1 Overview of power amplifier modelling

1.1 Introduction

This chapter presents an overview and comparative assessment of the various approaches to RF power amplifier (PA) modelling that have received widespread attention by the scientific community. The chapter is organised into three sections: power amplifier modelling basics, system-level power amplifier models and circuit-level power amplifier models.

Section 1.2 on power amplifier modelling basics provides the basic knowledge to support the subsequent PA model classification and analysis. First, physical and behavioural modelling strategies are addressed and then behavioural models are classified as either static or dynamic with varying levels of complexity. Then, a distinction is made between the heuristic and systematic approaches, hence creating a theoretical framework for comparing different behavioural model formats with respect to their formulation, extraction and, in most cases, predictive capabilities.

In Section 1.3, dedicated to system-level power amplifier models, PA representations intended to be used in system-level simulators are considered. These are analytic signal- or complex-envelope-based techniques; they do not represent the RF carrier directly and RF effects are not specifically included. They are single-input–single-output (SISO) low-pass equivalent models, whose input and output constitute the complex functions needed to represent the bidimensional nature of amplitude and phase modulation.

The final section, on circuit-level power amplifier models, provides an overview of behavioural models intended for use in conventional PA circuit simulators. These models handle the complete input and output RF modulated signals, which are real entities, at two different time scales, one, very fast, for the RF carrier and another, much slower, for the modulating envelope. So, in contrast with system-level models, they also take into account the signals' harmonic content and, possibly, the input and output mismatches. For that, they need to represent the voltage and current, or incident and reflected power waves, of the PA input and output ports, thus becoming two-input–two-output model structures.

Although there is an abundant literature on the various different PA behavioural modelling approaches, there are only a few works dedicated to their analysis and comparison. A widely known reference in this field is the book of Jeruchim *et al.* [1]. More recently, a book edited by Wood and Root [2] and the papers of Isaksson *et al.* [3] and of Pedro and Maas [4] have appeared. This introduction draws from all these four references but follows the last most closely.

1.2 Power amplifier modelling basics

Power amplifiers have a major effect on the fidelity of wireless communications systems, which justifies the large number of studies undertaken to understand their limitations and then to optimise their performance. Although some earlier studies simply consisted of empirical observations of PA input–output behaviour, later works have applied scientific theories to account for the observed behaviour and, hence, to justify the resulting PA models [1–8]. Seen from the more general context of system identification, PA models can be divided into two major groups according to the type of data needed for their extraction: physical models and empirical models [9].

Physical models require knowledge of the electronic elements that constitute the PA, their relationships and the theoretical rules describing their interactions. They use nonlinear models of the PA active device and of the other, passive, components (these models may themselves be of a physical or empirical nature) to form a set of nonlinear equations relating the terminal voltages and currents. Using an equivalent-circuit description (typically having an empirical nature) of the PA, these models are appropriate to circuit-level simulation and provide a result accuracy that is, nowadays, limited almost only by the quality of the active device model. Unfortunately, such precision has a high price in simulation time and the need for a detailed description of the PA internal structure.

When such a PA equivalent circuit is not available, or whenever a complete system-level simulation is desired, PA behavioural models are preferred. Since they are solely based on input–output (behavioural) observations, their accuracy is highly sensitive to the adopted model structure and the parameter extraction procedure. So, it is no surprise that distinct model topologies and different observation data sets may lead to a large disparity in model applicability and simulation results. In fact, though such a behavioural-modelling approach may guarantee the accurate reproduction of the data set used for its extraction, or, possibly, of some other set pertaining to the same excitation class, it is not obvious that it will also produce useful results for a different data set, a different PA of the same family or a PA based on a completely different technology. That is, in contrast with the physical-modelling alternative, the generalisation of the predictive capability of a behavioural model should always be viewed with circumspection.

1.2.1 Nonlinear system identification background

In order to establish a theoretical framework with which to analyse the various approaches to PA behavioural modelling, it is convenient to recall some basic results of system identification theory.

In that framework, our power amplifier is described either by a nonlinear function or a system operator; it is assumed to be either static or dynamic respectively. In the static case its output $y(t)$ can be uniquely defined as a function of the instantaneous input $x(t)$, and the model reduces to

$$y(t) = f(x(t)) \tag{1.1}$$

or

$$y = f(x), \tag{1.2}$$

since the dependence with time is, in this case, immaterial.

When the PA presents memory effects to either the modulated RF signal or the modulating envelope, it is said to be dynamic. The output can no longer be uniquely determined from the instantaneous input. It now depends also on the input past and/or the system state. The relation between $y(t)$ and $x(t)$ cannot be modelled simply by a function but becomes an operator that maps a function of time $x(t)$ onto another function of time $y(t)$. Thus the input–output mapping of our PA is represented by a forced nonlinear differential equation,

$$f\left(y(t), \frac{d\,y(t)}{d\,t}, \ldots, \frac{d^p\,y(t)}{d\,t^p}, x(t), \frac{d\,x(t)}{d\,t}, \ldots, \frac{d^r\,x(t)}{d\,t^r}\right) = 0. \tag{1.3}$$

This states that the output and its time derivatives (in general, the system state) may be nonlinearly related to the input and its time derivatives. Since our PA behavioural models have to be evaluated in a digital computer, i.e. a finite-state machine, it is convenient to adopt a discrete-time environment, in which the time variable becomes a succession of uniform time samples of convenient sampling period T_s; thus the time and the continuous time signals may be translated as $t \to sT_\mathrm{s}$, $x(t) \to x(s)$ and $y(t) \to y(s)$, $s \in Z$. In this way, the solution of the nonlinear differential equation in Equation (1.3) can be expressed in the following recursive form [10]:

$$y(s) = f_\mathrm{R}(y(s-1), \ldots, y(s-Q_1), x(s), x(s-1), \ldots, x(s-Q_2)). \tag{1.4}$$

Here $y(s)$, the present output at time instant sT_s, depends in a nonlinear way, dictated by f_R, the nonlinear function, on the system state (herein expressed by $y(s-q), q = 1, \ldots, Q_1$), the present input $x(s)$ and its past values, $x(s-q)$. This nonlinear extension of infinite impulse response digital filters [10] (nonlinear IIR) is assumed to be the general form for recursive PA behavioural models.

System identification results have shown that, under a broad range of conditions [10–12] (basically operator causality, stability, continuity and fading memory), such a system can also be represented with any desirable small error by a non-recursive, or direct, form, where the relevant input past is restricted to $q \in \{0, 1, 2, \ldots, Q\}$, the so-called system memory span [10]:

$$y(s) = f_\mathrm{D}(x(s), x(s-1), \ldots, x(s-Q)) \tag{1.5}$$

in which $f_\mathrm{D}(\cdot)$ is again a multidimensional nonlinear function of its arguments. This nonlinear extension of finite impulse response digital filters [10] (nonlinear FIR), is again the general form that a direct, or feedforward, behavioural model should obey.

Various forms have been adopted for the multidimensional functions $f_{\mathrm{R}}(\cdot)$ and $f_{\mathrm{D}}(\cdot)$, although two of these have received particular attention in nonlinear system identification. This is due to their formal mathematical support and because they lead directly to a canonical realisation and so to a certain model topology. These two forms are polynomial filters [10–14] and artificial neural networks (ANNs) [15–17].

In the first case, $f_{\mathrm{D}}(\cdot)$ is replaced by a multidimensional polynomial approximation, so that Equation (1.5) takes the form

$$\begin{aligned} y(s) &= P_{\mathrm{D}}\left(x(s), x(s-1), \ldots, x(s-Q)\right) \\ &= \sum_{q=0}^{Q} a_1(q)x(s-q) + \sum_{q_1=0}^{Q}\sum_{q_2=0}^{Q} a_2(q_1, q_2)x(s-q_1)x(s-q_2) + \cdots \\ &\quad + \sum_{q_1=0}^{Q} \cdots \sum_{q_N=0}^{Q} a_N(q_1, \ldots, q_N)x(s-q_1)\cdots x(s-q_N). \end{aligned} \tag{1.6}$$

This form shows that the nonlinear system is approximated by a series of multilinear terms. Although simple in concept, this ‘polynomial FIR’ model architecture is known for its large number of parameters.

The function $f_{\mathrm{R}}(\cdot)$ can also be replaced by a multidimensional polynomial leading to recursive polynomial IIR structures. These provide similar approximation capabilities for many fewer parameters than the direct topology. However, the polynomial IIR is significantly more difficult to extract than the direct topology; this has impeded its application in the PA modelling field.

Indeed, the comparative ease of extraction of the polynomial FIR, in comparison with other PA models, provides its particular and attractive advantage. Since the output is linear in respect of the model parameters, i.e. the kernels $a_n(q_1, \ldots, q_n)$, and dependent only on multilinear functions of the delayed versions of the input, it can be extracted in a systematic way using conventional linear identification procedures.

If $f_{\mathrm{D}}(\cdot)$ or $P_{\mathrm{D}}(\cdot)$ is approximated by a Taylor series then this FIR filter is known as a Volterra series or Volterra filter [10–14]. This Volterra series approximation is particularly interesting as it produces an optimal approximation (in a uniform-error sense) near the point where it is expanded. Therefore it shows good modelling properties in the small-signal, or mildly nonlinear, regimes. However, it shows catastrophic degradation under strong nonlinear operation.

In fact, $f_{\mathrm{D}}(\cdot)$ can be replaced by any other multidimensional polynomial. For example, the Wiener series is orthogonal for white Gaussian noise as an excitation signal [13, 14]; other orthogonal polynomials have been proposed for other excitations [10, 13, 18, 19]. In these cases, the respective series produce results that are optimal (in a mean-square-error sense) in the vicinity of the power level used and for the particular type of input used in the model extraction. These representations are, therefore, amenable to the modelling of strong nonlinear systems when the excitation bandwidth and statistics can be considered close to those used in extraction experiments. A presentation of Wiener series expansions and their orthogonality under white Gaussian noise excitation is given in Section 3.11.

Such polynomial FIR filters can be realised in the form indicated in Figure 1.1. The multiplicity of nth-order cross products between all delayed inputs may be noted; it is to these that the nonlinear filter owes its notoriously complex, although general, form. In a similar way, polynomial IIR filters can be realised. A bilinear, recursive, nonlinear IIR filter implementation is shown in Figure 1.2 [4].

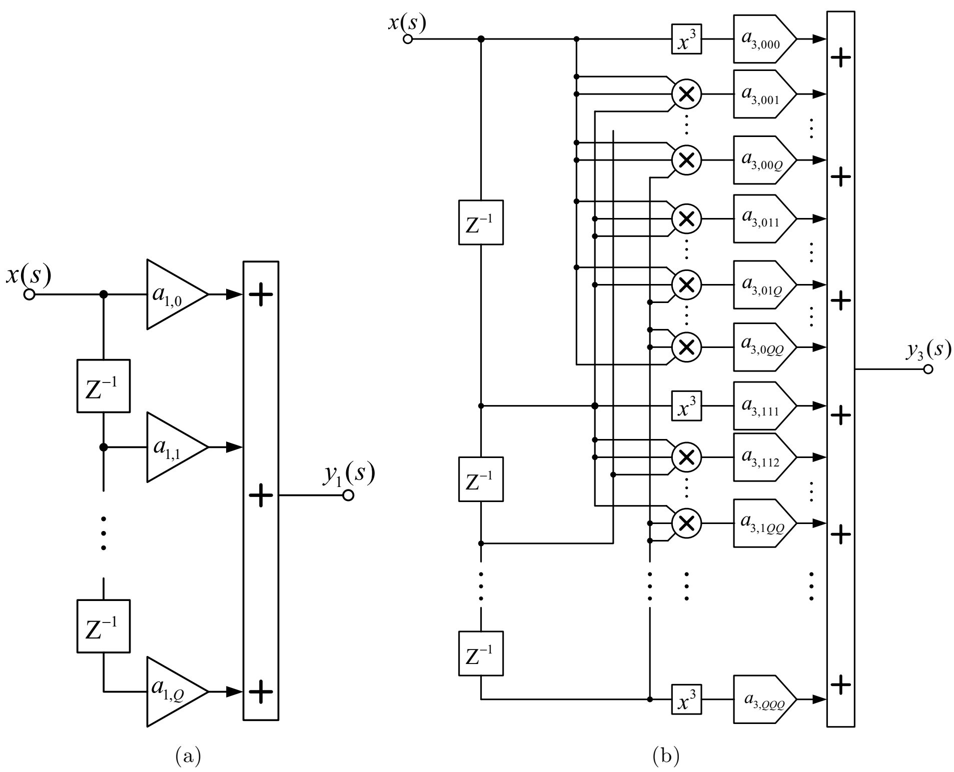

Figure 1.1 Examples of canonical forms of nonlinear FIR filters. (a) Canonical FIR filter of first order, (b) canonical FIR filter of third order. The operator Z^{-1} indicates a unit delay tap (see subsection 5.2.1).

When $f_R(\cdot)$ and $f_D(\cdot)$ are approximated by ANNs, Equations (1.4) and (1.5) take the following pairs of forms [15]:

$$u_k(s) = \sum_{q=1}^{Q_1} wy_k(q)y(s-q) + \sum_{q=0}^{Q_2} wx_k(q)x(s-q) + b_k,$$
$$y(s) = b_o + \sum_{k=1}^{K} wy_o(k)f(u_k(s)) \tag{1.7}$$

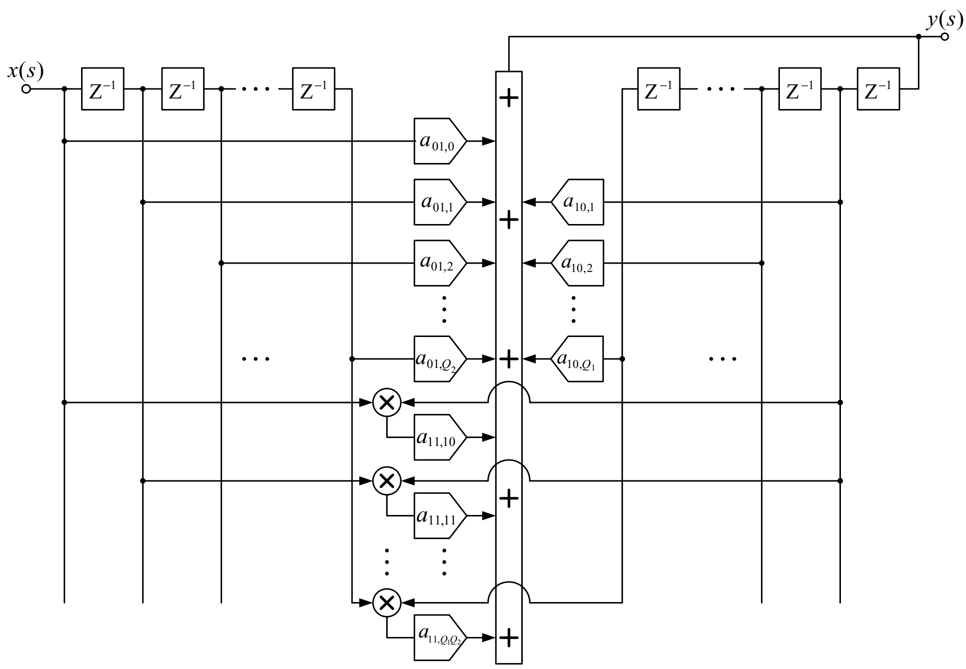

Figure 1.2 General structure of a bilinear recursive nonlinear filter.

and

$$u_k(s) = \sum_{q=0}^{Q} w_k(q)x(s-q) + b_k,$$
$$y(s) = b_o + \sum_{k=1}^{K} w_o(k)f(u_k(s)), \tag{1.8}$$

where $wy_k(q)$, $wx_k(q)$, $wy_o(k)$, $w_k(q)$ and $w_o(k)$ are weighting coefficients, b_k and b_o are bias parameters and $f(\cdot)$ is a predefined nonlinear function (the ANN activating function) of its argument [15]. As in the case of polynomial filters, these ANNs have universal approximation capabilities meaning that they are capable of an arbitrarily accurate approximation to arbitrary mappings [16, 17]. This aspect is dealt with in more detail in subsection 5.3.2.

These recursive and feedforward dynamic ANNs can be realised in the forms of Figures 1.3 and 1.4 respectively.

A close look at the feedforward ANN model of Equation (1.8) and Figure 1.4 shows that the model output is built from the addition of the activation functions $f(u_k(s))$ and the weighted outputs plus a bias and that the $u_k(s)$ are biased sums of the various delayed versions of the input, weighted by the coefficients $w_k(q)$. Each $u_k(s)$ can thus be seen as the biased output of a linear FIR filter whose input is the signal $x(s)$ and whose impulse response is $w_k(q)$. So the non-recursive ANN model is actually equivalent to a parallel connection of K branches of linear filters

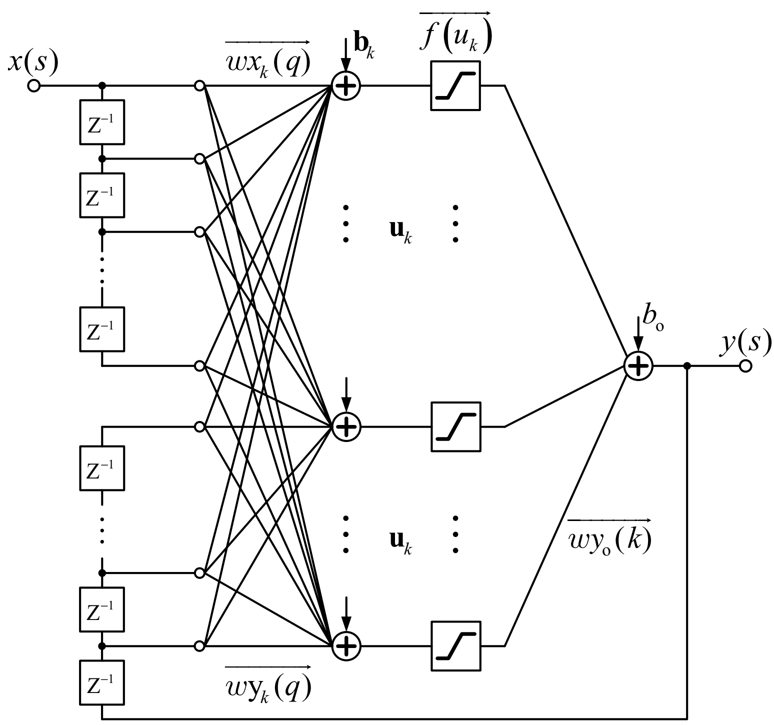

Figure 1.3 General structure of a recursive single-hidden-layer dynamic artificial neural network.

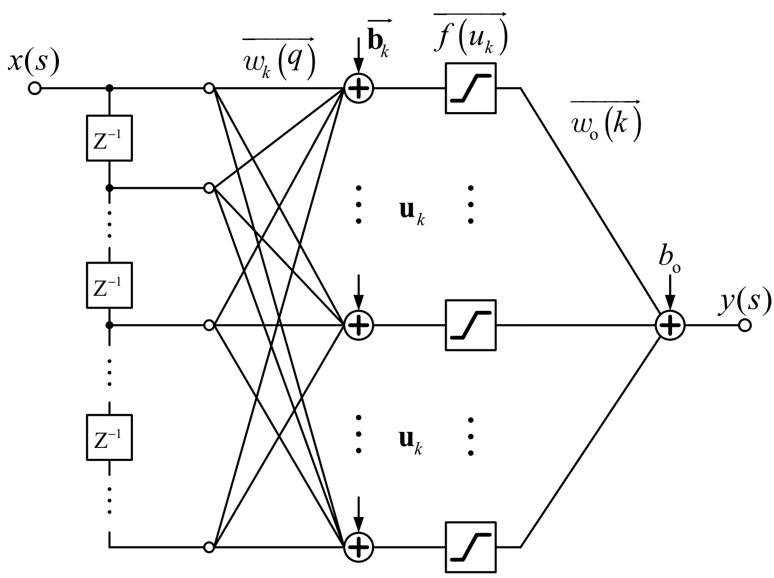

Figure 1.4 General structure of a feedforward single-hidden-layer dynamic artificial neural network.

followed by a memoryless nonlinearity, as shown in Figure 1.5.

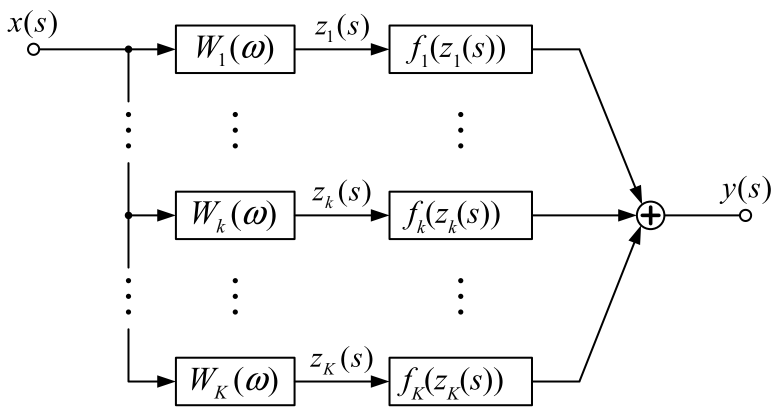

Figure 1.5 Equivalent structure of a feedforward single-hidden-layer perceptron ANN. Note that here the combinations of the branch biases b_k, the activation functions $f(u_k(s))$, the branch gains $w_o(k)$, and the final bias b_o are here represented by different branch memoryless nonlinearities $f_k(z_k(s))$.

If the branch memoryless nonlinearities were now approximated by polynomial functions we would end up again with a polynomial filter. This shows that there is essentially no distinction between a feedforward time-delay ANN and a non-recursive polynomial filter. They simply constitute two alternative ways of approximating the multidimensional function $f_D(\cdot)$, of Equation (1.5). There are, however, some slight differences in these two approaches that will be addressed below. These are worth mentioning because of their impact on PA behavioural modelling activities.

The series form of polynomial filters enables certain output properties to be related to each polynomial degree, and this can be used to guide the parameter extraction procedure. This is especially true if the polynomial series is orthogonal for the input used in the model identification process. For example, the relationship between the intermodulation content of the system's response to a multisine (a signal consisting of several sinusoidal tones) and the coefficients of an appropriate multidimensional orthogonal polynomial have recently been found [18, 19] (the structure and design of multisine signals will be discussed in subsection 2.5.6). However, since in an ANN all memoryless nonlinearities share a common form, there is no way to identify such relationships. Consequently, while polynomial filters can be extracted in a direct way, ANN parameters can be obtained only from some nonlinear optimisation scheme.

Moreover, despite the universal approximation properties of ANNs, there is no way of knowing *a priori* how many hidden neurons are needed to represent a specific system, nor is there any way of predicting the modelling improvement gained when this number is increased. It cannot even be ensured that the extracted ANN is unique or that it is optimal for a certain number of neurons. This can obviously pose some potential problems for the ANN's predictability, especially for inputs outside the signal class used for the identification, i.e. the ANN training process.

However, in contrast with the intrinsically local approximating properties of polynomials, ANNs behave as global approximates, an important advantage when one is modelling strongly nonlinear systems. Also, since the sigmoidal functions used in ANNs are bounded in output amplitude, ANNs are, in principle, better than polynomials at extrapolating beyond the zone where the system was operated during parameter extraction.

1.2.2 Nonlinear dynamic properties of microwave PAs

We now turn our attention to some typical nonlinear effects presented by practical microwave and wireless PAs. Considering the variety of available PA technologies, it is not easy to give a completely comprehensive view. Nevertheless, the technical literature in this subject indicates that a few effects at least are commonly observed in a fairly wide range of devices.

Both solid-state PAs (SSPAs) and travelling-wave tube PAs (TWTAs) have been frequently represented by cascade combinations of linear filters and a memoryless nonlinearity [20–22], the so-called two-box and three-box models. These structures introduce linear memory effects at the input and output that can be physically related to the PA's input and output tuned networks.

Beyond these linear memory effects, there are also some dynamic effects that show up only in the presence of nonlinear regimes. This is the case for the so-called long-term memory effects commonly attributed to the active device's low-frequency dispersion and electrothermal interactions and the interactions of the active device with the bias circuitry [23–29] (compare also subsection 2.4.1). Described by the dynamic interaction of two or more nonlinearities through a dynamic network, these long-term memory effects manifest nonlinear dynamics that cannot be modelled by any non-interacting linear filter and memoryless nonlinearity box models. Indeed, Pedro *et al.* [26] showed that such effects can be represented by a memoryless nonlinearity and a filter in a feedback path, as depicted in Figure 1.6, while Vuolevi *et al.* [25] and Vuolevi and Rahkonen [27] used a cascade connection of two nonlinearities with a linear filter in between.

As a common basis for the following behavioural-model discussion, we will assume that a general PA has the form shown in Figure 1.6. Through $H(\omega)$ and $O(\omega)$, this feedback model can account for linear memory effects not only in the carrier but also in the information envelope; these occur whenever the PA characteristic is not flat within the operating signal's bandwidth. In addition, the model is also capable of describing nonlinear memory effects in the carrier (AM–PM) and/or the envelope whenever the feedback filter $F(\omega)$ exhibits dynamic behaviour at the carrier frequency, the carrier harmonics frequencies or the demodulated envelope frequency [4, 26, 27].

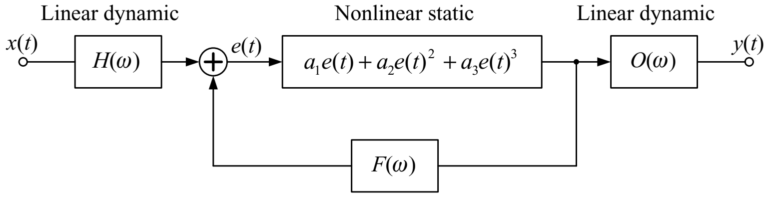

Figure 1.6 Typical nonlinear feedback structure of a microwave PA. Note the presence of the filters $H(\omega)$ and $O(\omega)$ representing linear memory effects related to the input and output matching networks; the feedback path represents nonlinear memory effects attributed to electrothermal and/or bias circuitry dynamics.

For reference, the first- and third-order Volterra nonlinear transfer functions of the dynamic feedback model of Figure 1.6 are [4, 26]:

$$S_1\,(\omega) = \;H(\omega)\,\frac{a_1}{D(\omega)}O(\omega) \tag{1.9}$$

and

$$\begin{aligned} S_3\,(\omega_1, \omega_2, \omega_3) &= \frac{H(\omega_1)\,H(\omega_2)\,H(\omega_3)}{D(\omega_1)\,D(\omega_2)\,D(\omega_3)}\;\frac{O(\omega_1+\omega_2+\omega_3)}{D(\omega_1+\omega_2+\omega_3)} \\ &\quad\times\left\{a_3 + \frac{2}{3}a_2^2\left[\frac{F(\omega_1+\omega_2)}{D(\omega_1+\omega_2)} + \frac{F(\omega_1+\omega_3)}{D(\omega_1+\omega_3)} + \frac{F(\omega_2+\omega_3)}{D(\omega_2+\omega_3)}\right]\right\}, \end{aligned} \tag{1.10}$$

where $D(\omega) = 1 - a_1 F(\omega)$.

On expanding Equation (1.10), input and output linear memory effects are described by the terms $H(\omega_1)H(\omega_2)H(\omega_3)$, $O(\omega_1+\omega_2+\omega_3)$, $F(\omega_1)F(\omega_2)F(\omega_3)$ and $F(\omega_1+\omega_2+\omega_3)$, while nonlinear memory can be seen to arise from the harmonics $F(\omega_j+\omega_k)$, $j,k = 1,2,3$ and the envelope dynamics $F(\omega_j-\omega_k)$.

1.3 System-level power amplifier models

System-level PA behavioural modelling employs low-pass equivalent PA models and thus processes only the complex-envelope information signal. Any specific effects related to or arising from the carrier frequency used must be individually incorporated. This distinguishes such models from circuit-level PA models, which maintain the full RF circuit's band-pass nature and information and work with the actual RF signal.

The RF signal may be written [1, 30]:

$$s(t) = \mathrm{Re}\left\{r(t)e^{j[\omega_0 t+\phi(t)]}\right\} = r(t)\,\cos[\omega_0 t + \phi(t)], \tag{1.11}$$

where an RF carrier of frequency ω_0 is modulated by the complex envelope:

$$\tilde{s}(t) = r(t)e^{j\phi(t)}. \tag{1.12}$$