Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6043391, 12 pages
https://doi.org/10.1155/2021/6043391

Research Article

A General Order Reduction Method of Wideband Digital Predistortion Model Using Attention Mechanism

Zhijun Liu, Xin Hu, and Weidong Wang

School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

Correspondence should be addressed to Weidong Wang; wangweidong@bupt.edu.cn

Received 9 September 2021; Revised 4 October 2021; Accepted 12 October 2021; Published 5 November 2021

Academic Editor: Junjuan Xia

In wireless networks, for the common in-phase and quadrature-phase (I/Q) imbalance in the transmitters, the I/Q branch models of digital predistortion (DPD) need to be identified separately, to improve the linearization effects. The existing order reduction methods of the predistorter are based on the contributions of the complex basis function terms, so as not to deal with the different contributions of I/Q components of the complex basis function terms caused by the separate identification of the I/Q branch models. The separate pruning of the I/Q branch models will increase the complexity. Aiming at this issue, this paper proposes a general order reduction method based on the attention mechanism for the predistortion of the power amplifiers (PAs). This method is suitable for pruning both the traditional models and neural network-based models. In this method, the attention mechanism is used to evaluate the contributions of the real basis function terms to the predistorted output's I/Q components through offline training, and the influence of the cross terms of the I/Q branch models is considered. The experimental results based on the comparison with other typical methods under 100 MHz Doherty PA and different I/Q imbalance levels show that this method has superior pruning performance and good linearization ability.

1. Introduction

With the rapid iteration of the fifth-generation (5G) wireless systems, wider signal bandwidth and more complex modulation modes are used, to satisfy the rapid growth of the requirement of the data service [1–3]. However, the wide signal bandwidth and efficient modulation make the transmitters, especially the power amplifiers (PAs), exhibit more complex nonlinear behavior characteristics [4], which leads to the difficulty of high-efficiency transmission in the transmitting system. To solve this problem, digital predistortion (DPD) is one of the most commonly used linearization techniques [5–7].

DPD techniques compensate for the nonlinear behaviors of the transmitter by constructing a nonlinear model that is opposite to the nonlinear characteristics of the transmitter [8]. At present, the most common and popular predistortion models are the full Volterra (FV) series models. Since these models' parameters are linear with respect to the output of the system, these models can be easily identified by the classical regression theory [9]. However, the complex nonlinear behaviors (including nonlinearity and memory effects [6]) caused by the increase of the signal bandwidth and complex modulation modes will lead to the curse of dimensionality of the FV models [9]. Therefore, the order reduction of the FV model has become an effective means to improve the availability of the model and reduce the cost [10, 11]. To this end, based on the general FV series model, various prior pruning models, such as the memory polynomial (MP) model [12] and the generalized MP (GMP) model [13], are proposed. These models are easy to be modeled in the field-programmable gate array (FPGA), such as through lookup table (LUT) [14–16], so they are commonly used engineering models at present. However, these models are pruned based on prior knowledge and are still general predistortion models [4]. For a specific PA, in order to meet the linearization requirements, these models still include many basic function terms with fewer contributions, leading to the complexity of the model.

For this reason, classical posterior pruning techniques are proposed to select the necessary terms based on the nonlinear behavior of the specific PA, to find the optimal

structure under a given PA [14, 17, 18]. The most typical method is the predistortion model pruning technique based on orthogonal matching pursuit (OMP) [11]. This method selects the term with the greatest correlation with the remaining output in each iteration [11] to determine the optimal predistortion structure. To solve the ill-condition of the equation system caused by the high correlation between the basis function terms, a doubly OMP (DOMP) algorithm uses Gram-Schmidt orthogonalization to eliminate the correlation between the selected and unselected basis function spaces after each iteration [17]. However, the pseudo inverse calculation and the Kronecker product calculation in the orthogonalization process lead to the high computational complexity of the algorithm [9]. To this end, the simplified sparse parameter identification DOMP (SSPI DOMP) algorithm is proposed to implement the pseudo inverse computation through the recursive process [19], which effectively reduces the computational complexity. Reference [9] also proposed to realize the pseudo inverse calculation by processing the covariance matrix by the orthogonal properties, to reduce the calculation cost further. In addition, a predistortion model pruning algorithm based on adaptive principal component analysis (PCA) was proposed in reference [14]. Reference [20] also proposed a pruning algorithm based on the projection of the residual vector. All the above pruning methods regard the complex basis function term as a whole and then achieve order reduction.

However, in real wireless communication systems, the nonideal behavior of the modulator will lead to the mismatch between the gain and the phase of the transmission signal and then cause the imbalance of in-phase and quadrature-phase (I/Q) components [21]. The modulator imperfections are interwoven with the nonlinear behavior of PA, which further reduces the transmission quality of the system [22, 23]. For this situation, the two branches (namely, I/Q components) of the transmitters can be compensated, respectively. In other words, the I/Q components of the compensator can be identified independently, to cope with the nonideal behavior of the modulator. For example, widely used artificial neural network (ANN) models, such as the neural network (NN) model [21] and convolutional NN (CNN) model [5], are predistortion models of I/Q separate identification. The traditional models can also be used for independent modeling of I/Q branches, which can be expressed as

$$\begin{cases} x_I(n) = \left[\bar{\mathbf{U}}_I(n) \quad \bar{\mathbf{U}}_Q(n)\right]\mathbf{w}_I', \\ x_Q(n) = \left[\bar{\mathbf{U}}_I(n) \quad \bar{\mathbf{U}}_Q(n)\right]\mathbf{w}_Q', \end{cases} \tag{1}$$

where $x_I(n)$ and $x_Q(n)$ represent the I/Q components of the predistorter and $\bar{\mathbf{U}}(n)$ is the predistortion model. Table 1 shows the comparison of the normalized mean square error (NMSE) performance between independent identification and combined identification of I/Q components of the predistorter under 100 MHz Doherty PA, which verifies the above idea. References [8, 23] also proposed the compensation models for I/Q imbalance, which are independent of the DPD model and resulting in the complexity of the design.

In this case, the I/Q components of the basis function terms have independent contributions to the linearization effects. If the I/Q branch models of the predistorter are pruned separately, such as using DOMP, the basis function terms of the I/Q branch models of the predistorter need to be constructed independently, which leads to the high design complexity in FPGA. It has become a difficult point to find the real basis function terms that are important to the I/Q components of the predistortion output.

To solve this issue, this paper proposes a general order reduction method of the predistortion model based on the attention mechanism. In reference [24], we have verified that this method can effectively prune the input items of the NN-based models. In this paper, we improve this method and apply it to the pruning of the traditional polynomial models, to prove its universality. This method firstly calculates the comprehensive contributions of the real basis function terms to the predistorted output's I/Q components using the attention mechanism through offline training, which considers the influence of the cross basis function terms of the I/Q branch models. Since the contributions of the real basis function terms to the predistorted output's I/Q components are calculated simultaneously, that is, the cross terms are evaluated, the I/Q branch models are consistent, which further reduces the design complexity of the model. The experimental results based on the comparison with other typical methods under 100 MHz Doherty PA and different I/Q imbalance levels show the effectiveness of the method.

The contributions of this paper are as follows:

(i) The traditional I/Q imbalance models are configured independently, which leads to high model complexity [8, 23]. In order to reduce the model complexity, the I/Q branch models of the predistorter are modeled separately, to compensate for the I /Q imbalance and PA's nonlinearity simultaneously

(ii) The existing order reduction methods are based on the contributions of the complex basis function terms, so as not to deal with the different contributions of I/Q components of the complex basis function terms [11, 17]. This paper distinguishes the different contributions of the I/Q components of the basis function terms to the I/Q branch models, to further reduce the complexity of the model

(iii) As a result of all the above contributions, this work achieves a good compromise between the model complexity and linearization effects to drive the 100 MHz Doherty PA. In addition, compared with the existing order reduction models, the proposed model has the lowest complexity of the model

The structure of the paper is organized as follows. In Section 2, the modeling and identification processes of the I/Q branch models of the predistorter are described, and the principle of the attention mechanism is analyzed. Section 3 describes in detail the proposed order reduction method of

Table 1: Performance comparison of I/Q independent identification and I/Q combined identification (GMP model).

	PA nonlinearity	PA nonlinear and I/Q imbalance
Combination identification of I/Q	NMSE = −32.67 dB	NMSE = −22.75 dB
Independent identification of I/Q	NMSE = -37.67 dB	NMSE = −34.47 dB

the predistortion model based on the attention mechanism and gives the specific training process. Section 4 introduces the test platform for validation of the proposed order reduction method. In Section 5, the measurement and validation results of the proposed method are described and analyzed. The conclusion is given in Section 6.

2. Digital Predistortion Based on I/Q Separate Identification

2.1. Predistortion Model of I/Q Separate Identification. Due to the nonideal behavior of the modulator, the nonlinear behaviors of PA are interleaved with the I/Q imbalance, which leads to more complex nonlinear characteristics of the transmitter [21]. Therefore, to improve the linearization effects, the compensators of the I/Q branches should be identified separately, to deal with the asymmetry of the I/Q branches of the transmitter. The predistortion structure of I/Q separate identification is shown in Figure 1. The I/Q branch models of the predistorter are modeled using the real basis function terms composed of the I/Q components of the traditional complex basis function terms and then identified separately. The indirect learning architecture (ILA) [7] is used to identify the predistorter. The I/Q branch models based on the GMP model can be expressed as follows [13]:

$$\begin{aligned} x_I(n) = & \sum_{k=0}^{Ka-1}\sum_{l=0}^{La-1} \alpha_{kl}^{II} y_I(n-l)|y(n-l)|^k \\ & + \sum_{k=1}^{Kb}\sum_{l=0}^{Lb-1}\sum_{m=1}^{Mb} \beta_{klm}^{II} y_I(n-l)|y(n-l-m)|^k \\ & + \sum_{k=1}^{Kc}\sum_{l=0}^{Lc-1}\sum_{m=1}^{Mc} \gamma_{klm}^{II} y_I(n-l)|y(n-l+m)|^k \\ & + \sum_{k=0}^{Ka-1}\sum_{l=0}^{La-1} \alpha_{kl}^{IQ} y_Q(n-l)|y(n-l)|^k \\ & + \sum_{k=1}^{Kb}\sum_{l=0}^{Lb-1}\sum_{m=1}^{Mb} \beta_{klm}^{IQ} y_Q(n-l)|y(n-l-m)|^k \\ & + \sum_{k=1}^{Kc}\sum_{l=0}^{Lc-1}\sum_{m=1}^{Mc} \gamma_{klm}^{IQ} y_Q(n-l)|y(n-l+m)|^k, \end{aligned} \tag{2}$$

where $y(n)$ is the output signal of the PA and $y_I(n)$ and $y_Q(n)$ represent the I/Q components of $y(n)$, respectively. $\{K_a, L_a, K_b, L_b, M_b, K_c, L_c, M_c\}$ are the parameters of the GMP model. $\{\alpha_{kl}^{II}, \beta_{klm}^{II}, \gamma_{klm}^{II}, \alpha_{kl}^{IQ}, \beta_{klm}^{IQ}, \gamma_{klm}^{IQ}\}$ are the model coefficients. $x_I(n)$ is the I component of the PA input. The Q component can be represented by the same model as Equation (2).

The input and output data of N groups of the predistortion model are collected, and then, the I/Q branch models of the predistortion can be written in matrix form.

$$\begin{cases} \mathbf{x}_I = \left[\bar{\mathbf{Y}}_I \quad \bar{\mathbf{Y}}_Q\right]\boldsymbol{\omega}_I, \\ \mathbf{x}_Q = \left[\bar{\mathbf{Y}}_I \quad \bar{\mathbf{Y}}_Q\right]\boldsymbol{\omega}_Q, \end{cases} \tag{3}$$

where $\mathbf{x}_I = [x_I(N), x_I(N-1), \cdots, x_I(1)]^T$,

$$\mathbf{x}_Q = \left[x_Q(N), x_Q(N-1), \cdots, x_Q(1)\right]^T, \tag{4}$$

$$\boldsymbol{\omega}_I = \left[\alpha_{00}^{II}, \alpha_{01}^{II}, \cdots, \gamma_{Kc(Lc-1)Mc}^{II}, \alpha_{00}^{IQ}, \cdots, \gamma_{Kc(Lc-1)Mc}^{IQ}\right]^T, \tag{5}$$

$$\boldsymbol{\omega}_Q = \left[\alpha_{00}^{QI}, \alpha_{01}^{QI}, \cdots, \gamma_{Kc(Lc-1)Mc}^{QI}, \alpha_{00}^{QQ}, \cdots, \gamma_{Kc(Lc-1)Mc}^{QQ}\right]^T, \tag{6}$$

$$\bar{\mathbf{Y}} = \left[\bar{\mathbf{Y}}(N), \bar{\mathbf{Y}}(N-1), \cdots, \bar{\mathbf{Y}}(1)\right]^T, \tag{7}$$

$$\bar{\mathbf{Y}}(n) = [\bar{y}_1(n), \bar{y}_2(n), \cdots, \bar{y}_S(n)]^T, \tag{8}$$

where $\bar{\mathbf{Y}}_I$ and $\bar{\mathbf{Y}}_Q$ are the I/Q component matrices of $\bar{\mathbf{Y}}$, respectively. $\bar{\mathbf{Y}}(n), (n = 1, 2, \cdots N)$ is a complex vector composed of the basis function terms corresponding to signal $y(n)$, and $S = K_aL_a + K_bL_bM_b + K_cL_cM_c$ is the number of complex basis function terms.

The I/Q branch models of the predistortion have the same structure but are identified separately, to cope with the I/Q imbalance. Equation (3) is solved by the least-squares (LS) algorithm [25, 26]; then, the coefficients of the I/Q branch models can be estimated.

$$\begin{cases} \widehat{\boldsymbol{\omega}}_I = \left(\left[\bar{\mathbf{Y}}_I \quad \bar{\mathbf{Y}}_Q\right]^H\left[\bar{\mathbf{Y}}_I \quad \bar{\mathbf{Y}}_Q\right]\right)^{-1}\left[\bar{\mathbf{Y}}_I \quad \bar{\mathbf{Y}}_Q\right]^H\mathbf{x}_I, \\ \widehat{\boldsymbol{\omega}}_Q = \left(\left[\bar{\mathbf{Y}}_I \quad \bar{\mathbf{Y}}_Q\right]^H\left[\bar{\mathbf{Y}}_I \quad \bar{\mathbf{Y}}_Q\right]\right)^{-1}\left[\bar{\mathbf{Y}}_I \quad \bar{\mathbf{Y}}_Q\right]^H\mathbf{x}_Q, \end{cases} \tag{9}$$

where $\widehat{\boldsymbol{\omega}}_I$ and $\widehat{\boldsymbol{\omega}}_Q$ are the estimations of $\boldsymbol{\omega}_I$ and $\boldsymbol{\omega}_Q$, respectively. In the calculation of the I/Q components of the predistortion in FPGA, the coefficients of the predistortion model are multiplied by the model in I/Q branches, respectively. Therefore, the different coefficients of the I/Q branch models do not complicate the predistortion process.

2.2. The Principle of the Attention Mechanism. The achievements of artificial intelligence in the field of communication provide us with ideas [27, 28]. The attention mechanism is an effective structure to focus on important features, which has been widely applied in the fields of speech recognition [29] and image processing [30]. Based on the importance of the input features to the generation of the output, the attention mechanism weights the input features, to

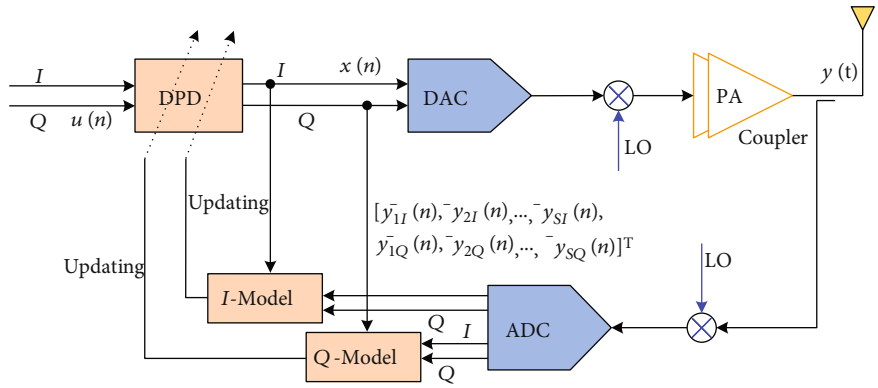

Figure 1: Digital predistortion structure based on I/Q separate identification.

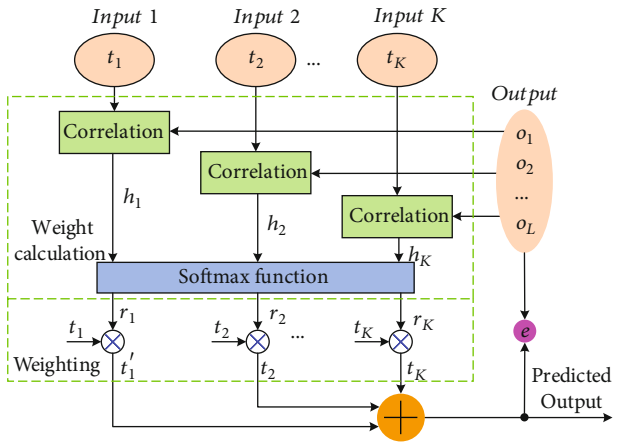

Figure 2: Implementation block diagram of the attention mechanism.

strengthen the important features and weaken the unimportant features, which can improve the fitting ability. The principle of the attention mechanism is shown in Figure 2. Let the input of the module be $\mathbf{t}$, which can be written as

$$\mathbf{t} = [t_1, t_2, \cdots, t_K]^T, \tag{10}$$

where K is the number of input nodes. The fitting output is $\mathbf{o}$, which can be written as

$$\mathbf{o} = [o_1, o_2, \cdots, o_L]^T, \tag{11}$$

where L is the number of output nodes.

First, the correlation between each input $t_i, (i = 1, 2, \cdots, K)$ and all outputs $\mathbf{o}$ is calculated. The commonly used method to calculate correlation is NN [28]. The correlation obtained can be written as

$$h_i = \mathrm{Cor}(t_i, \mathbf{o}), (i = 1, 2, \cdots, K), \tag{12}$$

where $\mathrm{Cor}(\cdot)$ is the function that calculates the correlation.

Then, the correlations are normalized and converted to the probability form through the Softmax function, to mean the weights of the input. The weights of the input can be written as

$$r_i = \frac{\exp(h_i)}{\sum_{j=1}^{K} \exp(h_j)}, (i = 1, 2, \cdots, K), \tag{13}$$

where $\exp(k)$ stands for e^k and $\sum(\cdot)$ represents the sum function.

Finally, the weights are used to weigh the corresponding input of the module. Using weighted inputs, the model outputs can be fitted. By weighting the inputs, the valid input features are emphasized, and the invalid ones are weakened, so as to improve the fitting performance. This paper does not use the attention mechanism to improve the modeling ability but embeds the attention mechanism into the predistortion model to obtain the weights of the basis function terms to evaluate the contributions of the basis function terms by weights.

3. The Proposed Order Reduction Method of the Predistortion Model

In this section, the structure of the proposed order reduction method of the predistortion model is given first, and each module is described in detail. Then, the training method and process of the proposed order reduction method are analyzed.

3.1. The Structure of the Proposed Order Reduction Method. In order to improve the linearization performance in the case of I/Q imbalance, the I/Q branch models of the predistorter are identified separately. At this point, the I/Q components of all complex basis function terms have separate model coefficients and independent contributions to the predistortion output, as shown in the analysis in Section 2, part 1. To further simplify the structure of the predistortion model, this paper proposes an order reduction method of the predistortion model based on the attention mechanism, as shown in Figure 3. In this method, all real basis function terms are distinguished, and their contributions to the predistortion output are given, so that I/Q branch models of the predistorter can be pruned. Meanwhile, to ensure the

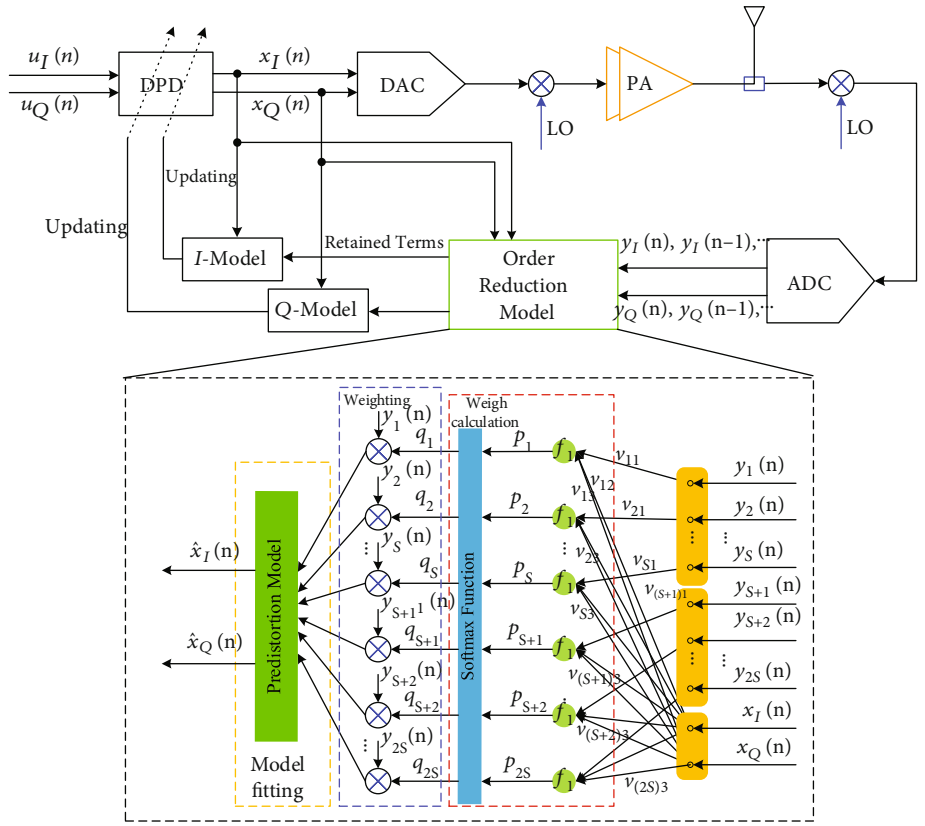

Figure 3: The structure of the proposed order reduction method of the predistortion model based on the attention mechanism.

consistency of the I/Q branch models of the predistorter, the contributions of the real basis function terms to the I/Q components of the predistortion output are calculated simultaneously, to reduce the predistortion model's design complexity in FPGA. The specific model structure is described as follows.

The input signal $x(n)$ is fed into PA after passing through the upconversion module and the digital-to-analog converter (DAC). In the feedback loop, the coupling output of the coupler passes through the downconversion module and the analog-to-digital converter (ADC) to obtain the digital baseband signal $y(n)$ of the PA output. We use the output signal $y(n)$ and input signal $x(n)$ of the PA to build the I/Q branch models of the predistorter based on the indirect learning architecture (ILA). Then, the proposed order reduction method is used to select important basis function terms in the I/Q branch models. Finally, the selected basis function terms are modeled on the main road using the lookup table (LUT) in the field-programmable gate array (FPGA), to achieve the PA's linearization.

The I/Q branch models can be constructed by the I/Q components of the traditional models or the ANN-based models. Let us take the GMP model as an example. The I/Q components of the complex basis function terms have independent contributions to the predistortion output, so the input data of the order reduction structure should contain all the real basis function terms, as shown in Equation (8), which can be written as

$$\begin{aligned}\bar{\mathbf{y}}(n) &= \left[\bar{y}_{1I}(n), \bar{y}_{2I}(n), \cdots, \bar{y}_{SI}(n), \bar{y}_{1Q}(n), \bar{y}_{2Q}(n), \cdots, \bar{y}_{SQ}(n)\right]^T \\ &= \left[y_1(n), y_2(n), \cdots, y_S(n), \cdots, y_{2S}(n)\right]^T,\end{aligned} \tag{14}$$

where $\bar{y}_{iI}(n)$ and $\bar{y}_{iQ}(n), (i = 1, 2, \cdots, S)$ are the I/Q components of the complex basis function term $\bar{y}_i(n)$, respectively. $\bar{\mathbf{y}}(n)$ is a vector with a dimension of $2S \times 1$, where $2S$ is the number of the real basis function terms. To facilitate numbering, $y_1(n), y_2(n), \cdots, y_{2S}(n)$ are used to represent these elements.

The output data of the structure contains the I/Q components of the output of the predistorter, which can be expressed as

$$\bar{\mathbf{x}}(n) = \left[x_I(n), x_Q(n)\right]^T, \tag{15}$$

where $x_I(n)$ and $x_Q(n)$ are the predistortion model output $x(n)$'s I/Q components.

To improve the fitting performance, a NN layer is used to calculate the correlation between each input and output data. Since each input needs to be calculated the correlation with all outputs, the i-th neuron in the NN layer is connected to the i-th input $y_i(n)$ and all outputs $\bar{\mathbf{x}}(n)$. The number of neurons is $2S$, corresponding to $2S$ inputs of the module. The NN layer's output can be written as

$$p_i = f_1(v_{i1}y_i(n) + [v_{i2}v_{i3}]\bar{\mathbf{x}}(n) + b_i), (i = 1, 2, \cdots, 2S), \quad (16)$$

where $\{v_{i1}, v_{i2}, v_{i3}\}$ are the weight coefficients of the i-th neuron and b_i is the bias coefficient of the i-th neuron. $f_1(\cdot)$ is the activation function, usually "tanh." The output p_i, $(i = 1, 2, \cdots, 2S)$ of the i-th neuron represents the correlation between the i-th input $y_i(n)$ and the output $\bar{\mathbf{x}}(n)$.

Then, the obtained correlation values p_i between the input and output data are converted numerically using the Softmax function. And the output of the Softmax function can be expressed as

$$q_i = \frac{\exp(p_i)}{\sum_{j=1}^{2S} \exp\left(p_j\right)}, (i = 1, 2, \cdots, 2S), \quad (17)$$

where $\sum_{i=1}^{2S} q_i = 1$. q_i, $(i = 1, 2, \cdots, 2S)$ is the form of probability, reflecting the spatial importance of the corresponding input $y_i(n)$ to the output, which is considered in this paper as the contribution of input (the basis function term) $y_i(n)$ to the generation of the output.

The inputs are weighted by the contributions of the inputs (the basis function terms), to emphasize the important inputs. The weighted inputs $y_i'(n)$ can be written as

$$y_i'(n) = q_i \times y_i(n), (i = 1, 2, \cdots, 2S). \quad (18)$$

By weighting the inputs (basis function terms) using the contributions, the important spatial details can be emphasized, and unimportant information can be weakened.

Finally, the weighted inputs $y_i'(n)$ are used to fit the output of the predistortion model. Since the I/Q branch models of the predistorter are identified separately, two coefficient vectors are used to fit the I/Q components of the predistorter. The predicted I/Q components of the predistorter can be expressed as

$$\begin{cases} \hat{x}_I(n) = \left[y_1'(n), y_2'(n), \cdots, y_{2S}'(n)\right]\mathbf{w}_I, \\ \hat{x}_Q(n) = \left[y_1'(n), y_2'(n), \cdots, y_{2S}'(n)\right]\mathbf{w}_Q, \end{cases} \quad (19)$$

where $\hat{x}_I(n)$ and $\hat{x}_Q(n)$ are the predicted I/Q components of the predistorter, respectively. $\mathbf{w}_I$ and $\mathbf{w}_Q$ are the coefficient vectors of the model, $\mathbf{w}_I = [w_{I1}, w_{I2}, \cdots, w_{\mathrm{IS}}, w_{I(S+1)}, \cdots, w_{I(2S)}]^T$, $\mathbf{w}_Q = [w_{Q1}, w_{Q2}, \cdots, w_{QS}, w_{Q(S+1)}, \cdots, w_{Q(2S)}]^T$.

The label data of the model training is output data $\bar{\mathbf{x}}(n)$ in Equation (15). By calculating the error between the predicted output and the label data, the order reduction structure can be trained. When the model converges, the contributions q_i of the inputs (the basis function terms) are obtained. Then, the real basis function terms can be sorted according to their contributions. Considering the trade-off between the model complexity and linearization effects, a contribution threshold q_0 is set. Then, the real basis terms with contributions greater than the threshold are retained, and the real basis terms with contributions less than the threshold are removed. According to the retained basis function terms, the I/Q branch models of the predistorter are modeled and identified.

3.2. Training of the Proposed Order Reduction Method. The input signal $y(n)$ and output signal $x(n)$ of the predistortion model are captured first. Then, the input data $\bar{\mathbf{y}}(n)$ of the order reduction structure is constructed according to Equation (14), and the output data $\bar{\mathbf{x}}(n)$ is obtained according to Equation (15). This paper uses 16,000 sets of input and output data to model the proposed method. The data is divided into training data and test data in a ratio of 1 : 1, which are used to train the method and test the method, respectively. The cost function of the training is set as the mean square error (MSE) function, which is written as

$$\mathrm{MSE} = \frac{1}{2N}\sum_{i=1}^{N}\left(\left(x\wedge_I(i) - x_I(i)\right)^2 + \left(x\wedge_Q(i) - x_Q(i)\right)^2\right), \quad (20)$$

where N is the number of data sets for training.

In this paper, the Adam optimization algorithm [31] is used to update the coefficients $\theta = \{v_{i1}, v_{i2}, v_{i3}, b_i, \mathbf{w}_I, \mathbf{w}_Q\}$ of the proposed structure. The updating process of coefficients can be expressed as

$$\begin{cases} \theta^{(k)} = \theta^{(k-1)} - \delta\dfrac{A^{(k)}/\left(1-\beta_1^k\right)}{\sqrt{B^{(k)}/\left(1-\beta_2^k\right)} + \varepsilon}, \\ \begin{cases} A^{(k)} = \beta_1 A^{(k-1)} + (1-\beta_1)\nabla, \\ B^{(k)} = \beta_2 B^{(k-1)} + (1-\beta_2)\nabla^2, \end{cases} \end{cases} \quad (21)$$

where ∇ is the gradient of the cost function to the coefficients and δ is the learning rate. β_1, β_2, and ε are constants.

The training process of the proposed order reduction method is shown in Algorithm 1. During the training, the attention module's output and the model output are calculated successively, and then, the cost function is calculated based on the model output and label data. According to the obtained cost function, the coefficients of the proposed method were updated using the Adam algorithm. In the next iteration, the attention module's output and the model output are calculated based on the updated model coefficients. Then, the cost function is calculated, and the coefficients are updated again. When the training times of the method reach the given iteration times, the training is finished.

When the model completes the training, the weight values q_i of the attention module output represent the contributions of the corresponding input. According to the contribution values of the basis function terms, the real basis function terms are filtered, and the retained basis function terms are obtained

$$\bar{\mathbf{y}}'(n) = \left[y_{d_1}(n), y_{d_2}(n), \cdots, y_{d_D}(n)\right]^T, \quad (22)$$

where D is the number of the retained basis function terms, and d_i, $(i = 1, 2, \cdots, D)$ satisfy $q_{d_i} \geq q_0$. Then, the

Definition:
1. Construct the structure of the order reduction method;
2. Obtain the training data, including the input and output data of the training;
3. Define MSE as the cost function of training.

Training of the method:
1. Initialization:
   1) Set the learning rate $\delta = 0.001$, and the training parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$;
   2) Initialize vector $A^{(0)} = 0$ and $B^{(0)} = 0$;
   3) Initialize coefficients $\theta^{(0)}$.
2. Training:

Loop: $k = 1, 2, \cdots$
   1) Calculate the output q_i of the attention module from Equation (17);
   2) Calculate model output from Equation (19);
   3) Calculate the cost function MSE from Equation (20);
   4) Calculate vector $A^{(k)} = \beta_1 A^{(k-1)} + (1-\beta_1)\nabla$ and $B^{(k)} = \beta_2 B^{(k-1)} + (1-\beta_2)\nabla^2$;
   5) Update coefficients $\theta^{(k)} = \theta^{(k-1)} - \delta(A^{(k)}/(1-\beta_1^k)/\sqrt{B^{(k)}/(1-\beta_2^k)} + \varepsilon)$.

End

Algorithm 1: Training of the order reduction method

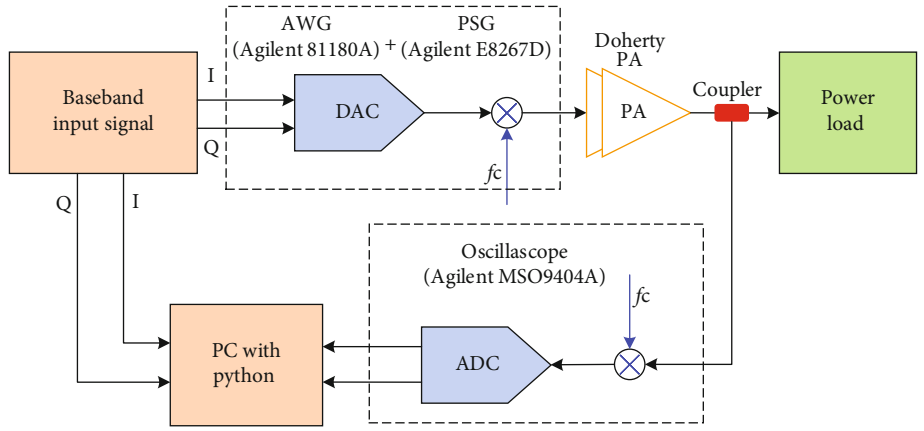

Figure 4: Experimental setup.

predistortion model is constructed using the retained basis function terms, and the predistorter coefficients are calculated by the LS algorithm.

$$\begin{cases} \widehat{\boldsymbol{\omega}}_I' = \left(\left(\bar{\mathbf{y}}'\right)^H \bar{\mathbf{y}}'\right)^{-1} \left(\bar{\mathbf{y}}'\right)^H \mathbf{x}_I, \\ \widehat{\boldsymbol{\omega}}_Q' = \left(\left(\bar{\mathbf{y}}'\right)^H \bar{\mathbf{y}}'\right)^{-1} \left(\bar{\mathbf{y}}'\right)^H \mathbf{x}_Q, \end{cases} \tag{23}$$

where $\bar{\mathbf{y}}' = [\bar{\mathbf{y}}'(N), \bar{\mathbf{y}}'(N-1), \cdots, \bar{\mathbf{y}}'(1)]^T$.

4. Experimental Setup

The experimental platform in Figure 4 is used to test the pruning effect of the order reduction method. The test signal used is an orthogonal frequency division multiplexing (OFDM) signal with a bandwidth of 100 MHz and a PAPR of 9.46 dB. In this OFDM signal, the symbol vector is modulated by 16 quadrature amplitude modulation (16-QAM). The OFDM signal is first transmitted to the Arbitrary Waveform Generator (AWG81180A), and then, the device is connected to the Performance Signal Generator (PSGE8267D) to transmit the generated baseband signal. PSG realizes digital to analog conversion and upconversion functions and then transmits the signal to Doherty PA. The PA has a center frequency of 2.14 GHz and a saturated power of 43 dBm, and the output backoff (OBO) is 6 dB. The output signal of PA is fed into the coupler.

In the feedback loop, the coupler's coupling output is connected to an oscilloscope (MSO9404A), to realize the sampling of the feedback signal. MSO9404A realizes the functions of downconversion and ADC. The sampling bandwidth is set to 500 MHz. Finally, the sampled digital baseband signal is downloaded to a personal computer (PC) to achieve predistortion design. The order reduction method of the predistortion model is constructed using the Python software's TensorFlow module on a PC. In order to verify the performance of the proposed method under different

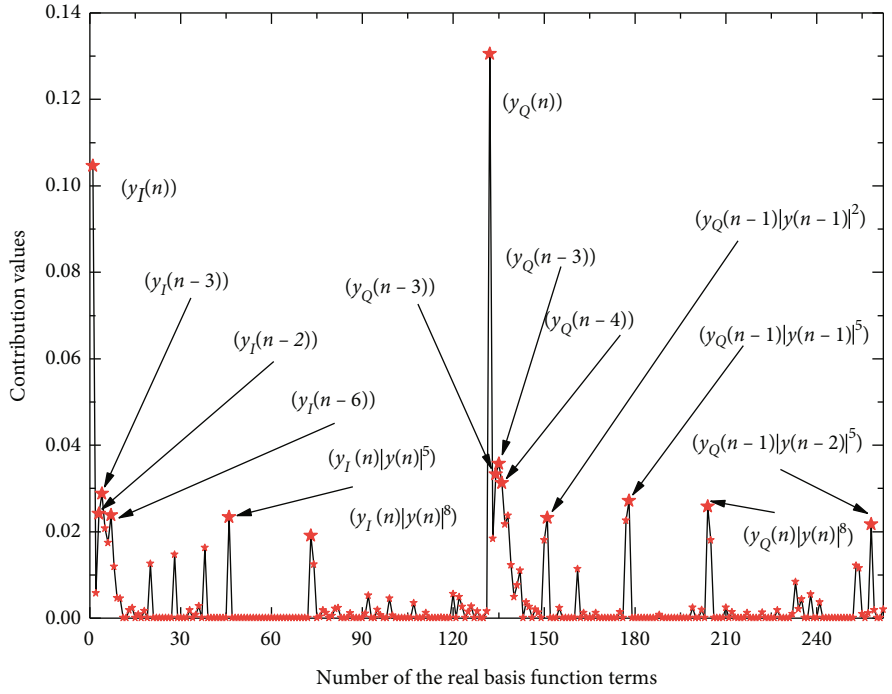

Figure 5: The contribution values of the real basis function terms.

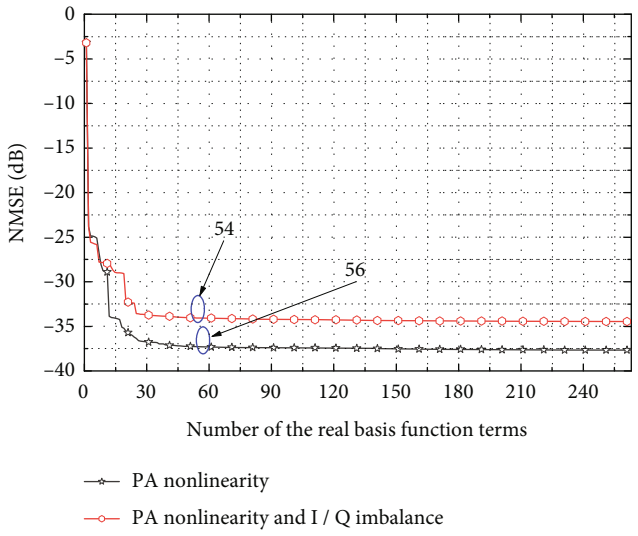

Figure 6: Comparison of NMSE performance between case A and case B at different order reduction levels.

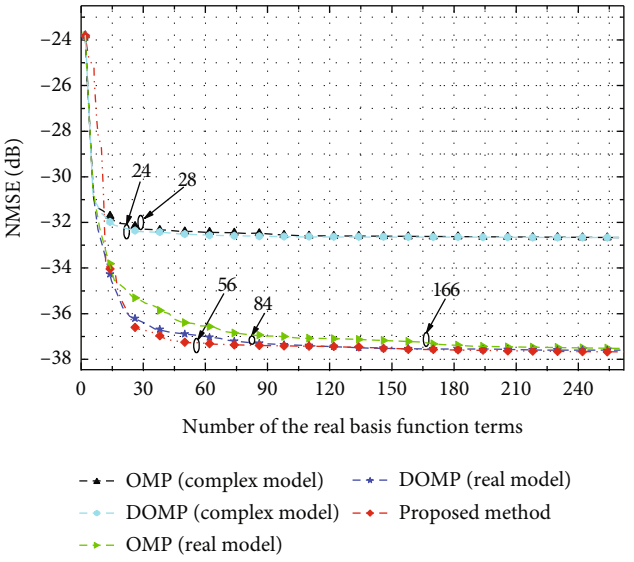

Figure 7: Comparison of NMSE performance between the proposed method and other typical methods at different order reduction levels.

conditions, we evaluate two cases of transmitter nonlinearity. Case A only contains PA's nonlinear distortion in the link, and case B contains nonlinear distortion of PA and I/Q imbalance in the link. In the I/Q imbalance, the amplitude imbalance is set to 1 dB, and the phase imbalance is set to 3 degrees.

5. Experimental Results

Figure 5 shows the contribution values of the real basis function terms to the generation of the predistortion output, where the predistortion model is modeled using the GMP model, and the model parameters $K_a = 9$, $L_a = 9$, $K_b = 5$, $L_b = 2$, $M_b = 5$, $K_c = 0$, $L_c = 0$, and $M_c = 0$. In Figure 5, the horizontal axis shows the number of the real basis function terms, in the order shown in Equation (10), and the vertical axis represents the corresponding contribution values of the real basis function terms. It can be seen from the figure that different basis function terms show different contribution values with a great difference, which is the basis of the effectiveness of the proposed method. Meanwhile, the I/Q components of some complex basis function terms all display

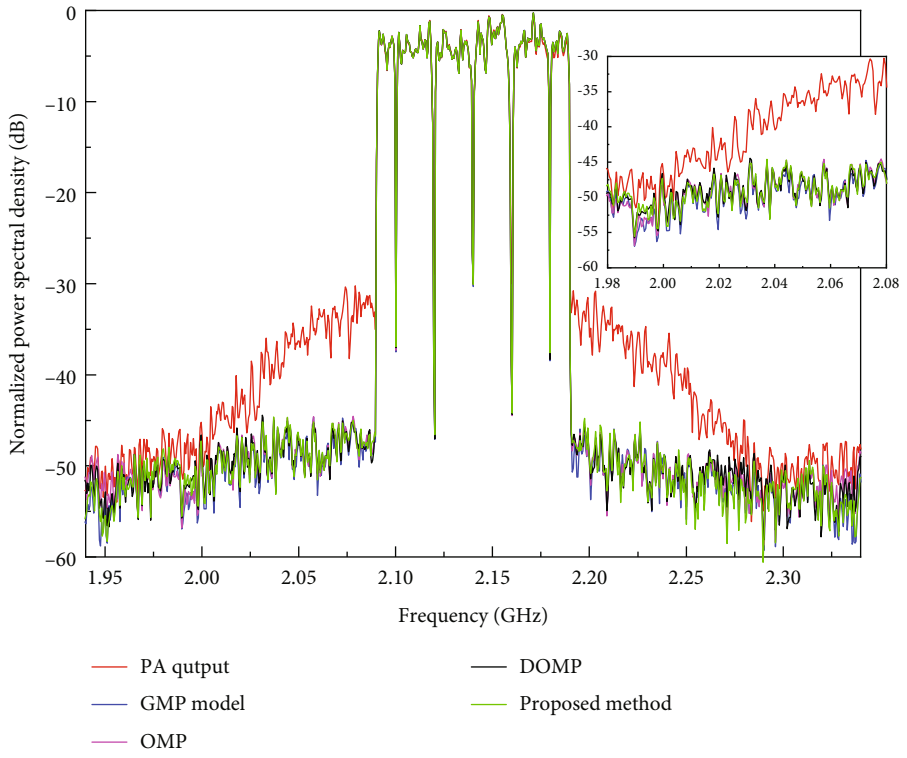

Figure 8: Comparison of the predistortion effects between the typical pruning models and the proposed pruning model.

Table 2: Comparison of linearization effects under different nonlinear conditions.

	PA nonlinearity			PA nonlinearity and I/Q imbalance		
	Num. of real basis function terms	NMSE (dB)	ACPR (±25 MHz) (dBc)	Num. Of real basis function terms	NMSE (dB)	ACPR (±25 MHz) (dBc)
No DPD	\	-8.68	-31.14/-33.24	\	-7.51	-29.75/-31.52
GMP (complex model)	262	-32.67	-44.63/-44.99	262	-22.75	-38.02/-39.94
OMP (complex model)	28	-32.30	-44.34/-44.95	8	-22.47	-38.14/-39.81
DOMP (complex model)	24	-32.30	-44.57/-44.87	6	-22.40	-38.07/-39.91
GMP (real model)	262	-37.67	-45.65/-46.97	262	-34.47	-41.29/-42.77
OMP (real model)	166	-37.28	-45.66/-46.74	188	-34.09	-41.20/-42.69
DOMP (real model)	84	-37.29	-45.53/-46.84	74	-34.09	-41.18/-42.75
Proposed pruning model	56	-37.28	-45.64/-46.81	54	-34.07	-41.21/-42.72

large contribution values, such as the I/Q components $y_I(n)$, $y_Q(n)$, $y_I(n-3)$, $y_Q(n-3)$, $y_I(n)|y(n)|^8$, $y_Q(n)|y(n)|^8$ of $y(n)$, $y(n-3)$, and $y(n)|y(n)|^8$. However, there are also some complex basis functions with only one component showing a larger contribution, such as $y_I(n-6)$, $y_I(n)|y(n)|^5$, and $y_Q(n-1)|y(n-1)|^5$, which suggests that distinguishing the contributions of the I/Q components of the complex basis function terms can further reduce model complexity.

Figure 6 compares the linearization performance of case A and case B at different order reduction levels. It can be found that in case A, the NMSE decreases rapidly with the increase of the number of the selected real basis function terms, when the number of the selected real basis function terms is less than 40. This is because the basis function terms with larger contributions are selected first and can generate most of the predistortion output. When the number of the selected real basis function terms is between 40 and 55, the NMSE decreases slowly. When the real basis function terms' number exceeds 56, the NMSE performance is barely improved. We select 56 real basis function terms, and the NMSE performance at this time can be maintained to -37.28 dB. Compared with 262 real basis terms of the GMP model, the number of the real function basis terms of the pruning method is reduced by 79%. The NMSE curve of case B shows the same trend as that of case A. However, the NMSE curve of case B has a faster decline rate as the number of the real basis function terms increases. According to the linearization performance, 54 real basis function terms with large contributions were selected to mitigate the PA's

nonlinearity and I/Q imbalance, and the NMSE at this time can be maintained to -34.07 dB. In case B, the number of the real basis function terms of the pruning model is reduced by 79%.

Figure 7 compares the NMSE of the proposed method with other methods at different order reduction levels. In OMP (real model) and DOMP (real model), the important real basis function terms for the I/Q branch models of the predistorter are calculated separately. It can be found that the linearization performance of the real model (the I/Q branch models of the predistorter) is obviously better than that of the complex model. In the real model, to achieve the same NMSE (-36.27 dB, which is 0.4 dB higher than the NMSE of the GMP model), the proposed method requires only 56 basis function terms and the basis function terms' number is reduced by 79%. However, OMP and DOMP require 166 and 84 basis function terms, respectively, which are 3 times and 1.5 times the number of the selected basis function terms of the proposed method, respectively. The reduced basis function terms represent the influence of the cross terms of the I/Q branch models.

Figure 8 compares the linearization effects of the typical pruning models and the proposed pruning model. The number of the real basis function terms selected for the proposed model is the abovementioned 56. It can be seen from the figure that the proposed pruning model reduces the adjacent channel power ratio (ACPR) of the PA output signal from -32 dBc to -46 dBc, which proves the superior pruning performance of this model. Compared with the full GMP model, the linearization effects of the proposed pruning model are almost no worse, which can be seen from the almost overlapping spectrum. Meanwhile, the linearization effects of the proposed model are almost the same as those of the OMP model and the DOMP model, but the complexity of the predistortion model is significantly reduced. The number of the real basis function terms of the proposed model is only 67% of the number of the real basis function terms of the OMP model and 34% of the number of the real basis function terms of the DOMP model.

Table 2 comprehensively compares the performance of the typical methods and the proposed pruning method in case A and case B. It can be found that in case A, the proposed pruning method achieved the NMSE of -36.28 dB (0.4 dB higher than the NMSE of the GMP model) with only 21% real basis function terms. The ACPR performance of the proposed pruning method is almost equal to that of the full GMP model. Meanwhile, the ACPR and NMSE of the proposed pruning method are almost the same as those of OMP and DOMP, using the least real basis function terms. In case B, the NMSE performance of the proposed pruning method reaches -34.07 dB (0.4 dB higher than the NMSE of the GMP model) using only 21% of the real basis function terms. Similarly, the proposed pruning method achieves nearly the same NMSE and ACPR performance as OMP and DOMP by using the least real basis function terms in case B.

Figure 9 shows the NMSE performance of the proposed order reduction model under different I/Q imbalance levels. It can be found that under different I/Q imbalance levels, the

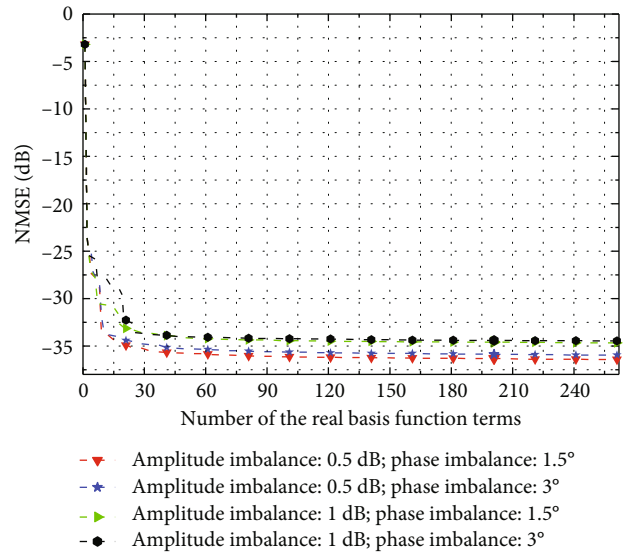

Figure 9: Comparison of NMSE performance under different I/Q imbalance levels.

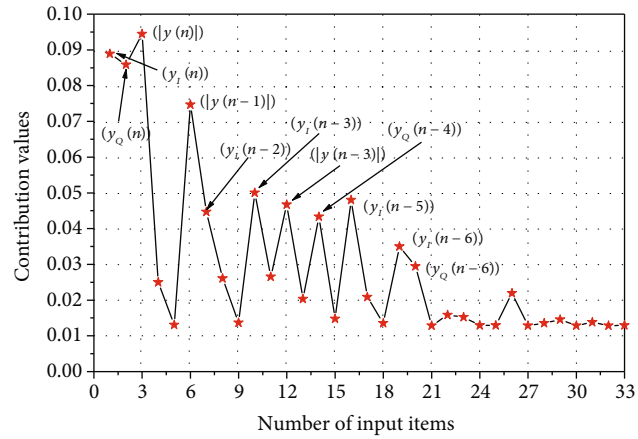

Figure 10: The contribution value of the input items in the NN-based predistortion model.

proposed order reduction model can quickly find the important real basis function terms, to construct the low-complexity predistortion model, which proves that the proposed order reduction model is suitable for different transmitter nonlinear conditions. The higher the level of the amplitude imbalance and phase imbalance, the worse the NMSE performance. However, the proposed order reduction model can almost achieve the optimal NMSE performance when the number of the selected real basis function terms is 30.

Figure 10 shows the contributions of the input items of the NN-based predistortion model, where the proposed method is used for the input term pruning of the NN-based predistortion model, and the pruning structure is described in literature [24]. It can be found that this method can get the contribution values of the input items. Then, the input items are sorted according to the contribution values. Figure 11 shows the NMSE performance under the different

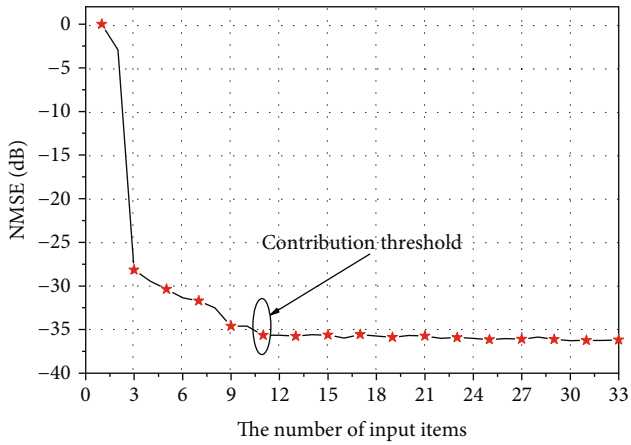

Figure 11: NMSE performance under the different number of the input items in the NN-based predistortion model.

number of input items. It can be found that with the increase of the input item's number, the NMSE performance improves rapidly. When the input item's number is 11, the NMSE performance is almost equal to that of the full model, which proves that this method is suitable for the pruning of the NN model-based model.

6. Conclusions

In this paper, an order reduction method of the predistortion model based on the attention mechanism is proposed. This method calculates the contributions of the real basis function terms to the I/Q components of the predistortion output using the attention mechanism, to select the important real basis function terms to build the I/Q branch models. The experimental results based on 100 MHz Doherty PA and I/Q imbalance verify the superior pruning performance of this method. In case A, the proposed method can prune the number of the real basis function terms to 21%, and the NMSE can be maintained to -36.3 dB. And in case B, the proposed method prunes the number of the real basis function terms to 21%, and the NMSE can be maintained to -34.1 dB. Meanwhile, to achieve almost the same ACPR and NMSE performance, the number of the basis function terms required by the proposed method is only 67% that of DOMP. In order to further reduce the complexity of the digital predistortion model in wideband systems, we will consider designing a fixed core suitable for most nonlinear transmitters and then pruning the I/Q branch models for the remaining basis function terms.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Low-orbit satellite undersampling broadband predistortion high-efficiency transmission technology (No. A2021023) and the BUPT Excellent Ph.D. Students Foundation (No. CX2020112).

References

[1] J. J. Xia, L. Fan, N. Yang et al., "Opportunistic access point selection for mobile edge computing networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 695–709, 2021.

[2] K. He, Z. Wang, D. Li, F. Zhu, and L. Fan, "Ultra-reliable MU-MIMO detector based on deep learning for 5G/B5G-enabled IoT," *Physical Communication*, vol. 43, no. 11, p. 101181, 2020.

[3] S. P. Tang, W. Zhou, L. Chen, L. Lai, J. Xia, and L. Fan, "Battery-constrained federated edge learning in UAV-enabled IoT for B5G/6G networks," *Physical Communication*, vol. 47, p. 101381, 2021.

[4] J. A. Becerra, M. J. Madero-Ayora, and C. Crespo-Cadenas, "Comparative analysis of greedy pursuits for the order reduction of wideband digital predistorters," *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, no. 9, pp. 3575–3585, 2019.

[5] X. Hu, Z. Liu, X. Yu et al., "Convolutional neural network for behavioral modeling and predistortion of wideband power amplifiers," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[6] X. Hu, Z. Liu, W. Wang, M. Helaoui, and F. M. Ghannouchi, "Low-feedback sampling rate digital predistortion using deep neural network for wideband wireless transmitters," *IEEE Transactions on Communications*, vol. 68, no. 4, pp. 2621–2633, 2020.

[7] Z. Wang, W. Chen, G. Su, F. M. Ghannouchi, Z. Feng, and Y. Liu, "Low feedback sampling rate digital predistortion for wideband wireless transmitters," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 11, pp. 3528–3539, 2016.

[8] L. Anttila, P. Handel, and M. Valkama, "Joint mitigation of power amplifier and I/Q modulator impairments in broadband direct-conversion transmitters," *IEEE Transactions on Microwave Theory and Techniques*, vol. 58, no. 4, pp. 730–739, 2010.

[9] J. A. Becerra, M. J. M. Ayora, J. Reina-Tosina, and C. Crespo-Cadenas, "Sparse identification of Volterra models for power amplifiers without pseudoinverse computation," *IEEE Transactions on Microwave Theory and Techniques*, vol. 68, no. 11, pp. 4570–4578, 2020.

[10] A. Abdelhafiz, A. Kwan, O. Hammi, and F. M. Ghannouchi, "Digital predistortion of LTE-A power amplifiers using compressed-sampling-based unstructured pruning of Volterra series," *IEEE Transactions on Microwave Theory and Techniques*, vol. 62, no. 11, pp. 2583–2593, 2014.

[11] J. Reina-Tosina, M. Allegue-Martinez, C. Crespo-Cadenas, C. Yu, and S. Cruces, "Behavioral modeling and predistortion of power amplifiers under sparsity hypothesis," *IEEE Transactions on Microwave Theory and Techniques*, vol. 63, no. 2, pp. 745–753, 2015.

[12] 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, 2001.

[13] 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, 2006.

[14] D. Lopez-Bueno, Q. A. Pham, G. Montoro, and P. L. Gilabert, "Independent digital predistortion parameters estimation using adaptive principal component analysis," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 12, pp. 5771–5779, 2018.

[15] A. Molina, K. Rajamani, and K. Azadet, "Digital predistortion using lookup tables with linear interpolation and extrapolation: direct least squares coefficient adaptation," *IEEE Transactions on Microwave Theory and Techniques*, vol. 65, no. 3, pp. 980–987, 2017.

[16] P. L. Gilabert, A. Cesari, G. Montoro, E. Bertran, and J. M. Dilhac, "Multi-lookup table FPGA implementation of an adaptive digital predistorter for linearizing RF power amplifiers with memory effects," *IEEE Transactions on Microwave Theory and Techniques*, vol. 56, no. 2, pp. 372–384, 2008.

[17] J. A. Becerra, M. J. Madero-Ayora, J. Reina-Tosina, C. Crespo-Cadenas, J. Garcia-Frias, and G. Arce, "A doubly orthogonal matching pursuit algorithm for sparse predistortion of power amplifiers," *IEEE Microwave and Wireless Components Letters*, vol. 28, no. 8, pp. 726–728, 2018.

[18] W. Chen, S. Zhang, Y. J. Liu, F. M. Ghannouchi, Z. Feng, and Y. Liu, "Efficient pruning technique of memory polynomial models suitable for PA behavioral modeling and digital predistortion," *IEEE Transactions on Microwave Theory and Techniques*, vol. 62, no. 10, pp. 2290–2299, 2014.

[19] J. A. Becerra, M. J. Madero-Ayora, J. Reina-Tosina, C. Crespo-Cadenas, J. Garcia-Frias, and G. Arce, "A reduced-complexity doubly orthogonal matching pursuit algorithm for power amplifier sparse behavioral modeling," in *2019 IEEE Topical Conference on RF/Microwave Power Amplifiers for Radio and Wireless Applications (PAWR)*, pp. 1–3, Orlando, FL, USA, January 2019.

[20] J. Peng, S. He, Z. Dai, and B. Wang, "A simplified sparse parameter identification algorithm suitable for power amplifier behavioral modeling," *IEEE Microwave and Wireless Components Letters*, vol. 27, no. 3, pp. 290–292, 2017.

[21] D. Wang, M. Aziz, M. Helaoui, and F. M. Ghannouchi, "Augmented real-valued time-delay neural network for compensation of distortions and impairments in wireless transmitters," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 242–254, 2019.

[22] S. Lajnef, N. Boulejfen, A. Abdelhafiz, and F. M. Ghannouchi, "Two-dimensional Cartesian memory polynomial model for nonlinearity and I/Q imperfection compensation in concurrent dual-band transmitters," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 63, no. 1, pp. 14–18, 2016.

[23] H. Y. Cao, A. Soltani Tehrani, C. Fager, T. Eriksson, and H. Zirath, "I/Q imbalance compensation using a nonlinear modeling approach," *IEEE Transactions on Microwave Theory and Techniques*, vol. 57, no. 3, pp. 513–518, 2009.

[24] Z. Liu, X. Hu, T. Liu, X. Li, W. Wang, and F. M. Ghannouchi, "Attention-based deep neural network behavioral model for wideband wireless power amplifiers," *IEEE Microwave and Wireless Components Letters*, vol. 30, no. 1, pp. 82–85, 2020.

[25] 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, pp. 594–603, 2012.

[26] Lei Ding, Zhengxiang Ma, D. R. Morgan, M. Zierdt, and J. Pastalan, "A least-squares/Newton method for digital predistortion of wideband signals," *IEEE Transactions on Communications*, vol. 54, no. 5, pp. 833–840, 2006.

[27] J. J. Xia, D. Deng, and D. Fan, "A note on implementation methodologies of deep learning-based signal detection for conventional MIMO transmitters," *IEEE Transactions on Broadcasting*, vol. 66, no. 3, pp. 744-745, 2020.

[28] J. Xia, L. Fan, W. Xu et al., "Secure cache-aided multi-relay networks in the presence of multiple eavesdroppers," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7672–7685, 2019.

[29] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Advances in Neural Information Processing Systems*, pp. 577–585, Cambridge, MA, USA, 2015.

[30] W. Xu, R. Chen, B. Huang, and Q. Zhou, "Enhanced context attention network for image super resolution," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11665–11673, 2021.

[31] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proc. 3rd International Conference for Learning Representations (ICLR 2015)*, pp. 7–9, San Diego, USA, 2015.