency and the envelope is the modulation at

that frequency. However, the models can still be excited at other frequencies simply by using a complex exponential as the envelope. The fundamental frequency in this section is assumed to be the center frequency of operation of the device, changes to this frequency are always included in the envelope.

In terms of DPD and behavioral modeling the models are typically converted into discrete time, for use with discrete time simulation and digital signals. Furthermore, the extraction of the models is also typically done in the discrete time domain. Converting (2.15) to discrete time, (2.16) is obtained.

$$\tilde{y}_{2p-1}(n) = \sum_{m_1=0}^{n} \dots \sum_{m_{2p-1}=0}^{n} \tilde{h}_{2p-1}(m_1, \dots, m_{2p-1})\tilde{x}(n - m_{2p-1}) \times \prod_{r=1}^{p-1} \tilde{x}(n - m_r) \prod_{k=p}^{2p-2} \tilde{x}^*(n - m_k) \quad (2.16)$$

$$\tilde{y}(n) = \sum_{p=1}^{P} \tilde{y}_{2p-1}(n)$$

In this section, a number of models is analyzed and compared to the previously explored effects to gain some insight into the relationship between the physical effects and the models' description capabilities. The models have been separated into three categories. Initially, the static formulation is studied to establish the basic formulation, dynamics are then progressively added to the model in several formulations found in the literature. Finally, a methodology that has recently been adopted to describe long term effects is explored.

2.2.1. Static Models

To begin the study of nonlinear models one of the approaches is to start in static models. Static models cannot describe any dynamic behavior, using the Volterra formalism from (2.16) this means that the Volterra kernels are impulses and (2.17) is obtained.

$$\tilde{h}_{2p+1}(\vec{m}) = \begin{cases} \tilde{h}_{2p+1}, \vec{m} = \vec{0} \\ 0 \quad\quad , \vec{m} \neq \vec{0} \end{cases}$$
$$\tilde{y}(n) = \sum_{p=0}^{P-1} \tilde{h}_{2p+1} x(n)|x(n)|^{2p} \quad (2.17)$$

As explained in section 2.1.1 in a static approximation the PA's input-to-output envelope conversion is reduced to a nonlinear AM/AM characteristic and AM/PM

characteristic, which means that the output amplitude and phase-shift are functions of the input amplitude. Separating (2.17) into phase and amplitude we can observe that this model can approximate this behavior, as shown in (2.18). Where the error of the approximation will depend on the number of coefficients and the precision with which they can be extracted.

$$\tilde{y}(n) = x(n)\sum_{p=0}^{P-1}\tilde{h}_{2p+1}|x(n)|^{2p} = \left(f_r(|x(n)|) + jf_i(|x(n)|)\right)x(n) \tag{2.18}$$

According to (2.18), the Volterra series static model can generate a complex gain where the real and imaginary parts are independently generated. Effectively this translates into an independent generation of AM/AM and AM/PM. The polynomials can be replaced by other basis functions, as long as the chosen set of functions is a global approximant to the group of continuous functions. A static model has the block diagram shown in Fig. 2.15.

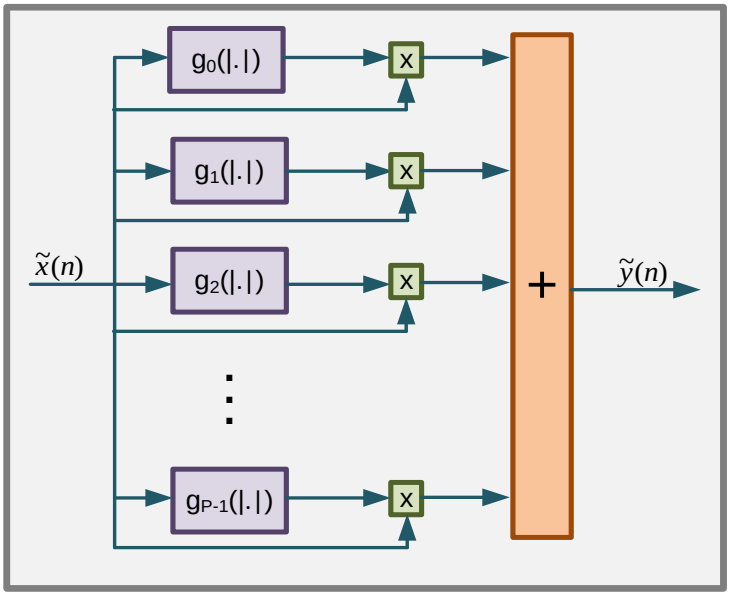


Fig. 2.15: Static model block diagram. The Volterra polynomials have been replaced with generic basis functions.

The static model is of limited application in modeling and DPD, but can be used as an initial approximation of the PA behavior. Observation of the static characteristic can also provide some insight into the required number of coefficients for the nonlinear part of the models. For instance, a PA showing steep variations of gain will need more coefficients for a good approximation.

2.2.2. Dynamic Models

Static models can be used to correct a PA's behavior in a very narrow band and/or when the linearization goal is not very demanding. Typically, a static model will

not be sufficient in modern communication scenarios, where bandwidths have grown considerably. In these cases, memory must be introduced in the DPD models. A general nonlinear model with memory would be a complete Volterra series, which is many times inappropriate for signal processing, due to a very high number of coefficients and low orthogonality (high condition number) for most signals.

A number of models with memory have been proposed in the literature to bypass the exceedingly high number of coefficients in a full Volterra series. These models are a truncation of the original series according to some rational. In this section, these models are explored and compared to the previously described device behavior.

Memory Polynomial

The memory polynomial was originally proposed in [10], and remains one of the most popular models with memory for modelling and DPD, greatly due to its simplicity and generally good performance. The memory polynomial model is defined as shown in (2.19).

$$\tilde{h}_{2p+1}(\vec{m}) = \begin{cases} \tilde{h}_{2p+1}(m) & , \vec{m} = (m, \dots, m) \\ 0 & , \vec{m} \neq (m, \dots, m) \end{cases}$$
$$\tilde{y}(n) = \sum_{p=0}^{P-1} \sum_{m=0}^{M-1} \tilde{h}_{2p+1}(m) \tilde{x}(n-m) |\tilde{x}(n-m)|^{2p} \tag{2.19}$$

As seen in (2.19), the memory polynomial adds memory in a simple way by reducing the multidimensional kernels to one-dimensional, considering only the same displacements along all dimensions.

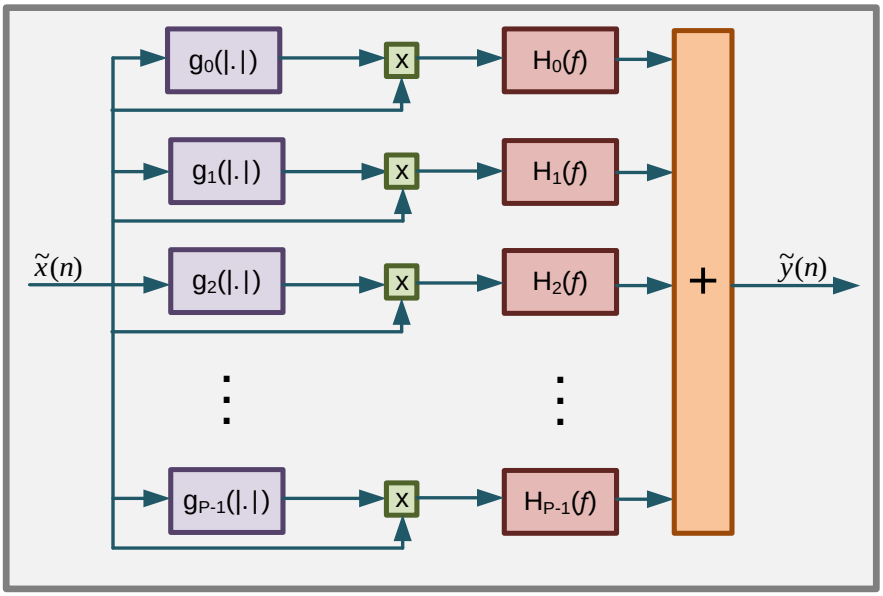


Fig. 2.16: Memory polynomial model block diagram. The Volterra polynomials have been replaced with generic basis functions.

The memory polynomial model has a block diagram as shown in Fig. 2.16, in this case, the polynomial functions have been replaced with generic basis functions, the conversion from (2.19(2.19) to the representation in Fig. 2.16 is included in annex a.2.

As seen in Fig. 2.16, the memory polynomial defines a set of P filters, one for each nonlinear branch, responsible for inducing the memory behavior of the model. The static and memory polynomial models can thus be easily related. Relating the memory polynomial to the phenomena studied for the transistor it can be seen that this model can describe AM/AM and AM/PM variations along a defined bandwidth for narrowband signals. To observe this, start by defining the input, $\tilde{x}(n)$, as in (2.20), where ω_d is a displacement of the center frequency in relation to the LPE fundamental frequency and T_s is the sampling time.

$$\tilde{x}(n) = \tilde{x}_{env}(n)e^{jn\omega_d T_s} \tag{2.20}$$

Replacing (2.20) in (2.19), assuming $\tilde{x}_{env}(n)$ is a narrowband signal, the transformation in (2.21) can be applied.

$$\begin{aligned}
&\sum_{m=0}^{M-1} \tilde{h}_{2p+1}(m)\tilde{x}(n-m)|\tilde{x}(n-m)|^{2p} \\
&\quad = \tilde{x}_{env}(n)|\tilde{x}_{env}(n)|^{2p} \sum_{m=0}^{M-1} \tilde{h}_{2p+1}(m)e^{j(n-m)\omega_d T_s} \\
&\quad = \tilde{H}_{2p+1}(\omega_d)\tilde{x}_{env}(n)|\tilde{x}_{env}(n)|^{2p}e^{jn\omega_d T_s} \\
&\tilde{y}(n) = \sum_{p=0}^{P-1} \tilde{H}_{2p+1}(\omega_d)\tilde{x}_{env}(n)|\tilde{x}_{env}(n)|^{2p}e^{jn\omega_d T_s}
\end{aligned} \tag{2.21}$$

From this point of view the Memory Polynomial model can be used to represent a PA in a wide frequency range, for narrowband signals. This model can be constructed by measuring AM/AM and AM/PM curves for different center frequencies. The description of the AM/AM and AM/PM characteristics along frequency, exhausts the representation capabilities of the memory polynomial model. Since other effects are neglected, this model will deviate from the PA behavior as the bandwidth of the excitation signal is increased.

In the context of DPD and behavioral modeling, for a particular signal, the approach to obtain the model is very different. A set of coefficients are extracted that best describe the behavior of the PA for the particular signal. In this light, the memory polynomial model is an attempt to describe the deviation from the static characteristic. Since we have observed that the filters in the memory polynomial must be related to

the RF networks of the PA, in particular, the filters are sensitive to the phase of the excitation signal, we can attempt to relate the memory polynomial to the memory induced by the matching networks.

As previously shown in section 2.1.2 the fundamental current generated from the transistor, assuming a small deviation to the static case due to the matching networks, is given by (2.22), where $a(t)$ is the amplitude of the excitation signal.

$$\begin{aligned} \tilde{\imath}_{ds_1}(t) &= A_1\big(a(t)\big)\exp\Big(j\Big(\phi(t)+P_1\big(a(t)\big)\Big)\Big)+\widetilde{\Delta\imath}_{ds_1}(t) \\ \widetilde{\Delta\imath}_{ds_1}(t) &= A_0^{\frac{\delta F}{\delta v_{gs}}}\big(a(t)\big)\left(\int_0^t \widetilde{\Delta h}_{i_1}(\tau)\tilde{x}_1(t-\tau)dt+\int_0^t \widetilde{\Delta h}_{c_{gd_1}}(\tau)\tilde{\imath}_{ds_1}(t-\tau)dt\right) \end{aligned} \tag{2.22}$$

Converting this description to the digital domain, (2.23) is obtained, where $z(n)$ is the generated fundamental current.

$$\begin{aligned} \tilde{z}(n) = &\sum_{p=0}^{P_1-1} a_p\tilde{x}(n)|\tilde{x}(n)|^{2p} \\ &+\sum_{p=0}^{P_2-1} b_p|\tilde{x}(n)|^{2p}\left(\sum_{m=0}^{M_x-1}\tilde{h}_x(m)\tilde{x}(n-m)+\sum_{m=0}^{M_z-1}\tilde{h}_z(m)\tilde{z}(n-m)\right) \end{aligned} \tag{2.23}$$

The problem in (2.23) is solving the nonlinear feedback. Assuming the PA gain is sufficiently smooth and the PA is not hardly driven into compression, the linear feedback will be dominant and (2.24) can be obtained.

$$\begin{aligned} \tilde{z}(n) \cong &\sum_{k=0}^{K}\tilde{h}_{inv}(k)\left(\sum_{p=0}^{P-1} a_p\tilde{x}(n-k)|\tilde{x}(n-k)|^{2p}\right. \\ &\left.+\sum_{m=0}^{M_x-1} b_1\tilde{h}_x(m-k)\tilde{x}(n-m-k)\right) \end{aligned} \tag{2.24}$$

$$\widetilde{H}_{inv}(\Omega) = \Big(1-b_1\widetilde{H}_z(\Omega)\Big)^{-1}$$

The fundamental current is then filtered through the output matching network to the load. Even though the expression does not exhaust the Memory Polynomial representation it is also not contained in a Hammerstein system.

Dynamic Deviation Reductions

The DDR model, first presented in [14], was developed from the Nonlinear Integral Model, earlier presented in [40, 82, 83, 13]. The reasoning behind this is that

the system settles into a mostly nonlinear static operation, with small perturbations induced by the dynamic circuitry. In RF operation, the systems are typically excited in a small fractional bandwidth, compared to the carrier frequency, supporting the nonlinear integral model approach. Previously, in section 2.1, when the PA operation was analyzed, this argument was also used to support the propagation of the perturbations induced by the dynamic components as a deviation from the static steady-state operation.

While the Nonlinear Integral Model is nonlinear in the parameters, the DDR model solves this problem with clever manipulation of the nonlinear terms. For applications in DPD and behavioral modelling, where linear extraction is favored, this advantage is very significant.

The DDR model suffers from a fast expansion in the number of terms when increasing the model order. The model order in the DDR model is connected to the number of considered deviation terms [84]. This fact has restricted the used DDR models to the first and simplified second orders [15]. Using higher orders DDR models is not common in the literature.

The DDR1 model is mathematically described as seen in (2.25).

$$\tilde{y}(n) = \sum_{p=0}^{P-1} |\tilde{x}(n)|^{2p} \sum_{m=0}^{M-1} h_{m,p}^{1} \tilde{x}(n-m) + \sum_{p=0}^{P-1} |\tilde{x}(n)|^{2p} \tilde{x}(n)^2 \sum_{m=1}^{M-1} h_{m,p}^{2} \tilde{x}^*(n-m) \qquad (2.25)$$

The DDR1 model applies filters to the unmodified signal. These filters introduce the dynamics of the model which will perturb the static functions. The DDR1 model also uses a filtered conjugate term which does not typically appear in other models.

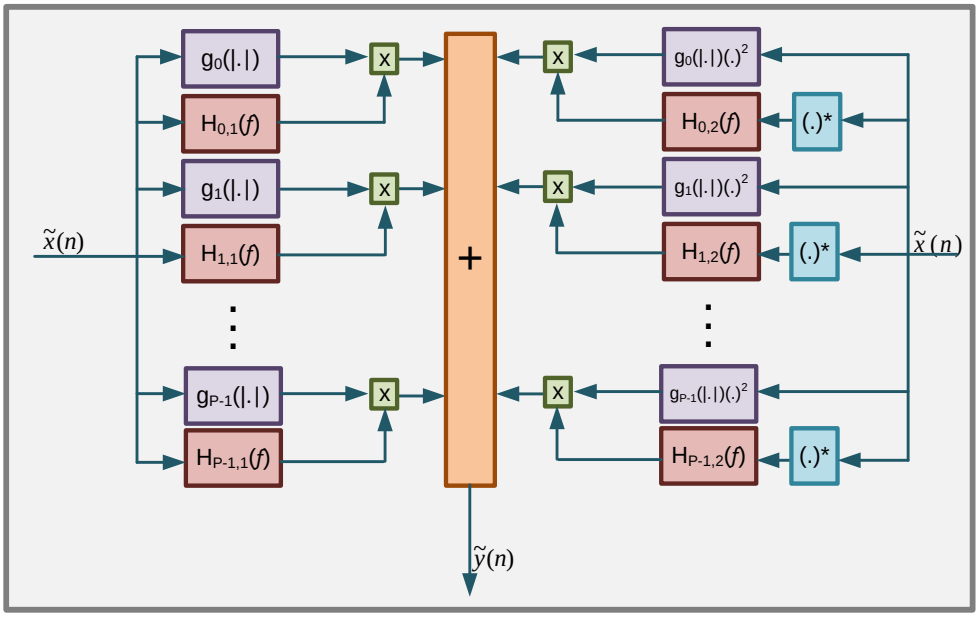


Fig. 2.17: DDR1 model block diagram. The Volterra polynomials have been replaced with generic basis functions.

The model has the block diagram shown in Fig. 2.17.

When analyzing the DDR1 model, it is noticeable that the filters are sensitive to the phase information of the envelope. Generally, this means that the dynamics should be connected to the matching networks.

Comparing the information described by the DDR1 model with the perturbations induced by the matching networks, (2.23), immediately there are some similarities. In fact, ignoring the feedback term, the DDR1 model is capable of exactly describing the rest of the expression. This is expected, since expression (2.23) is obtained under the same rationale as the DDR1 model. The main advantage of the DDR1 model over the Memory Polynomial model is that the elimination of the nonlinear dependence is not necessary. Looking at Fig. 2.5, this is beneficial for correct approximation in the saturation region.

To completely fit (2.23) with the DDR1 model, it is assumed that either the feedback is negligible or that $z(n) \cong x(n)$. Furthermore, the translation from $z(n)$ to the output, $y(n)$, should be memoryless.

The second part of the DDR1 model, $\sum_{p=0}^{P-1}|\tilde{x}(n)|^{2p}\tilde{x}(n)^2\sum_{m=1}^{M}h_{m,p}^2\tilde{x}^*(n-m)$, can also be related to the expressions developed in subsection 2.1.2. Going back to (2.7), taking the term $\sum_{n=2}^{+\infty}A_n b_{n-1}(t)\cos\big(\omega t+\phi(t)+P_n-\beta_{n-1}(t)\big)$, for $n=2$, and converting it to the LPE domain, (2.26) is obtained.

$$\Delta z(n)=\sum_{p=0}^{P-1}a_p\tilde{x}(n)^2|\tilde{x}(n)|^{2p}\left(\sum_{m=0}^{M_x}\tilde{h}_x(m)\tilde{x}(n-m)+\sum_{m=0}^{M_z}\tilde{h}_z(m)\tilde{z}(n-m)\right)^* \qquad (2.26)$$

This term of (2.7) reflects the impact of dynamic changes at the fundamental on the fundamental, translated through the second harmonic. Comparing (2.26) with the second term of the DDR1 model, similarities can be noticed. This term can be correctly approximated by the DDR1 model under the same assumptions as before.

Going back to section 2.1.2, looking at Fig. 2.5 it is noticeable that the impact of A_0 should be much superior to the impact of A_2 on the conversion of the perturbations induced by dynamics around the fundamental frequency. As so, it is expected that the impact of these terms of the DDR1 model should be small.

The SDDR2 has the mathematical formula shown in (2.27).

$$\tilde{y}(n) = \sum_{p=0}^{P-1} |\tilde{x}(n)|^{2p} \sum_{m=0}^{M-1} h_{m,p}^{1} \tilde{x}(n-m) + \sum_{p=0}^{P-1} |\tilde{x}(n)|^{2p} \tilde{x}(n)^{2} \sum_{m=1}^{M-1} h_{m,p}^{2} \tilde{x}^{*}(n-m) + \sum_{p=0}^{P-1} |\tilde{x}(n)|^{2p} \tilde{x}(n)^{*} \sum_{m=1}^{M-1} h_{m,p}^{2} \tilde{x}^{2}(n-m) + \sum_{p=0}^{P-1} |\tilde{x}(n)|^{2p} x(n) \sum_{m=1}^{M-1} h_{m,p}^{1} |\tilde{x}(n-m)|^{2} \quad (2.27)$$

The SDDR2 model introduces some new terms. The first one is still related to the matching networks and can be described as the impact of the dynamics at the second harmonic converted to the fundamental. The second term has a phase insensitive filter, according to the previously discussed theory, these effects are related to the bias networks. The SDDR2 model has the block diagram shown in Fig. 2.18, added to the DDR1 block model.

Following the same rationale as before, this can be related to (2.7), where the conversion term is A_1 and the perturbation term is b_2. Looking at Fig. 2.5 it is noticeable that A_1 is a high impact term, b_2 is the gate voltage perturbation at the second harmonic. Sinc it is considered that the input is linear, this is solely due to the feedback from the output, and so depends on the second harmonic current leaked to the gate of the transistor. To fit this with the SDDR2 model, it is necessary to assume that this

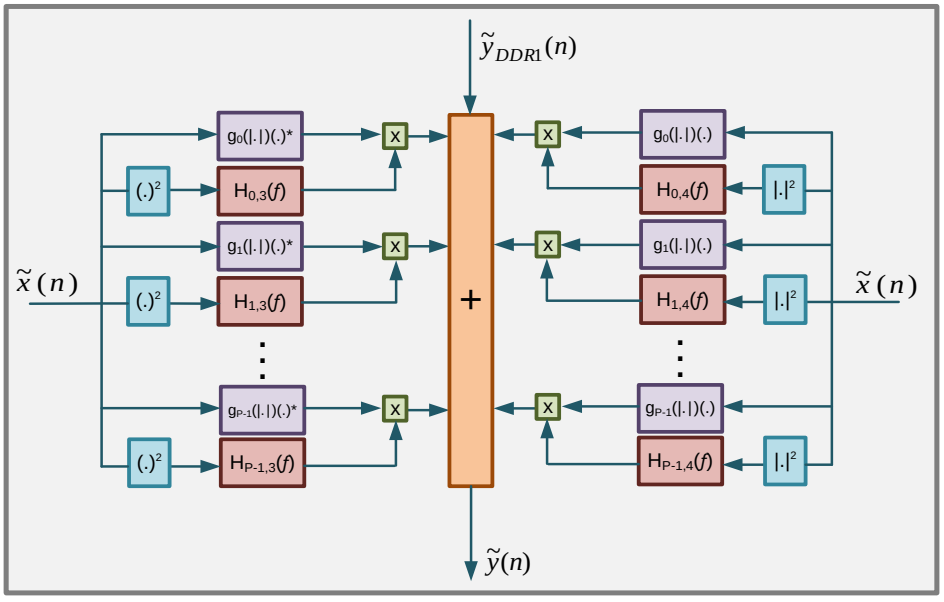


Fig. 2.18: SDDR2 model block diagram, adding to the DDR1 model. The Volterra polynomials have been replaced with generic basis functions.

current is nonlinearly related to the input by the square, which is considerably restrictive.

The last term introduced by the SDDR2 model is slightly different from the others in the sense that the filter will be insensitive to the phase of the carrier. This case is in line with the currents generated at baseband in the transistor that flow through the bias networks. According to section 2.1.3, the effects due to the matching networks are mainly the dynamic changes in the drain voltage due to the excitation of the bias networks by the drain current. To fit this in the SDDR2 model it is necessary to assume that the drain bias current can be represented as the square of the input envelope amplitude. Again, this is very restrictive.

Similarly to the DDR1 model, to fit the developed theory with the SDDR2 model the transistor current should be statically converted to the output voltage of the PA.

Generalized Memory Polynomial

The final model examined here is the GMP model, first proposed in [11]. The GMP model has the mathematical formulation in (2.28).

$$\tilde{y}(n) = \sum_{p=0}^{P-1}\sum_{m=0}^{M-1} \tilde{h}_{2p+1}(m,0)\tilde{x}(n-m)|\tilde{x}(n-m)|^{2p} + \sum_{p=1}^{P-1}\sum_{m=0}^{M-1}\sum_{l=1}^{L-1} \tilde{h}_{2p+1}(m,l)\tilde{x}(n-m)|\tilde{x}(n-m-l)|^{2p} + \sum_{p=1}^{P-1}\sum_{m=0}^{M-1}\sum_{l=1}^{L-1} \tilde{h}_{2p+1}(m,-l)\tilde{x}(n-m)|\tilde{x}(n-m+l)|^{2p} \quad (2.28)$$

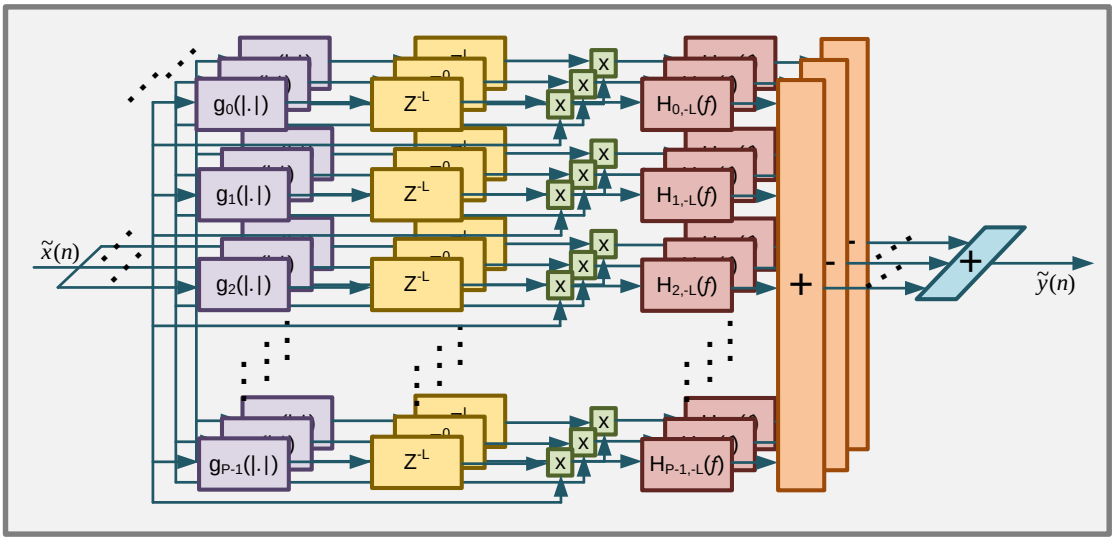


Fig. 2.19: GMP model block diagram. The Volterra polynomials have been replaced with generic basis functions.

The GMP model is an expansion of the MP model, it introduces a separate filtering, insensitive to the phase that modifies the original MP response. The block diagram for the GMP model is shown in Fig. 2.19.

As mentioned before, the GMP model has two separate zones of filtering. The first, is sensitive only to the amplitude of the envelope. Relating this back to the theory in section 2.1.3, these filters support the dynamics induced by the bias networks. Effectively, according to (2.12), the fundamental current generated by the transistor can be written in the LPE domain as shown in (2.29).

$$z(n) = \sum_{m=0}^{M} \sum_{p=0}^{P-1} \sum_{k=0}^{K-1} h_{p,k}(m)|x(n)|^{2k}x(n)|x(n-m)|^{2p} \tag{2.29}$$

Taking into account the matching networks, as explained in the MP section, the current will have an expression very close to the one proposed by the GMP.

2.2.3. Long Term Compensation

The presented models cannot typically accommodate very long term time constants. This is because the increase of the memory taps would be too high to allow the model extraction. To accommodate these phenomena, the use of auxiliary models has been proposed in the literature [79, 31, 30].

While powerful, this technique has several problems. Typically, it can be difficult to define the auxiliary model, as well as, a procedure to extract its coefficients. Furthermore, the auxiliary model generates a signal that must impact the output from the main model. The way the auxiliary model interacts with the main one can also be difficult to define. Nevertheless, this technique can be used to generate signals describing specific effects as temperature or trapping, when some physical knowledge already exists to guide the creation of the auxiliary model.

Typically, the control signals from the auxiliary models vary the coefficients from the main model. These coefficients can be expanded polynomially as functions of the control signal. Using a large polynomial expansion for this can rapidly lead to a high number of coefficients, however, many times a simple linear control can be sufficient and worth to implement. In this case, a controlled GMP would have the formulation in (2.30), where $c_k(n)$ are the control signals, $c_0(n) = 1$, and K is the number of control signals.

$$\tilde{y}(n) = \sum_{p=0}^{P-1}\sum_{m=0}^{M}\sum_{k=0}^{K} \tilde{h}_{2p-1,k}(m,0)c_k(n)\,\tilde{x}(n-m)|\tilde{x}(n-m)|^{2p}$$
$$+\sum_{p=1}^{P}\sum_{m=0}^{M}\sum_{l=1}^{L}\sum_{k=0}^{K} \tilde{h}_{2p-1,k}(m,l)c_k(n)\,\tilde{x}(n-m)|\tilde{x}(n-m-l)|^{2p} \quad (2.30)$$
$$+\sum_{p=1}^{P}\sum_{m=0}^{M}\sum_{l=1}^{L}\sum_{k=0}^{K} \tilde{h}_{2p-1,k}(m,-l)c_k(n)\,\tilde{x}(n-m)|\tilde{x}(n-m+l)|^{2p}$$

Furthermore, these long term memory effects are typically controlled by the amplitude of the signal and are mostly phase insensitive. This means that the control signal models are typically simpler than the main model and do not require complex number algebra.

This type of models has the block diagram shown in Fig. 2.20. In this case, the control models have not been defined, the way to define these models is explored further into this thesis.

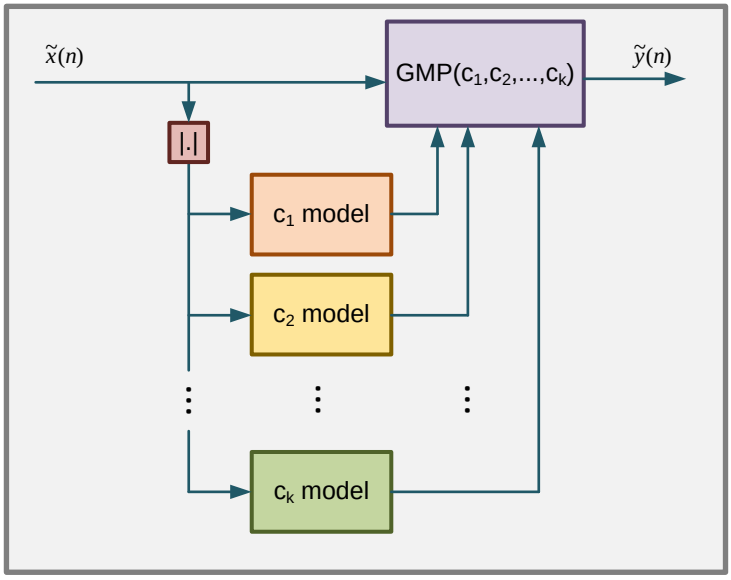


Fig. 2.20: Controlled model block diagram. The control models are fed with the amplitude of the envelope and their output changes the behavior of the main model.

2.3. Summary and Final Considerations

This section presents a theory for the analysis of both dynamic and static nonlinear behavior in PAs, as well as relating these behaviors with state-of-the-art models.

In terms of nonlinear static effects it was shown that, for sufficiently narrow-band signals, the PA settles into a steady-state defined by an amplitude and phase of a sinusoid and dependent on the amplitude of the excitation. Maintaining the bandwidth of the signal, it is shown that this steady-state is also controlled by the center frequency of the excitation. For narrowband signals, these effects are fully contained within the MP polynomial, extracted by evaluating the AM/AM and AM/PM of the PA at the frequencies of interest.

In another perspective, the dynamics of the PA are evaluated as a change from the static behavior as the bandwidth is increased. This approach is taken to describe the effect of the matching networks on the PA as the bandwidth is taken. It is shown that the effect of the matching networks for small dynamic effects is dictated by two types of terms: the conversion terms (A_k) – dependent on the amplitude of the envelope and convert perturbations on every frequency to the fundamental; and the perturbation terms (b_k) – dependent on the filter responses at each frequency. A higher impact is expected from the changes produced by these filters on the gate voltage. These dynamic effects are related to the formulations presented in the MP and DDR models. While the DDR model permits a higher malleability of the conversion terms (nonlinear conversion, linear filtered signal), the MP model allows a higher liberty in the filtered signals (linear conversion, nonlinear filtered signal).

Bias effects are also examined, two main differences from the matching networks effects are pointed out: the bias effects are insensitive to the phase modulation; and the bias effects are mainly due to the drain. The GMP model is pointed out as the model that is more adequate to describe the dynamic bias effects together with the dynamic matching effects.

Finally, long term memory effects are examined, namely thermal and trapping effects. Both effects are expected to be dependent on the amplitude modulation of the excitation but present different dynamics. It is pointed out that controlled models can be used to forcefully insert these effects into the behavioral or DPD model, since physical support is available to build the auxiliary models.

Taking into account the considerations developed in this chapter, the main model used for the work developed in this thesis is the GMP model. As mentioned, the GMP model shows strong correlation to the effects observed in the physical device and,

besides this, has also shown very good results in the literature and has been widely adopted.

3. Extraction Robustness of Digital Predistortion and Behavioral Models

Behavioral modelling and DPD of RFPAs based on the traditional polynomial Volterra series can suffer from extraction problems, not only due to the number of coefficients, but also due to the conditioning of the underlying basis functions. After understanding how the components and structure of the Volterra models relate to the behavior of the PA, one of the concerns when developing this work was making sure the models were correctly extracted and robust, from a signal processing point of view.

While the relationships found when comparing the models to the physics of the device guarantee the model structure is sound, the robustness from the numerical point of view guarantees that the extraction procedure is numerically sound and will yield good results.

While investigating the signal processing techniques used in Volterra based model extraction, a conclusion was reached that the conditioning problem inherent of these models in their original form, needed to be addressed. The approach taken in this thesis was to change the original model representation (in terms of monomials) into more robust representations (interpolated LUTs were the original target). A number of original contributions were made in the application of spline interpolated LUTs to DPD and behavioral modelling systems and how splines can be related with the polynomial formulation.

The proposed techniques have shown good results in significantly improving the conditioning of the models. Moreover, the developed techniques also led to a polynomial formulation that has a significantly better conditioning, but is equivalent to the original one. The results using these methods were satisfactory for guaranteeing the numerical stability of the extraction.

3.1. Improving Model Conditioning

While the Volterra representation is a general formulation of nonlinear systems, it suffers from a number of problems. One of these problems is the excess of coefficients, which has been broadly studied in the literature with several constrained coefficient formulations being proposed [10, 11, 12, 84]. One of the other problems commonly experienced in the extraction of Volterra based models is ill-conditioning of the regression matrix. This ill-conditioning of the Volterra series is due to the lack of orthogonality between the basis functions. In particular, the monomial basis, frequently used in model description ($[|x|^0, |x|^1, |x|^2, \dots]$) is particularly ill-conditioned because the basis functions become more and more indistinguishable, as shown in Fig. 3.1.

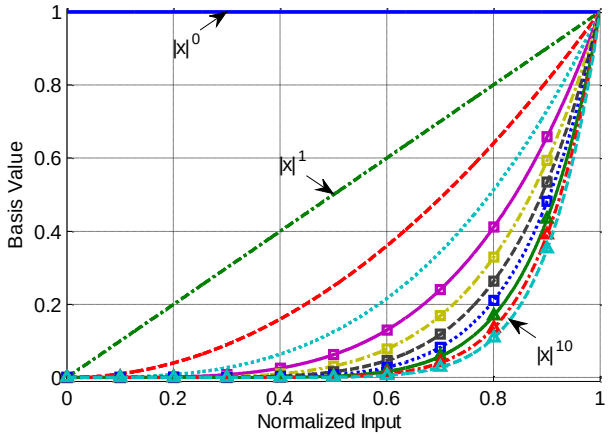


Fig. 3.1: Monomial basis functions from orders 0 to 10.

It's interesting to note that, according to Fig. 3.1 the polynomial basis becomes more and more ill-conditioned independently of the signal. In fact, the monomial basis is orthogonal to the complex exponential signal. However, this is not useful for LPE behavioral and DPD models, due to the signal phase conditions that have to be fulfilled [43]. These phase conditions, required for the physical consistency of the LPE models, generally lead to formulations that make use of amplitude functions whereas the signal phase only excites the memory. Since, we are mainly interested in the GMP, note that the GMP can be described as shown in (3.1), where the summation along p has been resolved to yield the $f_{m,l}$ functions.

$$\tilde{y}(n) = \sum_{m=0}^{M} \sum_{l=-L}^{L} \tilde{x}(n-m) f_{m,l}(|\tilde{x}(n-m-l)|) \tag{3.1}$$

The static and MP models also show this type of formulation, as well as the EMP model, and, to some extent, even the DDR based models. It is therefore noticeable that, while the memory and nonlinearity are not separable, these truncated models show a strong nonlinear core to which is subsequently added memory. In this sense,

the treatment of the nonlinear basis to become more orthogonal can lead to substantial decreasing of the condition number and an improved conditioning.

3.1.1. Interpolated Look-Up Tables and Local Polynomials

One of the directions that was investigated in this thesis was the representation of the DPD and behavioral models using Interpolated Look-Up Tables (LUTs), these LUTs are generally interpolated using spline functions. For instance, for the GMP model, taking into account the formulation shown in (3.1), a LUT based representation would describe the functions, $f_{m,l}(z)$ as a set of pairs of $\left[z_p, f_{m,l}(z_p)\right]$ points in a table. These points are then interpolated using the chosen method (generally, a low order spline). The model parameters are, in this case, converted into the LUT points.

The argument in favor of using LUTs for the representation of DPD and behavioral models is locality. Locality plays in favor of orthogonality, and, consequently, conditioning, as can be seen using expression (3.2), which measures the correlation between two basis functions. Looking at (3.2), as the basis functions become more local in x, independently of the probability distribution of x, they become more orthogonal.

$$O_{p,k} = \frac{1}{\left\|S_p(x)\right\|_2 \left\|S_k(x)\right\|_2} \int_{-\infty}^{+\infty} S_p(x) S_k^*(x) \mathrm{pdf}(x) dx$$
$$\|S(x)\|_2 = \int_{-\infty}^{+\infty} S(x) S^*(x) \mathrm{pdf}(x) dx \tag{3.2}$$

The LUTs represent the functions as their sampled versions at designated points. In this sense, it is *a priori* known that the function must have that value at that point and so, the basis functions will be restricted by this representation. In fact, when using spline interpolation, the functions can be expanded as seen in (3.3), as the work developed for this thesis shows in [78] ([C1] included in annex to this chapter).

$$f_{m,l}(z) = \sum_{p=0}^{P} f_{m,l}(z_p) S_p(z)$$
$$S_p(z_p) = 1 \tag{3.3}$$

The expansion in (3.3) shows that the spline basis functions $S_p(z)$ are restricted within a local zone and must vanish at least at each z_p point. The $S_p(z)$ functions depend on the interpolation points, z_p, and the spline order. However, the locality feature is almost always preserved, independently of these parameters. This can be seen when the spline basis functions are plotted, in the regular grid, for increasing

spline order, as shown in Fig. 3.2. Fig. 3.2 shows well the locality of the basis, however, another phenomenon appears when the interpolation order is increased – an oscillation of the basis functions close to the interval edges. This oscillation degrades the locality of the basis functions and, therefore the conditioning of the problem.

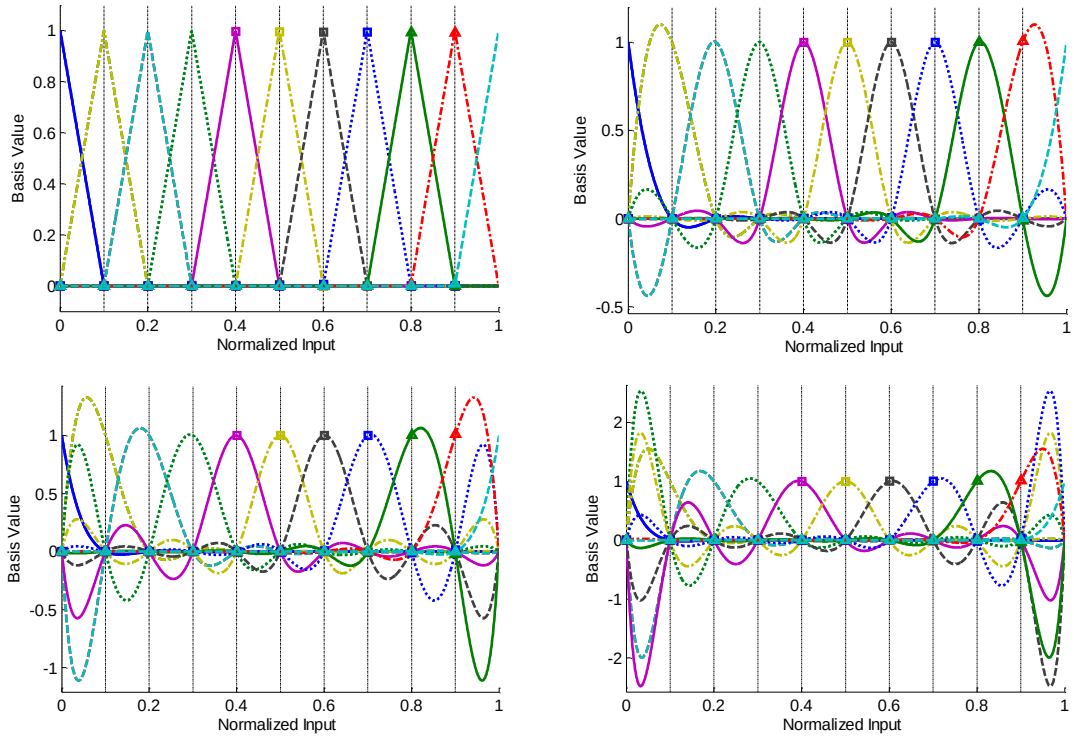


Fig. 3.2: Spline basis functions for orders 1, 3, 5 and 7 on eleven equally spaced nodes on the normalized interval.

The oscillatory behavior, also known as Runge's Phenomenon, observed in Fig. 3.2 is due to the distribution of the interpolation points, z_p, [85]. The distribution of these points is a problem in the LUT description because the basis functions and thus the output are nonlinearly dependent on them. The optimization problem considering the interpolation points is then nonlinear. Even though nonlinear fitting can be sometimes used in these types of large scale problems (high number of points), one example being the canonical piecewise approximation [86], it is preferably avoided in DPD systems when a coefficient update is required in relatively small time-frames.

In the DPD and behavioral modeling literature, the point distribution is usually approached for error minimization. For this case, the optimal point distribution depends on the interpolator functions and has been explored for linear interpolation in the literature [87, 88, 89]. In this thesis, the interpolation point distribution is not used to reduce the model error, instead this distribution is used to correct the oscillatory phenomenon, which will improve the conditioning of the problem. While this might appear counter-intuitive, note that, when these parameters are not used for the error minimization problem, the number of parameters is maintained equal to the polynomial

case. The use of these parameters in the minimization of the error, increases the number of coefficients in the overall problem. To minimize the correlation between the basis, we noted that, in interpolation theory, the use of Chebyshev point distribution is recognized as minimizing the Runge's Phenomenon [85]. Therefore, these points were chosen as interpolation points and the behavior shown in Fig. 3.3 was observed in the spline basis.

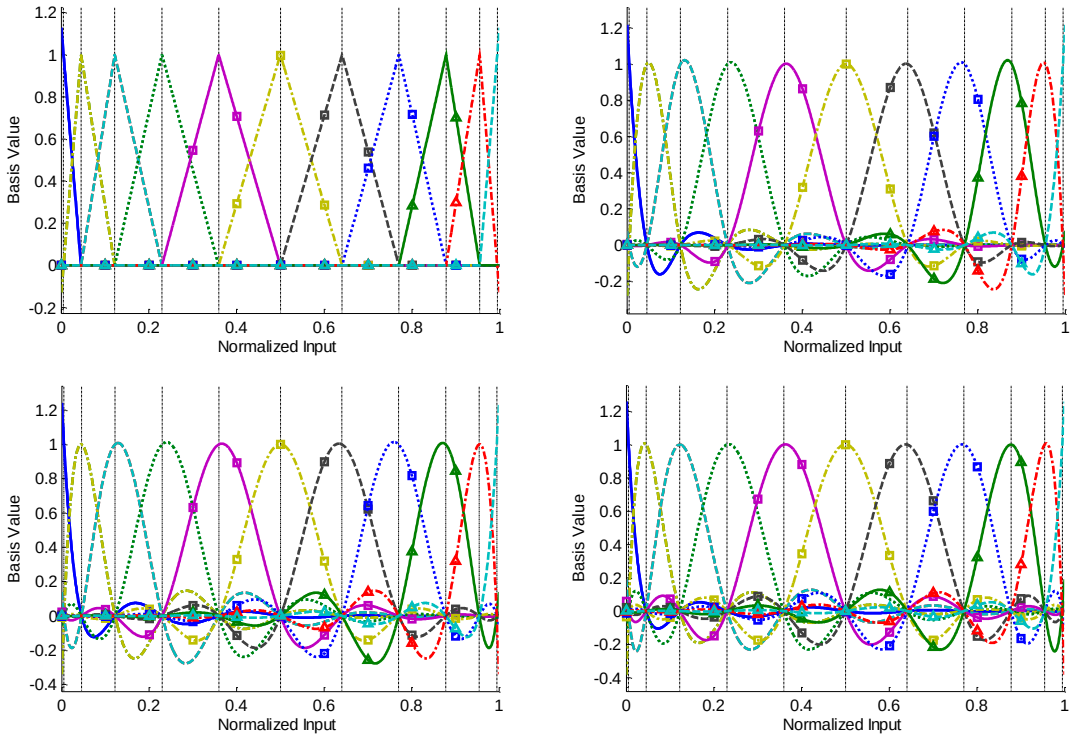


Fig. 3.3: Spline basis functions for orders 1, 3, 5 and 7 on eleven Chebyshev spaced nodes on the normalized interval.

As can be seen in Fig. 3.3 the oscillation of the basis at the interval edges has been completely suppressed. The spline interpolation order in these basis functions defines the smoothness of the functions. Previously, it was noted that, as this order was increased the oscillations became more severe and the basis less conditioned. However, when the interpolation points were changed, the increase in the spline order did not degrade the quality of the basis. Because of this, the order (smoothness) of the basis functions can be increased until the spline basis coincides with the polynomial basis. This happens when the order is increased sufficiently to maintain full continuity of all derivatives at the interpolation points. Evidently, the polynomial basis obtained in this way is not the monomials, shown in Fig. 3.3, but the Lagrange polynomials that also show a highly increased condition number when compared to the monomials, which was shown as part of the work for this thesis [32] ([C2] included in annex to this chapter).

Using the theory developed for improving the representation of the behavioral

and DPD models, this thesis reached the conclusion that the spline interpolated LUTs and the polynomials can be seen under the same light. In fact, the developed work was unified and extended to reflect this as part of the work for this thesis, presented in [75] ([J1] included in annex to this chapter). The polynomials can therefore be used similarly to LUTs in behavioral modelling and DPD. Previous analysis on this matter generally resulted with conclusions in favor of LUTs due to the conditioning of the polynomials or flexibility of the LUTs. However, in this work it was shown that these desired properties can be given to the polynomials as well.

3.2. Summary and Final Considerations

Volterra based models suffer from high condition numbers leading to low extraction robustness. In this work it was shown that this problem can be severely alleviated by using local amplitude basis functions. In order to explore locality, the GMP model was converted to a spline interpolated LUT formulation. It was shown that the spline formulation indeed corresponds to underlying local basis functions in the amplitude. However, as the spline order is increased the basis functions develop an oscillatory behavior close to the interval edges, degrading the conditioning of the problem. In the literature, this oscillation has been associated with the node spacing, Chebyshev spacing minimizes this phenomena and was used in this work to this effect. Using Chebyshev spacing allowed the increase of the interpolation order up to an equivalent polynomial representation. The polynomial basis functions obtained in this way are the Lagrange polynomials which preserve the locality features of the splines. It is then shown that the condition number of these polynomials is comparable to the spline basis functions and is argued that splines and polynomials can be regarded as similar for these behavioral modeling and DPD applications.

Paper C1: Using Spline Basis Functions in Volterra Series Based Models

Filipe M. Barradas, Telmo R. Cunha, Pedro M. Lavrador, and José C. Pedro

2014 International Workshop on Integrated Nonlinear Microwave and Millimetre-wave Circuits (INMMiC)

©2014 IEEE

Using Spline Basis Functions in Volterra Series Based Models

Filipe M. Barradas, Telmo R. Cunha, Pedro M. Lavrador, and José C. Pedro

DETI, Instituto de Telecomunicações, Universidade de Aveiro, 3810-193 Aveiro, Portugal

***Abstract* — The choice of nonlinear description used in Volterra based models strongly influences the robustness of the parameter's extraction.**

Traditionally, monomials are used as the nonlinear description. However, it has been shown that other polynomials provide easier extraction. Volterra-based models can also be described using spline interpolated look-up tables (LUTs). Splines have increased locality which improves conditioning, converging to polynomials as the spline order-1 tends to the number of LUT points. In this paper we first show how to describe Volterra models in terms of splines and then study the conditioning and error for increasing spline order in a practical example.

***Index Terms* — Predistortion, Least squares approximations, Power amplifiers.**

I. Introduction

The Volterra series is a polynomial expansion which represents nonlinear dynamic systems with fading memory. Many models used in power amplifier (PA) behavioral modelling and PA digital pre-distortion are Volterra-based [1]-[5]. These models are described using monomials as basis functions, this basis might incur in extraction problems while other formats have been shown to be easier to identify [7]-[9].

Extraction problems are due to the non-orthogonality of the model kernels, with respect to an input signal, which becomes more problematic with lower arithmetic precision.

In [8] orthogonality of the basis is improved by examining the excitation properties, namely its probability density function (PDF). This can also be achieved by making the terms of the model more local. Locality can be intuitively reached using interpolated LUT based models, where each LUT point is only related to a small range of input amplitudes.

Interpolated LUTs can use splines which brings problems in the model's smoothness as they are restricted to be continuous only up to some derivative order. On the other hand, splines are simple to create, evaluate, and use with LUT descriptions.

In this paper we first show how to transform typical Volterra series based models into spline interpolated LUT models. We then study the predistortion capabilities and conditioning of the model for an increasing spline order in a specific case.

II. Orthogonality

The Volterra series can be seen as a combination of multiple nonlinear kernels. A question then arises concerning the effects described by each of these kernels. If each kernel describes completely separate effects, then we say they are orthogonal. If this is not the case, then they are non-orthogonal.

We can measure the orthogonality of each kernel of the summation versus another by evaluating expression (1), where φ are the kernels of the model and x is the excitation signal.

$$\lim_{T\to+\infty}\frac{1}{T}\int_0^T \frac{\varphi_k(x(t))}{\|\varphi_k(x(t))\|}\frac{\varphi_l^*(x(t))}{\|\varphi_l(x(t))\|}dt \qquad (1)$$
$$\|\varphi_k(x(t))\| = \sqrt{\lim_{T\to+\infty}\frac{1}{T}\int_0^T \varphi_k(x(t))\varphi_k^*(x(t))dt}$$

If this integral yields zero then the terms are orthogonal. The further they stand from zero the less orthogonal they are, which means the described effects are very similar. Note that orthogonality is dependent on the signal as well as on the chosen description.

If we take a memoryless system then we can evaluate this integral in another form by making use of the signal's PDF as seen in (2) where a and b are the amplitude limits of the signal.

$$\frac{1}{b-a}\int_a^b \frac{\varphi_k(x)}{\|\varphi_k(x)\|}\frac{\varphi_l^*(x)}{\|\varphi_l(x)\|}w(x)dx \qquad (2)$$
$$\|\varphi_k(x)\| = \sqrt{\frac{1}{b-a}\int_a^b \varphi_k(x)\varphi_k^*(x)w(x)dx}$$

In this case we can see that locality of the memoryless basis functions, φ, provides higher orthogonality independently of the signal's PDF. This is because the overlap betweem the basis functions decreases.

Even though this is not as effective for models with memory, since we are not addressing memory orthogonality, we still expect to improve the conditioning of the extraction.

Orthogonality of the model terms is important because we need to separate each described effect to properly identify the system. If kernels are poorly orthogonal (highly correlated) then their separation becomes difficult to achieve.

III. Changing the Models

Since we are suggesting changes to the nonlinear basis, the traditional representation of the Volterra-based models is not in the most desirable form. We need to have the summation on the nonlinear orders first, to then modify those bases freely.

Take, for instance, the Generalized Memory Polynomial (GMP) model for the complex envelope, [3]. We can rewrite it in the form of (3) by changing the summation order and explaining some terms; this makes the nonlinear basis easier to manipulate. This procedure can be applied to many other models such as the memory and envelope memory

polynomials [1]-[2] and the dynamic deviation reduction of first and simplified second orders [4]-[5]. The summations in each *p* are actually descriptions of real functions of real variable for each *m* and *l*. We will denominate the functions generated by the summations in *p* as $f_{m,l}$, as shown in (4). Note that this change in the model description does not produce a model in which the basis of the functions $f_{m,l}$ is the basis of the model, since, as we pointed out, we are not addressing memory related orthogonality.

$$\begin{aligned} y(n) = & \sum_{l_a=0}^{L_a} x(n-l_a) \sum_{p_a=0}^{P_a} h_{p_a}(l_a) |x(n-l_a)|^{p_a} + \\ & + \sum_{l_b=0}^{L_b} \sum_{m_b=1}^{M_b} x(n-l_b) |x(n-l_b-m_b)| \sum_{p_b=0}^{P_b-1} h_{p_b}(l_b,m_b) |x(n-l_b-m_b)|^{p_b} + \\ & + \sum_{l_c=0}^{L_c} \sum_{m_c=1}^{M_c} x(n-l_c) |x(n-l_c+m_c)| \sum_{p_c=0}^{P_c-1} h_{p_c}(l_c,m_c) |x(n-l_c+m_c)|^{p_c} \end{aligned} \quad (3)$$

$$f_{m,l}(x(n-l-m)) = \sum_{p=0}^{P} h_p(l,m) |x(n-l-m)|^p \quad (4)$$

IV. Splines as Basis Functions

We would now like to describe the function in (4) using splines. To do this we first define the spline construction. A spline is a function composed of branches of *p'th* order polynomials. The branches are separated at specific point called knots. A *p'th* order spline maintains the function continuity up to the *p-1'th* order derivative.

These conditions are not enough to define a spline. In fact, using these restrictions we always miss *p-1* conditions to have a determined problem. To overcome this, we can apply the so-called not-a-knot conditions which impose continuity up to the *p'th* order derivative in *p-1* knots. In our case, for even *p* we impose the conditions on the edge knots and the center knot, for odd *p* we impose these conditions on the edge knots.

Using not-a-knot conditions, when the spline order is the number of knots minus one, we impose continuity up to the *p'th* order in every knot which means we fall back into the polynomials. However, the polynomials defined in this way are the Lagrange polynomials in (5) where x_p are the defined knots.

$$l_k(x) = \prod_{\substack{p=0 \\ p \neq k}}^{P} \frac{x - x_p}{x_k - x_p}, k \in [0, P], k \in \mathbb{Z} \quad (5)$$

To have a completely defined spline we must carefully select an adequate position for the knots. A good choice of knots are the Chebyshev nodes, which provide well behaved splines of higher orders in the sense of limiting Runge's phenomenon.

Changing to a spline representation can be seen as a change of basis functions, from the monomials to others depending on the spline order. Fig. 1 shows the spline basis of increasing order (excluding the Lagrange polynomials) for seven Chebyshev nodes.

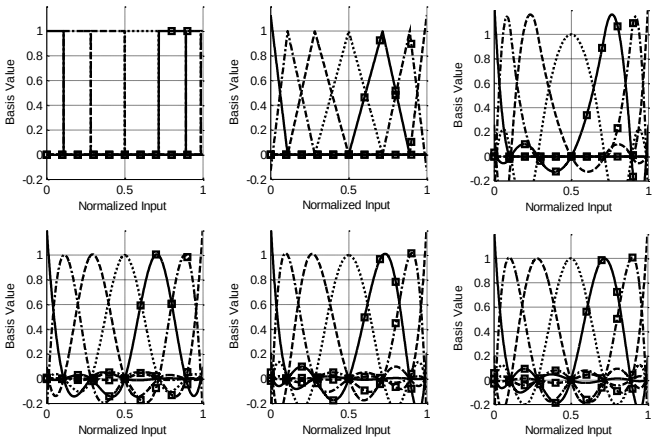


Fig. 1. Spline basis functions of increasing order (0 to 2 on top and 3 to 5 on bottom) for seven Chebyshev nodes. Note the locality of each element. The splines can be defined as summations of these basis functions.

The number of basis functions to describe the splines is always the same. This means that the spline order does not increase the number of parameters in the model, it only controls the shape of the basis functions.

V. Results

We are mainly interested in two results. The first is the capability of approximation of the defined model, and the second is the robustness of the model identification process with increasing spline order.

Note, however, that, in practical terms, the two results are entangled since a poor identification produces a poor approximation when the model is applied. To study these results the chosen metrics were the normalized mean squared error (NMSE) and the condition number of the regression matrix. The NMSE evaluates the approximation error and the condition number evaluates the robustness of the extraction.

The extracted model is the GMP described in (4). The regression is achieved using singular value decomposition (SVD). All data is represented in IEEE double precision floating-point format. The input-output data was obtained by exciting a 20W GaN Doherty PA with a 5MHz, single carrier, LTE signal (Fig. 2). The input and output signals are sampled at 100MHz.

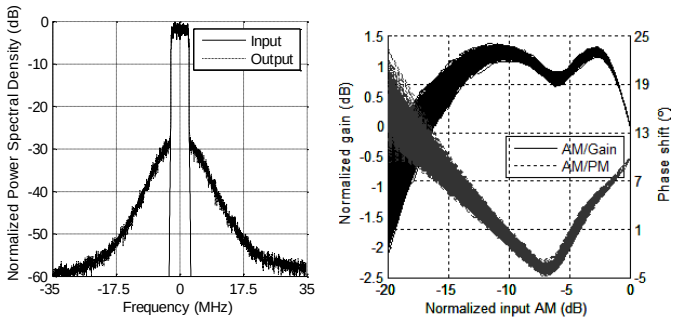


Fig. 2. Input and output spectrum, on the left, and the PA's gain and phase characteristics for the excitation, on the right.

A. Forward Model Identification

For this test the memory depth related parameters were fixed to $\{L_a, L_b, L_c, M_b, M_c\}=\{2, 2, 2, 1, 1\}$, nonlinear orders were set to $\{P_b,P_c\}=\{3,3\}$ and polynomials were used. The first half of the signal (25000 samples) was used for extraction and the second half for the calculation of the error.

Fig. 3 shows the variation of the condition number and NMSE for increasing spline orders on P_a, for 5 knots (P_a = 4), on the left, and 9 knots (P_a = 8), on the right. The condition number and error when using the traditional monomials is also represented for reference.

Note that using splines the condition number is lower allowing the use of less robust, and typically faster, identification techniques.

Since we are working in double precision we do not incur in numerical problems. This means that, in terms of NMSE, the splines converge to the monomials, as the models become equivalent.

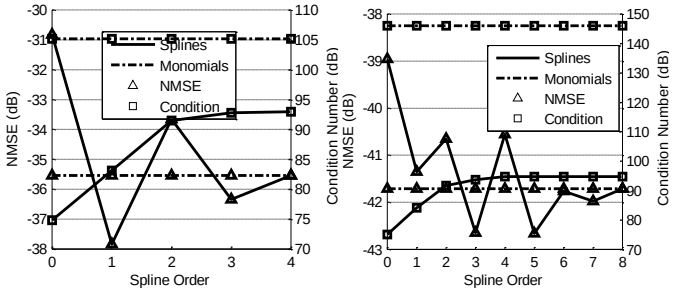


Fig. 3. Condition number and NMSE for increasing spline order and when using monomials. Left figure: 5 knots (P_a = 4); right figure 9 knots (P_a = 8).

B. DPD Identification Via Indirect Learning

For this test the GMP memory depth related parameters were kept, but nonlinear orders were set to $\{P_a ,P_b,P_c\}=\{9,3,3\}$. The models were extracted (through iterative indirect learning) and tested as digital predistorters, DPD, for the PA referred above.

Fig. 4 shows the variation of the condition number of the regression matrix and NMSE between the PA output and DPD input at each iteration of the indirect learning algorithm for zero, first order, and cubic splines, and Lagrange polynomials to interpolate in P_a.

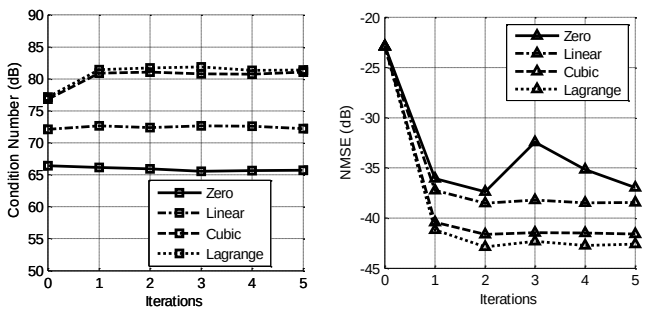


Fig. 4. Condition number, on the left, and NMSE, on the right, for several spline orders at each indirect learning iteration.

VI. Conclusion

This paper showed a simple procedure to transform common Volterra-based models, which allows the description of the model's nonlinearities using splines. The power of the technique is shown in terms of condition number reduction for an illustrative example.

The used basis functions are derived from spline interpolated LUTs and converge to polynomial basis functions creating a bridge between these two modelling approaches.

Even though good results can be obtained in forward modelling using low spline orders, in predistortion the polynomials proved superior, for the same number of terms.

Acknowledgement

The first author would like to acknowledge the financial support provided by Portuguese Science and Technology Foundation, FCT (Ref. SFRH/BD/90103/2012).

References

[1] J. Kim, and K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *Electronics Letters* , vol.37, no.23, pp.1417,1418, Nov 2001

[2] O. Hammi, F. M. Ghannouchi, and B. Vassilakis, "A Compact Envelope-Memory Polynomial for RF Transmitters Modeling With Application to Baseband and RF-Digital Predistortion," *Microwave and Wireless Components Letters*, IEEE , vol.18, no.5, pp.359-361, May 2008

[3] D. R. Morgan, Zhengxiang Ma, and Jaehyeong Kim; M. G. Zierdt; Pastalan, J., "A Generalized Memory Polynomial Model for Digital Predistortion of RF Power Amplifiers," *IEEE Transactions on Signal Processing*, vol.54, no.10, pp.3852-3860, Oct. 2006

[4] Anding Zhu, J. C. Pedro, and T. J. Brazil, "Dynamic Deviation Reduction-Based Volterra Behavioral Modeling of RF Power Amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, vol.54, no.12, pp.4323-4332, Dec. 2006

[5] Lei Guan, and Anding Zhu, "Simplified dynamic deviation reduction-based Volterra model for Doherty power amplifiers," 2011 Workshop on *Integrated Nonlinear Microwave and Millimetre-Wave Circuits (INMMIC)*, vol., no., pp.1-4, April 2011

[6] T. R. Cunha, E. G. Lima, and J. C. Pedro, "Validation and Physical Interpretation of the Power-Amplifier Polar Volterra Model," *IEEE Transactions on Microwave Theory and Techniques*, vol.58, no.12, pp.4012-4021, Dec. 2010

[7] Lei Guan, and Anding Zhu, "Optimized Low-Complexity Implementation of Least Squares Based Model Extraction for Digital Predistortion of RF Power Amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, vol.60, no.3, pp.594-603, March 2012

[8] R. Raich, Hua Qian, and G. T. Zhou, "Orthogonal polynomials for power amplifier modeling and predistorter design," *IEEE Transactions on Vehicular Technology*, vol.53, no.5, pp.1468-1479, Sept. 2004

[9] L. Aladren, P. Garcia, P. L. Carro, J. De Mingo, and C. Sanchez-Perez, "Digital predistortion based on Zernike polynomial functions for RF nonlinear power amplifiers," *2012 International Symposium on Wireless Communication Systems (ISWCS)*, vol., no., pp.865-869, Aug. 2012

Paper C2: Higher Locality Non-Linear Basis Functions of Volterra Series Based Models to Improve Extraction Conditioning

Filipe M. Barradas, Telmo R. Cunha, Pedro M. Lavrador, and José C. Pedro

2014 IEEE MTT-S International Microwave Symposium (IMS2014)

©2014 IEEE

Higher Locality Non-Linear Basis Functions of Volterra Series Based Models to Improve Extraction Conditioning

Filipe M. Barradas, Telmo R. Cunha, Pedro M. Lavrador, and José C. Pedro

DETI, Instituto de Telecomunicações, Universidade de Aveiro, 3810-193 Aveiro, Portugal

***Abstract* — Extraction of Volterra based nonlinear system's models can be hindered by conditioning problems. This is especially true when working with low precision arithmetic, typical of hardware implemented DPD systems.**

Traditionally, monomials are chosen as the basis functions because they are continuous and easy to manipulate, although other bases could be used. Spline bases, for example, have been chosen instead of polynomials for their increased locality. However, we now show that it is also possible to construct polynomials with increased locality and thus, keeping continuity but still improving the condition numbers of the regression matrix. This polynomial basis can be seen as a bridge between LUT and polynomial models.

***Index Terms* — Predistortion, Least squares approximations, Power amplifiers.**

I. Introduction

The Volterra series is a polynomial expansion which describes fading memory nonlinear dynamic systems. Many models used in PA behavioral modelling and PA digital pre-distortion are Volterra based [1]-[5]. However, since the underlying function basis are the monomials, of all orders for the amplitude [6], identification of the series through least squares can incur in conditioning problems [7]-[9].

The conditioning problem can be understood as the inability to differentiate between the bases' elements we are using for the system's description. It is, therefore, intrinsically connected to the choice of the bases used to build the model, and to the excitation signals used in the extraction.

In [8] orthogonality of the bases is improved by examining the excitation properties, namely the probability density function (PDF), and reach an orthogonal basis. Even though the orthogonality is not preserved for excitations with different PDF, the conditioning of the problem is significantly improved. In the present work, we show that a similar effect can be achieved using the bases' locality, regardless of the signals' PDF, and thus connect LUT and polynomial based models.

Identifying Volterra series based models is typically done through least squares. This means solving the problem in (1), where *X* contains the effect of each of the model's bases applied to the input signal *x*, and *y* is the system's output in vector form.

$$\arg\min_{\theta} \left\| y - X\theta \right\|^2 \qquad (1)$$

To solve (1) easily, *X* should be as close to orthogonality as possible, making the identification more separable and robust.

II. Understanding Conditioning

The conditioning of a problem is related to the ability of distinguishing between different effects we are trying to describe. For instance, a poorly conditioned problem is trying to identify a filter's frequency response in a frequency range we are not exciting at the input. Another example is trying to distinguish small perturbations on a very strong signal.

The conditioning problem is, therefore, related to both the excitation, as seen in [8], and the model format. In this paper we will focus on modifying the model format with no assumption on the excitation.

When identifying a system through least squares estimation we are doing the following:

1. Identifying the effect that each basis produces;
2. Eliminating similar effects from different basis (correlation and orthogonalization);
3. Identifying the remaining effects in the output;
4. Returning to the original domain (prior to orthogonalization).

These steps can be identified in the pseudo-inverse least squares solver (2), in which A^H is the complex transpose of A. The product $\left(X^H X\right)$ is the correlation matrix which must be Hermitian and positive semi-definite.

$$\theta = \left(X^H X\right)^{-1} X^H y \qquad (2)$$

First, we build the regression matrix *X* which identifies the effect of each basis of the model. Then, we pseudo-invert this matrix. This can be seen as a transformation to an orthogonal description (3), where Λ is the eigenvalues' diagonal matrix.

$$\left(X^H X\right)^{-1} = \left(U\Lambda U^H\right)^{-1} = U\Lambda^{-1}U^H,\ U^H U = I \qquad (3)$$

Multiplication of the regression matrix *X* by $U\sqrt{\Lambda^{-1}}$ produces an orthogonalized problem. The new basis is used to identify the system parameters which are then converted into the original basis by multiplication with $U\sqrt{\Lambda^{-1}}$.

The eigenvalues in Λ give us the power of each orthogonal effect we are describing. We have to be able to distinguish the lowest power effect on top of the highest power effect. If we cannot, typically we say the signal does not excite our basis. In the same way, we can say that our basis does not fit the signal.

The ill-conditioning is a problem of dynamic range in the numerical algorithm. Higher correlation between bases will

produce higher condition numbers and require higher numerical precision. If we create a basis with high locality, i.e., with confined support in the excitation axis, *xx*, we should reduce this effect and thus improve the conditioning of the problem. Locality can be seen when using LUT based models, where each point of the LUT is only related to a small range of input amplitudes.

A basis function is orthogonal from another if it verifies (4), where $w(x)$ is the signal PDF, φ are the basis functions, and a and b are the limits of the input amplitude, [0,1], for instance. If the functions have very restricted superposition, then the integral must be small. Furthermore, this reduction of the integral's value is insensitive to the PDF.

$$\int_a^b \varphi_k(x)\varphi_l(x)w(x)dx \tag{4}$$

III. CHANGING THE MODELS

Since we are suggesting changes to the nonlinear basis, the traditional representation of the Volterra based models is not in the most desirable form. We want to have the summation on the non-linear orders first, to modify those basis freely.

Take, for instance, the Generalized Memory Polynomial (GMP) for the complex envelope, [3]. We can rewrite it in the form of (4) by changing the summation order and explaining some terms; this makes the nonlinear basis easier to manipulate. This procedure can be applied to many other models such as the memory and envelope memory polynomials [1]-[2] and the dynamic deviation reduction of first and simplified second orders [4]-[5]. The summations in each p are actually descriptions of real functions of real variable for each m and l. We will call the functions generated by the summations in p the functions $f_{m,l}$. In this paper we will adjust the bases' elements of these functions to produce better conditioning. Note that this change in the model description does not produce a model in which the basis of the functions $f_{m,l}$ is the basis of the model. This is because we will multiply the functions by delayed versions of x. However, we expect a better condition number simply by adjusting the bases' elements for these functions.

$$\begin{aligned} y(n) = &\sum_{l_a=0}^{L_a} x(n-l_a)\sum_{p_a=0}^{P_a} h_p(l_a)\left|x(n-l_a)\right|^{p_a} + \\ &+\sum_{l_b=0}^{L_b}\sum_{m_b=1}^{M_b} x(n-l_b)\left|x(n-l_b-m_b)\right|\sum_{p_b=0}^{P_b-1} h_p(l_b,m_b)\left|x(n-l_b-m_b)\right|^{p_b} + \\ &+\sum_{l_c=0}^{L_c}\sum_{m_c=1}^{M_c} x(n-l_c)\left|x(n-l_c+m_c)\right|\sum_{p_c=0}^{P_c-1} h_p(l_c,m_c)\left|x(n-l_c+m_c)\right|^{p_c} \end{aligned} \tag{5}$$

IV. ADJUSTING THE BASES FUNCTIONS

As we made notice before, we would like our bases functions to have higher locality so that when one basis function has strong influence, the others do not. Effectively, what we accomplish is a better separation of bases, and, in this way, smaller numerical dynamic range is required to be able to distinguish them.

A basis with the characteristics we are looking for, is composed of square pulses which exist only in one zone of the amplitude interval (maximized locality). The orthogonality of this basis is the highest for any signal since the pulses do not overlap, i.e, (4) always yields zero for finite PDFs.

Unfortunately, using square pulses as bases functions produces a discontinuous model. In fact, these are the basis functions of a zero-order hold spline. To go back to the polynomials, we can enforce continuity up to a certain order, leading us to linear, quadratic, cubic, etc. splines. All of these bases show high locality.

For a high enough order spline, the continuity is preserved across all the derivatives and our spline tends to a polynomial. The spline order needed to converge to a polynomial basis is related to the number of divisions of the interval.

The polynomial basis obtained in this way is the Lagrange polynomial basis, (6), defined in the points x_p which delimit each zone in the interval.

$$l_k(x) = \prod_{\substack{p=0 \\ p\neq k}}^{P} \frac{x-x_p}{x_k-x_p}, k\in[0,P], k\in Z \tag{6}$$

The calculation of these bases elements does not need to be done explicitly. These functions can, for instance, be stored in look-up tables in memory, with the desired precision.

Increasing the polynomial order is now a question of increasing the number of divisions, or the number of x_p. This will cause the elements of the basis to have a stronger influence in an increasingly smaller span, and, if the points are not correctly chosen, to create large perturbations at the edges of the interval, leading to undesired correlation between the elements. A suitable choice of points are the Chebyshev nodes (7) which mitigate this behavior, Fig. 1. The Lagrange polynomials on the Chebyshev nodes form our chosen basis.

$$x_p = \frac{1+\cos\left(\frac{2(k+1)-1}{2(P+1)}\pi\right)}{2}, k\in[0,P], k\in Z \tag{7}$$

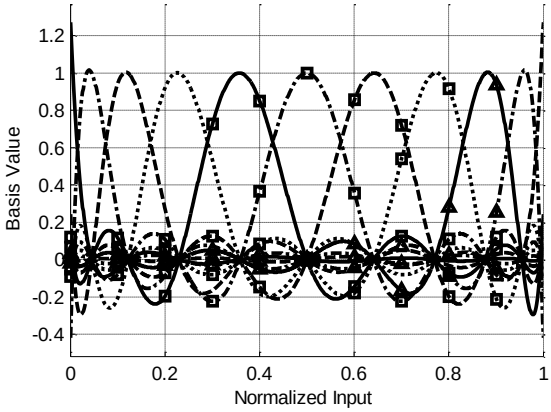


Fig. 1. Polynomial basis functions using Lagrange polynomials in the interval [0, 1], for 11 Chebyshev nodes. Note the locality of each element.

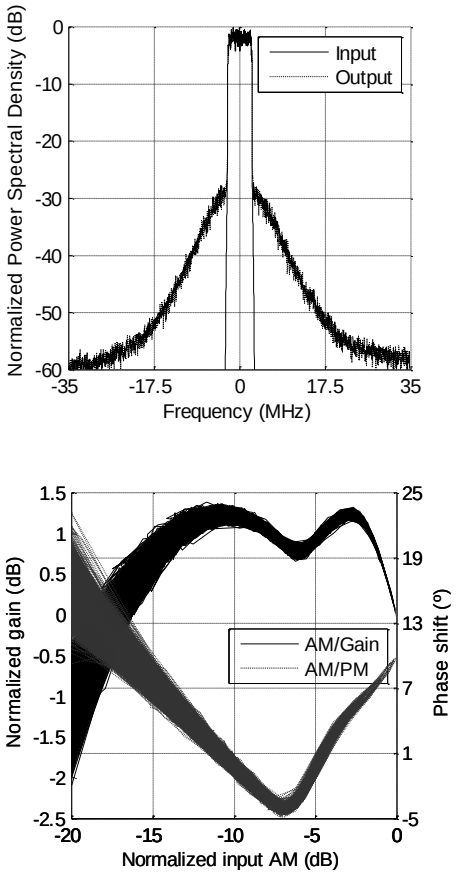


Fig. 2. Input and output spectrum, on top, and the PA's gain and phase characteristics for the excitation, on the bottom.

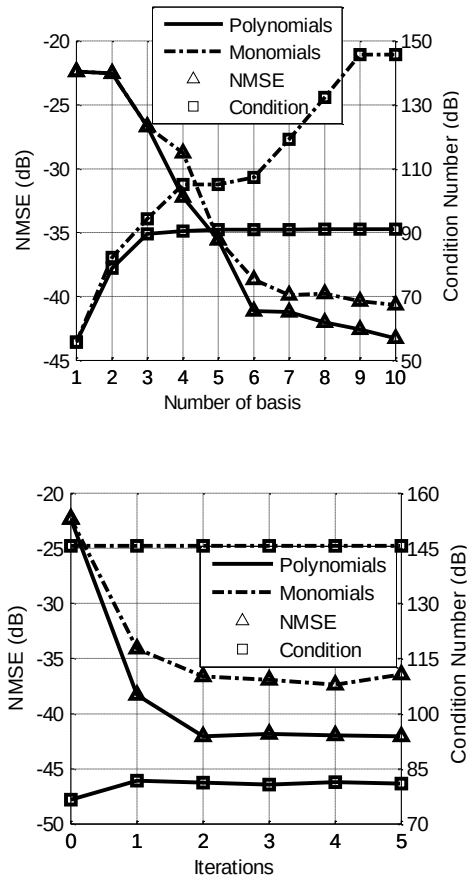


Fig. 3. Evolution of the condition number in dB (20log), and error, for increasing number of bases elements, in forward model identification on top, and in each iteration of the indirect learning procedure on the bottom.

V. Results

To evaluate the conditioning, the chosen metric is the condition number, which is herein defined as the ratio between the largest and smallest singular values of the regression matrix. To evaluate the obtained model, the parameters are extracted with the first and second half of the signals and compared using the NMSE.

The extracted model is the GMP described in (4). The regression is done using singular value decomposition (SVD). All data is represented in IEEE single precision floating-point format. The input-output data was obtained by exciting a 20W GaN Doherty PA with a 5MHz, single carrier, LTE signal (Fig. 2). The input and output signals are sampled at 100MHz, with 70MHz of usable bandwidth.

A. Forward Model Identification

For this test the number of bases elements to describe the functions f_{ml} of (4) was swept from 1 to 10 in L_a and from 0 to 3 in L_b and L_c. The memory depth related parameters were fixed to $\{L_a, L_b, L_c, M_b, M_c\}=\{2, 2, 2, 1, 1\}$. The first half of the signal (25000 samples) was used for extraction and the second half for the calculation of the error.

Fig. 3, on top, shows the variation of the condition number and error proving the smaller condition number and better model quality when using the proposed bases.

B. DPD Identification Via Indirect Learning

In this test the model was used as a pre-distorter of the PA, where the memory depth parameters were set to $\{L_a, L_b, L_c, M_b, M_c\}=\{2, 2, 2, 1, 1\}$. The number of bases functions was set to 10 in L_a and to 3 in L_b and L_c.

Fig. 3, on the bottom, shows the variation of the condition number and error of the PA output at each iteration of the indirect learning algorithm. A significantly better linearization performance is obtained with the model using the proposed polynomial bases.

To further test this technique, the indirect learning procedure was again run, but now for a wider bandwith signal: a two carriers LTE signal, where each carrier has 5MHz of bandwith. The carriers are separated by 10MHz in a 101 arrangement. The obtained linearization results are shown in Fig. 4.

VI. Conclusion

This paper showed a simple procedure to improve the conditioning of the least squares regression matrix. The

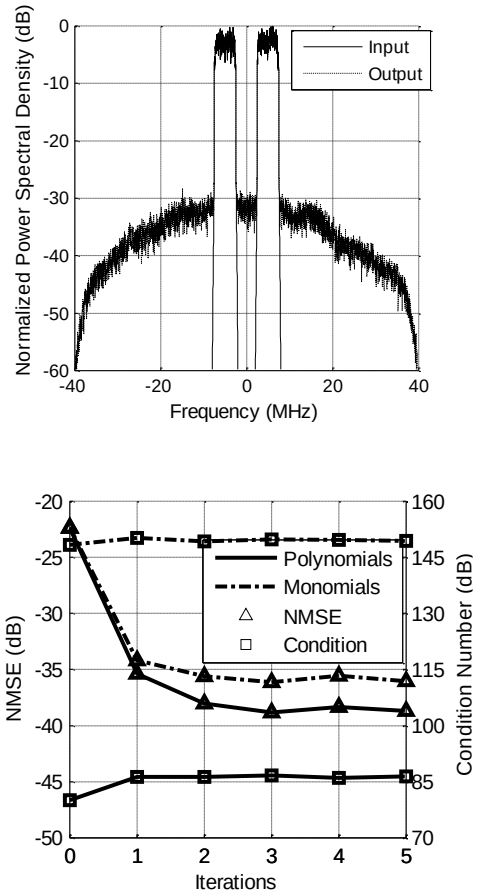


Fig. 4. Input and output spectrum, on top, and evolution of the condition number in dB (20log), and error in each iteration of the indirect learning procedure on the bottom.

introduced modification allows for the identification of the coefficients with 70dB lower condition number for the used signal. This results in an increase of 3-5dB in NMSE when trying to linearize the PA using single precision arithmetic.

The used basis functions can be derived starting from interpolated LUTs basis functions and thus create a bridge between these two modelling approaches.

ACKNOWLEDGEMENT

The first author would like to acknowledge the financial support provided by Portuguese Science and Technology Foundation, FCT (Ref. SFRH/BD/90103/2012).

The authors would like to thank the Huawei Sweden PA Design Team, namely Dr. Francesc Purroy, for interesting technical discussions regarding some of the observations dealt with in this work.

REFERENCES

[1] J. Kim; K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *Electronics Letters* , vol.37, no.23, pp.1417,1418, 8 Nov 2001

[2] O. Hammi; F. M. Ghannouchi; B. Vassilakis, "A Compact Envelope-Memory Polynomial for RF Transmitters Modeling With Application to Baseband and RF-Digital Predistortion," *Microwave and Wireless Components Letters*, IEEE , vol.18, no.5, pp.359,361, May 2008

[3] D. R. Morgan; Zhengxiang Ma; Jaehyeong Kim; M. G. Zierdt; Pastalan, J., "A Generalized Memory Polynomial Model for Digital Predistortion of RF Power Amplifiers," *IEEE Transactions on Signal Processing*, vol.54, no.10, pp.3852,3860, Oct. 2006

[4] Anding Zhu; J. C. Pedro; T. J. Brazil, "Dynamic Deviation Reduction-Based Volterra Behavioral Modeling of RF Power Amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, vol.54, no.12, pp.4323,4332, Dec. 2006

[5] Lei Guan; Anding Zhu, "Simplified dynamic deviation reduction-based Volterra model for Doherty power amplifiers," 2011 Workshop on *Integrated Nonlinear Microwave and Millimetre-Wave Circuits (INMMIC)*, vol., no., pp.1,4, 18-19 April 2011

[6] T. R. Cunha; E. G. Lima; J. C. Pedro, "Validation and Physical Interpretation of the Power-Amplifier Polar Volterra Model," *IEEE Transactions on Microwave Theory and Techniques*, vol.58, no.12, pp.4012,4021, Dec. 2010

[7] Lei Guan; Anding Zhu, "Optimized Low-Complexity Implementation of Least Squares Based Model Extraction for Digital Predistortion of RF Power Amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, vol.60, no.3, pp.594,603, March 2012

[8] R. Raich; Hua Qian; G. T. Zhou, "Orthogonal polynomials for power amplifier modeling and predistorter design," *IEEE Transactions on Vehicular Technology*, vol.53, no.5, pp.1468,1479, Sept. 2004

[9] L. Aladren; P. Garcia; P. L. Carro; J. De Mingo; C. Sanchez-Perez, "Digital predistortion based on Zernike polynomial functions for RF nonlinear power amplifiers," *2012 International Symposium on Wireless Communication Systems (ISWCS)*, vol., no., pp.865,869, 28-31 Aug. 2012

Paper J1: Polynomials and LUTs in PA Behavioral Modeling: A Fair Theoretical Comparison

Filipe M. Barradas, Telmo R. Cunha, Pedro M. Lavrador, and José C. Pedro

IEEE Transactions on Microwave Theory and Techniques

©2014 IEEE

Polynomials and LUTs in PA Behavioral Modeling: A Fair Theoretical Comparison

Filipe M. Barradas, *Student Member, IEEE*, Telmo R. Cunha, *Member, IEEE*, Pedro M. Lavrador, *Member, IEEE*, and José C. Pedro, *Fellow, IEEE*

***Abstract*— Traditionally, power amplifier behavioral modeling and digital predistortion have been based on polynomials or Volterra series. Recently, however, there has been a trend to substitute polynomial representations by interpolated look-up tables (LUTs) because of the known problems associated with the series' identification. These problems, commonly recognized as ill-conditioning, are generally related to the poor orthogonality verified in the polynomial basis functions for many common input signals.**

However, in this paper we will show that polynomials can be reformulated to achieve the identification quality of interpolated LUTs, while maintaining their beneficial properties. At the same time we will provide a unified view between polynomial and spline interpolation basis functions. By doing this we will present polynomials that share the benefits of spline interpolated LUTs and argue that both descriptions can be used interchangeably and with similar advantages.

***Index Terms*—Predistortion, Least squares approximations, Power amplifiers, Behavioral modeling,**

I. INTRODUCTION

VOLTERRA series is a polynomial expansion used to represent nonlinear systems of fading memory. It can, however, be interpreted as a particular instantiation of a system model composed of a one-to-many linear dynamic block followed by a many-to-one nonlinear static block. This is known as the Wiener architecture, and is shown in Fig. 1. In fact, the polynomial expansion is simply one of the possible representations for the multidimensional nonlinearity, while other representations such as LUTs also fit nicely into this theory.

In recent years, Volterra series has found applications in behavioral modeling and digital predistortion of radio-frequency (RF) power amplifiers (PAs). For this purpose, it is typically truncated according to some criteria, giving rise to several different models. The most widely used models are the memory polynomials [1]-[3], which are obtained by direct truncation of the series in its original form, and the dynamic deviation reduction based models [4]-[5], which were obtained through truncation of the series in the dynamic deviation format. These models can be presented in the traditional real and imaginary or polar [6] representations.

The first author would like to acknowledge the financial support provided by Portuguese Science and Technology Foundation, FCT (Ref. SFRH/BD/90103/2012). This paper is an expanded version from the IEEE MTT-S International Microwave Symposium, Tampa Bay, FL, USA, June 1-6 2014.

The authors are with DETI, Instituto de Telecomunicações, Universidade de Aveiro, Campus Universitário de Santiago, Aveiro 3810-193, Portugal (e-mail: filipebarradas@ua.pt; trcunha@ua.pt; plavrador@ua.pt; jcpedro@ua.pt).

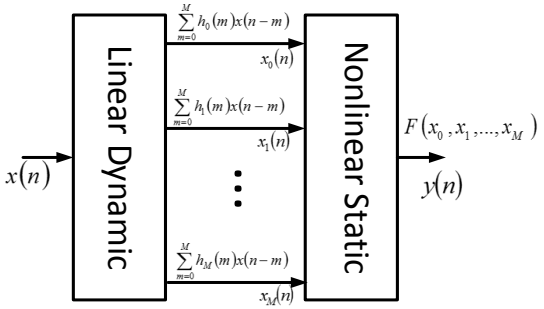


Fig. 1. Wiener architecture for the nonlinear dynamic (fading memory) system.

One of the main characteristics of Volterra based models, which has supported its widespread use, is that they can be seen has a linear combination of nonlinear, multidimensional basis functions. This means that, after evaluating these basis functions, identifying the system can be reduced to a linear problem of identifying θ, the parameter vector (the set of coefficients that multiply each basis function), so that it minimizes $\|y - X\theta\|^2$ where y is the measurements vector and the matrix X contains the chosen basis functions, evaluated with the tested excitation x.

One of the problems verified when trying to obtain the parameter vector θ, is that the matrix X may be ill-conditioned [7]-[9]. The conditioning problem makes the extraction more difficult and damages the quality of the solution. To escape from this problem, modifications on the nonlinearity representation have been suggested. These include polynomials orthogonal to some excitation probability distribution function [8], [9] or the use of interpolated LUT descriptions [10]-[12].

Moving from Volterra series approximations into spline-interpolated LUT descriptions can be seen as changing the basis functions from globally supported polynomials into splines, which are known to have a local support in the input amplitude space [10], [11].

We have shown in [13] how to use spline interpolated LUTs in Volterra based models. Furthermore, in [14] we pointed out that it is in fact the local support in the input amplitude that is the main strength of this basis. In both these works we have also stated that, by exploring the construction of the spline interpolated LUTs basis functions, a polynomial basis with

local support can be achieved, which can provide a link between these two approaches. Polynomial and interpolated LUTs are compared in two different scenarios: approximation capability and implementation. Usually, results using high-order polynomials point to poor approximation capabilities, generally attributed to ill-conditioning in the extraction process.

In fact, a comparison of LUTs and polynomials for PA behavioral modeling and linearization has already been performed in [15], which resulted in an opinion favorable to LUTs. However, we will now show that this comparison was biased in favor of the LUTs because the polynomials were described in their traditional non-local form. Actually, supported by the works we presented in [13] and [14], in this paper we will show that spline theory can be neatly unified with polynomial theory and that the polynomial basis can enjoy the locality of the spline basis as well. Using this unified theory we can compare the polynomials and the spline interpolated LUTs fairly and, contrary to what was presented and is widely believed, we will conclude that both representations can be used interchangeably and with similar characteristics.

Please note that this comparison is completely detached from the specific implementation of each of the models. It is out of the scope of this paper to compare the cost or speed of each solution as these will most likely depend on the actual hardware platform characteristics. The present analysis is restricted to the viewpoint of operator approximation theory.

This paper has the following structure. An introduction to the conditioning problem for linear identification is provided in Section II. This section presents the problems associated with the lack of orthogonality and poor extraction conditioning, as well as their underlying genesis. This exposition allows us to point out, in the following section, the reason why LUTs are preferred over polynomials. In Section III, a brief explanation of the two commonly used methods to improve conditioning is presented. Furthermore, Section III also explores the benefits commonly associated with spline representations. Then, a methodology for converting the nonlinear basis functions of common Volterra series based models is shown in Section IV. In addition, in Section IV, the spline basis is explored and the related polynomial basis is found, proving that they share equal approximation properties. Finally, some results of the application of this method to a particular power amplifier are shown in Section V and the conclusions are drawn in Section VI.

II. Orthogonality and Conditioning

The conditioning in the identification of a model is closely related to the orthogonality of its underlying basis functions.

The basis functions of a particular model can be thought of as descriptors of the effects observed at the output. Unfortunately, the effects described by distinct basis functions are, in general, not only signal dependent as they overlap. This makes orthogonality much more difficult to achieve as the signal properties, sometimes unknown, must be taken into account [8].

When we identify the model, i.e., find the model parameters, we must look at the effects that each basis function produces independently from the others. We can therefore imagine the linear solver as an orthogonalizer of the problem as it examines the correlation between the basis functions and separates them. It stands to reason that the closer our basis functions are to being independent from each other, the better the identification procedure will behave.

With this intuitive information in mind, we will now explore the concept of conditioning and orthogonality more deeply so that we can introduce the foundations of this work.

A. Basis functions and separable effects

The first thing that we will look at is what exactly is meant by basis functions and effects.

When identifying a system, we will have an input and output measurement data set which contains hints to the behavior of the system. In addition, we can imagine that this behavior is decomposable in a series of terms. These terms of the series are what are called the basis functions.

In most cases, each basis function does not model a particular separable effect, and the end results of distinct basis functions overlap.

A measure of the correlation or orthogonality of a basis element in relation to another is given by (1), where $w(x)$ is the probability distribution of the input, x, in the integration domain and $\varphi_k(x)$ and $\varphi_l(x)$ are the two evaluated basis functions.

$$O_{kl} = \int_{-\infty}^{+\infty} \frac{\varphi_k(x)}{\|\varphi_k(x)\|} \frac{\varphi_l^*(x)}{\|\varphi_l(x)\|} w(x)dx \tag{1}$$

In the following subsection, we will show that, when solving the identification problem, the algorithm will look at the chosen basis functions and generate an orthogonalized and normalized version of them. As the orthogonalization is performed through linear combination, the underlying basis remains the same.

B. Least squares solution

The orthogonality of the basis elements directly influences the selection of θ, i.e., the solution of the least squares problem ($\|y - X\theta\|^2$), which is conducted by finding the pseudo-inverse of the matrix X. This is the matrix containing the several basis functions evaluated in the particular excitation, as shown in (2), in which $x(n)$ is the excitation vector at time sample n.

$$X = \begin{bmatrix} \varphi_0(x(0)) & \varphi_1(x(0)) & \dots & \varphi_K(x(0)) \\ \varphi_0(x(1)) & \varphi_1(x(1)) & \dots & \varphi_K(x(1)) \\ \dots & \dots & \dots & .. \\ \varphi_0(x(N)) & \varphi_1(x(N)) & \dots & \varphi_K(x(N)) \end{bmatrix} \tag{2}$$

Let us now carefully analyze the pseudo-inverse procedure and find the orthogonalization mentioned above. The pseudo-inverse of a matrix X is given by (3), where $(.)^H$ represents the conjugate transpose and $(.)^{-1}$, represents the matrix inverse.

$$X^{-1} = C^{-1}X^H, \qquad C = X^H X \tag{3}$$

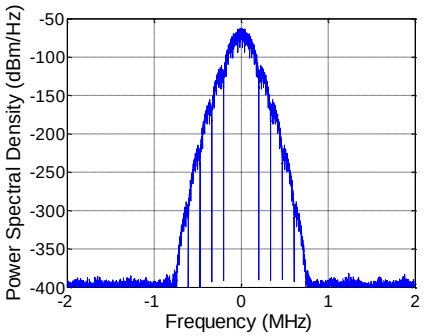


Fig. 2. GMSK signal spectrum.

Looking at (3) two different matrices are identifiable. The first one, C^{-1}, is the inverse of the correlation matrix, and the second, X^H, is the conjugate transpose of the original matrix.

Effectively, C contains the necessary information to separate the effects described by the basis. This can be more easily seen if C is decomposed. Equation (4) shows the eigendecomposition of C, in which U contains orthonormal eigenvectors and Λ is diagonal and contains the eigenvalues. Since C is Hermitian, the eigenvalues are real and non-negative. We will see later that these are the powers of the orthogonal effects described by our basis.

$$C = U\Lambda U^H \tag{4}$$

The eigenvectors in matrix U represent an orthonormal basis for the columns of X. X can thus be orthogonalized by projecting its columns over this basis as shown in (5), since Λ is diagonal (XU) must be orthogonal.

$$(XU)^H XU = U^H X^H XU = U^H U\Lambda U^H U = I\Lambda I = \Lambda \tag{5}$$

To correctly identify these orthogonal effects on the measurement y, they should be normalized so that no gain error is inserted. The orthonormalized problem and solution are then given by (6), and the identified parameters will correspond to the new orthonormalized space.

$$\left(XU\sqrt{\Lambda^{-1}}\right)\theta' = y, \qquad \theta' = \left(XU\sqrt{\Lambda^{-1}}\right)^H y = \sqrt{\Lambda^{-1}}U^H X^H y \tag{6}$$

Equation (6) is very similar to (3). Actually, the difference arises from the fact that the solutions exist in two different spaces. However, these spaces are linear transformations of each other, as is shown in (7). In (7), we first group the last three terms of the first part of (6) and realize that this must equal the original problem. Secondly, we use the second part of (6) to obtain the least squares solution for the original θ.

$$\begin{aligned}\left(XU\sqrt{\Lambda^{-1}}\right)\theta' &= X\left(U\sqrt{\Lambda^{-1}}\theta'\right) = X\theta = y \\ \theta &= U\sqrt{\Lambda^{-1}}\theta' = U\sqrt{\Lambda^{-1}}\sqrt{\Lambda^{-1}}U^H X^H y = C^{-1}X^H y\end{aligned} \tag{7}$$

In conclusion, the pseudo-inverse is a projection into, and a return from, an orthogonal and normalized description.

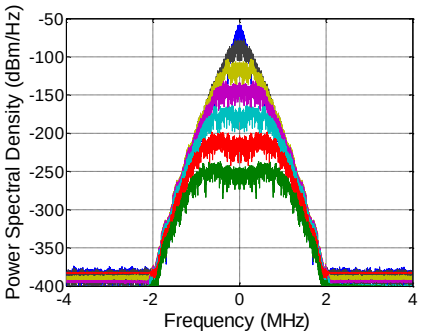


Fig. 3. Spectrum of the orthogonalized effects of the basis functions. Observe how the effects quickly fall to the noise floor showing that this is not a suitable basis for the signal. For instance, if we limit the dynamic range to 80 dB (-60 dB to -140 dB) only 4 out of 7 effects can be distinguished correctly.

C. Condition number

The conditioning in the linear least squares algorithm is measured using the condition number. The condition number is defined as the ratio of the highest to the lowest eigenvalue. Because this is a measure of the precision loss from the highest to the lowest power effect, it is also a measure of the sensitivity of the solution to perturbations of the correlation matrix. The condition number measures the worst case only, and a better evaluation could be obtained by looking at all the eigenvalues [16].

Let us first look at the condition number from the precision point of view. The numerical dynamic range of the IEEE double precision standard is around 313 dB (and 138dB for single precision), which is the typical numerical standard used in MatLab, the numerical platform used in this work. We can observe the degradation of the precision in each orthogonal effect with the following numerical experiment.

First, generate a cyclic signal in the sense that taking samples from the time sequence end and injecting them in the beginning, or vice-versa, does not change the signal's properties. In this case, the Discrete Fourier Transform (DFT) can be applied with no spectral leakage. Second, pick a static nonlinear basis and apply it to the signal, building matrix X, as in (2). Then, apply an orthogonal decomposition algorithm to X – singular value decomposition (SVD) can be used for numerical robustness. Finally, observe the orthogonalized effects in the frequency domain.

Figs. 2 and 3 show this process applied to a cyclic Gaussian Minimum Shift Keying (GMSK) signal sampled at 10 MHz and using the basis functions in (8).

$$\varphi(x) = \begin{bmatrix} x & x|x|^2 & x|x|^4 & x|x|^6 & x|x|^8 & x|x|^{10} & x|x|^{12} \end{bmatrix} \tag{8}$$

Note that odd orders of $|x|$ were avoided since they produce high order distortion which covers the noise floor. Fig. 2 shows the GMSK signal and Fig. 3 shows the orthogonalized effects of the basis, both in the frequency domain. The effects in Fig. 3 are linear combinations of the basis in (8), when applied to the GMSK signal, such that they are orthogonal to each other, starting from the highest power effect to the lowest.

The conditioning in this case degrades quickly with the

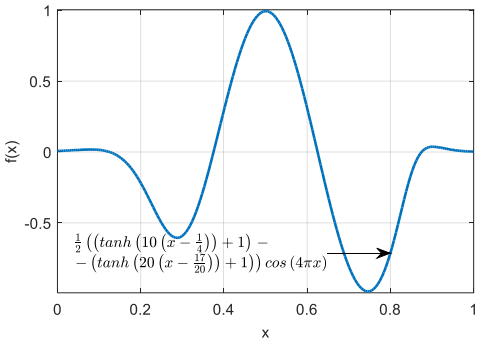


Fig. 4. Test function which requires a high polynomial order for correct approximation.

number of nonlinear basis functions because the GMSK signal is nearly constant in amplitude. Actually, the example was specifically designed in this way so that the effect is more perceptible in the frequency domain.

Let us now look at the condition number from the perspective of the sensitivity of the solution to small errors.

The orthogonalized vectors obtained from the matrix *X* contain the information of the separable effects which the chosen basis is able to describe. By projecting the output over these effects we will discover the power each effect has on the output. As we have seen, for high condition numbers the low power effects are masked by numerical noise. When we measure these effects in the output we will obtain a value that is corrupted as well. This means that it will be highly tuned for that particular measurement. This causes the solution this way obtained to be very sensitive to changes in the excitation signal, and to measurement and numerical noise [16].

Furthermore, please remember that the solution of the orthogonal problem is scaled by the inverse of the square root of the powers of each effect and that the coefficients are mixed to obtain a solution in the original problem. Actually, we are attributing an enormous gain to parameters we cannot correctly measure and then mixing them with all the others.

III. Improving The Extraction Conditioning

To alleviate conditioning problems several techniques can be used. In the case of Volterra series models used in predistortion and behavioral modeling, two strategies are usually employed. The first is regularization, typically Tikhonov regularization [19], and the second is changing the basis functions. Here we will present a brief description of this last method, which applies directly to our study.

A. Changing the description

Another way to avoid ill-conditioning is to change the basis that are used to describe the system. This means that we are changing the matrix *X* itself, and so, modifying the original problem. Unfortunately, this is not an easy task as it involves mathematical transformations of the original model to make them more suitable for identification. Nevertheless, changing the description is supported by strong mathematical arguments, providing a method of identification which is generally more robust.

Two different modifications have been performed to Volterra series based models. One is changing the memory

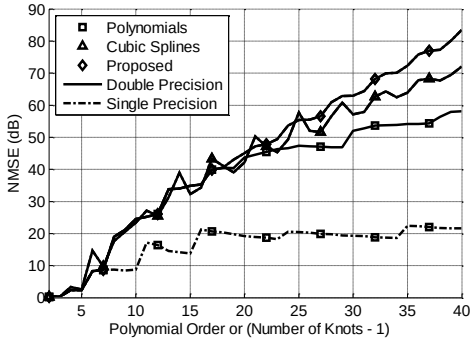


Fig. 5. Approximation error in NMSE for the typical polynomials, cubic splines, and the polynomials proposed in the next section. For both the splines and the proposed polynomials the error behaves similarly for single and double precision which shows good conditioning of the problem.

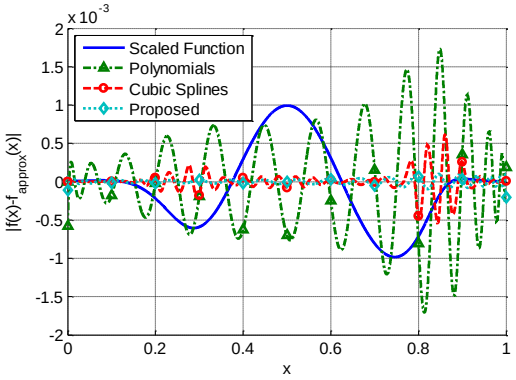


Fig. 6. Error behavior for the typical polynomials, cubic splines and the polynomials proposed in the next section. Both the cubic splines and the proposed polynomials show good representation capabilities.

descriptors, using different filter types instead of the typical delays [17]-[18]. The other consists of changing the nonlinearity type [8]-[12].

Since power amplifiers are typically regarded as nonlinear devices with a small memory span, changing the nonlinear systems basis functions is more appealing for the models of these devices.

A careful observation of (1) helps in deciding how exactly the models should be changed. Since there is no *a priori* information on the distribution of the signals, we have to work on reducing the integral independently from this parameter. One way to achieve this is to limit the influence of each basis function to a particular interval of the amplitude space. When doing this, the multiplication of one basis function by another will immediately be reduced because they exist in different intervals.

B. LUTs and locality

The common benefit associated with spline descriptions is locality. By focusing the approximation power of low order polynomials on increasingly smaller intervals of the functions we expect to get rid of Runge's phenomenon while maintaining an accurate representation of the underlying function. Because of the way we think of spline interpolated LUTs, it is straightforward and intuitive to understand that the concept of locality applies directly to them. Conversely, when thinking in terms of polynomials this possibility tends to be disregarded since the general assumption is that polynomials cannot possibly be local. We will show next that this

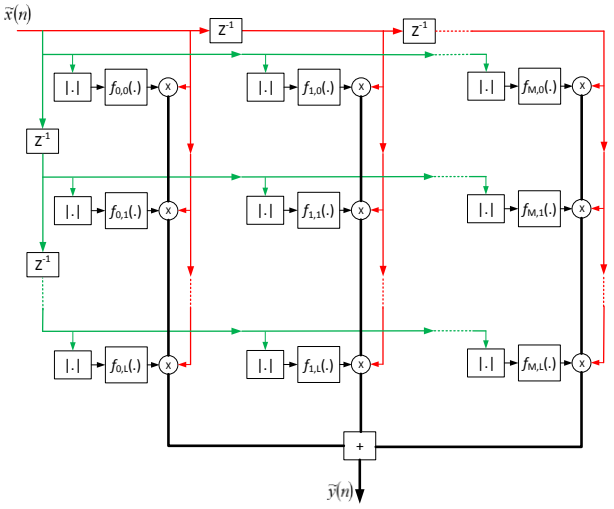


Fig. 7. Structure of the model presented in (13).

assumption is false and that the concept of locality is also present in a particular polynomial representation.

The implication of these local polynomials is that spline representations (interpolated LUTs) then lose their main advantage when compared against them.

As an example, consider the least-squares approximation, to the function in Fig. 4, using the typical polynomials, cubic splines and the polynomials we will describe in the next chapter. The approximation is done using 10000 equally spaced function points while the maximum number of coefficients is 41, to ensure a well posed least squares problem. The results are shown in Fig. 5 for increasing polynomial order, or number of spline partitions, and two different numerical precision standards. Finally, in Fig. 6 the behavior of the error in the interval, for the best approximation, is plotted.

From this experiment we understand that, while the classical polynomials suffer from numerical problems, both the spline interpolated LUTs and the polynomials we will present later do not. This comes from the fact that the locality we observe in spline approximations will also be present in the proposed polynomial basis functions.

IV. Transforming Volterra Series Models

The general Volterra series is very rarely used for RF PA modeling and predistortion and, when it is, it is generally used with small nonlinear orders. This is because the number of parameters grows with M^P, where M is the number of memory taps and P is the nonlinear order. This number of parameters quickly becomes difficult to manage.

Two main model families are used in modeling and predistortion: the memory polynomials (memory polynomial (MP [1]), envelope memory polynomial (EMP [2]) and generalized memory polynomial (GMP [3])); and the dynamic deviation reduction's (DDR [4]) family (first (DDR1) and simplified second order (SDDR2 [5])). Even for these truncated cases the number of parameters must be correctly managed.

Since the Volterra series is itself a polynomial expansion, the original form of the truncated models is also in a polynomial form. It turns out that this description is typically

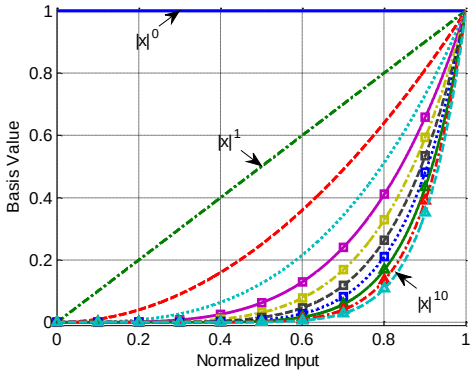


Fig. 8. Monomial basis functions from order zero to order ten. Note how each basis becomes more and more similar to the previous one.

not well suited for identification because the basis functions exhibit high correlation for most input signals. To understand why this happens, let us first observe the GMP and describe what exactly is meant by polynomial basis.

The GMP model was presented as shown in (9) where $\widetilde{(.)}$ means that the signal is a complex envelope signal. Note that it is equivalent to the model presented in (10), being this last form used for readability reasons.

$$\begin{aligned}\tilde{y}(n)=&\sum_{p=0}^{P}\sum_{m=0}^{M}a(p,m)\tilde{x}(n-m)\left|\tilde{x}(n-m)\right|^{p}\\&+\sum_{p=1}^{P}\sum_{m=0}^{M}\sum_{l=1}^{L}b(p,m,l)\tilde{x}(n-m)\left|\tilde{x}(n-m-l)\right|^{p}\\&+\sum_{p=1}^{P}\sum_{m=0}^{M}\sum_{l=1}^{L}c(p,m,l)\tilde{x}(n-m)\left|\tilde{x}(n-m+l)\right|^{p}\end{aligned}\tag{9}$$

$$\tilde{y}(n)=\sum_{p=0}^{P}\sum_{m_1=0}^{M}\sum_{m_2=m_1-L}^{m_1+L}h(p,m_1,m_2)\tilde{x}(n-m_1)\left|\tilde{x}(n-m_2)\right|^{p}\tag{10}$$

Taking (10) we can now switch the summations' order so that the nonlinear terms are summed first. However, the summation of terms in a polynomial is a function. In fact, we can assume that there is an underlying function, $f_{m_1,m_2}(\cdot)$, which we approximated using a polynomial. Therefore, there is no loss of generality when we say that the GMP has the form of (11), and the structure shown in Fig. 7.

$$\tilde{y}(n)=\sum_{m_1=0}^{M}\tilde{x}(n-m_1)\sum_{m_2=m_1-L}^{m_1+L}f_{m_1,m_2}\left(\left|\tilde{x}(n-m_2)\right|\right)\tag{11}$$

The functions f are complex functions of a real variable. In principle, we could use several different approximators. What we mean when we say that the polynomials were used, is, actually, that the monomial basis of (12) was used to expand the functions.

$$\left[\left|\tilde{x}\right|^{0}\quad\left|\tilde{x}\right|^{1}\quad\ldots\quad\left|\tilde{x}\right|^{p}\right]\tag{12}$$

If we observe the behavior of this basis as the order is raised, we immediately understand why the correlation

increases. This is illustrated in Fig. 8 where it is shown that, regardless of the input signal distribution, this basis presents poor orthogonalization.

After reaching the description of (11) for the GMP, we can change the basis functions used for the approximation to whichever we want. We can therefore generally write the GMP equations as shown in (13) where $B_p(.)$ are the chosen basis functions. In particular, we would like to use functions which are easily separable. As discussed before this can be achieved using $B_p(.)$ that exhibit local behavior in amplitude.

$$\tilde{y}(n)=\sum_{m_1=0}^{M}\tilde{x}(n-m_1)\sum_{m_2=m_1-L}^{m_1+L}\sum_{p=0}^{P}h(p,m_1,m_2)B_p(|\tilde{x}(n-m_2)|) \qquad (13)$$

Even though it is commonly assumed that spline interpolated LUTs will work better than polynomials, we will show here that, generally, such comparison is unfair for the polynomials. To show this, we will first describe the method for obtaining the spline basis, we will then demonstrate that this method converges to a polynomial basis (different from the monomials) under certain conditions, and finally we will compare spline and polynomial basis derived from the same methodology. We believe that this comparison is fair and it shows that polynomials, although originally suffering from orthogonality problems, can be reformulated into an identification friendly basis, which is as convenient as the spline interpolated LUTs' basis.

A. Spline interpolated LUTs

The spline basis is a particular class of piecewise polynomial basis. Specifically, a spline basis is defined over a partitioned interval; the points which separate the several interval partitions are called knots. Also, a spline has an associated order – this order dictates up to which order the spline must maintain continuity of the derivatives. An order 3 spline, also called cubic spline, maintains continuity of the function it approximates up to the second order. In general, an order P spline will maintain continuity up to order P-1. Furthermore, an order P spline is also built of order P piecewise polynomials. After we define the knots and the spline order, we can almost fully describe a spline basis; depending on the spline order, a few more conditions are needed, as we will see.

For N+1 knots we can define N, order P, piecewise polynomials. These polynomials require $N(P+1)$ coefficients to be fully defined. However, because of the continuity conditions, the coefficients are not independent. In fact, because of this dependence the spline can also be fully described by a set of N+1 basis functions (for any spline order) instead of the $N(P+1)$ coefficients.

A spline function is defined by a set of piecewise polynomials as shown in (14), where x_i are the knots. However, if the coefficients were independent we would be defining a piecewise polynomial which may not be a spline. The continuity conditions restrict the solutions for the coefficients. By exploiting these restrictions we can define a different basis of the spline space.

$$y=q_i(x),\ \forall x\in[x_i,x_{i+1}],\qquad q_i(x)=\sum_{p=0}^{P}a_{i,p}(x-x_i)^p \qquad (14)$$

where $i\in[0,N-1]$.

1) Changing the LUTs' basis

The basis we will define for the spline space has a number of advantages over using the polynomial coefficients:

- The number of parameters is reduced;
- Any combination of the basis leads to a spline;
- The basis functions are orthogonal at the knots.

This basis can be generated using a linear map from the coefficients of the piecewise polynomials to the function values at the knots. To do this we use the spline conditions shown in (15), which can be defined at all knots but the last [20].

$$\begin{cases} y_i=q_i(x_i) \\ 0=q_i(x_{i+1})-q_{i+1}(x_{i+1}) \\ \dots \\ 0=q_i^{(P-1)}(x_{i+1})-q_{i+1}^{(P-1)}(x_{i+1}) \end{cases} \qquad (15)$$

The conditions in (15) can be put in a matrix form using the closed expression for the derivatives of q_i shown in (16).

$$q_i^{(k)}=\sum_{p=k}^{P}a_{i,p}\frac{p!}{(p-k)!}(x-x_i)^{p-k} \qquad (16)$$

The matrix form for these conditions can be built as shown in (17). The resulting matrix is a sparse block matrix which is easy to manage with simple algorithms. The problem is, however, incomplete as the defined matrix has size $N(P+1)$ by $(N-1)(P+1)$ because we cannot define (15) at the last edge knot, since q_{i+1} does not exist. Actually, we are missing P+1 conditions to fully define a spline.

$$\begin{bmatrix} \vec{b}_0 \\ \vec{b}_1 \\ \vec{b}_2 \\ \dots \\ \vec{b}_{N-3} \\ \vec{b}_{N-2} \end{bmatrix}=\begin{bmatrix} A_0 & I_P & [0] & \dots & [0] & [0] \\ [0] & A_1 & I_P & \dots & [0] & [0] \\ [0] & [0] & A_2 & \dots & [0] & [0] \\ \dots & \dots & \dots & \dots & \dots & \dots \\ [0] & [0] & [0] & \dots & I_P & [0] \\ [0] & [0] & [0] & \dots & A_{N-2} & I_P \end{bmatrix}\begin{bmatrix} \vec{a}_0 \\ \vec{a}_1 \\ \vec{a}_2 \\ \dots \\ \vec{a}_{N-2} \\ \vec{a}_{N-1} \end{bmatrix} \qquad (17)$$

Where,

$$A_i=\begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & d_i & d_i^2 & d_i^3 & \dots & d_i^P \\ 0 & 1 & 2d_i & 3d_i^2 & \dots & Pd_i^{P-1} \\ 0 & 0 & 2 & 6d_i & \dots & \frac{P!}{(P-2)!}d_i^{P-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & P!d_i \end{bmatrix} \qquad (17.a)$$

$$d_i=(x_{i+1}-x_i)$$

together with,

$$I_P = \begin{bmatrix} 0 & 0 & 0 & ... & 0 & 0 \\ -1 & 0 & 0 & ... & 0 & 0 \\ 0 & -1 & 0 & ... & 0 & 0 \\ 0 & 0 & -2 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & -(P-1)! & 0 \end{bmatrix} \quad (17.b)$$

represent the continuity information from eq. (15) at the knots. And,

$$\vec{b}_i = \begin{bmatrix} y_i \\ 0 \\ 0 \\ 0 \\ ... \\ 0 \end{bmatrix} \qquad \vec{a}_i = \begin{bmatrix} a_{i,0} \\ a_{i,1} \\ a_{i,2} \\ a_{i,3} \\ ... \\ a_{i,P} \end{bmatrix} \quad (17.c)$$

represent the function values and coefficient domains which we are relating.

The missing conditions can be defined in a number of ways. Since we are looking for a methodology which allows seamless use of any spline order, the chosen method was the not-a-knot conditions [21]. These were integrated in the following fashion in the problem. First, the conditions in (18), which can be defined for the last knot, were integrated in the original problem. The remaining (*P*-1) conditions are obtained by forcing the continuity of one more derivative on the edge knots, for odd spline orders, or for one center knot (or one of them, for even knot numbers) and the edge knots for even spline orders.

$$\begin{cases} y_{N-1} = q_{N-1}(x_{N-1}) \\ y_N = q_{N-1}(x_N) \end{cases} \quad (18)$$

The only exception to this approach is the zero order spline. For this case, the first condition in (18) is implemented but one more spline basis is added to meet the second condition. This is done so that the number of basis functions is kept consistent with the other spline orders.

To add a continuity condition to the system, the matrices A_i and I_P need to be changed for the knots at which these conditions are implemented. At these knots (15) is expanded to (19) and, A_i and I_p are changed to (20). The only difference with respect to what we have seen before is the last continuity equation, which is translated to the last row of the matrixes.

$$\begin{cases} y_i = q_i(x_i) \\ 0 = q_i(x_{i+1}) - q_{i+1}(x_{i+1}) \\ ... \\ 0 = q_i^{(P-1)}(x_{i+1}) - q_{i+1}^{(P-1)}(x_{i+1}) \\ 0 = q_i^{(P)}(x_{i+1}) - q_{i+1}^{(P)}(x_{i+1}) \end{cases} \quad (19)$$

$$A'_i = \begin{bmatrix} & A_i & \\ 0 & ... & P! \end{bmatrix}$$

$$I'_P = \begin{bmatrix} & I_p & \\ 0 & ... & P! \end{bmatrix} \quad (20)$$

Adding this continuity condition, forces the piecewise polynomial defined in one interval to be continuous with the one defined in the adjacent interval, for all derivatives. This condition is as if the knot ceased to exist since it no longer defines a break-point (not-a-knot). Following this rationale, as we add more and more not-a-knot conditions we tend to a polynomial approximation.

In a cubic spline the changed matrices would be A_0 and A_{N-2}, imposing continuity conditions on the knot 1 and *N*-1, which means that the first two, and the last two, piecewise polynomials are now continuous at all derivatives. In a fifth order spline it would be A_0, A_1, A_{N-3} and A_{N-2}, forcing the aforementioned continuity on the first three and last three piecewise polynomials, and so forth for higher orders. Since, as we explained, there is no discontinuity for any order of the derivative, for the modified knots, for a sufficiently high spline order we are actually defining a polynomial. This happens when the *P*-1 missing conditions, after (18) is included, match the *N*-1 knots where the continuity conditions can be defined; in other words, when the number of knots exceeds the spline order by one.

After we fully define the problem we have the $N(P+1)$ polynomial coefficients on one side which are linearly transformed into the $N+1$ y_i values on the other. This means that the actual problem can be defined in both ways. Since the number of unknowns is significantly lower, and insensitive to the spline order, when defined in terms of the y_i this basis is advantageous. This discrepancy in terms of model parameters is due to the fact that the polynomial coefficients describe any piecewise polynomial function and not only splines (to obtain a spline we must force the continuity conditions). On the other hand, the description in terms of y_i values can only describe a spline, since the continuity conditions have been used in its genesis. In general, this means that picking any set of polynomial coefficients may not lead to a spline, while any set of y_i values will always lead to a spline. The problem now is how we transform the problem from one description to the other, since we start with the coefficients' description. In fact, what we want to find are the basis functions which describe the spline in terms of y_i.

We do not need to explicitly calculate the underlying basis functions. Using the matrix defined in (17) with the added conditions we can start in the original, coefficients' based, description and switch to the new one as shown in (21). This is true for any well-defined linear transformation.

$$\begin{bmatrix} \vec{b}_0 \\ \vec{b}_1 \\ \vec{b}_2 \\ ... \\ \vec{b}_{N-2} \\ \hat{b}_{N-1} \end{bmatrix} = A \begin{bmatrix} \vec{a}_0 \\ \vec{a}_1 \\ \vec{a}_2 \\ ... \\ \vec{a}_{N-2} \\ \vec{a}_{N-1} \end{bmatrix} \qquad A^{-1} \begin{bmatrix} \vec{b}_0 \\ \vec{b}_1 \\ \vec{b}_2 \\ ... \\ \vec{b}_{N-2} \\ \hat{b}_{N-1} \end{bmatrix} = \begin{bmatrix} \vec{a}_0 \\ \vec{a}_1 \\ \vec{a}_2 \\ ... \\ \vec{a}_{N-2} \\ \vec{a}_{N-1} \end{bmatrix} \quad (21)$$

where,

$$\hat{A}_{N-1}=\begin{bmatrix}1 & 0 & 0 & ... & 0\\ 1 & d_i & {d_i}^2 & ... & {d_i}^P\end{bmatrix} \quad \hat{b}_{N-1}=\begin{bmatrix}y_{N-1}\\ y_N\end{bmatrix} \qquad (21.a)$$

are the conditions taken from eq. (18), now in matrix form. And,

$$A=\begin{bmatrix}A_0 & I_P & [0] & ... & [0] & [0]\\ [0] & A_1 & I_P & ... & [0] & [0]\\ [0] & [0] & A_2 & ... & [0] & [0]\\ ... & ... & ... & ... & ... & ...\\ [0] & [0] & [0] & ... & A_{N-2} & I_P\\ [0] & [0] & [0] & ... & [0] & \hat{A}_{N-1}\end{bmatrix} \qquad (21.b)$$

is the matrix with the complete set of continuity conditions.

In general, the dimensionality of A and its inverse in (21) should be the same. However, the elements in b are mostly zero. In fact, exactly N+1 elements are different from zero, which means that we can discard most of the columns of A^{-1}. When we eliminate these columns from A^{-1}, a transformation matrix, of dimension $N(P+1)$ by N+1, from the coefficient space to the y_i space is obtained and (22) can be written:

$$\begin{bmatrix}\vec{a}_0\\ \vec{a}_1\\ \vec{a}_2\\ ...\\ \vec{a}_{N-2}\\ \vec{a}_{N-1}\end{bmatrix}=\Gamma\begin{bmatrix}y_0\\ y_1\\ ...\\ y_N\end{bmatrix} \qquad (22)$$

Using the transformation matrix Γ the problem can be described in the original representation and transformed into the new one. A closed form expression for the new basis functions can be obtained, or the basis functions can be sampled to the desired precision and stored. These choices tradeoff simplicity, speed and accuracy.

An issue that is frequently raised in the comparison of LUTs and polynomials is the implementation cost. We have shown here that the required number of basis functions for any description – spline interpolated LUTs or polynomials – is the same. In terms of implementation cost, only the hardware cost for the calculation of the particular functions associated with the given spline order or polynomial could make a difference. However, this is a problem ultimately related to the choice of implementation which might favor different types of calculation.

2) Choosing the knots

When we defined the spline basis, showed that it can be defined with the same number of basis independently of the spline order and described how we can shift between representations, we based the definition of the spline in the existence of the knots x_i. Actually, we even defined the new space making use of the corresponding spline value at the knot y_i, without exactly defining these knots.

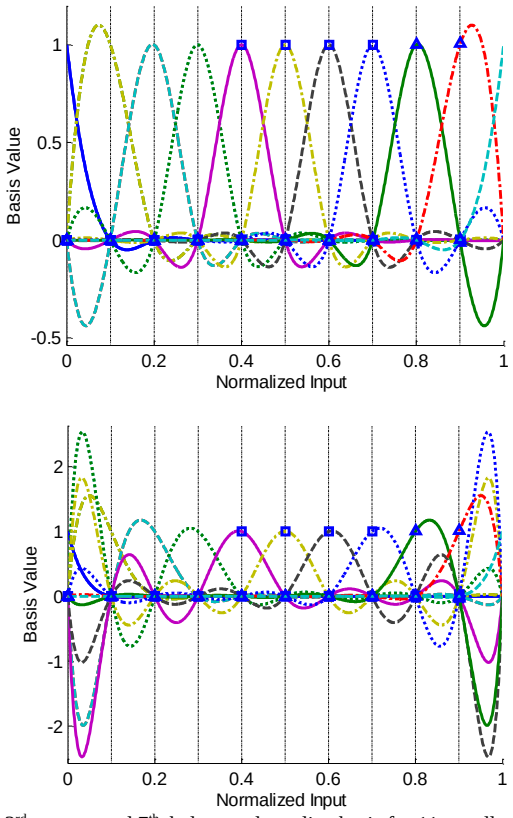


Fig. 9. 3rd, on top, and 7th, below, order spline basis for 11 equally spaced knots. Note how the basis functions exhibit a poor behavior at the edges of the interval. This behavior gets worse with the spline order.

In order to explore the possibilities for the definition of the knots we will work on the interval from zero to one. This interval can then be scaled and shifted to cover the desired region.

A good definition of the knots is important for the performance of the basis. To understand this, let us observe what happens when we choose linearly spaced knots. Fig. 9 shows the cubic spline basis, on top, and the seventh order spline basis, below, defined over eleven equally spaced knots.

When no care is taken in selecting the knots, the behavior of the spline basis close to the limits of the considered interval tends to become highly oscillatory, more so for higher order spline basis. This is a manifestation of Runge's Phenomenon.

Looking at the obtained basis functions we can observe that on the knots only one of them is one while all the others are zero. The undesired behavior occurs in between knots and close to the edge of the intervals. It seems reasonable that by compressing the distance between the knots at the interval's edge the basis will necessarily be smoother. A knot distribution that fits this criteria is the Chebyshev knot distribution, which was created to suppress Runge's Phenomenon in polynomial interpolation [22]. For this interval, the Chebyshev knots are given by (23).

$$x_i=\frac{1}{2}+\frac{1}{2}\cos\left(\frac{2(i+1)-1}{2(N+1)}\pi\right) \qquad (23)$$

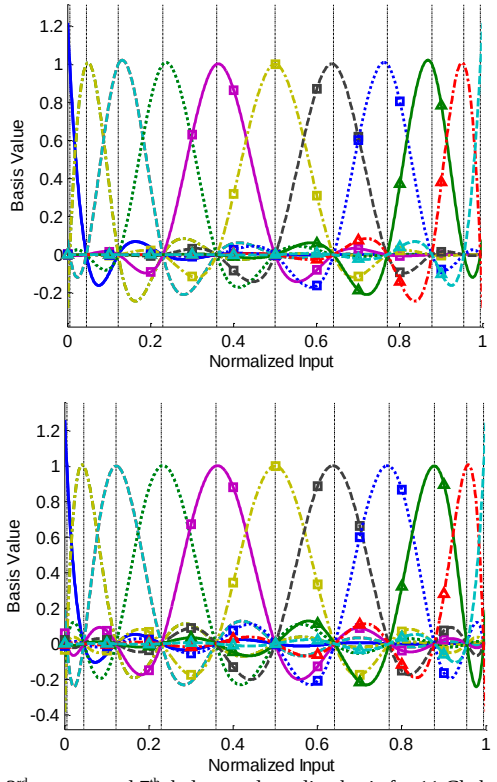


Fig. 10. 3rd, on top, and 7th, below, order spline basis for 11 Chebyshev knots. Note how the basis functions exhibit a controlled behavior throughout the interval. The previously observed oscillation at the edges has been eliminated.

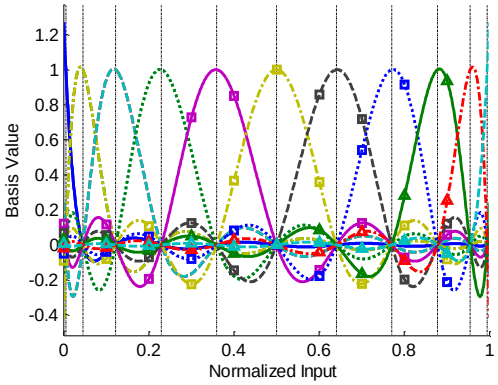


Fig. 11. Lagrange polynomials for 11 Chebyshev knots. The similarities to the spline basis functions shows their unified genesis.

If we create the spline basis over the Chebyshev knots the basis plotted in Fig. 10 are obtained. As we can see, the detrimental effect on error at the edges has been eliminated as intended.

Other types of knot distribution can be used. However, using the Chebyshev knots leads to a basis with good characteristics as we will show in the results. Observe, in particular, that every basis exhibits a high degree of locality in the sense that its value is low in almost every excitation interval except in the vicinity of the particular knot they are attached to.

B. Polynomial models with Lagrange basis

As we made notice before, for a sufficiently high spline order the basis we obtain is actually a polynomial basis. However, the basis obtained in this way is different from the monomials. Looking at Fig. 10 we understand that the basis functions are similar for different spline orders. This being the case, it would be very strange to converge to the monomials

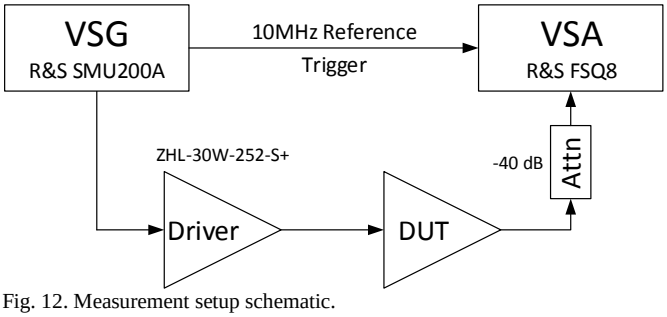


Fig. 12. Measurement setup schematic.

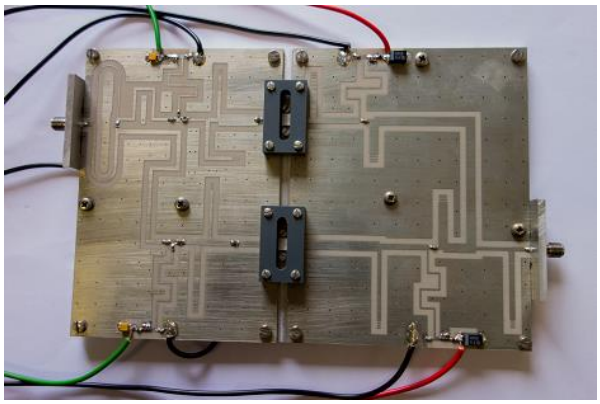


Fig. 13. The DUT used to perform the measurements in this work.

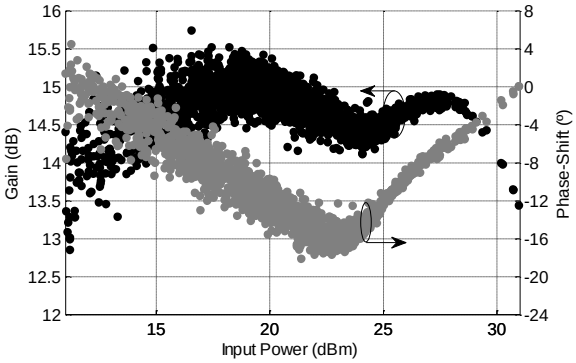


Fig. 14. Gain (in black) and phase-shift (in grey) behavior of the DUT with input power.

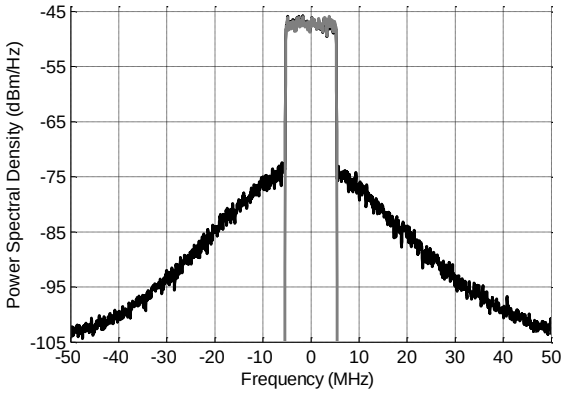


Fig. 15. Output (in black) and scaled input (in gray) spectrum of the DUT.

TABLE I
MODELING RESULTS

Spline Order	NMSE floating (dB)	NMSE fixed (dB)	Condition Number (dB)
1	42.0	41.9	66.4
3	44.3	44.1	67.4
5	44.4	44.0	67.6
7	44.0	43.5	67.4
Lagrange Polynomials	43.4	42.9	67.4

Modeling results for the validation signal.
Initially, the models are extracted using a different signal.

TABLE II
PREDISTORTION RESULTS

Spline Order	NMSE (dB)	ACPR (L/R) (dB)	Condition Number (dB)
1	39.5	48.0 / 47.0	55.0
3	40.9	49.9 / 48.7	57.7
5	41.5	50.9 / 49.7	58.3
7	41.5	50.9 / 49.6	57.6
Lagrange Polynomials	41.2	50.6 / 49.2	57.5

Predistortion results for the validation signal.
Initially, the models are trained using a different signal through the indirect learning technique.

Model parameters: P = 9, M = 3, when $m_1 = m_2$, or P = 3, M = 1, L = 1, when $m_1 \neq m_2$

which differ greatly from the spline basis.

The polynomial basis obtained from raising the spline order to the P-1 limit is actually the Lagrange polynomial basis, given by (24).

$$l_i(x) = \prod_{\substack{p=0 \\ p \neq i}}^{P} \frac{x - x_p}{x_i - x_p}, i \in [0, P] \subset Z \qquad (24)$$

Plotting the Lagrange polynomials for the Chebyshev knots immediately reveals their connection to the spline basis, as is shown in Fig. 11. Using this theory we can therefore unite the spline based and polynomial based models.

Since the polynomial solution is unique, the Lagrange polynomials and the monomials are equivalent. However, the Lagrange polynomials will provide much better conditioning than the monomials since they share the locality properties of the spline basis. Therefore, a fair modeling comparison between the spline-based and polynomial-based models must consider that the latter is built over a set of basis functions, such as the Lagrange polynomials, that show a similar locality to those of the spline basis functions.

V. RESULTS

In order to test the spline-based and Lagrange polynomial-based descriptions, two different tests were considered. The first one consists of modeling the response of an RF PA. The second consists of digitally predistorting the same system. In both tests the system is trained using one signal realization and validated using another. We are interested in several metrics: the condition number (ρ), which will measure the conditioning of the extraction problem; the normalized mean squared error (NMSE), which will measure the performance of the models; and, for the predistortion test, the achieved adjacent channel power ratio (ACPR). For the ACPR the out-of-band power is integrated in the same bandwidth as the in-band power shifted to the left or to the right by the integration bandwidth. All processing was done using MatLab, and all numbers were represented using the IEEE double precision floating point standard.

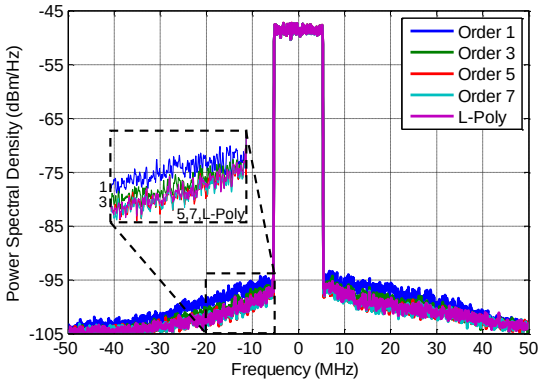


Fig. 16. Output signal spectrum showing the obtained linearization for the different spline orders.

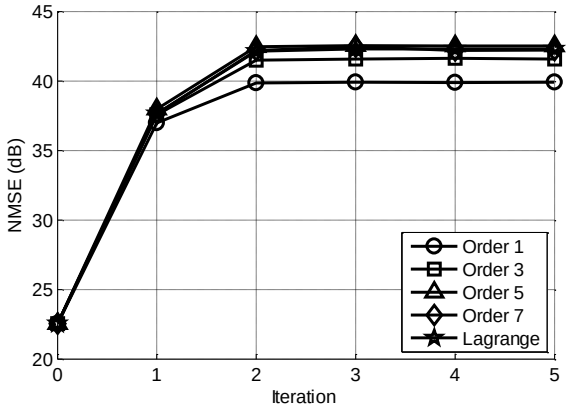


Fig. 17. Convergence behavior, in NMSE, during predistorter training.

The measurements were executed using a VSG to translate the digital signal to the RF domain and a VSA to translate the RF signal to the digital domain, as shown in Fig. 12. The device under test (DUT), shown in Fig. 13, was a 25 W Doherty (DHT) PA, built using CGH35015 Gallium Nitride (GaN) High Electron Mobility transistors (HEMT) from Cree, operating at 900 MHz. The PA presents an output power of approximately 44 dBm and a gain of 14 dB at the 1 dB compression point. The DUT was excited with a 10 MHz bandwidth LTE signal sampled at 100 MHz. Figs. 14 and 15 show the response of the amplifier. Fig. 14 shows the AM to Gain (AM/AM) and AM to Phase-Shift (AM/PM) characteristics and Fig. 15 the output spectrum.

The used PA and DPD model was the GMP with nonlinear order (number of knots - 1) 9 for the main terms and nonlinear order (number of knots - 1) 3 for the cross-terms, the memory-depth is 4 taps for the main terms and 1 tap for the cross terms, yielding a total of 52 (causal) parameters. In both tests the

models are trained and validated using different realizations of the signal.

A. Behavioral modeling results

The PA modeling results, referring to the validation signal are shown in Table I, for both floating- and fixed-point scenarios. In the fixed-point case, the signals were stored in 16 bits and the coefficients in 24 bits, intermediate calculations were run in 32 bits for multiplications and 48 bits in accumulation. The results for the modelled system are similar for most spline orders and in both number formats. Only the linear (order 1) spline shows a slightly worse result.

B. Predistortion results

For this test, the digital predistorter is first trained using one signal through the indirect learning technique [23] and then tested for a different signal.

The predistortion results for the validation signal are summarized in Table II, and the corresponding output spectra are plotted in Fig. 16. Note how the models behave similarly almost independently from the spline order. The worst results are, in fact, for the lower spline orders.

For reference, the convergence behavior for each spline order is plotted in Fig. 17.

VI. CONCLUSIONS

The major conclusion of this work is that, contrary to the widely spread belief, using polynomials or look-up tables in behavioral models will lead to fundamentally the same results. In fact, we demonstrated that, substituting the traditional monomials, as primary basis functions, by adequate polynomials bases, the major argument previously favoring the LUT description – which was locality – could now also be applied to polynomials.

We could furthermore conclude that any high errors observed in polynomial models are simply due to numerical constraints, namely, conditioning, which we have eliminated. This theoretical derivation was then validated by the application of both LUT-based and polynomial-based models to the same practical example, and then observing that they, indeed, provide similar capabilities.

In summary, arguments in favor, or against, look-up table descriptions with respect to polynomials should neither be based on the approximation capabilities of each of these formulations, nor on the capacity to accurately extract each of their coefficients. They could, at most, be based on the specific details of their hardware implementation.

ACKNOWLEDGMENT

The authors would like to thank the Huawei Sweden PA Design Team, namely Dr. Francesc Purroy and Dr. Richard Hellberg, for interesting technical discussions regarding some of the observations dealt with in this work.

REFERENCES

[1] J. Kim and K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *Electronics Lett.*, vol.37, no.23, pp.1417-1418, Nov. 2001.

[2] O. Hammi, F. M. Ghannouchi, and B. Vassilakis, "A Compact Envelope-Memory Polynomial for RF Transmitters Modeling With Application to Baseband and RF-Digital Predistortion," *IEEE Microwave and Wireless Components Lett.*, vol.18, no.5, pp.359-361, May 2008.

[3] D. R. Morgan, Z. Ma, J. Kim, M. G. Zierdt, and J. Pastalan, "A Generalized Memory Polynomial Model for Digital Predistortion of RF Power Amplifiers," *IEEE Trans. on Signal Processing*, vol.54, no.10, pp.3852-3860, Oct. 2006.

[4] A. Zhu, J. C. Pedro, and T. J. Brazil, "Dynamic Deviation Reduction-Based Volterra Behavioral Modeling of RF Power Amplifiers," *IEEE Trans. on Microwave Theory and Techn.*, vol.54, no.12, pp.4323-4332, Dec. 2006.

[5] L. Guan and A. Zhu, "Simplified dynamic deviation reduction-based Volterra model for Doherty power amplifiers," *2011 Workshop on Integrated Nonlinear Microwave and Millimetre-Wave Circuits (INMMIC)*, pp.1-4, Apr. 2011.

[6] T. R. Cunha, E. G. Lima, and J. C. Pedro, "Validation and Physical Interpretation of the Power-Amplifier Polar Volterra Model," *IEEE Trans. on Microwave Theory and Techn.*, vol.58, no.12, pp.4012-4021, Dec. 2010.

[7] L. Guan and A. Zhu, "Optimized Low-Complexity Implementation of Least Squares Based Model Extraction for Digital Predistortion of RF Power Amplifiers," *IEEE Trans. on Microwave Theory and Techn.*, vol.60, no.3, pp.594-603, Mar. 2012.

[8] R. Raich, H. Qian, and G. T. Zhou, "Orthogonal polynomials for power amplifier modeling and predistorter design," *IEEE Trans. on Vehicular Technology*, vol.53, no.5, pp.1468-1479, Sep. 2004

[9] L. Aladren, P. Garcia, P. L. Carro, J. De Mingo, and C. Sanchez-Perez, "Digital predistortion based on Zernike polynomial functions for RF nonlinear power amplifiers," *2012 International Symposium on Wireless Communication Systems (ISWCS)*, pp.865-869, Aug. 2012.

[10] N. Naraharisetti, C. Quindroit, P. Roblin, S. Gheitanchi, V. Mauer, and M. Fitton, "2D cubic spline implementation for concurrent dual-band system," *2013 IEEE MTT-S Intern. Microwave Symposium Dig.*, pp.1-4, Jun. 2013.

[11] N. Safari, P. Fedorenko, J. S. Kenney, and T. Roste, "Spline-Based Model for Digital Predistortion of Wide-Band Signals for High Power Amplifier Linearization," *2007 IEEE MTT-S Intern. Microwave Symposium Dig.*, pp.1441-1444, Jun. 2007.

[12] Y. Ma, Y. Yamao, Y. Akaiwa, and C. Yu, "FPGA Implementation of Adaptive Digital Predistorter With Fast Convergence Rate and Low Complexity for Multi-Channel Transmitters," *IEEE Trans. on Microwave Theory and Techn.*, vol.61, no.11, pp.3961-3973, Nov. 2013.

[13] F. M. Barradas, T. R. Cunha, P. M. Lavrador, and J. C. Pedro, "Using spline basis functions in Volterra series based models," *2014 International Workshop on Integrated Nonlinear Microwave and Millimetre-wave Circuits (INMMiC)*, pp.1-3, Apr. 2014.

[14] F. M. Barradas, T. R. Cunha, P. M. Lavrador, and J. C. Pedro, "Higher locality non-linear basis functions of Volterra series based models to improve extraction conditioning," *2014 IEEE MTT-S International Microwave Symposium (IMS)*, pp.1-4, Jun. 2014.

[15] O. Hammi, S. Boumaiza, and F. M. Ghannouchi, "On the Robustness of Digital Predistortion Function Synthesis and Average Power Tracking for Highly Nonlinear Power Amplifiers," *IEEE Trans. on Microwave Theory and Techn.*, vol.55, no.6, pp.1382-1389, Jun. 2007.

[16] E. Cheney and D. Kincaid, "Additional Topics Concerning Systems of Linear Equations," in *Numerical Mathematics and Computing*, 6th ed., Stamford, CT, USA: Cengage Learning.

[17] M. Isaksson and D. Ronnow, "A Kautz-Volterra Behavioral Model for RF Power Amplifiers," *2006 IEEE MTT-S Intern. Microwave Symposium Dig.*, pp.485-488, Jun. 2006.

[18] A. Zhu and T. J. Brazil, "RF power amplifier behavioral modeling using Volterra expansion with Laguerre functions," *2005 IEEE MTT-S Intern. Microwave Symposium Dig.*, pp.4-7,, Jun. 2005.

[19] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-posed Problems*, Washington, Winston & Sons, 1977

[20] J. H. Mathews and K. D. Fink, "Curve Fitting," in *Numerical Methods Using MatLab*, 3rd ed., USA: Prentice Hall.

[21] C. de Boor, "Piecewise Cubic Interpolation," in *A Practical Guide to Splines*, Revised ed., New York, USA: Springer-Verlag.

[22] C. de Boor, "Limitations of Polynomial Approximation," in *A Practical Guide to Splines*, Revised ed., New York, USA: Springer-Verlag.

[23] C. S. Eun and E. Powers, "A new Volterra predistorter based on the indirect learning architecture", *IEEE Trans. on Signal Processing*, vol.

45, no. 1, pp. 223-227, 1997.

Filipe M. Barradas (S'13) was born in Évora, Portugal, in July 1989. He received the M.Sc degree in electronics and telecommunications engineering from the Universidade de Aveiro, Aveiro, Portugal, in 2012, and is currently working toward the Ph.D. degree in electrical engineering at the the Universidade de Aveiro.

He was one of the recipients of the first prize in the power amplifier linearization through digital predistortion student design competition in 2014, which took place during the International Microwave Symposium.

His main interests include digital predistortion and behavioral modelling of RF PA, as well as signal processing with applications on telecommunications.

Telmo R. Cunha (M'05) was born in Porto, Portugal, in 1973. He received the Diploma and Ph.D. degrees in electronics and computer engineering from the Universidade do Porto, Porto, Portugal, in 1996 and 2003, respectively. Before 2004, he was involved with the Astronomical Observatory, University of Porto, and, afterward, he was a Technical Director and Research Engineer with Geonav Lda., a private company near Porto. Since 2004, he has been an Assistant Professor with the Department of Electronics, Telecommunications and Informatics, University of Aveiro, Aveiro, Portugal, and also a Research Engineer with the Institute of Telecommunications, University of Aveiro. He has been lecturing in the areas of control theory and electronics, and he has been involved in several national and international research projects. His current research interests include behavioral modeling and linearization applied to radio frequency and microwave devices and also integrated-circuit signal integrity analysis.

Dr. Cunha has been a reviewer for several IEEE journals.

Pedro M. Lavrador received his Diploma and PhD degrees in Electronics and Telecommunications Engineering from the Universidade de Aveiro, Aveiro, Portugal in 2001 and 2007 respectively.

He is with the Telecommunications Institute (Aveiro) since October 2000 and as Assistant Professor with the Department of Electronics, Telecommunications and Informatics, University of Aveiro. His main research interests are studying the impact of nonlinear effects in communication systems, and efficient simulation methods and models for the same kind of systems. He is now an invited assistant professor at Universidade de Aveiro.

Mr. Lavrador had received three times, during his graduation studies, the prize for the students with the best grades in engineering courses at the University of Aveiro.

José C. Pedro (S'90–M'95–SM'99–F'07) was born in Espinho, Portugal, in 1962. He received the diploma, doctoral and habilitation degrees in electronics and telecommunications engineering, from the Universidade de Aveiro, Portugal, in 1985, 1993 and 2002, respectively.

From 1985 to 1993 he was an Assistant Lecturer at Universidade de Aveiro, and a Professor since 1993. Currently he is a Full Professor at the same university, and a Senior Research Scientist at the Institute of Telecommunications.

His main scientific interests include active device modeling and the analysis and design of various nonlinear microwave circuits. He is the leading author of Intermodulation Distortion in Microwave and Wireless Circuits (Artech House, 2003), has authored or co-authored more than 200 papers in international journals and symposia, and served the IEEE in the Portuguese MTT/AP/ED Joint Chapter, the MTT-11 Technical Committee and as a reviewer and Associate Editor for the MTT Transactions and reviewer for the MTT-IMS and the EuMC.

Prof. Pedro has served his university department as the Coordinator of the Scientific Council and as the Department Head.

Prof. Pedro received the Marconi Young Scientist Award in 1993 and the 2000 Institution of Electrical Engineers (IEE) Measurement Prize. In 2007 he was elected Fellow of the IEEE for his contributions to the nonlinear distortion analysis of microwave devices and circuits. Currently, he is an IEEE MTT-S Distinguished Microwave Lecturer.

4. Characterization and Modeling of Radio Frequency Power Amplifiers

After first addressing the relationships of the Volterra based models with the physics of the device, and then investigating the signal processing aspects of the parameter extraction; a significant amount of know-how has been generated in both these areas that provide resources to characterization, modeling and DPD of RFPAs.

This chapter is dedicated to the characterization techniques developed throughout this work. Obtaining the static characterization of a device at a particular frequency is commonly achieved through the AM/AM and AM/PM measurement at that frequency. However, methods to characterize and represent the memory behavior of the RF PAs are still under research. Since two main types of memory were identified in the second chapter of this thesis, when the physics of the device were analyzed, the developed methods also attempt to capture these two different memory types.

The methods presented here tradeoff approximation in a broadband, for accuracy of the model. This modelling approach is intended for understanding the behavior of the PA and providing a crude approximation to its output for several operating conditions for system level simulation.

This chapter is divided into two main sections, the first section being dedicated to the direct memory and the second to the cross memory.

4.1. Direct Memory Characterization of Radio Frequency Power Amplifiers

In this work, the direct memory of a PA is the memory responsible for changing the response of the device when it is excited with CW signals at different frequencies. As explained in Chapter 2 of this thesis, this memory is evidenced by a change of AM/AM and AM/PM as the operating frequency of the device is varied.

Also in Chapter 2, it was noted that the MP model could represent this behavior. In fact, for a CW signal, it was shown that the MP model is effectively a frequency dependent AM/AM and AM/PM, (4.1).

$$\begin{aligned}&\sum_{m=0}^{M}\tilde{h}_{2p-1}(m)\tilde{x}(n-m)|\tilde{x}(n-m)|^{2p-2}\\&= a|a|^{2p-2}\sum_{m=0}^{M}\tilde{h}_{2p-1}(m)e^{j(n-m)\omega_d T_s},\quad x(n)=ae^{jn\omega_d T_s}\\&= \widetilde{H}_{2p-1}(\omega_d)a|a|^{2p-2}e^{jn\omega_d T_s}\\&\tilde{y}(n)=\sum_{p=1}^{P}\widetilde{H}_{2p-1}(\omega_d)a|a|^{2p-2}e^{jn\omega_d T_s}\end{aligned} \tag{4.1}$$

Taking (4.1) into account, the MP model is ideal to represent this direct memory behavior. It is interesting to note that the MP filters represent the variation of the parameter described by the corresponding nonlinear function through frequency. For instance, when in the original representation, the filter $\widetilde{H}_1(\omega_d)$ represents the small signal gain through frequency. This gain is then changed by the following filters when the amplitude is sufficiently high.

One possible way to extract this model is to characterize the RF PA in terms of AM/AM and AM/PM at different frequencies and compile this information into the MP model. The main problem of this approach is the phase coherence between different measurements to maintain a consistent AM/PM measurement at different frequencies.

As shown in Chapter 3, the MP model can be represented more generally than its polynomial form, (4.2).

$$\tilde{y}(n)=\sum_{p=1}^{P}\widetilde{H'}_p(\omega_d)f_p(a)ae^{jn\omega_d T_s} \tag{4.2}$$

From (4.2) and the previous interpretation, by changing the functions, $f_p(a)$, used to describe the model, the meaning of the filters change. For instance, using the Lagrange polynomials described in Chapter 3, the output depends exclusively on one of the MP branches for the input powers corresponding to the nodes where the polynomials are defined, (4.3).

$$f_p(a) = L_p(a) = \prod_{\substack{k=0 \\ k \neq p}}^{K} \frac{a - a_k}{a_p - a_k}, i \in [0, K]$$

$$\tilde{y}(n)|_{a=a_p} = \sum_{p=1}^{P} \tilde{H'}_p(\omega_d) L_p(a) a e^{jn\omega_d T_s} = \tilde{H'}_p(\omega_d) a_p e^{jn\omega_d T_s} \tag{4.3}$$

$$\Rightarrow \tilde{H'}_p(\omega_d) = \frac{\tilde{y}(n)|_{a=a_p}}{a_p e^{jn\omega_d T_s}}$$

From (4.3), when defining the MP with the Lagrange polynomials the model can be compiled from transmission measurements at different input powers. In particular, if the grid for the Lagrange polynomials is defined for the measurement powers each branch of the MP model is orthogonal to the other, and the measurements are directly translated to the model. This particular interpretation of the MP model is particularly useful because it allows extraction using VNA measurements where the phase relationships are more easily maintained than in VSG + VSA setups.

From these relationships it is clear that the MP model is very versatile in describing the direct memory of an RF PA and can easily be used to condense measurement information into a useful model. One particular measurement information that is typically used to quantify the linearity of the device, and that can also be used to generate an MP model, is the IMD for two-tone measurements at a particular input power. In this case, the input excitation can be defined as seen in (4.4), where ω_s is the frequency spacing between the two-tones, and ω_d is the distance from the center operating frequency, as before.

$$\tilde{x}(n) = a \cos(\omega_s n T_s)\, e^{j\omega_d n T_s} \tag{4.4}$$

One advantage of the excitation defined in (4.4) is that it will sweep the input power up to an amplitude defined by a. With this in mind, the MP functions should no longer represent different power levels. Instead, note that the output of an RF PA excited with this input signal can be described as shown in (4.5).

$$\tilde{y}(n) = e^{j\omega_d nT_s} \sum_{k=1}^{K} a_{2k-1} e^{j\theta_{2k-1}} \cos\big((2k-1)\omega_s nT_s + \phi_{2k-1}\big) \tag{4.5}$$

In (4.5), the parameters a_{2k-1}, θ_{2k-1}, and ϕ_{2k-1} depend on the excitation, ω_d, and spacing, ω_s, frequencies. Taking (4.5) into account, to obtain an orthogonal expansion for the two tone case, functions with the properties shown in (4.6) are required.

$$f_{2p-1}\big(a\cos(\omega_s nT_s)\, e^{j\omega_d nT_s}\big) a\cos(\omega_s nT_s)\, e^{j\omega_d nT_s} = a\cos\big((2k-1)\omega_s nT_s\big)\, e^{j\omega_d nT_s} \tag{4.6}$$

Fortunately, the required functions are widely known in the signal processing community. The functions with the properties shown in (4.6) are the Chebyshev polynomials, which are orthogonal for cosine excitations. For this application, due to the LPE restrictions, only the odd order polynomials are used.

As mentioned, when using this type of excitation, the output is sensitive to both the frequency spacing of the two-tones and their center frequency. However, the MP model only supports variation according to one of these two parameters. In fact, varying one or the other will generate sufficient diversity to fully utilize the MP resources. It then becomes apparent that a choice must be made when using two-tone measurements to characterize and model the RFPA into an MP model.

When looking into the signal processing literature, an interesting method, used in audio systems for nonlinear characterization, was found [90, 91]. This method uses one single exponential chirp measurement to obtain consistent information on all the IMDs as the frequency varies.

Using the exponential chirp technique and the previous model transformation, a method for the characterization and modelling of RF PAs with one single measurement was created [92], ([C4] included in annex to this chapter). The advantage of this technique is precisely that one measurement is able to fully contain all the required data for the extraction. After the extraction is concluded, the MP model can be used to obtain initial estimates of the RF PA behavior and also contains easily interpreted useful information. In particular, the filters of the MP model extracted in this way show the equivalent linear gain of the PA for two-tones in the measured frequency spacing, as well as the IMD variation over the same interval.

The problem in this approach is that the spacing of the tones is swept to obtain the data. This means that the model will describe the system for increasing signal

bandwidths. Therefore, this extraction procedure is breaking the initial requirement specified for the direct memory. The increase in signal bandwidth will produce excitations in the baseband that may impact the response of the RF PA, which is related to the cross memory behavior of the device. In this sense, the parameter that should be varied for obtaining a direct memory representation is the center frequency, while maintaining an infinitesimal separation between the two-tones. This realization also leads to the conclusion that the two-tone excitation may be rich enough to characterize the cross memory of the device.

4.2. Cross Memory Characterization of Radio Frequency Power Amplifiers

After realizing that the two-tone excitation has sufficient generality to characterize both the direct memory and, at least to some extent, the cross memory. The initial characterization using the two-tone measurements was extended to further include this memory behavior.

The cross memory behavior characterization is more interesting than the direct memory characterization because it is unique of nonlinear systems. The general problem in characterizing this memory behavior is separating its effects from the direct memory. Using the two-tone signal gives a simple possibility to achieve this separation. First, the direct memory is extracted by changing the center frequency of the excitation and using a very small spacing frequency. After this direct memory is characterized, the system is excited with two-tone signals maintaining the center frequency and increasing the tone spacing. The difference between the two measurements can be related to the cross memory and compiled under the GMP representation.

The GMP model is more general than these two different (center frequency and spacing) two-tone sweeps. In fact, the GMP model can be represented as shown in (4.7).

$$\begin{aligned}\tilde{y}(n) &= \sum_{p=1}^{P}\sum_{m_1=0}^{M_1}\sum_{m_2=-M_2}^{M_2}\tilde{h}_{2p-1}(m_1,m_2)\tilde{x}(n-m_1)|\tilde{x}(n-m_1-m_2)|^{2p-2} \\ &= \sum_{p=1}^{P}\sum_{m_1=0}^{M_1}\tilde{x}(n-m_1)\sum_{m_2=-M_2}^{M_2}\tilde{h}_{2p-1}(m_1,m_2)f_p(|\tilde{x}(n-m_1-m_2)|)\end{aligned} \tag{4.7}$$

From (4.7), the memory corresponding to m_2 is insensitive to the center frequency of the two-tones, but sensitive to the bandwidth or frequency spacing of the tones. In fact, selecting the $f_p(.)$ functions to be the Chebyshev polynomials of even order (in contrast to the previously chosen odd order) and selecting a two-tone excitation yields (4.8).

$$\sum_{m_2=-M_2}^{M_2} \tilde{h}_{2p-1}(m_1,m_2) f_p(|\tilde{x}(n-m_1-m_2)|), \qquad \tilde{x}(n) = a\cos(\omega_s n T_s) e^{j\omega_d n T_s}$$

$$= a \sum_{m_2=-M_2}^{M_2} \tilde{h}_{2p-1}(m_1,m_2) \cos(2(p-1)\omega_s(n-m_1-m_2)T_s) \tag{4.8}$$

$$= a\left(\tilde{h'}_{2p-1}(m_1,\omega_s) + \tilde{h'}_{2p-1}(m_1,-\omega_s)\right)\cos(2(p-1)\omega_s(n-m_1)T_s)$$

$$+ja\left(\tilde{h'}_{2p-1}(m_1,\omega_s) - \tilde{h'}_{2p-1}(m_1,-\omega_s)\right)\sin(2(p-1)\omega_s(n-m_1)T_s)$$

Equation (4.8) shows that the response of the m_2 memory is indeed only dependent on the spacing, however, the $h(m_1,m_2)$ kernels are fully bi-dimensional, which means that for a complete extraction the spacing needs to be excited for each state of the m_1 memory. This memory component is sensitive to both the spacing and the center frequency, which means that a complete extraction would require spacing sweeps for each center frequency. Furthermore, it also shows that the measurements performed to extract this model are not orthogonal to the GMP.

The lack of orthogonality in this description regarding the related measurements has a negative impact on both the model extraction and its usefulness. First, because each component of the model is not uniquely related to a model component, it is difficult to identify the relationships between what is measured and the model parameters. Second, the model itself becomes more difficult to extract from the measurements.

To limit the number of measurements to only the two initial sweeps, low bandwidth with varying center frequency and varying spacing with fixed center frequency, one possible assumption is (4.9).

$$\tilde{h}(m_1,m_2) \cong \tilde{h}_1(m_1)\tilde{h}_2(m_2) \tag{4.9}$$

Equation (4.9) severely restricts the variation of the kernels in the multidimensional memory space and will therefore reduce the number of required measurements. This assumption solves the number of measurements problem, however, the orthogonality is not achieved. In fact, transforming the model into an orthogonal representation for both the cross and direct memories and the two-tone excitation proved too difficult.

Using the even Chebyshev polynomials yields a quasi-orthogonal representation

of the problem and, even though the utility is reduced, the method was explored as a means to produce a cross memory capable model in a broadband around a particular center frequency in [93] ([C5] included in annex to this chapter).

4.3. Summary and Final Considerations

This chapter was dedicated to the characterization and modelling of RF PAs. The main intentions in this chapter was exploring the possibilities of using typically measured RFPA data to generate system level models for simulation. On the one hand, it is expected that the models represent the RFPA with a reasonable degree of accuracy. On the other hand, the measurements themselves represent useful information for the system designer.

In this sense, the first approach to the problem was made using the MP model. The MP model has a simple structure that allows high malleability in both the memory and nonlinear components. Using transformations of this model, it is possible to generate representations that can be extracted from typical relevant measurements of RFPAs, S-parameters at several powers or IMD measurements. Furthermore, a method was developed to extract this model in a single shot measurement.

After investigating the direct memory characterization, the representation of the cross memory was undertaken. The cross memory is sensitive to the signal bandwidth only, unlike the direct memory which was excited by varying the center frequency of excitation. The cross memory was more difficult to properly represent in a way that is both usable from the simulation point of view and also represents relevant information. Nonetheless, a quasi-orthogonal representation was found using the GMP model. Furthermore, a method to verify if the device has cross-memory was proposed using two two-tone sweeps, one varying the center frequency and another the frequency spacing.

Paper C4: RF PA modeling with one chirp measurement

Filipe M. Barradas, Telmo R. Cunha, Pedro M. Lavrador, and José C. Pedro

2015 European Microwave Conference (EuMC)

©2015 IEEE

RF PA Modeling With One Chirp Measurement

Filipe M. E. Barradas, Pedro M. Lavrador, Telmo R. Cunha, José C. Pedro
DETI, Instituto de Telecomunicações
Universidade de Aveiro
Aveiro, Portugal
filipebarradas@ua.pt

a. ***Abstract*—In this paper we present a technique for obtaining a behavioral model for a power amplifier using a single exponential chirp measurement. The obtained model is equal to the memory polynomial. However, the process relies on a particular excitation that enables orthogonal extraction, making the identification quick and reliable. The model presents good approximation results for different excitations, and can be used for close inspection of the harmonic frequency responses of a power amplifier.**

Keywords—Behavioral modelling, power amplifiers, chirp, parallel Hammerstein.

I. INTRODUCTION

Power amplifiers (PA) are complex systems which are overall difficult to physically model. Oftentimes a top level behavioral modelling approach can be used to gain some insight into the PA's response to certain stimulus.

In some cases, only the operation characteristic around a certain carrier frequency is required. To model this carrier centered frequency band, the low pass equivalent (LPE) domain can be used.

In this paper we present a method which allows broadband characterization of a PA using a single, and fast, measurement. The method characterizes the frequency response of the PA around its fundamental operating frequency. To do this, we adapted the method from [1]-[2] to work in the low pass equivalent domain (LPE). In addition, we made use of particular nonlinear functions which orthogonalize the extraction process, making it robust and fast. Furthermore, we have slightly modified the process to allow chirps of any time length, contrary to what was presented in [2].

This method allows for the extraction of a parallel Hammerstein system, similar to the one in Fig. 1 which is equal to the memory polynomial [4]. It uses an exponential chirp excitation and measures all the harmonic responses of the nonlinear system. In our case, the chirp is designed in the LPE domain and modulated to the fundamental carrier frequency (we obtain two tones, progressively more separated in frequency). The measured harmonic responses are the ones around the fundamental frequency as well, corresponding to the intermodulation distortion generated by the system.

Thus, the method here purposed generates a model capable of reproducing the intermodulation distortion variation through frequency, for the chosen amplitude.

Filipe M. Barradas, the first author, would like to acknowledge the financial support provided by Portuguese Science and Technology Foundation, FCT (Ref. SFRH/BD/90103/2012).

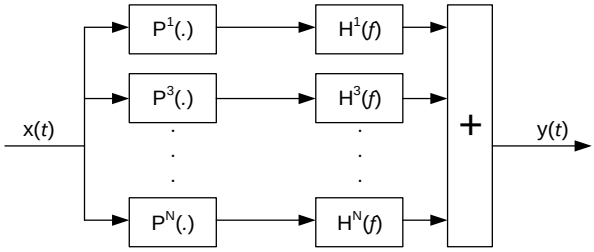


Fig. 1. Parallel Hammerstein model topology, consisting of several branches of a nonlinear static block followed by a linear dynamic system.

II. METHOD OVERVIEW

This method makes use of the exponential chirp properties to allow the direct extraction of a Parallel Hammerstein model without making use of any regression algorithm. In fact, it only requires the frequency response, (1), obtained with the exponential chirp excitation.

$$H(f)=\frac{Y(f)}{X(f)} \tag{1}$$

The fact that a nonlinear system can be completely identified by inspection of a supposedly linear response is remarkable. This section provides the qualitative knowledge to understand the process, which is detailed in the following sections.

The exponential chirp starts at a particular frequency which is increased as time passes. This means that, if we pick one frequency, then, after some time, the chirp will be at some multiple of that frequency. Because of this property, the nonlinear harmonic responses are interpreted in the linear extraction of *H(f)* as non-causal linear responses, as if the system were responding to a part of the signal which is only applied to it further ahead in time.

Since it is an exponential chirp, the time the signal takes to go from one frequency to another is the same it takes to go from a multiple of this first frequency to the same multiple of the second. Because of this property, the nonlinear responses, identified as non-causal linear responses, do not disperse through time.

These two characteristics make the exponential chirp one of the very particular, if not the only, class of signals that can be used in this process.

The model defined in this way has the structure shown in Fig. 1, where the functions $P^n(.)$ are responsible for generating the harmonic content. Polynomials can be used for this purpose, as shown in prior art [2].

However, by defining the polynomials as shown here we orthogonalize the extraction for the chirp excitation and make the identification of each branch immediate, which greatly simplifies the method. Furthermore, since we are working in the LPE domain, only odd orders are used, as shown in (2).

$$P^n(x)=\sum_{p=1}^{n} c_p |x|^{p-1} x, \qquad n=2k+1 \qquad k\in \mathrm{N}_0 \tag{2}$$

The polynomials we must find are those that generate the *nth* harmonic from a sinusoidal input. We can define this condition as seen in (3), which gives the generation rule presented in (4), and proved in (5), using odd orders only.

$$P^n(\sin(\theta))=\sin(n\theta) \tag{3}$$

$$\begin{aligned} P^n(x) &= -2\left(2|x|^2 P^{n-2}(x)-P^{n-2}(x)\right)-P^{n-4}(x) \\ P^1(x) &= x \\ P^3(x) &= -4|x|^2 x+3x \end{aligned} \tag{4}$$

$$\begin{aligned} P^n(\sin(\theta)) &= 2\left(1-2\sin(\theta)^2\right)\sin((n-2)\theta)-\sin((n-4)\theta)= \\ &= -2\cos(2\theta)\sin((n-2)\theta)-\sin((n-4)\theta)= \\ &= \sin(n\theta) \end{aligned} \tag{5}$$

These $P^n(.)$ functions can be adjusted for any sine amplitude by dividing x by the amplitude of the sine.

Since the sinusoids at each frequency are orthogonal to each other, the overall regression is orthogonalized by using the polynomials defined in this way.

Through this simple process of finding these orthogonal nonlinear functions, we have simplified the problem and avoid using any regression algorithm.

III. Defining the Input Signal

The correct definition of the input signal is very important in this method, since, as explained before, the extraction procedure relies on its properties.

The common exponential chirp equation, (6), depends on three parameters, *A*, *K* and *L*, which are defined by setting the start and stop frequency of the chirp, its duration and amplitude. These parameters define the region of validity of the extracted model in terms of bandwidth, amplitude response and accuracy.

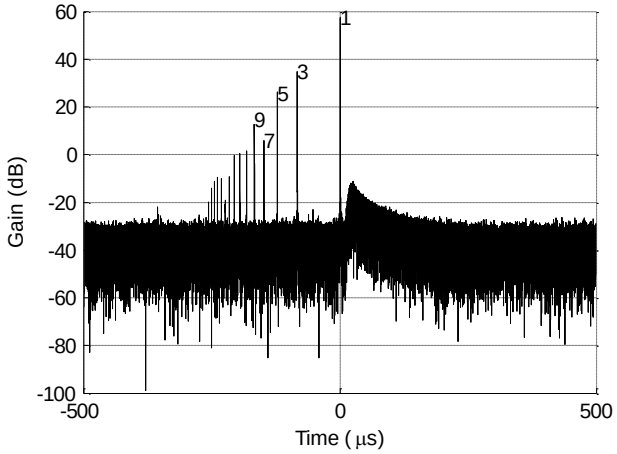


Fig. 2. Linear impulse response. Observe the causal response and the set of anti-causal responses corresponding to the several harmonics.

$$\begin{aligned} x(t) &= A\sin(\phi(t)) \\ \phi(t) &= 2\pi K\left[\exp\left(\frac{t}{L}\right)-1\right] \end{aligned} \tag{6}$$

The amplitude of the chirp, *A*, sets the amplitude limit of the behavior model we are extracting. The other parameters, *K* and *L*, define the bandwidth of the extracted model. They can be determined by defining the instantaneous frequency at two different times (typically the beginning and end of the chirp), (7).

$$f(t)=\frac{1}{2\pi}\frac{d\phi(t)}{dt}=\frac{K}{L}\exp\left(\frac{t}{L}\right)$$

$$\frac{K}{L}=f(0) \qquad \frac{K}{L}\exp\left(\frac{T}{L}\right)=f(T) \tag{7}$$

$$L=\frac{T}{\ln(f(T))-\ln(f(0))} \qquad K=f(0)L$$

Typically, the chirp would have to excite up to the *Nth* harmonic of the maximum model frequency, where *N* is the highest considered harmonic, so that identification of the *Nth* filter through (1) is possible.

However, if we consider a limited bandwidth system, responses above a certain frequency do not show up at the output. This means we can generate a chirp with frequencies up to the system's maximum frequency only.

The chirp's time duration, *T*, is related to the time distance between the several non-linear responses. The *nth* nonlinear response is displaced to negative time by the time distance from one frequency to its *nth* harmonic in the chirp. We can define this time displacement as shown in (8).

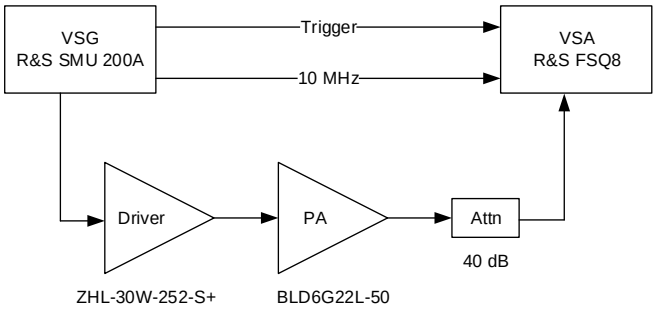


Fig. 3. Measurement setup

$$\frac{K}{L}\exp\left(\frac{t_n}{L}\right) = nf(0) \Rightarrow t_n = L\ln(n) \tag{8}$$

Ideally *T* would be set such that the time distance between the harmonic *N* and *N-1* allows for the separation of their two responses. Since we do not know the length of these impulse responses a priori, this is typically a hard parameter to set.

IV. Extraction

After the input exponential chirp is defined, the extraction procedure is straightforward. Since we have defined the orthogonal nonlinear functions, the filter impulse responses we measure are the model filters directly. However, there are some particularities which should be accounted for and have not been described in the prior literature if we want to fully control the time length of the chirp.

As previously mentioned, the first step is to obtain the linear impulse response. The typical procedure for this is to first obtain the linear frequency response, (1), and transform it to time domain.

A causal and a series of anti-causal responses can be observed in this linear impulse response, Fig. 2. The start time of each harmonic response can be calculated using (8). The harmonic responses are then separated. These impulse responses are the impulse responses of the filters $H^n(f)$.

Due to a mismatch between the phase of each harmonic in the exponential chirp and the phase of the harmonics generated by the functions $P^n(.)$, a phase correction is required in the filters. This phase correction can be calculated in the frequency domain as seen in (9).

In [2], the time length of the chirp is controlled so that this phase is always zero. However, this can lead to very large chirp lengths, dependent on the lowest chirp frequencies.

$$\phi_c(f) = \exp\left(j\,\mathrm{sign}(f)\phi(t_n)\right) \tag{9}$$

Typically the identification of this model is performed on the digital domain. In this case, each of the impulse responses cannot be sampled exactly at each of their starting points tn, as there is no sample at that exact time. This creates a phase mismatch between the several harmonic orders which should also be corrected, through (10), where tsn is the sample corresponding to the sampled time immediately before tn.

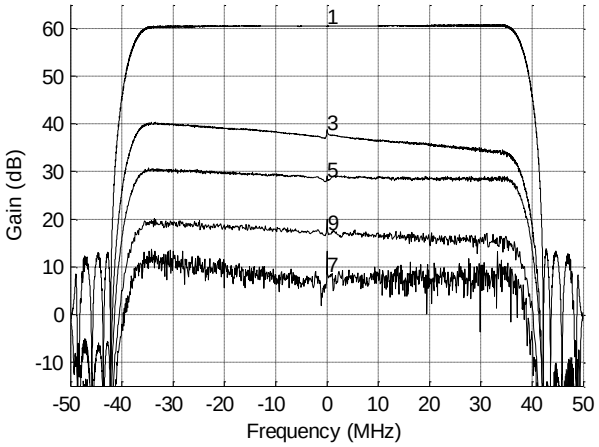


Fig. 4. Measured frequency responses

$$\phi_t(f) = \exp\left(jf\left(t_n - t_{sn}\right)\right) \tag{10}$$

After applying these corrections, all necessary data to build the model of Fig. 1 is available. Further signal processing can be used to smooth the data and eliminate some noise.

V. Results

This type of modelling can be applied for the characterization of a DPD platform. Our test setup is composed of a R&S SMU200A VSG which feeds a driver and PA cascade, the output is sampled using a R&S FSQ8 VSA. The tested PA was a BLD6G22L-50 from NXP, driven with a ZHL-30W-252-S+ from mini-circuits, as shown in Fig. 3.

In this setup there are two major memory generation mechanisms: the PA itself and the VSG and VSA which impose a lowpass characteristic to the system.

The measured frequency responses at each harmonic component are plotted in Fig. 4 (here up to the *9th* order, although the system was modelled up to the *23rd*). Note that, since we are working on the LPE domain, the frequency responses may not be complex conjugate about the origin.

The responses shown in Fig. 4 show a low pass characteristic, which is imposed by the generator and analyzer. The frequency characteristic from -35 MHz to 35 MHz is mainly imposed by the amplifier.

After the model is extracted it can be used to obtain the system response to a wide variety of signals. We tested the model for multi-carrier GSM and multi-carrier LTE signals. The spacing between each carrier is 3 MHz, and each LTE carrier has a bandwidth of 2 MHz. Using 4 LTE carriers the occupied bandwidth is about 11 MHz.

The results are summarized in Table I, in NMSE. Fig. 5

TABLE I
MODEL NSME

Carriers	Bandwidth	GSM	LTE
2	3 MHz	-46.0 dB	-31.8 dB
3	6 MHz	-38.5 dB	-30.4 dB
4	9 MHz	-35.0 dB	-30.4 dB

Model NMSE for different excitations.

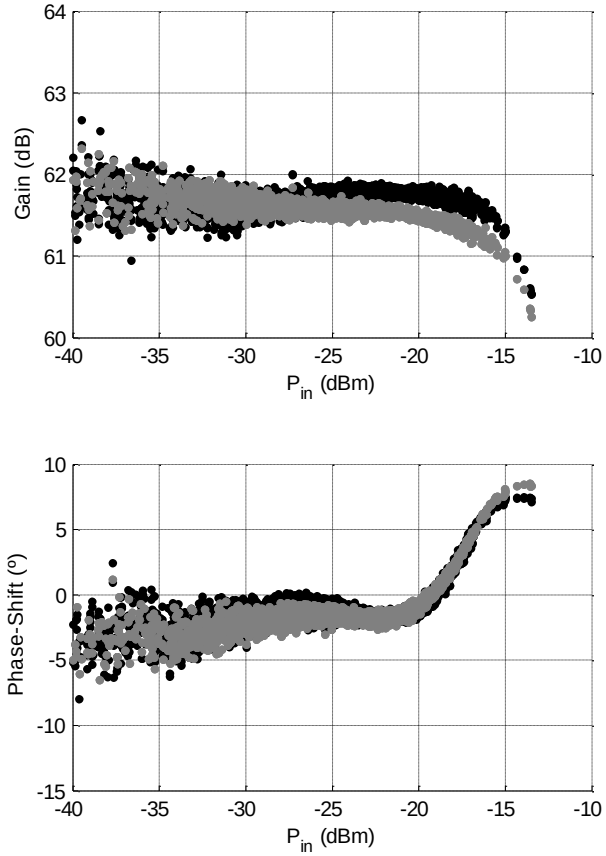


Fig. 5. AM/AM and AM/PM responses in the worst case. Measured is in black and modelled in grey.

shows the AM/AM and AM/PM obtained in the worst case (4 LTE carriers). Furthermore, Fig. 6 shows the error in the spectrum domain. The main source of error is possibly electro-thermal effects, not contemplated in the model.

VI. CONCLUSION

We have adapted the exponential chirp characterization procedure to work on the LPE domain and have provided a methodology for its swift extraction. This methodology relies on a set of nonlinear functions which orthogonalize the extraction for the chirp excitation.

The extracted model is straightforward to implement but can give accurate behavioral information on the overall system. Furthermore, the model is robust to input signal changes, maintaining a good approximation capability (NMSE below -30 dB) for several input signals with distinct characteristics.

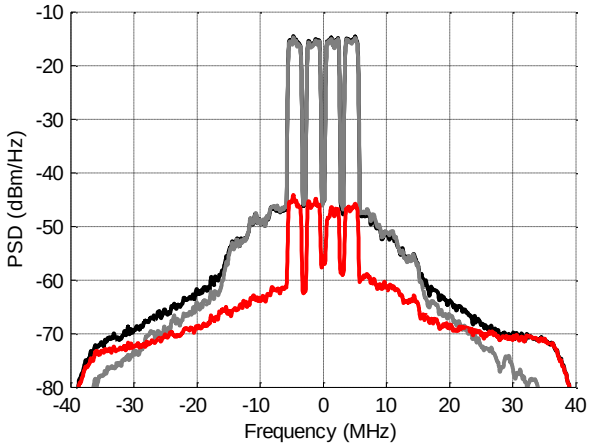


Fig. 6. Power spectral density of the modelled (in grey) and measured (in black) responses and the error (in blue) for the worst case.

ACKNOWLEDGMENT

The authors would like to thank the Huawei Sweden PA Design Team, namely Dr. Francesc Purroy and Dr. Richard Hellberg, for interesting technical discussions regarding some of the observations dealt with in this work.

REFERENCES

[1] A. Farina, “Simultaneous measurement of impulse response and distortion with a swept-sine technique,” in Proc. AES 108th Conv., Paris, France, Feb. 2000.

[2] A. Novák, L. Simon, F. Kadlec and P. Lotton, "Nonlinear System Identification Using Exponential Swept-Sine Signal," IEEE Trans. on Instr. and Meas., vol. 59, no. 8, pp. 2220-2229, Aug. 2010.

[3] J. Kim; K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," Electronics Letters , vol. 37, no. 23, pp. 1417-1418, Nov. 2001

[4] J. C. Pedro; S. A. Maas, "A comparative overview of microwave and wireless power-amplifier behavioral modeling approaches," IEEE Trans. on Microwave Theory and Tech., vol. 53, no. 4, pp. 1150-1163, April 2005

Paper C5: Characterizing Direct and Cross Memory in RF Nonlinear Systems Using Simple Two Tone Measurements

Filipe M. Barradas, Telmo R. Cunha, Pedro M. Lavrador, and José C. Pedro

2016 European Microwave Conference (EuMC)

©2016 IEEE

Characterizing Direct and Cross Memory in RF Nonlinear Systems Using Simple Two Tone Measurements

Filipe M. Barradas, Pedro M. Lavrador, Telmo R. Cunha and José C. Pedro
DETI, Instituto de Telecomunicações
Universidade de Aveiro
Aveiro, Portugal
filipebarradas@ua.pt

***Abstract*—Because RF nonlinear systems exhibit dynamic behaviors that are not only dependent on the central frequency (direct memory) but also on the envelope dynamics (cross memory), characterization metrics have been an object of research for many years. Unfortunately, as the response of these systems depends on the excitation, a general metric requires complex stimulus and cumbersome measurements. This work uses the fact that a standard two-tone excitation is already sufficiently complex to excite the most important nonlinearity and memory in nonlinear RF systems, as a step towards the development of such a metric. As an example, simple two-tone excitation measurements are used to evaluate both the direct and cross memory of an RF PA in a large bandwidth around a carrier frequency, and are then used to build a quasi-orthogonal GMP-like model.**

Keywords—Behavioral modeling, memory polynomials, power amplifiers, two-tone measurements

I. INTRODUCTION

In the past, two tone measurements have been extensively used to characterize the nonlinear behavior of RF systems using the n'th order IMD intersect point metric. Unfortunately, as such a metric is only capable of identifying a static mild nonlinearity, it was found inappropriate for hardly driven RF systems. Instead, modern characteristics include the static AM/AM and AM/PM at different carrier frequencies, which fail to capture the nonlinear cross memory of the devices. To expose cross memory, the system should be excited with complex modulated stimuli, which prevents the derivation of any simple standardized characterization procedure or performance metric. To achieve those goals, we realized that a model extracted from common two-tone tests of variable central frequency, ω_c, and tone separation, $2\omega_d$, is already sufficient to identify most of the main nonlinearity direct and cross memory effects [1, 2]. In fact, the recognition that the best PA behavioral models – as the generalized memory polynomial (GMP) model [3] or the 2nd order dynamic-deviation reduction (DDR-2) model [4] – share the same bi-dimensionality of the two-tone stimulus, constitutes sufficient evidence that such a signal is indeed able to expose the most important PA dynamics.

This work is funded by National Portuguese Funds through FCT - Fundação para a Ciência e a Tecnologia.

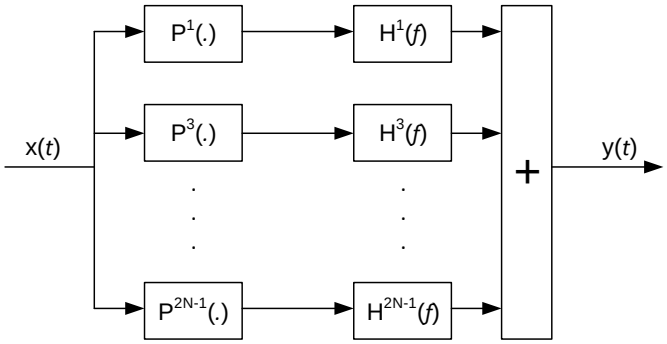


Fig 1. Parallel Hammerstein model topology, consisting of several branches of a nonlinear static block followed by a linear filter.

However, to be useful as metrics of nonlinearity and memory, the kernels of such a model should be capable of resolving (i.e., characterizing separately) each of the desired effects, i.e., they must be selected so that they can be extracted in an orthogonal way. The work that will be presented in the following sections was conceived as an important step to fulfill this goal.

Among the many potential applications of this method, one of the most obvious is the characterization of PA circuits whose cross-memory effects may impose linearizability restrictions, potentially even more severe than the ones arising from the direct memory effects. Therefore, to test the method, we picked up an RF PA and characterized it. As shown in the results, the method shows promising capabilities in terms of cross-memory measurement.

II. METHOD OVERVIEW

The proposed method, which is conceived in the low-pass complex envelope domain, relies on measurements made with two sets of equal amplitude two-tone stimuli described as

$$x[\omega_d, \omega_c](t) = \sin(\omega_d t)\exp(j\omega_c t) \quad (1)$$

where, in each set, ω_d or ω_c are swept.

In the first set of measurements, the system is excited with

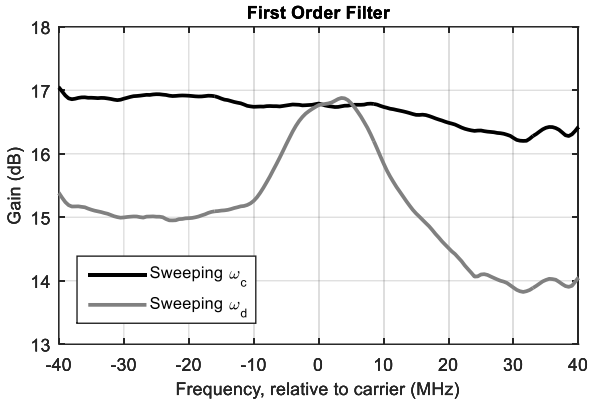


Fig. 2. Measured output filter of the parallel Hammerstein model when sweeping the carrier frequency and the bandwidth. The system under test is a PA MRFE6S9045N test board from Freescale Semiconductors.

a very narrow bandwidth two-tone signal (the tones are closely spaced) of sufficient amplitude and variable center frequency, ω_c, to sweep the whole input amplitude and frequency domains. These measurements provide information only on the direct memory, since the excitation has a very small and constant bandwidth. This is equivalent to static AM/AM and AM/PM measurements at various carrier frequencies.

In the second set of measurements the center frequency, ω_c, is kept fixed while the two-tone frequency separation, $2\omega_d$, is varied. This excites both the direct and cross memory effects of the device since we are using a signal with increasing bandwidth. Because in [5] it was shown that an exponential chirp signal can be used to quickly obtain information on the behavior of a PA for two-tones of increasing bandwidth, we will use this method to reduce the number of measurements.

After measuring the RF system with the two sets of stimuli, we build two parallel Hammerstein models (see Fig. 1), which contain all the measured information. Since the measurements excite the system differently, the two models will have significant differences. We explore these differences to gain insight into the cross-memory behavior and translate this behavior into a GMP-like model [3].

A. Building the Hammerstein Models

As shown in [5], when using the odd-degree Chebyshev polynomials, (2),

$$p^1(x) = x$$
$$p^3(x) = -4|x|^2 x + 3x \tag{2}$$
$$p^{2n-1}(x) = -2(2|x|^2 p^{2(n-1)-1}(x) - p^{2(n-1)-1}(x)) - p^{2(n-2)-1}(x)$$

as static kernels' nonlinearities for the parallel Hammerstein system, the model becomes orthogonal for two-tone excitations of a particular amplitude (now normalized to 1). When this is the case, each branch will generate a particular intermodulation product order. The filters are then responsible for setting the phase and amplitude of these products which are summed at the output.

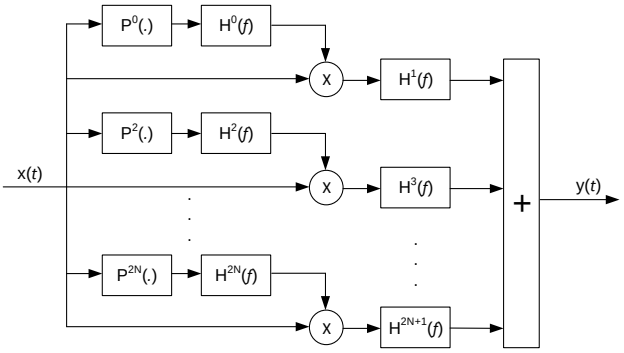


Fig. 3. GMP-like model structure consisting of branches with two filters. One is sensitive only to the tone separation while the second is sensitive to the center frequency and the frequency separation.

Taking (1) and (2) we see that in each branch of the parallel Hammerstein system we have the signal described in (3).

$$p^{2n-1}[x(t)] = \sin[(2n-1)\omega_d t]\exp(j\omega_c t) \tag{3}$$

For each two tone measurement the output filters are evaluated at $\omega_c+(2n–1)\omega_d$ and $\omega_c–(2n–1)\omega_d$ in each branch. This means we can measure the output filters of this model by sweeping either ω_c or ω_d. If the system would not have cross memory this would yield the same result. However, typically this is not the case as shown in Fig. 2. In fact, for bandwidths over a few MHz the models diverge. This divergence of the two models is a clear indication of the cross memory behavior.

B. Introducing Cross Memory

In order to introduce cross memory behavior into the model, we used a simplified version of the generalized memory polynomial, Fig. 3. This model has the transfer function in (4), which is close to the GMP model but lacks the complete two-dimensional description of the kernels.

$$y(t) = \sum_{n=1}^{N}\sum_{m_1=0}^{M_1}\sum_{m_2=0}^{M_2} h_{2n+1}(m_1)h_{2(n-1)}(m_2)\cdot x(n-m_1)\,p^{2(n-1)}[x(n-m_2)] \tag{4}$$

Now we can use the even-order Chebyshev polynomials as the static kernels' nonlinearities to obtain a quasi-orthogonal model for the two tone measurements.

$$p^0(x) = 1$$
$$p^2(x) = -2|x|^2 + 1 \tag{5}$$
$$p^{2(n-1)}(x) = -2(2|x|^2 p^{2(n-2)}(x) - p^{2(n-2)}(x)) - p^{2(n-3)}(x)$$
$$p^{2(n-1)}[x(t)] = \cos[2(n-1)\omega_d t]$$

When using the even-order Chebyshev polynomials, each branch produces two intermodulation products, except the first branch that produces only the fundamental. In fact, each branch of the quasi-GMP model produces the output shown in (6), when excited with a sine (and disregarding the filters).

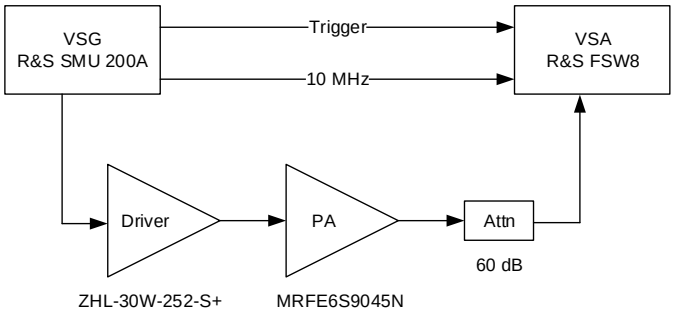


Fig. 4. Functional diagram of the two-tone measurement setup.

$$y_{2n-1} = \cos[2(n-1)\omega_d t]\sin[\omega_d t]\exp[j\omega_c t]$$
$$y_{2n-1} = \frac{1}{2}\left(\sin[(2n-1)\omega_d t] - \sin[(2n-3)\omega_d t]\right)\exp[j\omega_c t] \quad (6)$$

Comparing (6) with (3) we realize that the output of the quasi-GMP model can be expressed in terms of the output of the parallel Hammerstein model (when disregarding the filters), for each sine excitation, with the help of a conversion matrix as shown in, *W*, (7). There, y_{GMP} is the vector containing the output of each branch of the quasi-GMP system and y_{PH} is a similar vector, but now for the parallel Hammerstein system.

$$y_{GMP} = W y_{PH}$$
$$\text{intermodulation} \rightarrow \quad (7)$$
$$W = \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ -0.5 & 0.5 & & ... & 0 \\ 0 & -0.5 & 0.5 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 0.5 \end{bmatrix} \downarrow \text{branch}$$

In this quasi-GMP topology from Fig. 3, the even-order filters are only sensitive to the tone separation, while the odd-order filters are sensitive to the carrier frequency as well. For small bandwidths, the even-order filters are transparent, which means that, in this operating region, expression (7) should hold even when taking the output, odd-order, filters into account. This being the case, these odd-order filters can be directly extracted from the parallel Hammerstein filters, measured when sweeping the center frequency with the help of the conversion matrix from (7), as shown in (8).

$$H_{\mathrm{GMP}}{}^{T} y_{GMP} = H_{\mathrm{PH}}{}^{T} y_{PH} \Leftrightarrow$$
$$\Leftrightarrow H_{\mathrm{GMP}}{}^{T} W y_{PH} = H_{\mathrm{PH}}{}^{T} y_{PH} \quad (8)$$

$$H_{GMP}{}^{T} = H_{PH}{}^{T} W^{-1} \qquad H_x = \left[H_x^1 \quad ... \quad H_x^{2n-1}\right]^T$$

With the definition in (8), the models from Fig. 1 and 3 are equivalent for small bandwidths, when the even-order filters are not excited. When the bandwidth is increased, the even-order filters modify the overall response of the system. More precisely, for a two-tone signal with a given frequency spacing, the frequency response of each branch of the model in Fig. 3 comes multiplied by the response of the even-order filter for that spacing. Unfortunately, since our description is not fully two-dimensional we cannot independently accommodate the changes in the two intermodulation products at the output of each branch. In this work, to prove the concept, we chose to design the even-order filters to correct the higher order intermodulation of each branch.

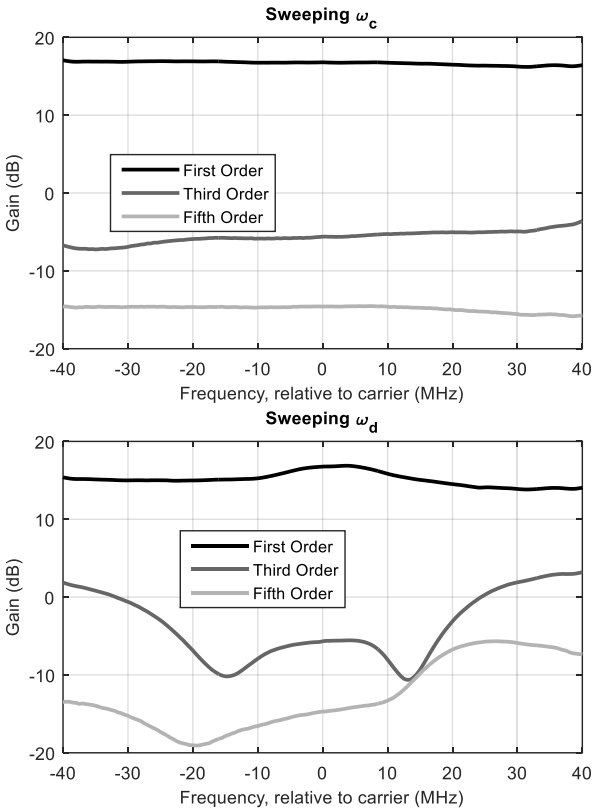


Fig. 5. Measured parallel Hammerstein filters obtained sweeping the center frequency (on top) and the frequency separation (on the bottom).

Taking the previous restriction into account, to calculate the even order filters we first obtain the parallel Hammerstein filters when sweeping ω_d and convert them using (8) to the quasi-GMP equivalent filters. We can then calculate the even-order filters according to (9).

$$H_{GMP}^{2(n-1)}\left(\frac{2(n-1)}{2n-1}\omega\right) = \frac{B_{GMP}^{2n-1}(\omega)}{H_{GMP}^{2n-1}(\omega)} \quad (9)$$

where B_{GMP} are the filters calculated when sweeping ω_d and H_{GMP} when sweeping ω_c. The frequency compression on the baseband filters happens because the harmonics that go through the even-order filters are then shifted upwards in the multiplier.

For systems with reduced cross memory effects, the calculated even-order filters will have small variations in frequency and remain close to unity throughout the frequency span. For systems with strong cross memory these even-order filters will show a large variation in frequency.

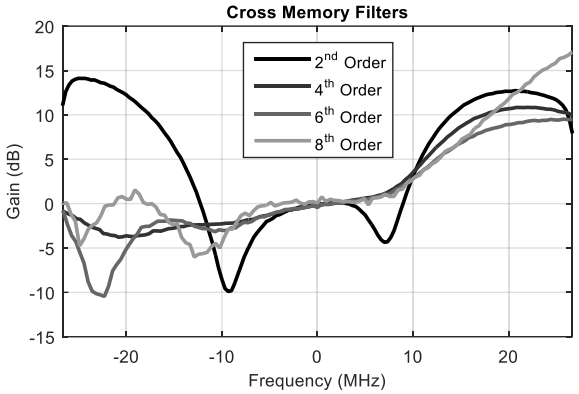


Fig. 6. Extracted cross memory filters for the GMP-like model at several orders. Near zero bandwidth the filters are unitary.

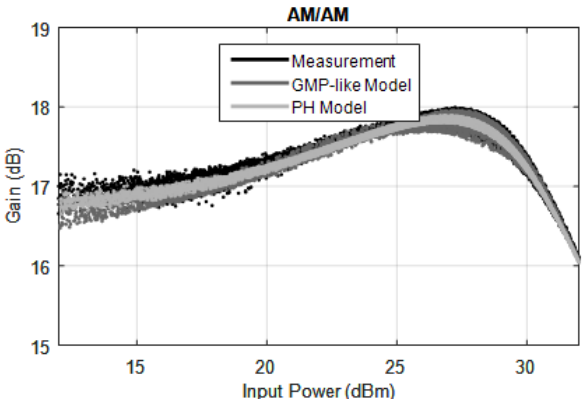


Fig. 7. AM/AM conversion for a 4cGSM signal and the direct and cross memory two-tone model predictions.

III. Results

In order to test the developed theory and show that the two-tone tests can, indeed, capture the most important nonlinearity and memory effects of RF systems, we measured an RF PA – the MRFE6S9045N test board from Freescale Semiconductors – using the setup shown in Fig. 4. The carrier frequency is kept constant, at 900 MHz, in the generator and analyzer and the measurements are performed by varying the digital baseband signal. The sampling rate is set to 100 MHz with 80 MHz of usable bandwidth. The automatic level control (ALC) at the generator is turned off to avoid changes in the gain of the system for each measurement.

A set of two-tone signals with varying center frequency is initially measured. The spacing of these tones is selected to be 200 kHz which is a narrow band for this PA. The center frequency is varied in 1 MHz steps. The extracted parallel Hammerstein filters for this set of measurements is shown in Fig. 5, on top.

Following the technique in [5], a chirp was measured to obtain the parallel Hammerstein system for increasing tone spacing. The extracted filters are shown in Fig. 5, on the bottom.

The tested amplifier shows a flat behavior when excited with a narrow band two-tone across the tested bandwidth. However, when excited with increasing bandwidths, the amplifier shows a significant frequency variation. We associated this frequency behavior to the cross memory and extracted the cross memory filters as shown before. The extracted filters are shown in Fig. 6. As expected, the cross memory filters can only be considered unitary in a considerably small bandwidth around zero.

To prove the extracted two-tone model capabilities, the PA was then excited with a four carrier GSM (4cGSM) signal with a 5 MHz bandwidth. The measured AM/AM conversion for this signal is shown in Fig. 7 along with the predictions from the parallel Hammerstein and GMP-like model. The initial parallel Hammerstein model is incapable of reproducing the increased memory region for higher powers, showing that this increase in memory is due to the cross memory of the PA. However, the proposed GMP-like model indeed captures the exposed cross memory, proving the two-tone characterization capability.

IV. Conclusions

In a way to find a metric for RF system nonlinearity and memory, we proposed a method to separately analyze its direct and cross memory using simple two tone measurements. We showed that it is possible to translate this information into a system level behavior model, which can then be used to track and diagnose possible problems with the RF systems' nonlinear frequency response in a large bandwidth around the carrier frequency.

Acknowledgment

The authors would like to thank the Huawei Sweden PA Design Team, namely Dr. Francesc Purroy and Dr. Richard Hellberg, for interesting technical discussions regarding some of the observations dealt with in this work.

This work is funded by National Portuguese Funds through FCT - Fundação para a Ciência e a Tecnologia under the project UID/EEA/50008/2013 and by the PhD grant given to the first author (Ref. SFRH/BD/90103/2012).

References

[1] J. Kenney and P. Fedorenko, "Identification of RF Power Amplifier Memory Effect Origins using Third-Order Intermodulation Distortion Amplitude and Phase Asymmetry," in IEEE MTT-S International Microwave Symposium Digest, San Francisco, 2006.

[2] J. Martins, P. Cabral, N. Carvalho and J. Pedro, "A Metric for the Quantification of Memory Effects in Power Amplifiers," IEEE Transactions on Microwave Theory and Techniques, vol. 54, no. 12, pp. 4432-4439, December 2006.

[3] D. Morgan, Z. Ma, J. Kim, M. Zierdt and J. Pastalan, "A Generalized Memory Polynomial Model for Digital Predistortion of RF Power Amplifiers," IEEE Transactions on Signal Processing, vol. 54, no. 10, pp. 3852-3860, October 2006.

[4] A. Zhu, J. Pedro and T. Brazil, "Dynamic Deviation Reduction-Based Volterra Behavioral Modeling of RF Power Amplifiers," IEEE Transactions on Microwave Theory and Techniques, vol. 54, no. 12, pp. 4323-4332, December 2006.

[5] F. M. Barradas, P. M. Lavrador, T. R. Cunha and J. C. Pedro, "RF PA modeling with one chirp measurement," in 10th European Microwave Integrated Circuits Conference (EuMIC), Paris, 2015.

5. Nonlinear Compensation of Radio Frequency Power Amplifiers

One of the main application of Volterra Models is in the nonlinear compensation of RFPAs. DPD supported by these models has been widely used and shown good results for a number of applications. However, two problems have recently arisen for DPD: the first problem is related to new, GaN based, transistor technology, while the second is due to the future communication standard 5G.

GaN transistors have shown resilience to typical DPD methods. In fact, GaN transistors have shown slow shifts of their behavior along time, which are faster than the coefficient update time, but much slower than what is described in the typical DPD FIR structure. These slow shifts, associated with electron trapping and detrapping phenomena in GaN, and also with thermal behavior, prevent accurate predistortion since the DPD model can only account for the “average” behavior of the PA. This is a problem related to the description capability of the used DPD models.

The second problem is tied to a change of architecture for future communication systems. In the 5G approach to telecommunications, one of the possibilities is the use of a high number of parallel transmitting branches, each with an analog transmission chain. This structure significantly reduces the power requirements for each PA and, consequently, the power used by the DPD becomes more significant. Furthermore, the complexity of typical DPD systems makes massive parallel deployment of such platforms a complex problem. 5G architectures thus require simpler DPD algorithms that can run in cheaper, smaller digital systems and on lower power budgets.

5.1. Compensation of Long-Term Memory Effects

Volterra series based compensation models have shown impressive results in linearizing a large number of devices. However, recent technology has shown some new effects that have been resilient to these linearization techniques. These effects, typically called long-term memory, are evidenced by slow shifts of the PAs characteristic along time, according to the input signal. Because this shift is so slow, when compared to the sampling rate or the signal bandwidth, it typically escapes the time span of the models. In fact, Volterra based models are traditionally supported by a FIR topology. This type of topology has problems approximating long-term phenomena because it would require a high number of coefficients, since the memory length in FIRs is limited to the number of used taps. Evidently, FIR topologies also have numerous advantages: immunity to instability, linearity in the parameters, simpler implementations, etc., which make their use appealing.

In this work, [94] ([J2] included in annex to this chapter), the flexibility of the Volterra based models is preserved, while the model is modified to include the long-term dynamics required for successful linearization of the devices. To do this, it is assumed that the long-term dynamics are sufficiently slow, so that, for each time-span of the short-term dynamics the system is successfully linearized by a typical GMP model. However, the GMP model required at each time interval is slightly different, according to the changes in the PA due to the long-term effects. In this sense, the DPD to correct the PA could be described as shown in (4.9), where $\vec{\alpha}(n)$ is the state vector describing the long term state.

$$\begin{aligned}\tilde{y}(n) = &\sum_{p=0}^{P}\sum_{m=0}^{M}\tilde{h}_p\big(m,0,\vec{\alpha}(n)\big)\tilde{x}(n-m)|\tilde{x}(n-m)|^p \\ &+\sum_{p=1}^{K}\sum_{m=0}^{M}\sum_{l=1}^{L}\tilde{h}_p\big(m,l,\vec{\alpha}(n)\big)\tilde{x}(n-m)|\tilde{x}(n-m-l)|^k \\ &+\sum_{p=1}^{K}\sum_{m=0}^{M}\sum_{l=1}^{L}\tilde{h}_p\big(m,-l,\vec{\alpha}(n)\big)\tilde{x}(n-m)|\tilde{x}(n-m+l)|^k\end{aligned} \tag{5.1}$$

In this approach, it is also assumed that the device will be in a quasi-settled state mode of operation where the long-term vector is almost constant with small variations. In this case, the coefficients of the model can be expanded as seen in (5.2).

$$\tilde{h}_p\big(m,l,\vec{\alpha}(n)\big) \cong \tilde{h}_p(m,l,\vec{\alpha}_0) + \frac{\Delta\tilde{h}_p(m,l,\vec{\alpha}(n))}{\Delta\vec{\alpha}(n)}\vec{\alpha}(n) \tag{5.2}$$

Both the settled state coefficients and the variation coefficients can be extracted using regular least-squares techniques as long as the long-term state vector is known. To find the state vector, in this work, a physics based approach was taken. Two sources of long-term memory were taken into account: thermal effects and electron trapping effects. These effects are known to produce significant changes in the behavior of GaN transistors [29, 63, 80, 95]. Taking these effects into account, auxiliary models that output signals mimicking the internal temperature variation and charge trapping are built. To do this, the models found in the literature to describe these effects are converted for application in the envelope domain.

The final model used for predistortion of the RFPA is shown in Fig. 5.1.

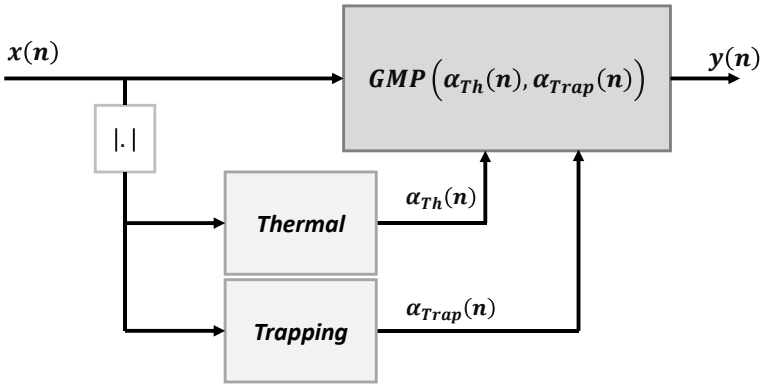


Fig. 5.1: Controlled GMP model used for predistortion of GaN HEMT based power amplifiers afflicted with long-term memory effects.

5.1.1. Thermal Model

The thermal model is obtained from the model shown in Fig. 5.2, [96].

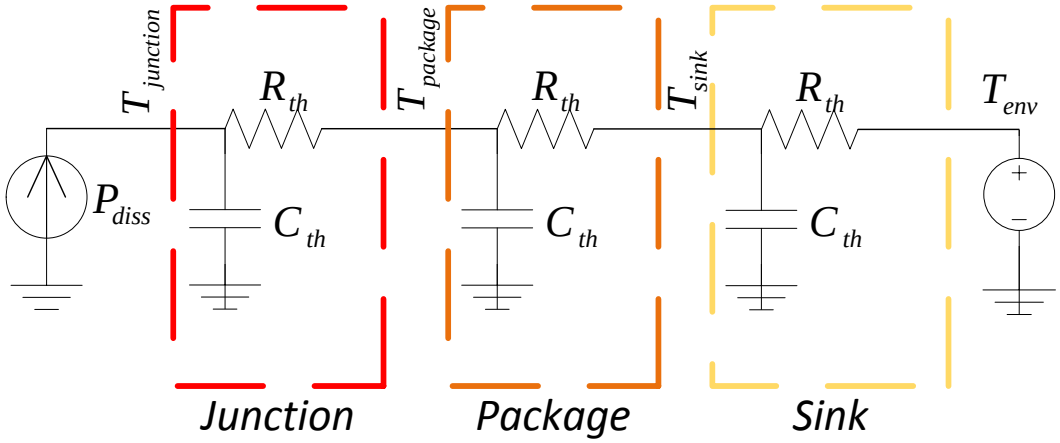


Fig. 5.2: Thermal circuit showing how the dissipated power impacts the temperature of the device, the temperature will then impact how the transistor generates current.

For a single-ended transistor the dissipated power can be calculated according to the conduction angle. Furthermore, the dissipated power is directly dependent on the amplitude of the envelope. This being the case, this model can be directly applied to the envelope signal. This model can be further simplified by noticing that the package

and sink should work at an approximately constant temperature. If this is the case, the time-constant of the junction will have the highest impact in the temperature variation inside the device. This limits the auxiliary model responsible for generating the temperature control signal to an Hammerstein system. First, a nonlinear function is applied to the input signal to obtain the dissipated power. Then, this dissipated power is filtered through a first order filter to obtain the temperature control signal [94] ([J2] included in annex to this chapter).

5.1.2. Trapping Model

In GaN HEMTs two main sources of trapping effects have been reported, one controlled by the gate to source voltage (gate lag), and the other by the drain to source voltage (drain lag). The gate lag has been greatly reduced in recent transistor technology, [81], and thus the main source of trapping effects in GaN HEMT devices is due to the drain side.

To obtain the auxiliary model for the trapping state, the model shown in Fig. 5.3, from [80], was adopted. In circuit level modeling, the excitation of the trapping circuit in Fig. 5.3 is the drain to source voltage of the transistor. In this case, the circuit is excited by an RF voltage and should be converted to the envelope domain for application in DPD. Furthermore, the excitation is made by the output of the device which is initially distorted. Nonetheless, as the linearization progresses the PA output becomes similar to the input signal to the predistorter. To simplify the DPD model, the input signal to the DPD can be used as the input to the trapping auxiliary model as well.

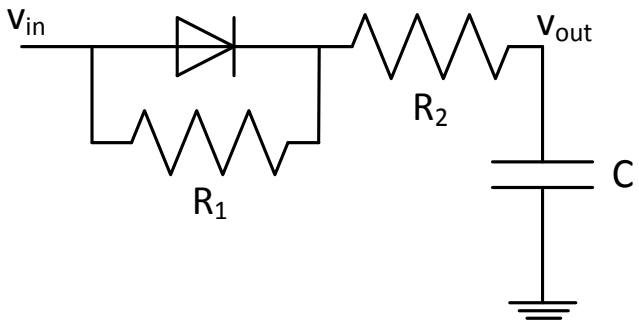


Fig. 5.3: Trapping circuit equivalent, the diode is responsible for generating a different charging and discharging behavior of the capacitor, the voltage in the capacitor impacts the current generation of the transistor. Typically the voltage is considered to cause a change of the threshold voltage.

In [94] ([J2] included in annex to this chapter), the model in Fig. 5.3 is directly used with the signal envelope. However, as pointed out, this model is actually excited by the RF voltage signal at the drain side. To correctly convert this model into the LPE domain, note that the model is essentially defined by two time-constants one for charging (representing electron trapping) and another for discharging (representing electro detrapping). Furthermore, each of these time-constants is much higher than the

RF time cycle. This means that the changes at the output in one RF cycle are very small. In fact, the following two behaviors can be assumed. The first type of response happens when the charge in the capacitor produces a sufficiently low voltage that the excitation RF wave rises above this voltage. In this case, at the beginning of the RF cycle the output capacitor presents some voltage V_0; during this RF cycle, the capacitor will charge whenever the RF wave goes above V_0 and will discharge whenever is goes below V_0; at the end of the RF cycle, the capacitor will hold a charge $V_1 = V_0 + \frac{dV}{dt}T$, where T is the RF cycle time. This behavior can only be assumed if the voltage changes very little along the RF cycle, which allows calculating the charging and discharging times, approximately, from the initial charge. This behavior is exemplified in Fig. 5.4.

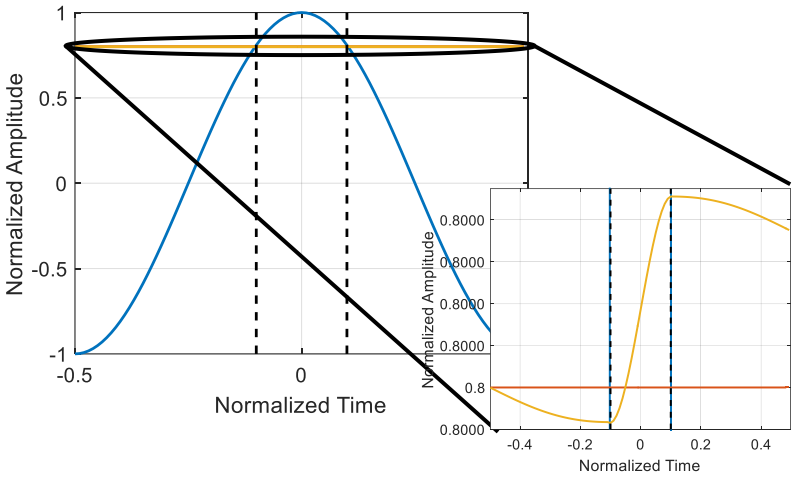


Fig. 5.4: Example of the trapping equivalent circuit behavior during one RF cycle. As can be seen, assuming the time constants of the system are much higher than the RF cycle time allows the calculation of the charging and discharging times within the cycle, from the initial capacitor voltage.

The second response type happens when the voltage stored in the capacitor is higher than the voltage reached by the excitation RF wave. In this case, the circuit is always discharging and is a linear first order filter, which means the output will tend towards the average of the input wave.

For these two response types it is assumed that the charging time constant is lower than the discharging time constant. If the inverse is true the circuit will have a similar behavior but will tend towards a negative voltage. If the constants are the same, the circuit is linear and will tend to the average input value.

Clearly the first response type is more complex, since the second response type is simply a discharging linear filter. Examining the first response type, looking at Fig. 5.4, the charging and discharging times are given by (5.3), assuming a sinusoidal excitation, where A is the excitation amplitude and ω is the excitation frequency.

$$T_c = \frac{2}{\omega} cos^{-1}\left(\frac{V_0}{A}\right)$$
$$T_d = \frac{2}{\omega}\left(\pi - cos^{-1}\left(\frac{V_0}{A}\right)\right) \quad (5.3)$$

During the charging time, the circuit is a linear first order filter with an input much beyond its cutoff frequency. In this case, the system will charge towards the average value of the excitation and the final voltage is given by (5.4), where $\tau_c = R_2 C$ is the charging time constant.

$$\bar{V} = \frac{1}{T_c}\int_{-\frac{T_c}{2}}^{\frac{T_c}{2}} A cos(\omega t) dt = \frac{2A}{\omega T_c} sin\left(\omega \frac{T_c}{2}\right) = \frac{2A}{\omega T_c}\sqrt{1-\left(\frac{V_0}{A}\right)^2} = \frac{\sqrt{1-\left(\frac{V_0}{A}\right)^2}}{cos^{-1}\left(\frac{V_0}{A}\right)}$$
$$V_c = V_0 + \left(\frac{\sqrt{1-\left(\frac{V_0}{A}\right)^2}}{cos^{-1}\left(\frac{V_0}{A}\right)} - V_0\right)\frac{T_c}{\tau_c} \quad (5.4)$$

In the discharging cycle, the circuit is again a linear first order filter excited much beyond cutoff. Again, the system will tend towards the average value of the excitation, and the final voltage is then given by (5.5), where $\tau_d = (R_1 + R_2)C$ is the discharging time constant.

$$\bar{V} = \frac{1}{T_d}\left(\int_{-\pi}^{-\frac{T_c}{2}} A cos(\omega t)\, dt + \int_{\frac{T_c}{2}}^{\pi} A cos(\omega t)\, dt\right) =$$
$$= -\frac{2A}{\omega T_d} sin\left(\omega \frac{T_c}{2}\right) = -\frac{2A}{\omega T_d}\sqrt{1-\left(\frac{V_0}{A}\right)^2} = -\frac{A\sqrt{1-\left(\frac{V_0}{A}\right)^2}}{\pi - cos^{-1}\left(\frac{V_0}{A}\right)} \quad (5.5)$$
$$V_d = V_0 + \left(-\frac{\sqrt{1-\left(\frac{V_0}{A}\right)^2}}{\pi - cos^{-1}\left(\frac{V_0}{A}\right)} - V_0\right)\frac{T_d}{\tau_d}$$

The overall response of the system in one RF cycle can then be given by (5.6).

$$V_1 = V_0 + \left(-\frac{A\sqrt{1-\left(\frac{V_0}{A}\right)^2}}{\pi - cos^{-1}\left(\frac{V_0}{A}\right)} - V_0\right)\frac{T_d}{\tau_d} + \left(\frac{A\sqrt{1-\left(\frac{V_0}{A}\right)^2}}{cos^{-1}\left(\frac{V_0}{A}\right)} - V_0\right)\frac{T_c}{\tau_c} \quad (5.6)$$

In fact, the same theory can be used to calculate the approximate response of the trapping circuit within the RF cycle. Overlapping this approximated response and

the response of the circuit model with a sinusoidal excitation is shown in Fig. 5.5.

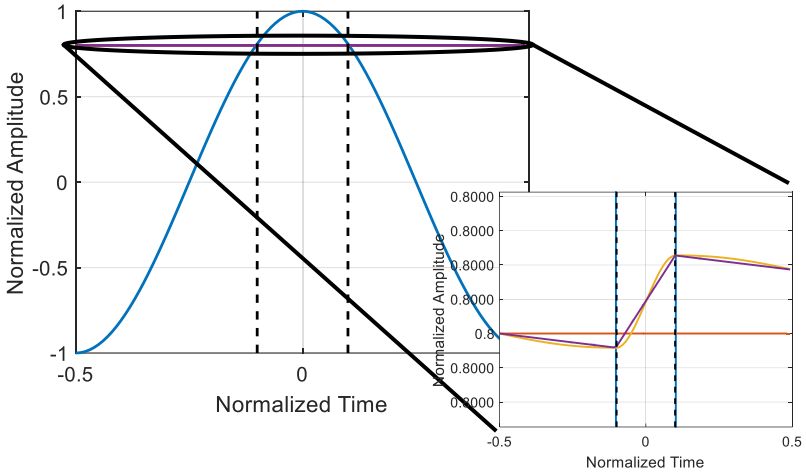


Fig. 5.5: Example of the trapping equivalent circuit behavior during one RF cycle (in yellow), with the approximated behavior overlapped (in purple). As can be seen, the approximated response is consistent with the simulated circuit response.

The dynamic LPE model for the charging can be calculated from (5.6) by noting that $V_1 = V_0 + \frac{dV}{dt}T$, this dynamic LPE model is then given by (5.7). The model for the discharging is a simple linear first order filter without any input.

$$\frac{dV}{dt} = \frac{\omega}{2\pi}\left(-\frac{A\sqrt{1-\left(\frac{V}{A}\right)^2}}{\pi - cos^{-1}\left(\frac{V}{A}\right)} - V\right)\frac{T_d}{\tau_d} + \left(\frac{A\sqrt{1-\left(\frac{V}{A}\right)^2}}{cos^{-1}\left(\frac{V}{A}\right)} - V\right)\frac{T_c}{\tau_c} = \tag{5.7}$$

$$\frac{dV}{dt} = \frac{1}{\pi}\left(\frac{1}{\tau_c} - \frac{1}{\tau_d}\right)\left(A\sqrt{1-\left(\frac{V}{A}\right)^2} - V\,cos^{-1}\left(\frac{V}{A}\right)\right) - \frac{V}{\tau_d}$$

Modifying the signals to the original nomenclature yields the mathematical model shown in (5.8), for the trapping auxiliary model, which can then be converted to the digital domain using the Euler approximation.

$$\frac{d\alpha_{trap}(t)}{dt} = \tag{5.8}$$

$$\frac{1}{\pi}\left(\frac{1}{\tau_u} - \frac{1}{\tau_d}\right)\left(|\tilde{x}(t)|\sqrt{1-\left(\frac{\alpha_{trap}(t)}{|\tilde{x}(t)|}\right)^2} - \alpha_{trap}(t)\,cos^{-1}\left(\frac{\alpha_{trap}(t)}{|\tilde{x}(t)|}\right)\right)$$

$$-\frac{\alpha_{trap}(t)}{\tau_d} \quad , \alpha_{trap}(t) \geq |\tilde{x}(t)|$$

$$\frac{d\alpha_{trap}(t)}{dt} = -\frac{\alpha_{trap}(t)}{\tau_d} \quad , \alpha_{trap}(t) < |\tilde{x}(t)|$$

Comparing this model with the one in [94] ([J2] included in annex to this chapter),

where the envelope's amplitude directly excites the RF circuitry, and the results from exciting the circuitry with the RF signal, shows that this model is indeed more accurate, at the cost of added complexity. Nonetheless, using the RF circuit model directly with the envelope already provides a good approximation of the output; this comparison is shown in Fig. 5.6 for a four carrier GSM (4cGSM) signal. Note that, for this application, the scaling of the signals is irrelevant, since the extraction procedure will adjust the sensitivity of the coefficients for better performance.

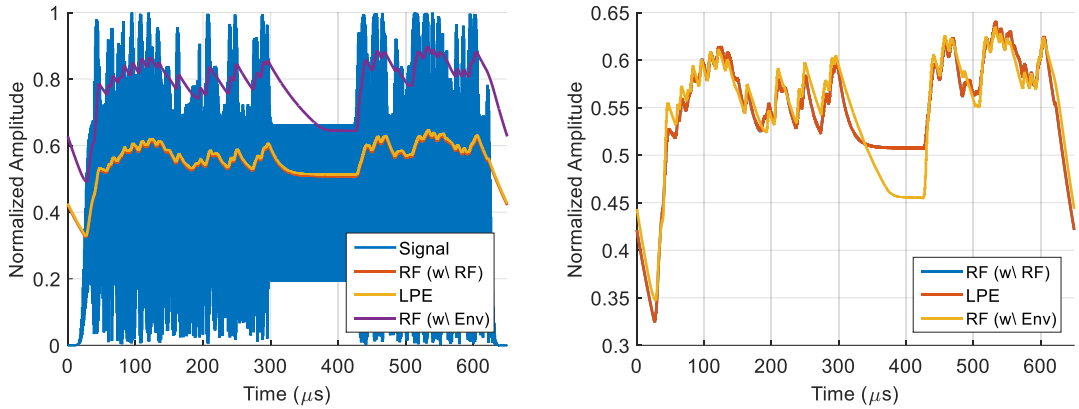


Fig. 5.6: Comparison between the auxiliary trapping model responses, non-scaled on the right and linearly scaled for better fitting on the left. When the signals are scaled both models provide a good approximation of the output, with the developed LPE model exactly fitting the RF model.

The related paper is here included for further information and results of the method.

5.2. Power Amplifier Linearization Methods for Future Communication Systems

One possibility for future transmitters is the use of a high number of parallel analog paths to feed an antenna array. This type of transmitter structure reduces the power of each RFPA and increases the significance of the power consumed by the baseband processing unit. In this sense, the complex DPD units that are used today may not be interesting for future transmitters.

One possibility to make DPD appealing in these architectures is to reduce the complexity, even at the cost of the achievable linearity, and make them simpler and, therefore, less resource intensive. To do this, relying in statistical descriptions of the signals has several advantages: first, the signal observation can simply be done in the minimal statistical sense required for the model to be calculated; second, in the case of a static model (possibly others) the statistics can be directly used to infer the direct and inverse relations from the input to the output.

In [97] ([C7] included in annex to this chapter) these statistical methods were used for static PA linearization, showing good performance without using any least squares solver and with minimal signal processing.

5.3. Summary and Final Considerations

This chapter was dedicated to novel methodologies applicable to PA linearization. First, a new approach for the correction of specific effects was presented. In particular, a DPD was designed from the physics based models and used for the linearization of GaN HEMT based PAs. Using this methodology, IMR improvements of up to 10 dB, with respect to standard linearization techniques, were achieved for 4cGSM signals in a number of amplifiers. Second, a methodology for fast, simple correction using statistical signal data was shown and applied to the correction of an LDMOS PA excited with several, different bandwidth, input signals.

The first approach, is intended for application in current DPD technologies, and extends the current state-of-the-art with a technique that allows compensation of the long-term phenomena that is observed mainly in GaN HEMT based PAs. This technique builds upon the typical DPD approach and adds auxiliary models to allow the description and compensation of the long term variations. To the author's knowledge, this proposed technique is the first to successfully linearize GaN PAs suffering from long-term memory effects due to trapping/detrapping.

The second approach is meant for application in future transmitter topologies. In this sense, the technique is meant to be cheap, possibly at the cost of effectiveness versus typical predistortion. The technique uses a statistical approach to avoid accurate representation of the signals mainly in the observation path. Furthermore, using statistical information, the model is directly given by the statistical measurements, with minimal signal processing, allowing cheap, fast predistortion.

Paper J2: Compensation of Long-Term Memory Effects on GaN HEMT Based Power Amplifiers

Filipe M. Barradas, Luís C. Nunes, Telmo R. Cunha, Pedro M. Lavrador, Pedro M. Cabral, and José C. Pedro

IEEE Transactions on Microwave Theory and Techniques

(Under Review)

Compensation of Long-Term Memory Effects on GaN HEMT Based Power Amplifiers

Filipe M. Barradas, *Student Member, IEEE*, Luis C. Nunes, *Student Member, IEEE*, Telmo R. Cunha, *Member, IEEE*, Pedro M. Lavrador, *Member, IEEE*, Pedro M. Cabral, *Member, IEEE* and José C. Pedro, *Fellow, IEEE*

***Abstract*—The long-term memory effects of Gallium Nitride (GaN) transistors have prevented its use in situations where the modulated envelope signal has a wide amplitude variation over time, such as in time division duplex (TDD) systems. These long term memory effects are generally attributed to electron trapping in GaN high electron mobility transistors (HEMT), which have shown to be very difficult to compensate, especially in cellular base station transmitters known to be subjected to highly restrictive linearity specifications. On top of the electron trapping effects, we show that thermal effects can also induce long term memory behaviors, which should also be accounted for when linearizing these devices.**

Because the conventional behavioral modeling approach has been incapable to compensate these long term memory effects on GaN HEMT based power amplifiers (PAs), we started by investigating the physical mechanisms responsible for these semiconductor impairments in GaN devices. This physics-based knowledge was then used to design new predistorter models that could effectively compensate those PAs subjected to GaN trapping and thermal effects. In this paper, we describe the new predistortion models for PA linearization, as well as the characterization methods used to determine their parameters.

To validate the linearization effectiveness of the proposed model, several high power GaN-based PAs are tested with multicarrier GSM signals, and their linearization results are compared against other state-of-the-art models, evidencing a clear and significant improvement. In fact, to the authors' knowledge, the proposed approach is the first one to reduce the PA distortion effects due to GaN long term memory effects to such low levels, allowing a comfortable compliance with the imposed linearity masks.

***Index Terms*— Digital predistortion, electron trapping, GaN, GaN HEMT, high power amplifiers, long-term memory effects, predistortion linearizers, power amplifier linearization.**

I. Introduction

DIGITAL predistortion (DPD) is now widely used for compensating the nonlinear characteristics of radio frequency (RF) PAs. In fact, modern RF PA design focuses mainly on efficiency, *a priori* assuming that the DPD can afterwards easily handle the nonlinear behavior [1].

For PA nonlinear compensation, DPD models based on the Volterra series have been largely favored because of their solid theoretical foundation and straightforward extraction methodology. Namely, the Volterra based models are universal approximators, whose structure is linear in the parameters, thus allowing the use of least squares solvers to extract the model parameters. Such models, as the memory polynomial [2] (MP), the generalized memory polynomial [3] (GMP) and the dynamic deviation reduction [4], [5] (DDR) have been successfully used to obtain a high level of nonlinear compensation of RF PAs. However, the recent GaN HEMT technology has introduced challenging behaviors in the PAs, which have prevented this previously obtained level of compensation to be reliably achieved.

These so called long term memory effects are characterized by a slow shift of the PAs' characteristics along the stimulus duration and have been attributed to the charge trapping and detrapping effects observed in the GaN technology [6]. Besides the trapping effects, our, and others, observations led us to conclude that thermal behaviors should also play a role in these slow dynamic behaviors, even if they are hardly observable in other technologies due to their associated higher time-constants [7]. It is the higher thermal conductivity of GaN substrates that now brings these effects to light.

TDD applications are particularly sensitive to these effects since the envelope has a wide range of amplitudes, ramping up and down slowly, which generates significant changes on the state of the PA. Thus, long term compensation would lead to a significant increase in the achievable linearity required to fulfill the stringent specifications in modulation schemes such as the multicarrier GSM.

Typical Volterra based DPD models can hardly be used to accommodate these long term memory effects due to their inherent feedforward structure. In this type of finite impulse response (FIR) like topology, long term memory effects would use an unbearably large number of memory coefficients [6], making the model evaluation computationally expensive and the coefficient extraction almost impossible. Moreover, as we shall see in the following sections, electron trapping effects have a different behavior from what is typically observed in electronic systems.

In this paper, we propose a novel approach to achieve unprecedented compensation of GaN based RF PAs. The technique is based on a mixture of typical behavioral modeling

Manuscript received October 4, 2016, revised December 22, 2016.

Filipe M. Barradas would like to acknowledge the financial support provided by the Portuguese Science and Technology Foundation, FCT (Ref. SFRH/BD/90103/2012).

This work was financed by Huawei Technologies.

The authors are with DETI, Instituto de Telecomunicações, Universidade de Aveiro, Campus Universitário de Santiago, Aveiro 3810-193, Portugal (e-mail: filipebarradas@ua.pt; lcnunes@ua.pt; trcunha@ua.pt; plavrador@ua.pt; pcabral@ua.pt; jcpedro@ua.pt).

Luís C. Nunes is now with Huawei Technologies Sweden.

signal processing techniques and physics based models. The physics based models are used as auxiliary models that control the coefficients of a typical DPD model, a technique that has been previously used in DPD [8-10].

This article is divided in three main sections. First, in Section II, we develop and support the DPD model structure that we will use for nonlinear distortion compensation. In Section III, we develop the auxiliary model structure to be used in the developed DPD. Finally, Section IV shows how the developed model was used for the compensation of several state-of-the-art RF PAs and compares the achieved results with the ones obtained with state-of-the-art DPD models.

II. DPD Model Structure

To develop the DPD model structure, let us initially assume an RF PA without any long-term memory behaviors. In this case, any typical DPD model can be used to achieve the required linearity. From experience, the GMP model [3], (1), typically fulfills this role with very good results and so we will use it as an example.

$$\begin{aligned}\tilde{y}(n) = &\sum_{p=0}^{P}\sum_{m=0}^{M}\tilde{h}_p(m,0)\tilde{x}(n-m)|\tilde{x}(n-m)|^p \\ &+\sum_{p=1}^{K}\sum_{m=0}^{M}\sum_{l=1}^{L}\tilde{h}_p(m,l)\tilde{x}(n-m)|\tilde{x}(n-m-l)|^k \\ &+\sum_{p=1}^{K}\sum_{m=0}^{M}\sum_{l=1}^{L}\tilde{h}_p(m,-l)\tilde{x}(n-m)|\tilde{x}(n-m+l)|^k\end{aligned} \quad (1)$$

In (1), the tilde is used to identify complex variables. $\tilde{x}(n)$ is the complex envelope at the input to the model and $\tilde{y}(n)$ is the complex envelope at the output of the model, $\tilde{h}$ are the model parameters, the trio (K, M, L) defines the model structure, and n is the sampled time.

Since the focus of this paper is on the compensation of the long-term memory behaviors, let us assume that the GMP model in (1) can indeed linearize the RF PA in the absence of these effects.

When the long-term behaviors are excited, the RF PA suffers a slow change of its characteristics through time, and the long-term behaviors can be characterized by a set of state variables, $\vec{\alpha}(n)$, changing in time. For each value of this state vector, the RF PA has changed slightly. In fact, as long as the usual PA time constants and the time-constants associated with the long-term memory effects are widely separated, we can assume that, for each state of the long-term state vector, we have a slightly different PA that could be corrected with the GMP model in (1), with a slight change in its coefficient values. So, assuming we have access to this hidden state vector we could describe a predistorter using (2).

$$\tilde{y}(n) = \sum_{p=0}^{P}\sum_{m=0}^{M}\tilde{h}_p\big(m,0,\vec{\alpha}(n)\big)\tilde{x}(n-m)|\tilde{x}(n-m)|^p \quad (2)$$

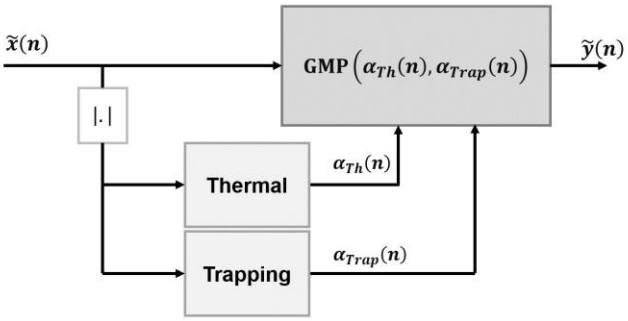


Fig. 1. DPD Model Structure, the model consists of a typical GMP model, controlled by two auxiliary signals, generated from the auxiliary models. These two models are developed from physical considerations.

$$\begin{aligned}&+\sum_{p=1}^{K}\sum_{m=0}^{M}\sum_{l=1}^{L}\tilde{h}_p\big(m,l,\vec{\alpha}(n)\big)\tilde{x}(n-m)|\tilde{x}(n-m-l)|^k \\ &+\sum_{p=1}^{K}\sum_{m=0}^{M}\sum_{l=1}^{L}\tilde{h}_p\big(m,-l,\vec{\alpha}(n)\big)\tilde{x}(n-m)|\tilde{x}(n-m+l)|^k\end{aligned}$$

In (2) the *a priori* knowledge of the long-term behavior is used to avoid modifying the underlying compensation model. Two questions must be answered when using the formulation in (2): first, how do the coefficients of the original model change with the new variables?; and second, how to access the hidden long term state vector?

Several types of nonlinear expansions can be used to address the first problem as shown in [8-11]. Care must be taken to make sure that the number of coefficients is not excessively increased. As an initial approximation, the coefficients can be expanded as linear functions of the state vector, as seen in (3). This expansion increases the number of coefficients by KN, where K is the number of long-term state variables, and N is the number of coefficients.

$$\tilde{h}_p\big(m,l,\vec{\alpha}(n)\big) \cong \left(\tilde{h}_p(m,l) + \frac{\Delta\tilde{h}_p(m,l)}{\Delta\vec{\alpha}(n)}\vec{\alpha}(n)\right) \quad (3)$$

In our approach, we have limited the number of the long-term state variables to two: one describing the changes due to the thermal effects, $\alpha_{Th}(n)$; and the other to the electron trapping effects, $\alpha_{Trap}(n)$. In order to have access to these state variables, we studied the behavior of these effects at the device level and represented them as digital models dependent on the input envelope.

Using the described approach, the DPD model structure is the one shown in Fig. 1, where the GMP coefficients depend linearly on the control signals. Please note that it is not necessary that the control signals track the absolute value of the temperature and trapping charge. In fact, since the model is updated regularly, only the deviations from the average value need to be tracked. Looking at (3), it can be seen that the average value can be accommodated in the $\tilde{h}_p(m,l)$ term. Furthermore, any scaling of the control signals is irrelevant since the corresponding coefficient is also extracted in the regression process.

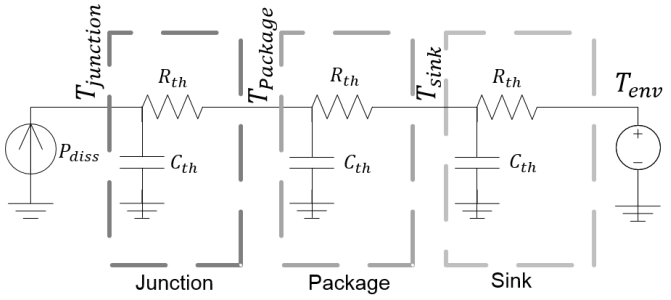


Fig. 2. Simple thermal model of the transistor from the junction of the device to the sink of the RF PA.

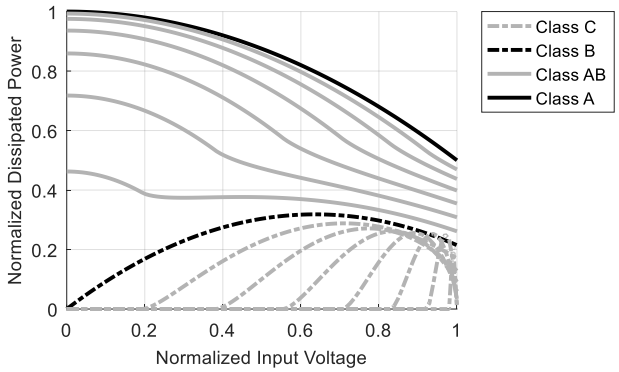


Fig. 3. Normalized (to $P_{DC,max}$) dissipated power, depending on the normalized (to $v_{in,max}$) input voltage, for several conduction angles of the transistor.

A. Coefficient Extraction

Assuming the control signals, $\alpha_{Th}(n)$ and $\alpha_{Trap}(n)$, are known, the model maintains linearity in the parameters, as shown in [9]. Hence, typical least squares extraction methods can be applied to the new regression matrix, shown in (4), where X_{GMP} is the original GMP regression matrix, A_{Th} is the thermal control signal in vector form, A_{Trap} is the trapping control signal in vector form, and $\otimes$ is the Kronecker product.

$$X = [X_{GMP} \quad \mathrm{A}_{Th} \otimes X_{GMP} \quad \mathrm{A}_{Trap} \otimes X_{GMP}] \tag{4}$$

III. Physics Based Auxiliary Models

From the previous section, the model is effectively determined, as long as the control signals are known. The main difficulty in this type of systems is in determining these auxiliary models. In fact, the global output of the system is nonlinearly dependent on the control signals and, therefore, on the parameters of the control models. Moreover, a direct measurement of the state variables themselves proves impossible as these are deep level trap charges and temperature of the GaN HEMT active layer. Instead of using a general nonlinear model, we determined the auxiliary models' structure by exploring the physical mechanisms that are known to be behind these phenomena. In this way, we expect to limit the model complexity and maintain a reasonable approximation. We then developed a method to extract the model parameters from the input-output measurements using real signals.

A. Thermal Model

The temperature of the device depends on the power dissipated on the transistor. This thermal mechanism can be interpreted as a heat flow source that will dissipate through the junction, to the package and the sink to the environment, as shown in Fig. 2, [12]. Each of the temperature interfaces will impose a time-constant on the thermal system. However, the package of the device and the sink of the RF PA present a significantly high time-constant and tend to settle at a constant temperature when the RF PA is in operation; this leaves only the junction of the device to consideration. The temperature at the junction interface is dependent on the thermal capacitance and conductivity of the transistor materials.

For GaN HEMTs the impact of the thermal effects has been noted to be of significance in the device modelling and for RF PA design, not only in terms of the static temperature, but also the associated dynamics [13].

The model for the junction temperature can be written as shown in (5), for continuous time. Since it is assumed that the package temperature settles to some constant, the $T_{package}$ term is irrelevent for our purposes.

$$T_{junction} + R_{th}C_{th}\frac{dT_{junction}}{dt} = R_{th}P_{diss} + T_{package} \tag{5}$$

In eq. (5), the variables relate to Fig. 2, $T_{junction}$ is the junction temperature, $T_{package}$ is the package temperature, R_{th} is the thermal resistance and C_{th} is the thermal capacitance.

Looking in the literature, [14], the dissipated power can be approximated as a function of the input envelope, typically not its squared value, depending on the class of operation.

Equation (6) shows the dissipated power formula for a normalized input envelope, depending on the conduction angle θ, for the optimum power load. Fig. 3, plots the dissipated power curve, depending on input amplitude, for the different conduction angles.

$$\frac{P_{DC}}{P_{DC,max}} = \frac{1}{\pi}\left(\frac{v_{in}}{v_{in,max}} sin\left(\frac{\xi}{2}\right) - cos\left(\frac{\theta}{2}\right)\frac{\xi}{2}\right)$$

$$\frac{P_L}{P_{DC,max}} = \frac{r_L}{2\pi^2}\left(\frac{v_{in}}{2v_{in,max}}(\xi + sin(\xi)) - 2\,cos\left(\frac{\theta}{2}\right) sin\left(\frac{\xi}{2}\right)\right)^2$$

$$\frac{P_{diss}}{P_{DC,max}} = \frac{P_{dc}}{P_{out,max}} - \frac{P_L}{P_{out,max}} \tag{6}$$

$$r_L = \frac{2\pi}{\theta - sin(\theta)}, \qquad normalized\ to\ 1\Omega$$

$$\xi = \begin{cases} 2\,cos^{-1}\left(\dfrac{1}{\frac{v_{in}}{v_{in,max}}} cos\left(\dfrac{\theta}{2}\right)\right) & \dfrac{v_{in}}{v_{in,max}} \geq \left|cos\left(\dfrac{\theta}{2}\right)\right| \\ \pi\big(1 + sign(\theta - \pi)\big) & \dfrac{v_{in}}{v_{in,max}} < \left|cos\left(\dfrac{\theta}{2}\right)\right| \end{cases}$$

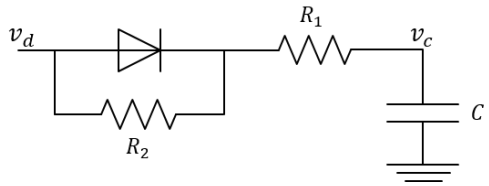


Fig. 4. Electron trapping model, the charge is controlled by the drain voltage. The diode is introduced to allow different charging and discharging time-constants.

In (6), v_{in} is the input voltage, P_L is the output power, P_{DC} is the consumed power, P_{diss} is the dissipated power, θ is the conduction cangle at full power, and ξ is the effective conduction angle.

Typically, transistors in single-ended topologies of RF PAs will be biased close to Class B operation. In this case, the normalized power dissipation curve can be written as seen in (7), significantly simplifying the dissipated power expression.

$$\frac{P_{diss}}{P_{DC,max}} = \frac{v_{in}}{v_{in,max}} - \frac{\pi}{4}\left(\frac{v_{in}}{v_{in,max}}\right)^2 \tag{7}$$

Taking all this information, the models can be translated to the digital low-pass equivalent (LPE) domain, using the bilinear z-transform to accommodate the dynamic effects [15]. This yelds the thermal auxiliary model, shown in (8). For simplicity, equation (8) has been scaled and the constant terms have been removed. As previously explained, this can be done since the regression process will restore these parameters.

$$\begin{aligned}(1-2f_s\tau_{Th})\alpha_{Th}(n-1) + (1+2f_s\tau_{Th})\alpha_{Th}(n) \\ = f(|x(n-1)|) + f(|x(n)|) \\ f(|x(n)|) = |x(n)| - \frac{\pi}{4}|x(n)|^2\end{aligned} \tag{8}$$

In (8), f_s is the sampling frequency and $\tau_{th} = R_{th}C_{th}$ is the thermal time constant.

The thermal auxiliary model depends on one single parameter, which is the thermal time-constant, τ_{Th}.

B. Electron Trapping Model

Trapping in GaN HEMTs has been observed at the gate, where it is known as gate lag, and at the drain, known as drain lag. Of these two effects, the gate lag has mostly been solved or significantly reduced in current GaN HEMT generations [16]. The drain lag effect is still observed and has a clear impact on the transistor behavior, the most significant being self-biasing when in operation [17].

The electron trapping effects are characterized by having two very different associated time-constants. One related to the charge trapping and another to the de-trapping. These different time-constants impose a different behavior than what is traditionally observed in common physical systems.

One of the proposed systems to track the trapped charge in simulation models was proposed in [18] and is shown in Fig. 4. The diode acts as an ideal device, presenting a short circuit when the voltage across is positive and an open circuit when it is negative. The two different resistors allow setting different charge and discharge time-constants. The charge model is described by (9).

$$\begin{cases} R_2 C\dfrac{dv_c(t)}{dt} = \big(v_d(t) - v_c(t)\big) & v_c(t) \le v_d(t) \\ (R_1+R_2)C\dfrac{dv_c(t)}{dt} = \big(v_d(t) - v_c(t)\big) & v_c(t) > v_d(t) \end{cases} \tag{9}$$

In (9), the variables relate to Fig. 4, v_d is the transistor drain voltage and v_c is the voltage in the capacitor, which directly relates to the accumulated charge.

The input to the trapping charge model is the real drain voltage, not the envelope amplitude. However, directly using the envelope amplitude in this model yields a good approximation of the real result. For simplicity, in this work we have used this expression directly converted to the digital LPE domain (using the bilinear transform) as shown in (10).

$$\begin{cases} (1-2f_s\tau_d)\alpha_{Trap}(n) + (1+2f_s\tau_d)\alpha_{Trap}(n-1) \\ = |x(n-1)| + |x(n)| & \alpha_{Trap}(n) > |x(n)| \\ (1-2f_s\tau_u)\alpha_{Trap}(n) + (1+2f_s\tau_u)\alpha_{Trap}(n-1) \\ = |x(n-1)| + |x(n)| & \alpha_{Trap}(n) \le |x(n)| \end{cases} \tag{10}$$

In (10), $|x(n)|$ is the envelope amplitude of the input signal, $\alpha_{Trap}(n)$ is the trapped charge, $\tau_{up} = R_1 C$ is the charging time-constant, $\tau_d = R_2 C$ is the de-charging time-constant, and f_s is again the sampling frequency. This model has two parameters to be found, the charging and de-charging time-constants.

Please note that the input to the trapping model should, in fact, be the output from the transistor, not its input, which would require a feedback model structure. So, initially, there will be a significant difference between the output of the PA and its input. However, as the PA becomes linearized the output becomes closer and closer to the input, making this a valid approximation.

C. Auxiliary Model Fitting

As mentioned before, the output of the DPD module depends nonlinearly on the control signals from the auxiliary models. Furthermore, the auxiliary models have infinite impulse response filters, which increases the difficulty of parameter extraction.

To extract the parameters for these models, we developed a new technique based on the observation of residuals (defined further ahead). The parameters are only extracted once and used for multiple different signals. This is an advantage of using physically meaningful models, as the model parameters cannot change, unless the device itself is changed.

In order to extract the auxiliary model parameters, a representation of what they should produce is required. However, as previously explained, we cannot directly measure the internal temperature of the device and the trapped charge. Nonetheless, an indirect measure of these signals can be obtained. As mentioned before, the RF PA characteristics shift with the change of temperature and trapped charge. This slow shift can be observed at the output fundamental envelope when the PA is excited with a signal. However, it will be mixed with the much faster changes produced by the input signal, which are clearly dominant. The output signal must then be treated to extract the required information. The output

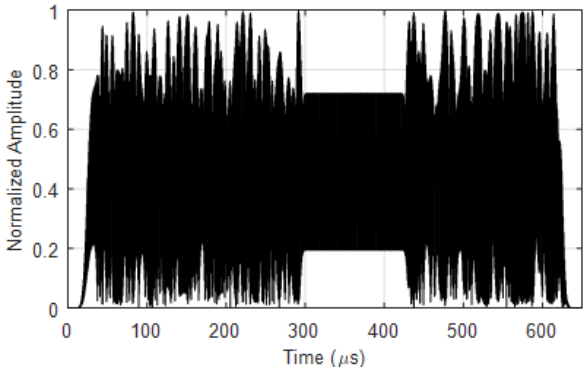


Fig. 5. 4cGSM normalized amplitude envelope.

signal can be written in the form shown in (11), where $a(.)$ is the AM/AM and $\phi(.)$ is the AM/PM.

$$\tilde{y}(n) = a\left(\tilde{x}(n), \tilde{x}(n-1), \dots, \alpha_{Trap}(n), \alpha_{Th}(n)\right) \exp\left(j\angle x(n)\right) \exp\left(j\phi\left(\tilde{x}(n), \tilde{x}(n-1), \dots, \alpha_{Trap}(n), \alpha_{Th}(n)\right)\right) \quad (11)$$

To extract the influence of the thermal and trapping signals on the output, we begin by approximating the output signal with a static function of the input signal, (12). To do this, we use a least squares extraction of a memoryless polynomial.

$$\tilde{y}_a(n) = a_a(|\tilde{x}(n)|) \exp\left(j\phi_a(|\tilde{x}(n)|)\right) \exp\left(j\angle x(n)\right) \quad (12)$$

In eq. (12), $\tilde{y}_a$ is the memoryless approximation, $a_a(.)$ is the memoryless AM/AM and $\phi_a(.)$ is the memoryless AM/PM.

After obtaining this memoryless approximation, we calculate a residual as shown in (13).

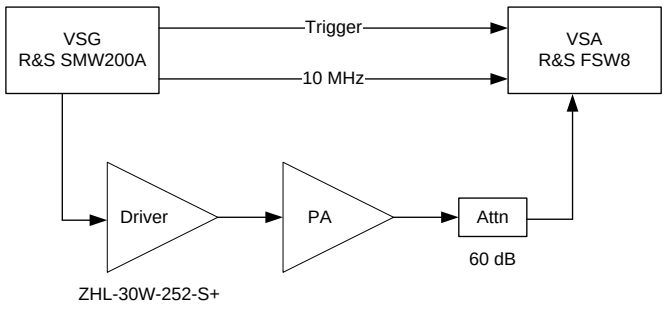
$$r(n) = a\left(\tilde{x}(n), \tilde{x}(n-1), \dots, \alpha_{Trap}(n), \alpha_{Th}(n)\right) - a_a\left(\tilde{x}(n)\right) \quad (13)$$

$$G_r(n) = \frac{r(n)}{a_a\left(\tilde{x}(n)\right)}$$

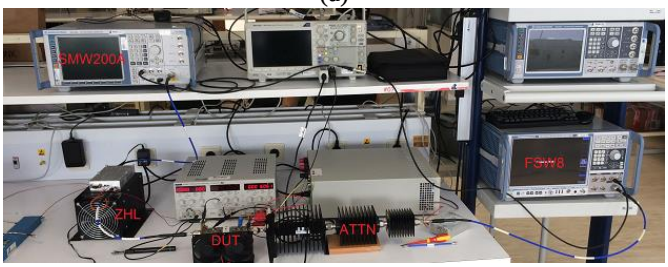
The residual formulation in (13) measures the deviation from the static gain. Therefore, we expect it to have the slow thermal and trapping behaviors, but also much faster PA dynamic components that should be captured by the GMP model. To remove these fast components, a low-pass filter is used. The resulting filtered residual is dependent on the thermal and trapping states and provides the required insight into these hidden variables. The time-constants are then found by using a nonlinear fitting method to minimize the error shown in (14), where k_{th}, k_{Trap} and C are extracted by least squares for each attempted solution of the time-constants.

$$\sum_n \left|G_{r,f}(n) - k_{th}a_{Th}(n) - k_{Trap}a_{Trap}(n) - C\right|^2 \quad (14)$$

In (14), $G_{r,f}$ is the aforementioned low pass filtered residual signal. Typically, the effects produced by the thermal model and the trapping model are very different: the thermal changes



(a)



(b)

Fig. 6. Measurement setup schematic (a) and photograph (b), the DUT in the photo is the LDMOS test board.

cause a slower variation across the full range of the signal while the trapping produces faster variations mainly related to the peaks of the signals. Thus, we have found that the optimization can be split into two different problems: first, the thermal time-constant is found by optimizing (15),

$$\sum_n \left|G_{r,f}(n) - k_{th}a_{Th}(n) - C_{th}\right|^2 \quad (15)$$

and then the thermal impact is subtracted from the total residue, so that, finally, the trapping time-constants are found by optimizing (16).

$$\sum_n \left|G'_{r,f}(n) - k_{Trap}a_{Trap}(n) - C_{Trap}\right|^2$$
$$G'_{r,f}(n) = G_{r,f}(n) - k'_{th}a_{Th}(n) - C'_{Th} \quad (16)$$

In (16), k'_{th} and C'_{th} are the specific constants extracted using least-squares for the optimal thermal time-constant.

This second approach with two optimization problems was largely favored and provides a number of benefits. First, the reduced number of optimization variables in each problem expedites the process. Second, the trapping and thermal models have similar topologies. In fact, when the charging and discharging time-constants are very close to each other, the trapping model is equal to the thermal model. This problem can lead to local minima in the error and other optimization problems.

IV. Results

Since our approach is based on the observation and fitting of a residual signal, we first verified that this residual indeed represents what we expected. After the method based on the residual was validated, we applied it to the described GaN HEMT based PA predistortion methodology and compared the

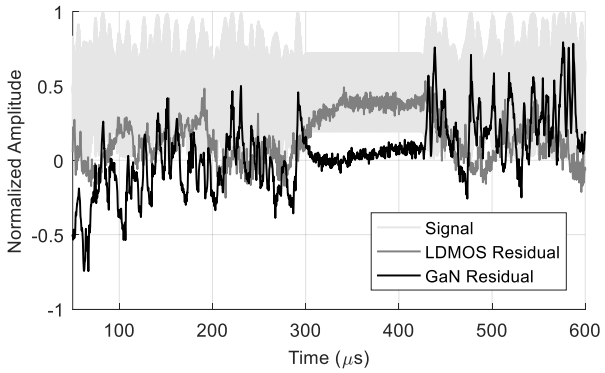


Fig. 7. Residual signal produced by a measurement of an LDMOS PA and a GaN HEMT PA, for a 4cGSM signal with 5 MHz bandwidth.

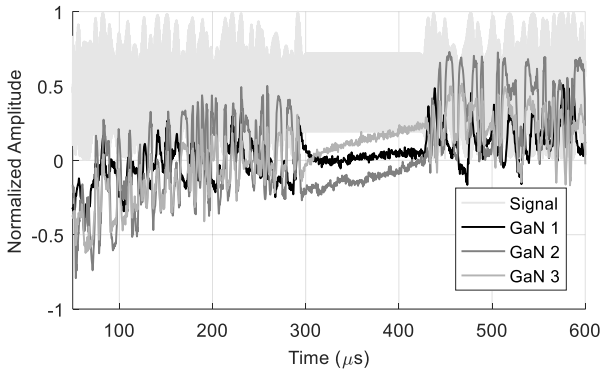


Fig. 8. Residual signals produced by measurements of several GaN HEMT based PAs, for a 4cGSM signal with 5 MHz bandwidth.

reached linearity with the typical DPD approach.

For these tests we used 4cGSM signals with several bandwidths (5 MHz, 10 MHz and 15 MHz). The bandwidth of this signal type is defined by the carrier separation, since each carrier has a constant bandwidth of 200 kHz. The amplitude envelope for a 5 MHz 4cGSM signal is shown in Fig. 5. As can be observed, the signals have a training sequence, as well as a ramp up and ramp down sections. These signal parts further excite the long term dynamics of the device. Multicarrier GSM signals are very stable in terms of PAPR, for the four carrier case the typical PAPR value is around 6.4 dB. This is true for all test cases.

All the results here presented were measured on a typical DPD test bed shown in Fig. 6. It consists of a SMW200A Vector Signal Generator (VSG) and an FSW8 Vector Signal Analyzer (VSA) – both from Rohde & Schwarz –, which excite and measure the response of the devices under test (DUT). A driver PA, a ZHL-30W-252+ from Mini-Circuits, is used to reach the required power levels. The output from the DUT is attenuated by a 60 dB attenuator.

For simplicity and convergence speed, the DPD is trained using an indirect learning approach [19], the sampling frequency is set at 100 MHz, each set of data is 650 µs long, corresponding to 65 thousand points. In all studied cases, the DPD model converges in 3~4 iterations.

A. Validation of the Residual Signal

To validate the residual signal we measured a GaN HEMT based RF PA in class AB and an LDMOS based RF PA, also in class AB. We expect that the long-term residual generated by the measurement of the GaN PA is higher than the residual generated by the LDMOS based RF PA.

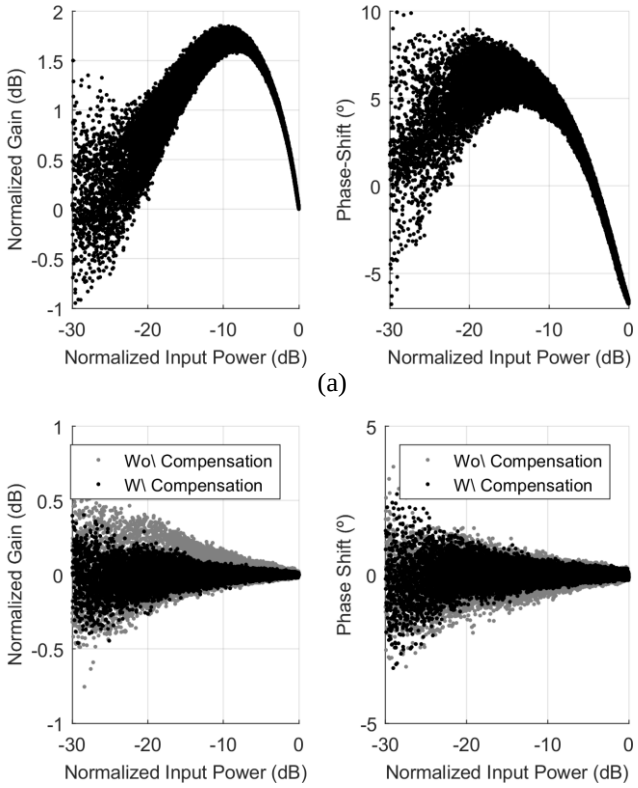


Fig. 9. Original AM/AM and AM/PM (a); and AM/AM and AM/PM after linearization with and without long-term compensation (b). The DUT was the GaN 3 based, class AB, RF PA.

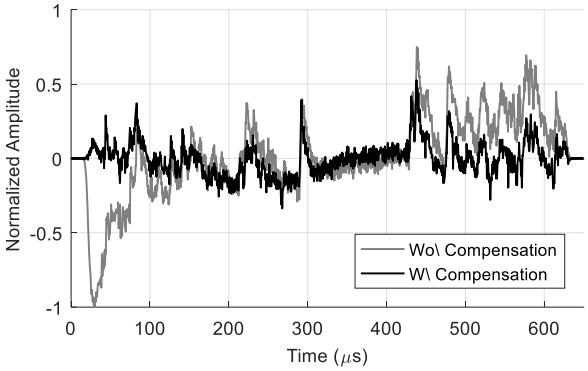


Fig. 10. Residual signal after linearization with long-term compensation and without compensation. The DUT was the GaN 3 based, class AB, RF PA.

The results for this test are shown in Fig. 7. The LDMOS PA produces a residual around zero with slow fluctuations, while the GaN PA produces a residual with a slow upward trend and high amplitude faster fluctuations. We have associated the slow trend with a temperature variation and the faster variations with the trapping effects.

This procedure was repeated for other GaN transistors with the same results, as shown in Fig. 8. This shows that this residual can indeed be used to obtain some information on the long-term memory of the device. However, it also shows that there is still room for improving this metric.

B. PA Linearization

TABLE I
LINEARIZATION RESULTS FOR THE GAN TRANSISTOR 1

Frequency (@34 dBm)	NMSE / IMR (dB)*			Power (@2 GHz)	NMSE / IMR (dB)*		
	5 MHz	10 MHz	15 MHz		5 MHz	10 MHz	15 MHz
1.8 GHz	48.2 / 60.8 (45.0 / 58.4)	45.2 / 60.1 (43.1 / 56.8)	45.2 / 62.7 (43.1 / 59.5)	31 dBm	51.4 / 65.8 (47.7 / 61.6)	48.7 / 65.0 (46.2 / 60.2)	49.0 / 65.3 (46.2 / 61.3)
2 GHz	50.2 / 64.2 (45.7 / 59.9)	47.4 / 62.6 (44.0 / 58.0)	47.7 / 63.2 (44.0 / 60.6)	32 dBm	51.8 / 66.2 (47.8 / 61.5)	48.6 / 64.1 (45.9 / 59.7)	48.8 / 65.8 (46.1 / 61.1)
2.2 GHz	51.3 / 65.3 (46.6 / 60.4)	48.0 / 63.0 (45.2 / 57.3)	48.0 / 64.4 (45.4 / 59.8)	33 dBm	51.3 / 65.2 (47.1 / 60.7)	48.4 / 63.8 (45.8 / 59.4)	48.6 / 65.7 (45.6 / 60.3)
2.4 GHz	50.0 / 63.0 (45.6 / 59.5)	46.8 / 61.7 (44.2 / 57.3)	47.3 / 62.3 (44.1 / 59.2)	34 dBm	50.8 / 64.8 (46.5 / 59.8)	48.0 / 63.4 (45.3 / 57.5)	48.3 / 63.9 (45.4 / 59.6)

*The results within brackets are from the model without long term compensation.

TABLE II
LINEARIZATION RESULTS FOR THE GAN TRANSISTOR 2

Frequency (@35 dBm)	NMSE / IMR (dB)*			Power (@1.8 GHz)	NMSE / IMR (dB)*		
	5 MHz	10 MHz	15 MHz		5 MHz	10 MHz	15 MHz
1.7 GHz	48.8 / 62.1 (43.7 / 51.5)	46.2 / 59.6 (42.3 / 50.4)	45.8 / 61.0 (42.3 / 51.9)	32 dBm	52.2 / 67.4 (47.2 / 64.2)	49.5 / 66.4 (46.0 / 63.2)	49.1 / 63.7 (45.7 / 62.6)
1.8 GHz	51.6 / 68.4 (46.3 / 62.5)	49.3 / 65.9 (45.3 / 61.3)	49.3 / 64.7 (45.0 / 60.8)	33 dBm	52.2 / 67.6 (46.6 / 63.4)	49.5 / 66.9 (45.8 / 61.9)	49.2 / 64.1 (45.5 / 62.0)
1.9 GHz	51.4 / 65.2 (45.6 / 58.0)	48.9 / 63.0 (44.4 / 55.9)	48.8 / 64.4 (44.3 / 57.4)	34 dBm	52.5 / 68.6 (46.5 / 62.7)	49.5 / 66.4 (45.4 / 61.9)	49.3 / 64.6 (45.4 / 61.5)
				35 dBm	51.6 / 68.7 (46.3 / 62.5)	49.3 / 65.9 (45.3 / 61.3)	49.3 / 64.7 (45.0 / 60.8)

*The results within brackets are from the model without long term compensation.

We performed linearization tests in four class AB PAs based on four different transistors. These tests were conducted for several peak powers and center frequencies. Throughout these tests we maintained the initially extracted long-term time constants. The time constants were extracted using one maximum power measurement for the center design frequency of the specific device.

After the time-constants were extracted, a series of linearization tests were run to verify the method. Linearization was attempted for each PA at the design frequency for several different input powers and at the maximum power for several frequencies. These tests reflect both the increased linearity obtainable when using the auxiliary models and the robustness of the extracted time-constants, since the long-term time-constants were *a priori* extracted and are maintained throughout the testing. For each test, the device was predistorted by a typical GMP with and without the long-term auxiliary models. First, the number of parameters for the GMP model without long-term memory effects was obtained. This is done by manually increasing the number of GMP parameters until the NMSE does not suffer a sensible improvement. For this, the polynomial order, *p*, was initially increased. Then the same was done with the direct memory delays, *m*; and finally the cross nonlinear and memory parameters, *k* and *l*. The final parameter set characterizing the used GMP model are (P=9, K=3, M=3, L=3). After these parameters are specified they are used in both the model with and without long-term compensation.

The results of the linearization tests are shown in Table I, II, III and IV (one for each device), for intermodulation ratio (IMR) and normalized mean square error (NMSE), with and without the proposed long-term compensation. The power is specified as input power. The IMR is measured between the

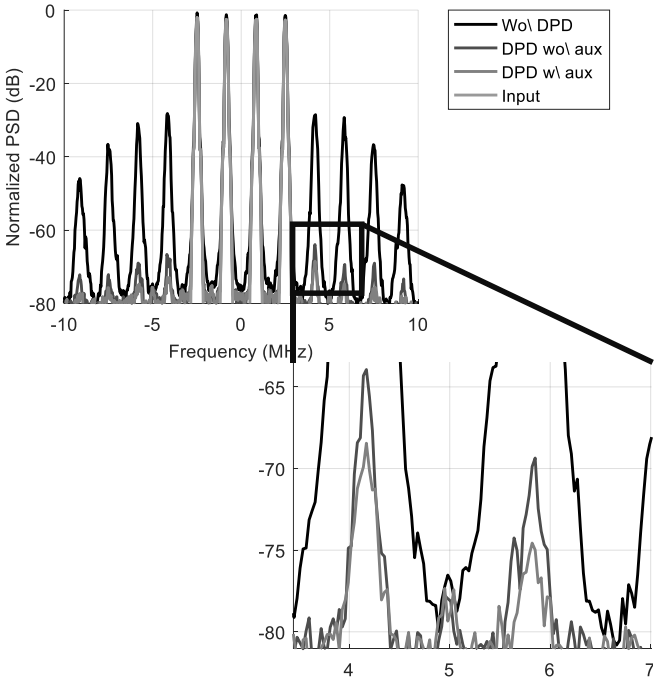


Fig. 11. Example of the signal spectra at the output of the device, before and after DPD correction, both with and without the auxiliary models; as well as, a close up showing the improvement of the output spectra when the auxiliary model is used.

TABLE III
LINEARIZATION RESULTS FOR THE GAN TRANSISTOR 3

Frequency (@35 dBm)	NMSE / IMR (dB)*			Power (@2.2 GHz)	NMSE / IMR (dB)*		
	5 MHz	10 MHz	15 MHz		5 MHz	10 MHz	15 MHz
2.1 GHz	48.1 / 59.1 (43.0 / 55.3)	46.0 / 59.8 (42.5 / 55.6)	46.3 / 60.3 (42.6 / 57.2)	32 dBm	52.9 / 68.0 (50.0 / 66.7)	49.6 / 66.4 (48.3 / 63.4)	48.2 / 66.0 (49.6 / 68.1)
2.2 GHz	52.5 / 66.2 (48.8 / 64.0)	49.4 / 64.9 (47.2 / 62.3)	49.4 / 65.5 (47.6 / 63.0)	33 dBm	53.0 / 67.9 (49.9 / 65.7)	49.7 / 65.5 (47.8 / 63.2)	49.7 / 67.4 (48.0 / 65.0)
2.3 GHz	49.2 / 62.5 (45.0 / 54.9)	46.6 / 61.8 (43.8 / 54.9)	46.7 / 61.7 (43.9 / 54.9)	34 dBm	52.6 / 66.2 (49.6 / 64.9)	49.6 / 65.6 (47.8 / 62.7)	49.7 / 66.8 (47.8 / 64.5)
				35 dBm	52.5 / 66.5 (49.0 / 64.3)	49.2 / 66.0 (47.5 / 62.5)	49.3 / 66.4 (47.9 / 63.5)

*The results within brackets are from the model without long term compensation.

TABLE IV
LINEARIZATION RESULTS FOR THE GAN TRANSISTOR 4

Frequency (@35 dBm)	NMSE / IMR (dB)*			Power (@2.2 GHz)	NMSE / IMR (dB)*		
	5 MHz	10 MHz	15 MHz		5 MHz	10 MHz	15 MHz
1.8 GHz	51.4 / 64.1 (45.7 / 60.4)	49.4 / 64.5 (44.8 / 59.0)	49.4 / 65.1 (45.2 / 61.3)	32 dBm	51.7 / 65.9 (46.7 / 61.3)	48.8 / 65.9 (45.6 / 59.8)	48.7 / 66.1 (45.4 / 62.7)
2.0 GHz	51.1 / 63.3 (44.7 / 59.6)	49.4 / 64.9 (48.2 / 60.8)	48.8 / 65.6 (44.0 / 59.6)	33 dBm	51.8 / 65.5 (46.4 / 61.6)	48.9 / 65.4 (45.2 / 60.0)	48.7 / 65.7 (45.0 / 62.2)
2.2 GHz	51.3 / 64.2 (45.5 / 60.0)	48.2 / 64.3 (44.6 / 58.9)	48.3 / 66.0 (44.5 / 61.4)	34 dBm	51.4 / 64.3 (46.1 / 61.1)	48.9 / 65.2 (44.9 / 59.7)	48.9 / 65.8 (44.8 / 62.0)
2.4 GHz	49.8 / 64.1 (44.1 / 59.8)	47.4 / 64.6 (43.4 / 58.7)	47.7 / 64.0 (43.7 / 59.8)	35 dBm	51.3 / 64.2 (45.8 / 60.3)	48.5 / 64.4 (44.6 / 58.6)	48.6 / 65.5 (44.6 / 61.2)

*The results within brackets are from the model without long term compensation.

lowest in-band carrier power and the highest out-of-band intermodulation distortion (IMD) power. The linearity specification for a multicarrier GSM signal is 70 dBc to the noise floor and 60 dBc where an intermodulation product is expected [20].

In each table the results are presented for several center frequencies, on the left, and for several input peak powers on the right. The frequencies were selected around the design frequency depending on the operation bandwidth of the PA. The maximum input power was selected to take the device into a gain compression of 2 dB to 3 dB. Each PA is tested from 3 dB backoff up to the maximum input power.

The results are presented as 2 numbers, the NMSE, on the left, in dB, and the IMR, on the right, also in dB, for each test case. For comparison the same results are shown below and in brackets for the model without the long-term variation.

Remarkable improvements in IMR figures from some 3 dB to up to 10 dB in IMR clearly attest the practical usefulness of the proposed linearization method.

Fig. 9 shows an example of the original and linearized AM/AM and AM/PM with and without this long-term model and Fig. 10 shows an example of the residuals after linearization with and without the long-term modelling. This example is taken from the 5 MHz case at full power and at the center design frequency. Finally, Fig. 11 shows the spectra, before and after DPD correction with and without long-term compensation.

V. CONCLUSION

In this paper, we showed how to solve the DPD linearization problem of GaN HEMT based PAs exhibiting long-term trapping and thermal phenomena. For this reason, a newly introduced residual signal enabled the extraction of an appropriate physically-based long-term memory description that was then used to generate the slowly-varying control variables of an otherwise common DPD model. With that, remarkable linearity improvements of up to 10 dB in IMR and by 5 dB in NMSE were demonstrated.

ACKNOWLEDGMENT

The authors would like to thank the Huawei Sweden PA Research and Development Team, namely Dr. Francesc Purroy and Mr. Richard Hellberg, for interesting technical discussions regarding some of the observations dealt with in this work.

REFERENCES

[1] A. Katz, J. Wood, and D. Chokola, "The Evolution of PA Linearization: From Classic Feedforward and Feedback Through Analog and Digital Predistortion," *IEEE Microwave Magazine*, vol. 17, no. 2, pp. 32-40, Feb. 2016.

[2] J. Kim and K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *Electronics Letters*, vol. 37, no. 23, pp. 1417-1418, Nov. 2001

[3] D. Morgan, Z. Ma, J. Kim, M. Zierdt, and J. Pastalan, "A Generalized Memory Polynomial Model for Digital Predistortion of RF Power Amplifiers," *IEEE Trans. on Signal Processing*, vol. 54, no. 10, pp. 3852-3860, Oct. 2006.

[4] A. Zhu, J. Pedro, and T. Brazil, "Dynamic Deviation Reduction-Based Volterra Behavioral Modeling of RF Power Amplifiers," *IEEE Trans. on*

Microwave Theory and Techn., vol. 54, no. 12, pp. 4323-4332, Dec. 2006.

[5] L. Guan and A. Zhu, "Simplified dynamic deviation reduction-based Volterra model for Doherty power amplifiers," *2011 Workshop on Integrated Nonlinear Microwave and Millimetre-Wave Circuits (INMMIC)*, Vienna, Austria, pp.1-4, Apr. 2011

[6] J. C. Pedro, P. M. Cabral, T. R. Cunha, and P. M. Lavrador, "A Multiple Time-Scale Power Amplifier Behavioral Model for Linearity and Efficiency Calculations," *IEEE Trans. on Microwave Theory and Techn.*, vol. 61, no. 1, pp. 606-615, Jan. 2013.

[7] G. Mouginot, R. Sommet, R. Quéré, Z. Ouarch, S. Heckmann, and M. Camiade, "Thermal and trapping phenomena assessment on AlGaN/GaN microwave power transistor," *5th European Microwave Integrated Circuits Conference*, Paris, Sep. 2010, pp. 110-113.

[8] S. Afsardoost, T. Eriksson, and C. Fager, "Digital Predistortion Using a Vector-Switched Model," *IEEE Trans. on Microwave Theory and Techn.*, vol. 60, no. 4, pp. 1166-1174, Apr. 2012.

[9] A. S. Tehrani, T. Eriksson, and C. Fager, "Modeling of long term memory effects in RF power amplifiers with dynamic parameters," *2012 IEEE MTT-S International Microwave Symposium Digest*, Montreal, QC, Canada, pp. 1-3, Jun. 2012.

[10] Y. Guo, C. Yu, and A. Zhu, "Power Adaptive Digital Predistortion for Wideband RF Power Amplifiers With Dynamic Power Transmission," *IEEE Trans. on Microwave Theory and Techn.*, vol. 63, no. 11, pp. 3595-3607, Nov. 2015.

[11] "*Operation and Performance of the ISL5239 Pre-distortion Linearizer*", Intersil, Appl. Note 1022, Jul. 2002.

[12] P. Aaen, J. A. Plá, and J. Wood, "Thermal Characterization and Modeling," in *Modeling and Characterization of RF and Microwave Power FETs*, Cambridge: Cambridge University Press, 2007

[13] J. B. King and T. J. Brazil, "Nonlinear Electrothermal GaN HEMT Model Applied to High-Efficiency Power Amplifier Design," *IEEE Trans.. on Microwave Theory and Techn.*, vol. 61, no. 1, pp. 444-454, Jan. 2013.

[14] J. C. Pedro and N. B. Carvalho, "Highly Linear Circuit Design," in *Intermodulation Distortion in Microwave and Wireless Circuit*, Norwood: Artech House, 2002

[15] A. V. Oppenheim and R. W. Schafer, "Filter Design Techniques," in *Discrete-Time Signal Processing*, New Jersey: Pearson, 2010.

[16] N. Ramanan, B. Lee, and V. Misra, "Device Modeling for Understanding AlGaN/GaN HEMT Gate-Lag," *IEEE Trans. On Electron Devices*, vol. 61, no. 6, pp. 2012-2018, Jun. 2014.

[17] R. Vetury, N. Q. Zhang, S. Keller, and U. K. Mishra, "The impact of surface states on the DC and RF characteristics of AlGaN/GaN HFETs," *IEEE Trans. On Electron Devices*, vol. 48, no. 3, pp. 560-566, Mar. 2001.

[18] O. Jardel *et al*., "An Electrothermal Model for AlGaN/GaN Power HEMTs Including Trapping Effects to Improve Large-Signal Simulation Results on High VSWR," *IEEE Trans. on Microwave Theory and Techn.*, vol. 55, no. 12, pp. 2660-2669, Dec. 2007.

[19] Changsoo Eun and E. J. Powers, "A new Volterra predistorter based on the indirect learning architecture," *IEEE Trans. on Signal Processing*, vol. 45, no. 1, pp. 223-227, Jan. 1997.

[20] European Telecommunications Standards Institute, "Digital cellular telecommunications system (Phase 2+) (GSM); Radio transmission and reception", *European Telecommunications Standards Institute*, ETSI TS 145 005, Aug. 2016. [Online] Available: http://www.etsi.org/deliver/etsi_ts/145000_145099/145005/13.01.00_60/ts_145005v130100p.pdf [Accessed: Dec. 2016].

Paper C7: Compensation of Long-Term Memory Effects on GaN HEMT Based Power Amplifiers

Filipe M. Barradas, Pedro M. Lavrador, Telmo R. Cunha, and José C. Pedro

2017 IEEE Topical Conference on Power Amplifiers for Wireless and Radio Applications (PAWR)

©2017 IEEE

Using Statistical Information for Fast Static DPD of RF PAs

Filipe M. Barradas, Pedro M. Lavrador, Telmo R. Cunha and José C. Pedro

DETI, Instituto de Telecomunicações, Universidade de Aveiro, Aveiro, Portugal,
filipebarradas@ua.pt

***Abstract* — Modern radio-frequency transmitters require digital predistortion to fulfil linearity requirements while operating the RF PA in highly efficient modes. A least-squares process is typically employed to calculate the predistorter coefficients. The least-squares method requires complex mathematical operations and relies on the observation, at the input and output of the PA, of a specific realization of a telecommunication signal. With the increase of signal bandwidths faster DPD methods are required. In this paper, we use statistical measures to create an inverse of the PA. The proposed method avoids the least-squares formulation potentiating a much faster DPD extraction solution.**

***Index Terms* — Power Amplifiers, Digital Predistortion, Probability Density Function.**

I. Introduction

Digital predistortion has been adopted as the main compensation method for nonlinear behavior in RF transmitters.

In recent years, DPD is mainly based in truncated Volterra series models, which use linear regression to obtain the coefficients. This process is many times solved using least-squares, which involves complex calculations that are not easily implemented in hardware. Furthermore, these complex platforms pose limitations on DPD viability, because of their power consumption.

With the advent of new generation communication standards using a high number of low power small cells, cheap DPD solutions can be interesting. In this sense, the current methods are not alluring to this new communication paradigm. An alternative to these methods is using statistical measures to obtain inverses of the PA's characteristics. This was already achieved for the AM/AM in the digital signal processing field [1]. However, not inverting the AM/PM characteristic prevents the use of this technique in the DPD linearization context.

In this paper we revisit AM/AM linearization using a histogram based inversion and complement it by presenting a method for the AM/PM inversion. The calculations involved in these methods are straightforward and avoid the use of least-squares solutions.

Furthermore, the solutions achieved by this method are tuned for a set of statistical quantities, instead of specific realizations of the input signals. We believe that this is more suited to typical communication scenarios where the input signal will vary but maintain its statistics.

II. Method Overview

For a static nonlinear device, the relationship between the input and output in the digital low pass equivalent domain (LPE) is given by (1), where $F(|x(n)|)$ is the AM/AM conversion of the amplifier and $\psi(|x(n)|)$ is the AM/PM conversion.

$$\begin{aligned} x(n) &= |x(n)|e^{j\phi(n)} \\ y(n) &= F(|x(n)|)e^{j\phi(n)+j\psi(|x(n)|)} \end{aligned} \tag{1}$$

The statistical inverse method is divided in two parts. Initially, the inverse AM/AM is found. For this purpose, the method presented in [1] is followed. This method consists of measuring the input and output amplitude probability density functions (p.d.f.s), $p_{\mathrm{x}}(|x|)$ and $p_{\mathrm{y}}(|y|)$, and relate them using (2).

$$\begin{aligned} &P_{\mathrm{x}}(|x|) = \int_0^{|x|} p_{\mathrm{x}}(w)dw \quad P_{\mathrm{y}}(|y|) = \int_0^{|y|} p_{\mathrm{y}}(w)dw \\ &\forall_{|x|,|y|}: P_{\mathrm{x}}(|x|) = P_{\mathrm{y}}(|y|) \Rightarrow |x| = F^{-1}(|y|) \end{aligned} \tag{2}$$

In (2), $p_{\mathrm{x}}(|x|)$ is the input amplitude p.d.f., $p_{\mathrm{y}}(|y|)$ is the output amplitude p.d.f. and $P_{\mathrm{x}}(|x|)$, $P_{\mathrm{y}}(|y|)$ are their respective cumulative density functions (c.d.f.s). The c.d.f.s represent the probability of an amplitude being smaller or equal to some value, for instance, $P_{\mathrm{x}}(|x|)$ is the probability of $|x|$ being smaller or equal to x.

The equation that supports inversion through statistical measures is the c.d.f. equality shown in (2). Equation (2) is valid for static, monotonically increasing, relations between the variables $|x|$ and $|y|$. It can be intuitively understood with the following explanation: defining a threshold in x where all values are below that threshold with some probability P_{x}; is equivalent to defining a threshold in $|y| = F(|x|)$ where all values are below that threshold with the same probability. This is true because the relationship between variables is monotonically increasing, meaning that there is no value above the threshold in the input (x) that will increase the probability of the values below the threshold in the output (y).

Using (2) an amplitude mapping from input to output and from output to input can be found, since the c.d.f.s are also monotonically increasing. The latter can be used as the inverse AM/AM of the amplifier, a process illustrated in Fig. 2.

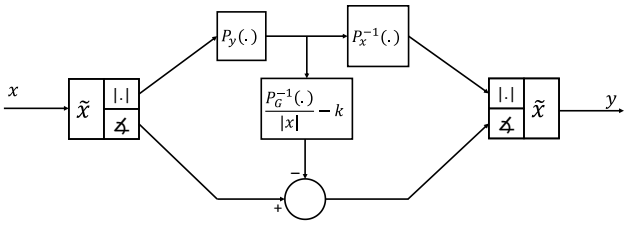


Fig. 1: Functional diagram of the statistical based DPD.

To compensate the phase-shift of the amplifier a measure of $\psi(|x(n)|)$ is required. The main problem, in this case, is that it is impossible to observe $\psi(|x(n)|)$ directly. This is because the AM/PM generation in the amplifier is perturbated by the original phase modulation of the signal, $\phi(t)$.

Some methods can be found in the literature using higher-order statistical moments to identify phase rotations [2]. These methods can be used when the modulation standards have well defined phase points (low order QAM, M-PSK, etc.), but tend to fail or give poor results when there are no reference points (high order QAM, OFDM, etc.).

The main problem is that the phase distribution in some signals is very close to being uniform, and rotations of this phase distribution yields the same distribution, meaning that the AM/PM has no statistical impact.

In order to obtain a good quality measure of the AM/PM we eliminate the phase modulation using the input information (this requires alignment of the input and output signals) and obtain $\psi(|x(n)|)$ directly.

Since the AM/PM is a function of the amplitude, similarly to the AM/AM, it is in our interest to apply the formalism from before to obtain an inverse. In order to do this, two problems must be solved. First, the AM/PM is not (or may not be) monotonically increasing. Second, small amounts of noise will cause high AM/PM variation for low amplitudes. To solve these problems with the AM/PM and use the same formalism as before, we instead use (3), where k is chosen sufficiently high to guarantee monotonicity.

$$G(|x|) = (\psi(|x(n)|) + k)|x| \tag{3}$$

For k to guarantee a strictly monotonous behavior of the function in (3), it must be higher than the highest negative slope of the AM/PM, so that (4) is verified.

$$\frac{dG(|x|)}{d|x|} = \frac{d\psi(|x(n)|)}{d|x|} + k > 0 \tag{4}$$

Equation (3) both resolves the dispersion at low amplitudes and the monotonicity problems observed in measured AM/PM characteristics. While k controls the slope of the function (allowing for a forced monotonous curve, (4)), for small amplitudes, the multiplication by the amplitude will limit the dispersion. The variable k should be set to the smallest possible value, since high values of k will hide the variation of $G(|x|)$ due to the AM/PM, desensitizing the extraction process.

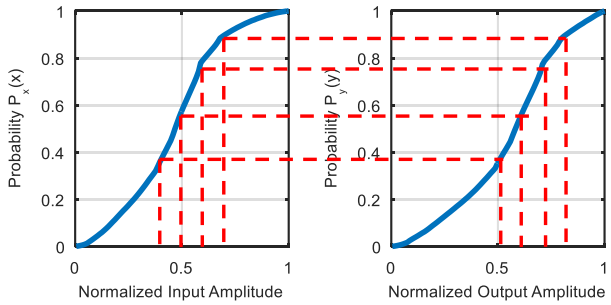


Fig. 2: Mapping between input (on the left) and output (on the right) amplitudes using c.d.f..

Similarly as done to the amplitude, the p.d.f. of (3) can be measured and the c.d.f. calculated. Since (3) was made monotonic a relationship similar to (2) is established. Using this relationship the predistorter in Fig. 1, which is only dependent on the several calculated c.d.f.s, can be built.

III. Details on the Statistical Model

The several required p.d.f.s are approximated by measuring the number of points that fall within a certain range, as seen in (5) where N is the total number of points and N_o is the number of points in the specified interval. In this sense, the specified number of bins for measuring the p.d.f.s gives the number of model coefficients.

$$p_{x_k}(x) \cong \frac{n_k}{N}$$
$$n_k = N_o\left\{\frac{x_{k-1} + x_k}{2} \le x_i < \frac{x_k + x_{k+1}}{2}\right\} \tag{5}$$

Similarly to least squares problems, this modelling approach can also suffer from ill-conditioning, for these models it is easily detected, if the resolution for the p.d.f.s is very fine, some bins will have no points, this will break the strict monotonicity of the c.d.f.s and cause them to become non-invertible. At this point the problem is ill-conditioned and the number of bins, coefficients, should be reduced.

Since the c.d.f.s will be approximated in a discretized fashion, an interpolation between each point is required. In our work, we have used a linear interpolation, other types of interpolation can be used to smooth the c.d.f. curve.

IV. Results

In order to test the proposed method and access its performance, a typical DPD test setup was used, as shown in Fig. 3.

A 20W RF PA based on the CGH35015 GaN transistor from Cree, operating at 900 MHz, was used as the nonlinear device in this experiment. The AM/AM and

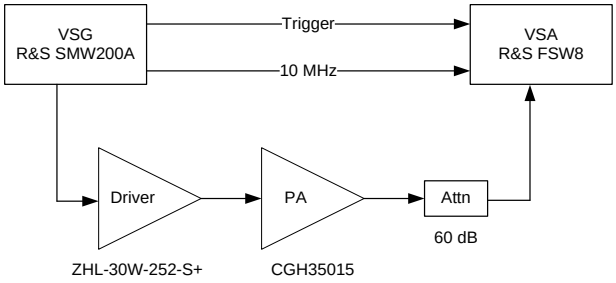


Fig. 3: Functional diagram of the measurement setup.

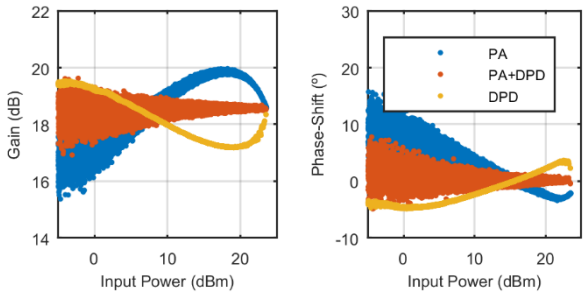


Fig. 4: AM/AM and AM/PM of the CGH35015 PA.

AM/PM of the PA, for a 5MHz OFDM signal, before and after linearization, are shown in Fig. 4.

An OFDM signal of 5, 10 and 20 MHz was used as the input excitation, for each test case scenario the predistorter is trained for 10 iterations, always changing the input signal. The resulting model is then tested with yet another different realization. Each realization has the same peak power, to simplify the application of the DPD. The results, in terms of ACPR and NMSE, are shown in table I. The spectra for the several test cases before and after linearization are shown in Fig. 5. The ACPR is taken as the minimum between the left and right ACPR.

TABLE I

Bandwidth	NMSE	ACPR
5 MHz	-41.2 dB	50.6 dB
10 MHz	-40.3 dB	47.4 dB
20 MHz	-36.8 dB	44.0 dB

V. Conclusions

We have revisited a method for AM/AM extraction and presented a method for AM/PM extraction using statistical measurements.

These methods are thought to give cheap, low-level, predistortion suitable for applications where high-level predistortion is too power hungry or too costly to be justifiable.

Despite providing only static compensation, the presented methods reached an NMSE of -37 dB and an ACPR of 44 dB, in a single-ended 20W GaN PA under 20MHz OFDM excitation.

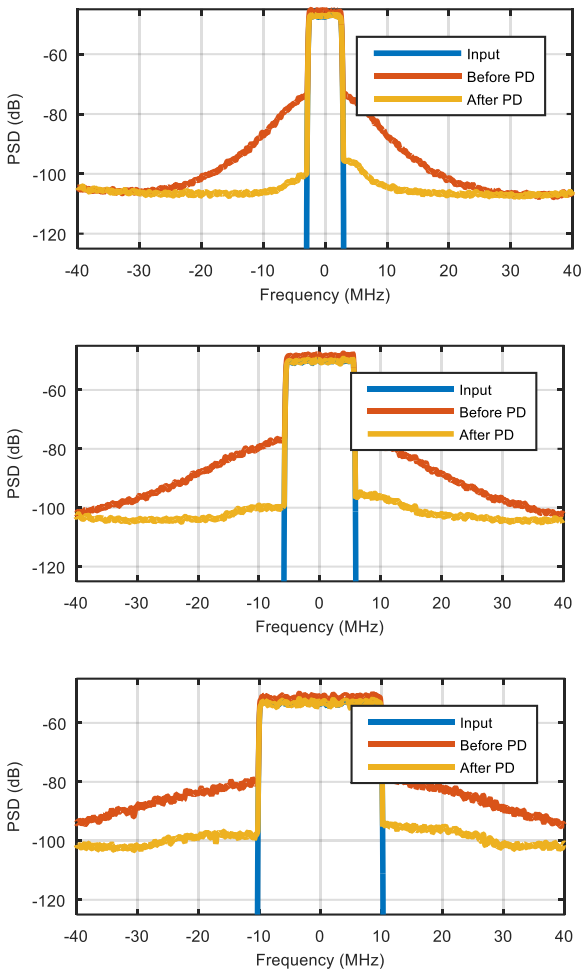


Fig. 5: Spectrum before and after application of the statistical predistorter, for the 5, 10 and 20 MHz case.

Acknowledgement

This work is funded by National Portuguese Funds through FCT - Fundação para a Ciência e a Tecnologia under the project UID/EEA/50008/2013 and by the PhD grant given to the first author (Ref. SFRH/BD/90103/2012).

References

[1] D. Huang, X. Huang and H. Leung, "Nonlinear Compensation of High Power Amplifier Distortion for Communication Using a Histogram-Based Method," in *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4343-4351, Nov. 2006.

[2] Z. Zhu, X. Huang, M. Caron and H. Leung, "A Blind AM/PM Estimation Method for Power Amplifier Linearization," in *IEEE Signal Processing Letters*, vol. 20, no. 11, pp. 1042-1045, Nov. 2013.

6. Conclusions and Future Work

6.1. Conclusions

The main motivation for this work was the development of novel techniques to understand the behavior of wireless RF transmitters, mainly the RFPA, for the information envelope of the excitation signals. The study was to be anchored in an understanding of the physics of the devices and supported on this knowledge.

Initially, in Chapter 2, a study of the PAs behavior and the relationships found to these characteristics in typical Volterra based models are described. This study is performed using a system level quasi-equivalent model of the RFPA and allows the attribution of specific effects in the PA to specific components of the digital models. Therefore, this initial study can be used to support the choice of the DPD or behavioral model to describe or compensate the RFPA.

While Chapter 2 supports the model topology choice, Chapter 3 is dedicated to understanding the limitations on the identification of such a model. Chapter 3 is then mostly dedicated to the signal processing aspects of the Volterra based models. In this aspect, this work has contributed to the state-of-the-art by: first, showing how simple transformations of the original model can lead to severely improved conditioning of the extraction procedure; second, showing that the polynomials are not disadvantaged over the LUTs from a conditioning point of view, when proper transformations are used.

In Chapters 2 and 3, support was built for the usability of models in a physics oriented approach while maintaining a solid signal processing friendly structure. Chapter 4 dedicates itself to applicability of these concepts, showing how these concepts can be applied for RFPA characterization and memory type differentiation. In this sense, Chapter 4 has contributed to the state-of-the-art by providing initial steps into the disentanglement of the different memory types of the RFPA and in fast characterization methods with physically meaningful measurements, the IMDs.

While Chapter 4 is dedicated to the characterization of the RFPA, Chapter 5 is dedicated to the compensation. The most challenging aspect of RFPAs that still required compensation were the long-term memory effects. These long-term effects,

induced by electron trapping/detrapping and thermal phenomena, are not compensated by current state-of-the-art linearizers, being this a strong barrier in the linearization of GaN PAs in highly demanding applications, such as those handling multicarrier GSM and time slotted signals. This work has proposed a method that provides excellent results in the successful linearization of these effects. This is evidenced by the large improvement of the measurement metrics, sometimes on the order of 10 dB (over standard linearization techniques) when the proposed methodology was applied. Again, to achieve these results, the methodology is heavily inspired on the physics of the devices that are being excited. Chapter 5 then proceeds to propose a simpler predistortion type for future transmitters, based on simpler correction models and using simpler measurements. This correction methodology is to be applied when the requirement is a low power and cheap system. Chapter 5 therefore extends the state-of-the-art in two aspects. First, it provides the first high-accuracy linearization of RFPAs afflicted with long-term memory; showing furthermore an atypical methodology based on the equivalent circuit models for the effects responsible for the underlying effects. Second, it proposes a change of paradigm for future communication systems, if a massive PA deployment approach is taken, where the focus is on cheap and simple predistortion. This work proposes a statistical based approach where the required measurements can be limited to filing the statistics of the input and output signals.

6.2. Future Work

This thesis was dedicated to two different aspects of LPE model usage, characterization and compensation.

In terms of characterization, an interesting theme is continuing the large signal characterization methods where the several memory types in the PA are disentangled. This allows a number of interesting knowledge and applications to be developed. First, the separate description and direct relationship between the model components and the physics of the PA allows diagnosis of the PA. Second, this characterization obtains a reasonable model for the RFPA in a large bandwidth that is useful for system level simulation under diverse conditions. Third, this characterization procedure can help in deciding where to invest compensation resources.

The current work has shown an initial step to diagnosing the several memory types of the RFPA. Continuing this work might reveal even more insight into the workings of the RFPA. Furthermore, the development of these types of models goes in the direction of building an analog to the S-parameters for nonlinear systems. Effectively, this is initially specifically developed with PAs in mind and only the PA transmission has been studied. However, as future work, the complete exploration of this characterization methodology could include measuring input reflection, isolation and output reflection. Another important step is moving the characterization into the frequency domain, using phase synchronous measurement equipment, like the VNA, for instance.

The current work has initially shown that the characterization procedure generates reasonable, broadband models of the measured system. It is further shown that the mapping of the measures to the model is very simple. Studying how these models do in system level simulations and if they can be used for reasonably predicting the behavior of a full system where the RFPA has been introduced would be interesting.

Since DPD is a mature technology with a number of techniques at its disposition, using the characterization procedure to obtain insight into the required compensation has less significance. Nonetheless, the characterization procedure gives information on how the PA responds in terms of the different dynamics and nonlinearity. This information can be useful to decide if cross memory is required in the compensation, and for which bandwidth is it required, as well as how aggressive is the nonlinearity.

In terms of DPD, classic methods for nonlinear compensation of RFPAs have already reached a very mature state from the model development point of view. In fact, the most difficult effects that had escaped compensation were the long-term memory

effects, which were compensated with the technique proposed in this work. In this sense, further pursuits in this theme have reduced interest. In terms of compensation, the most interesting investigation is in DPD for future applications. For this, this work presented initial work on development of a new DPD paradigm which could lead to interesting results and applicable solutions for future transmitters.

Evidently, there are many approaches for the future of nonlinear compensation in new generation transmitters; the continuation of digital compensation as a competitive technique is far from guaranteed. Nonetheless, the predicted reduction of the power budget for the predistorter will force the transition into simpler, cheap methods that achieve less accurate predistortion. The use of these limited predistorters may force a change in paradigm also at the training level, where a compromise between the PA efficiency, or the efficiency of the whole arrangement, should be found under the limitations of the compensator. Overall, development of work for the digital compensation in next generation transmitters looks promising.

A. Annex

a.1. Transistor Equations

β	8.08
V_T	-3.4
V_{ST}	0.1
Δ	0.004
V_L	0.09
V_K	8.7
λ	0.001
α	0.28
α_{lim}	1.49
α_{soft}	0.27
p_{lin}	0.97

$$v_{gs_1} = v_{gs} - V_T$$

$$v_{gs_2} = v_{gs_1} - \frac{1}{2}\left(v_{gs_1} + \sqrt{\left(v_{gs_1} - V_K\right)^2 + \Delta^2} - \sqrt{(V_K)^2 + \Delta^2}\right)$$

$$v_{gs_3} = V_{ST} \log\left(1 + exp\left(\frac{v_{gs_3}}{V_{ST}}\right)\right)$$

$$v_{ds_1} = 1 + \lambda v_{ds}$$

$$v_{ds_2} = 1 - \frac{\alpha_{lim}}{\alpha_{lim} + \frac{\sinh(\alpha v_{ds})}{\alpha_{soft}}} \tanh(20 v_{vds})$$

$$i_{ds} = \beta \left(\frac{\left(v_{gs_3}\right)^2}{1 + \frac{\left(v_{gs_3}\right)^{p_{lin}}}{V_L}} \right) v_{ds_1} v_{ds_2}$$

a.2. Model Formulation Conversion

Given the memory polynomial model in (a.1) and a set of basis functions ($g_k(.)$) that can represent the space of the polynomials of order $2P$, (a.2) can be written.

$$\tilde{y}(n) = \sum_{p=0}^{P-1} \sum_{m=0}^{M-1} \tilde{h}_{2p+1}(m) \tilde{x}(n-m) |\tilde{x}(n-m)|^{2p} \tag{a.1}$$

$$\tilde{y}(n) = \sum_{p=0}^{P-1} \sum_{m=0}^{M-1} \tilde{h}_{2p+1}(m) \tilde{x}(n-m) \sum_{k=0}^{K-1} c_{k,p} g_k(|\tilde{x}(n-m)|) \tag{a.2}$$

where $c_{k,p}$ is the set of coefficients such that the condition (a.3) is fulfilled.

$$|\tilde{x}(n-m)|^{2p} = \sum_{k=0}^{K-1} c_{k,p} g_k(|\tilde{x}(n-m)|) \tag{a.3}$$

Under this condition, the model can be equivalently written as shown in (a.4).

$$\tilde{y}(n) = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} \sum_{p=0}^{P-1} c_{k,p} \tilde{h}_{2p+1}(m) \tilde{x}(n-m) g_k(|\tilde{x}(n-m)|) \tag{a.4}$$

Resolving the summation for p a new set of parameters is obtained, note that only the parameters are dependent on p. Equation (a.5) can then be written.

$$\begin{aligned} \tilde{y}(n) &= \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} \widetilde{h'}_k(m) \tilde{x}(n-m) g_k(|\tilde{x}(n-m)|) \\ \widetilde{h'}_k(m) &= \sum_{p=0}^{P-1} c_{k,p} \tilde{h}_{2p+1}(m) \end{aligned} \tag{a.5}$$

Equation (a.5) maps exactly to Fig. 2.16, where each filter has been described in the frequency domain equivalent.

Using more relaxed constraints, even if the chosen set of basis functions does not exactly describe the polynomial space, but can be used as an approximate basis set of the function space, then it can be used to represent the memory polynomial model, since it fulfils the same role as the polynomials.

7. References

[1] S. Roy, "Energy Logic for Telecommunications," Emerson Network Energy Systems, 2008.

[2] B. Debaillie and C. Desset, "Power Modeling of Base Stations," 5GrEEn Summerschool, Stockholm, 2014.

[3] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P. M. Grant, H. Haas, J. S. Thompson, I. Ku, C.-X. Wang, T. A. Le, M. R. Nakhai, J. Zhang and L. Hanzo, "Green radio: radio techniques to enable energy-efficient wireless networks," *IEEE Communications Magazine,* vol. 49, no. 6, pp. 46-54, 2011.

[4] P. M. Lavrador, T. R. Cunha, P. M. Cabral and J. C. Pedro, "The Linearity-Efficiency Compromise," *IEEE Microwave Magazine,* vol. 11, no. 5, pp. 44-58, 2010.

[5] P. B. Kenington, High-Linearity RF Amplifier Design, Norwood: Artech House, 2000.

[6] S. C. Cripps, Power Amplifier Linearization Techniques, Norwood: Artech House, 2006.

[7] J. C. Pedro, "Técnicas de Linearização de Amplificadores de Potência em Microondas," Universidade de Aveiro, Aveiro, 1993.

[8] M. Schetzen, The Volterra and Wiener Theories of Nonlinear Systems, John Wiley & Sons, 1980.

[9] V. J. Mathews and G. L. Sicuranza, Polynomial Signal Processing, John Wiley & Sons, 2000.

[10] J. Kim and K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *Electronics Letters,* vol. 37, no. 23, pp. 1417-1418, 8 November 2001.

[11] D. R. Morgan, Z. Ma, J. Kim, M. G. Zierdt and J. Pastalan, "A Generalized Memory Polynomial Model for Digital Predistortion of RF Power Amplifiers," *IEEE*

Transactions on Signal Processing, vol. 54, no. 10, pp. 3852-3860, October 2006.

[12] O. Hammi, F. M. Ghannouchi and B. Vassilakis, "A Compact Envelope-Memory Polynomial for RF Transmitters Modeling With Application to Baseband and RF-Digital Predistortion," *IEEE Microwave and Wireless Components Letters,* vol. 18, no. 5, pp. 359-361, 2008.

[13] D. Mirri, G. Luculano, F. Filicori, G. Pasini, G. Vannini and G. P. Gabriella, "A modified Volterra series approach for nonlinear dynamic systems modeling," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications,* vol. 49, no. 8, pp. 1118-1128, 2002.

[14] A. Zhu, J. Pedro and T. Brazil, "Dynamic Deviation Reduction-Based Volterra Behavioral Modeling of RF Power Amplifiers," *IEEE Transactions on Microwave Theory and Techniques,* vol. 54, no. 12, pp. 4323-4332, December 2006.

[15] L. Guan and A. Zhu, "Simplified dynamic deviation reduction-based Volterra model for Doherty power amplifiers," in *2011 Workshop on Integrated Nonlinear Microwave and Millimetre-Wave Circuits (INMMIC)*, Vienna, 2011.

[16] C. Crespo-Cadenas, J. Reina-Tosina and M. J. Madero-Ayora, "Volterra behavioral model for wideband RF amplifiers," *IEEE Transactions on Microwave Theory and Technology,* vol. 55, no. 3, p. 449–457, 2007.

[17] C. Crespo-Cadenas, J. Reina-Tosina, M. J. Madero-Ayora and J. Munoz-Cruzado, "A New Approach to Pruning Volterra Models for Power Amplifiers," *IEEE Transactions on Signal Processing,* vol. 58, no. 4, pp. 2113-2120, 2010.

[18] G. Montoro, P. L. Gilabert, E. Bertran, A. Cesari and D. D. Silveira, "A New Digital Predictive Predistorter for Behavioral Power Amplifier Linearization," *IEEE Microwave Wireless Components Letters,* vol. 17, no. 6, p. 448–450, 2007.

[19] X. Wu, J. Shi and H. Chen, "On the numerical stability of RF power amplifier's digital predistortion," in *15th Asia-Pacific Conference on Communications*, Shanghai, 2009.

[20] C. Yu, Y. Liu and S. Li, "Triangular Memory Polynomial Predistorter," in *5th International Conference On Wireless Communications, Networking and Mobile Computing*, Beijing, 2009.

[21] P. Landin, "Digital Baseband Modeling and Correction of Radio Frequency Power Amplifiers," KTH Royal Institute of Technology, Stockholm, 2012.

[22] M. Herman, B. Miller and J. Goodman, "The cube coefficient subspace architecture for nonlinear digital predistortion," in *42nd Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, 2008.

[23] A. Zhu and T. J. Brazil, "Behavioral modeling of RF power amplifiers based on pruned volterra series," *IEEE Microwave and Wireless Components Letters,* vol. 14, no. 12, pp. 563-565, 2004.

[24] C. Eun and E. J. Powers, "A new Volterra predistorter based on the indirect learning architecture," *IEEE Transactions on Signal Processing,* vol. 45, no. 1, pp. 223-227, 1997.

[25] R. N. Braithwaite, "Wide bandwidth adaptive digital predistortion of power amplifiers using reduced order memory correction," in *2008 IEEE MTT-S International Microwave Symposium Digest*, Atlanta, 2008.

[26] L. Guan, "High Performance Digital Predistortion for Wideband RF Power Amplifiers," University College Dublin, Dublin, 2012.

[27] C. Yu, L. Guan, E. Zhu and A. Zhu, "Band-limited volterra series-based digital predistortion for wideband RF power amplifiers," *IEEE Transactions on Microwave Theory and Techniques,* vol. 60, no. 12, pp. 4198-4208, 2012.

[28] A. Grebennikov, "High-Efficiency Doherty Amplifier Architectures," Workshop on Unconventional Power Amplifier Architecture with High Efficiency, 2012.

[29] J. C. Pedro, P. M. Cabral, T. R. Cunha and P. M. Lavrador, "A Multiple Time-Scale Power Amplifier Behavioral Model for Linearity and Efficiency Calculations," *IEEE Transactions on Microwave Theory and Techniques,* vol. 61, no. 1, pp. 606-615, 2013.

[30] Y. Guo, C. Yu and A. Zhu, "Power Adaptive Digital Predistortion for Wideband RF Power Amplifiers With Dynamic Power Transmission," *IEEE Transactions on Microwave Theory and Techniques,* vol. 63, no. 11, pp. 3595-3607, 2015.

[31] A. S. Tehrani, T. Eriksson and C. Fager, "Modeling of long term memory effects in RF power amplifiers with dynamic parameters," in *2012 IEEE MTT-S International Microwave Symposium Digest (MTT)*, Montreal, 2012.

[32] F. M. Barradas, T. R. Cunha, P. M. Lavrador and J. C. Pedro, "Higher locality non-linear basis functions of Volterra series based models to improve extraction conditioning," in *2014 IEEE MTT-S International Microwave Symposium (IMS2014)*, Tampa, 2014.

[33] R. N. Braithwaite, "Reducing estimator biases due to equalization errors in adaptive digital predistortion systems for RF power amplifiers," in *2012 IEEE/MTT-S International Microwave Symposium Digest*, Montreal, 2012.

[34] T. R. Cunha, F. M. Barradas and J. C. Pedro, "DPD tuning with frequency selective Distortion Minimization," in *2015 IEEE MTT-S International Microwave Symposium*

(IMS2015), Phoenix, 2015.

[35] D. E. Root, J. Verspecht, D. Sharrit, J. Wood and A. Cognata, "Broad-Band, Poly-Harmonic Distortion (PHD) Behavioral Models from Fast Automated Simulations and Large-Signal Vectorial Network Measurements," *IEEE Transactions on Microwave Theory and Techniques,* vol. 53, no. 11, pp. 3656-3664, 2005.

[36] J. C. Pedro and S. A. Maas, "A comparative overview of microwave and wireless power-amplifier behavioral modeling approaches," *IEEE Transactions on Microwave Theory and Techniques,* vol. 53, no. 4, pp. 1150-1163, 2005.

[37] N. Naraharisetti, P. Roblin, C. Quindroit, M. Rawat and S. Gheitanchi, "Quasi-exact inverse PA model for digital predistorter linearization," in *2013 82nd ARFTG Microwave Measurement Conference*, Columbus, 2013.

[38] W. Pan, Y. Liu and Y. Tang, "A Predistortion Algorithm Based on Accurately Solving the Reverse Function of Memory Polynomial Model," *IEEE Wireless Communications Letters,* vol. 1, no. 4, pp. 384-387, 2012.

[39] M. Younes, O. Hammi, A. Kwan and F. M. Ghannouchi, "An Accurate Complexity-Reduced “PLUME” Model for Behavioral Modeling and Digital Predistortion of RF Power Amplifiers," *IEEE Transactions on Industrial Electronics,* vol. 58, no. 4, pp. 1397-1405, 2011.

[40] F. Filicori, G. Vannini and V. A. Monaco, "A nonlinear integral model of electron devices for HB circuit analysis," *IEEE Transactions on Microwave Theory and Techniques,* vol. 40, no. 7, pp. 1456-1465, 1992.

[41] J. C. Pedro, N. B. Carvalho and P. M. Lavrador, "Modeling nonlinear behavior of band-pass memoryless and dynamic systems," in *2003 IEEE MTT-S International Microwave Symposium Digest*, Philadelphia, 2003.

[42] A. Zhu, J. C. Pedro and T. R. Cunha, "Pruning the Volterra Series for Behavioral Modeling of Power Amplifiers Using Physical Knowledge," *IEEE Transactions on Microwave Theory and Techniques,* vol. 55, no. 5, pp. 813-821, 2007.

[43] T. R. Cunha, E. G. Lima and J. C. Pedro, "Validation and Physical Interpretation of the Power-Amplifier Polar Volterra Model," *IEEE Transactions on Microwave Theory and Techniques,* vol. 52, no. 12, pp. 4012-4021, 2010.

[44] M. Isaksson and D. Ronnow, "A Kautz-Volterra Behavioral Model for RF Power Amplifiers," in *2006 IEEE MTT-S International Microwave Symposium Digest*, San Francisco, 2006.

[45] J. C. Pedro and L. C. Nunes, "A single-ended power amplifier behavioral model for AM/AM and AM/PM predictions," in *2014 IEEE Topical Conference on Power*

Amplifiers for Wireless and Radio Applications (PAWR), Newport Beach, 2014.

[46] J. C. Pedro and L. C. Nunes, "A single-ended power amplifier behavioral model for AM/AM and AM/PM predictions," in *2014 IEEE Topical Conference on Power Amplifiers for Wireless and Radio Applications (PAWR)*, Newport Beach, 2014.

[47] T. Cunha, J. Pedro, P. Cabral and A. Zhu, "General Nonlinear Feed-forward RF Model for Power Amplifiers," in *2007 IEEE/MTT-S International Microwave Symposium*, Honolulu, 2007.

[48] T. R. Cunha, J. C. Pedro and P. M. Cabral, "Design of a Power-Amplifier Feed-Forward RF Model With Physical Knowledge Considerations," *IEEE Transactions on Microwave Theory and Techniques,* vol. 55, no. 12, pp. 2747-2756, 2007.

[49] T. Cunha, J. Pedro and E. Lima, "Low-pass equivalent feedback topology for Power Amplifier modeling," in *2008 IEEE MTT-S International Microwave Symposium Digest*, Atlanta, 2008.

[50] J. C. Pedro and N. B. Carvalho, Intermodulation Distortion in Microwave and Wireless Circuits, Norwood: Artech House, 2002.

[51] L. C. Nunes, P. M. Cabral and J. C. Pedro, "AM/AM and AM/PM Distortion Generation Mechanisms in Si LDMOS and GaN HEMT Based RF Power Amplifiers," *IEEE Transactions on Microwave Theory and Techniques,* vol. 62, no. 4, pp. 799-809, April 2014.

[52] R. Darraji, F. M. Ghannouchi and O. Hammi, "A Dual-Input Digitally Driven Doherty Amplifier Architecture for Performance Enhancement of Doherty Transmitters," *IEEE Transactions on Microwave Theory and Techniques,* vol. 59, no. 5, pp. 1284-1293, 2011.

[53] C. J. Clark, C. P. Silva, A. A. Moulthrop and M. S. Muha, "Power-amplifier characterization using a two-tone measurement technique," *IEEE Transactions on Microwave Theory and Techniques,* vol. 50, no. 6, p. 1590–1602, 2002.

[54] D. Schreurs, M. O'Droma, A. A. Goacher and M. Gadringer, RF power amplifier behavioral modeling, Cambridge: Cambridge University Press, 2009.

[55] M. Franco, A. Guida, A. Katz and P. Herczfeld, "Minimization of Bias-Induced Memory Effects in UHF Radio Frequency High Power Amplifiers with Broadband Signals," in *2007 IEEE Radio and Wireless Symposium*, Long Beach, 2007.

[56] S. Rumery and B. Noori, "A new technique for measuring the resonant behavior of power amplifier bias circuits," in *2007 69th ARFTG Conference*, Honolulu, 2007 .

[57] S. Boumaiza and F. M. Ghannouchi, "Thermal memory effects modeling and

compensation in RF power amplifiers and predistortion linearizers," *IEEE Transactions on Microwave Theory and Techniques,* vol. 51, no. 12, pp. 2427-2433, 2003.

[58] A. Prejs, S. Wood, R. Pengelly and W. Pribble, "Thermal analysis and its application to high power GaN HEMT amplifiers," in *IEEE MTT-S International Microwave Symposium Digest, 2009*, Boston, 2009.

[59] S. S. Meena, C. Baylis, L. Dunleavy and M. Marbell, "Duty cycle dependent pulsed IV simulation and thermal time constant model fitting for LDMOS transistors," in *2009 74th ARFTG Microwave Measurement Conference*, Broomfield, 2009.

[60] M. Mahalingam, E. Mares, W. Brakensiek, K. Burger and C. L. Hsu, "High power microwave device temperature measurement - Methodology and applications for pulsed devices," in *2007 IEEE/MTT-S International Microwave Symposium*, Honolulu, 2007.

[61] J. C. Pedro, L. C. Nunes and P. M. Cabral, "Soft compression and the origins of nonlinear behavior of GaN HEMTs," in *2014 44th European Microwave Conference (EuMC)*, Rome, 2014.

[62] T. R. Cunha and J. C. Pedro, "Short and Long-Term Memory Modelling via Generic FIR Filtering," in *2006 International Workshop on Integrated Nonlinear Microwave and Millimeter-Wave Circuits*, Aveiro, 2006.

[63] J. B. King and T. J. Brazil, "Nonlinear Electrothermal GaN HEMT Model Applied to High-Efficiency Power Amplifier Design," *IEEE Transactions on Microwave Theory and Techniques,* vol. 61, no. 1, pp. 444-454, 2013.

[64] Y. Liu, W. Pan, S. Shao and Y. Tang, "A New Digital Predistortion for Wideband Power Amplifiers With Constrained Feedback Bandwidth," *IEEE Microwave and Wireless Components Letters,* vol. 23, no. 12, pp. 683-685, 2013.

[65] L. Ding, F. Mujica and Z. Yang, "Digital Predistortion using Direct Learning with Reduced Bandwidth Feedback," in *2013 IEEE MTT-S International Microwave Symposium Digest (IMS)*, Seattle, 2013.

[66] M. Schwartz, W. R. Bennet and S. Steyn, Communication Systems and Techniques, IEEE Press, 1995.

[67] D. Zhou and V. E. DeBrunner, "Novel Adaptive Nonlinear Predistorters Based on the Direct Learning Algorithm," *IEEE Transactions on Signal Processing,* vol. 55, no. 1, pp. 120-133, 2007.

[68] Y. H. Lim, Y. S. Cho, I. W. Cha and D. H. Youn, "An adaptive nonlinear prefilter for compensation of distortion in nonlinear systems," *IEEE Transactions on Signal*

Processing, vol. 46, no. 6, pp. 1726-1730, 1998.

[69] F. M. Barradas, L. C. Nunes, J. C. Pedro, T. R. Cunha, P. M. Lavrador and P. M. Cabral, "Accurate Linearization with Low-Complexity Models Using Cascaded Digital Predistortion Systems," *IEEE Microwave Magazine,* vol. 16, no. 1, p. 94–103, 2015.

[70] R. N. Braithwaite and S. Carichner, "An Improved Doherty Amplifier Using Cascaded Digital Predistortion and Digital Gate Voltage Enhancement," *IEEE Transactions on Microwave Theory and Technology,* vol. 57, no. 12, p. 3118–3126, 2009.

[71] H. Huang, A. Islam, J. Xia, P. Levine and S. Boumaiza, "Linear filter assisted envelope memory polynomial for analog/radio frequency predistortion of power amplifiers," in *2015 IEEE MTT-S International Microwave Symposium*, Phoenix, 2015.

[72] L. Guan and A. Zhu, "Optimized Low-Complexity Implementation of Least Squares Based Model Extraction for Digital Predistortion of RF Power Amplifiers," *IEEE Transactions on Microwave Theory and Techniques,* vol. 60, no. 3, p. 594–603, 2012.

[73] R. Raich, H. Qian and G. T. Zhou, "Orthogonal Polynomials for Power Amplifier Modeling and Predistorter Design," *IEEE Transactions on Vehicular Technology,* vol. 53, no. 5, p. 1468–1479, 2004.

[74] L. Aladren, P. Garcia, P. L. Carro, J. d. Mingo and C. Sanchez-Perez, "Digital predistortion based on Zernike polynomial functions for RF nonlinear power amplifiers," in *2012 International Symposium on Wireless Communication Systems (ISWCS)*, Paris, 2012.

[75] F. M. Barradas, T. R. Cunha, P. M. Lavrador and J. C. Pedro, "Polynomials and LUTs in PA Behavioral Modeling: A Fair Theoretical Comparison," *IEEE Transactions on Microwave Theory and Techniques,* vol. 62, no. 12, pp. 3274-3285, 2014.

[76] L. Guan and A. Zhu, "Low-Cost FPGA Implementation of Volterra Series-Based Digital Predistorter for RF Power Amplifiers," *IEEE Transactions on Microwave Theory and Techniques,* vol. 58, no. 4, pp. 866-872, 2010.

[77] Y. Ma, Y. Yamao, Y. Akaiwa and C. Yu, "FPGA Implementation of Adaptive Digital Predistorter With Fast Convergence Rate and Low Complexity for Multi-Channel Transmitters," *IEEE Transactions on Microwave Theory and Techniques,* vol. 61, no. 11, pp. 3961-3973, 2013.

[78] F. M. Barradas, T. R. Cunha, P. M. Lavrador and J. C. Pedro, "Using spline basis functions in Volterra series based models," in *2014 International Workshop on Integrated Nonlinear Microwave and Millimetre-wave Circuits (INMMiC)*, Leuven, 2014.

[79] S. Afsardoost, T. Eriksson and C. Fager, "Digital Predistortion Using a Vector-Switched Model," *IEEE Transactions on Microwave Theory and Techniques,* vol. 60, no. 4, pp. 1166-1174, 2012.

[80] O. Jardel, F. D. Groote, T. Reveyrand, J.-C. Jacquet, C. Charbonniaud, J.-P. Teyssier, D. Floriot and R. Quere, "An Electrothermal Model for AlGaN/GaN Power HEMTs Including Trapping Effects to Improve Large-Signal Simulation Results on High VSWR," *IEEE Transactions on Microwave Theory and Techniques,* vol. 55, no. 12, pp. 2660-2669, 2007.

[81] N. Ramanan, B. Lee and V. Misra, "Device Modeling for Understanding AlGaN/GaN HEMT Gate-Lag," *IEEE Transactions on Electron Devices,* vol. 61, no. 6, pp. 2012-2018, 2014.

[82] F. Filicori and G. Vannini, "Mathematical approach to large-signal modelling of electron devices," *Electronics Letters,* vol. 27, no. 4, pp. 357-359, 1991.

[83] D. Mirri, F. Filicori, G. Iuculano and G. Pasini, "A nonlinear dynamic model for performance analysis of large-signal amplifiers in communication systems," *IEEE Transactions on Instrumentation and Measurement,* vol. 53, no. 2, pp. 341-350, 2004.

[84] A. Zhu, J. Dooley and T. J. Brazil, "Simplified Volterra Series Based Behavioral Modeling of RF Power Amplifiers Using Deviation-Reduction," in *2006 IEEE MTT-S International Microwave Symposium Digest*, San Francisco, 2006.

[85] C. d. Boor, A Practical Guide to Splines, Revised ed., New York: Springer-Verlag, 1978.

[86] L. Chua and A.-C. Deng, "Canonical piecewise-linear modeling," *IEEE Transactions on Circuits and Systems,* vol. 33, no. 5, pp. 511-525, 1986.

[87] J. K. Cavers, "Optimum table spacing in predistorting amplifier linearizers," *IEEE Transactions on Vehicular Technology,* vol. 48, no. 5, pp. 1699-1705, 1999.

[88] C.-H. Lin, H.-H. Chen, Y.-Y. Wang and J.-T. Chen, "Dynamically optimum lookup-table spacing for power amplifier predistortion linearization," *IEEE Transactions on Microwave Theory and Techniques,* vol. 54, no. 5, pp. 2118-2127, 2006.

[89] S. N. Ba, K. Waheed and G. T. Zhou, "Optimal Spacing of a Linearly Interpolated Complex-Gain LUT Predistorter," *IEEE Transactions on Vehicular Technology,* vol.

59, no. 2, pp. 673-681, 2010.

[90] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proceedings of the AES 108th Convention*, Paris, 2000.

[91] A. Novak, L. Simon, F. Kadlec and P. Lotton, "Nonlinear System Identification Using Exponential Swept-Sine Signal," *IEEE Transactions on Instrumentation and Measurement,* vol. 59, no. 8, pp. 2220-2229, 2010.

[92] F. M. Barradas, P. M. Lavrador, T. R. Cunha and J. C. Pedro, "RF PA modeling with one chirp measurement," in *2015 European Microwave Conference (EuMC)*, Paris, 2015.

[93] F. M. Barradas, P. M. Lavrador, T. R. Cunha and J. C. Pedro, "Characterizing Direct and Cross Memory in RF Nonlinear Systems Using Simple Two Tone Measurements," in *2016 European Microwave Conference (EuMC)*, London, 2016.

[94] F. M. Barradas, L. C. Nunes, T. R. Cunha, P. M. Lavrador, P. M. Cabral and J. C. Pedro, "Compensation of Long-Term Memory Effects in GaN HEMT based Power Amplifiers," *IEEE Transactions on Microwave Theory and Techniques,* (submitted 2016).

[95] R. Vetury, N. Q. Zhang, S. Keller and U. K. Mishra, "The impact of surface states on the DC and RF characteristics of AlGaN/GaN HFETs," *IEEE Transactions on Electron Devices,* vol. 48, no. 3, pp. 560-566, 2001.

[96] P. Aaen, J. A. Plá and J. Wood, Modeling and Characterization of RF and Microwave Power FETs, Cambridge : Cambridge University Press, 2007.

[97] F. M. Barradas, P. M. Lavrador, T. R. Cunha and J. C. Pedro, "Using Statistical Information for Fast Static DPD of RF PAs," in *2017 IEEE Topical Conference on Power Amplifiers for Wireless and Radio Applications (PAWR)*, Phoenix, 2017, (accepted).