Mixture of Experts Approach for Piecewise Modeling and Linearization of RF Power Amplifiers

Alberto Brihuega, *Graduate Student Member, IEEE*, Mahmoud Abdelaziz, *Member, IEEE*, Lauri Anttila, *Member, IEEE*, Yue Li, *Member, IEEE*, Anding Zhu, *Senior Member, IEEE*, and Mikko Valkama, *Senior Member, IEEE*

***Abstract*— Piecewise behavioral models are commonly adopted for modeling and linearization of RF power amplifiers (PAs) that exhibit strong amplitude-dependent nonlinear distortion characteristics, as global polynomial approximations tend to underperform in such scenarios. In this article, we consider a new piecewise model for PAs based on the mixture of experts (ME) approach, which builds on a probabilistic model that allows the different submodels to cooperate—as opposed to operating in an independent fashion that is commonly the case in existing reference methods. We first introduce the ME framework theory while also extend it such that it can be applied to model complex baseband signals and nonlinearities. Then, we show how the ME model allows overcoming some of the intrinsic shortcomings that existing piecewise behavioral models commonly exhibit, which translates into improved modeling accuracy and improved linearization performance. Furthermore, the extension of the ME approach to a tree-structured regression model, referred to as the hierarchical ME model, is also introduced and shown to provide further performance improvements over the basic ME approach. The proposed solutions are validated with extensive RF measurements, covering both PA direct modeling and digital predistortion (DPD)-based linearization, on a gallium nitride (GaN) load-modulated balanced PA, on a GaN Doherty PA, and on a class AB GaN high electron mobility transistor PA, while being compared against several state-of-the-art piecewise methods. The results demonstrate that the ME framework-based models outperform the state of the art.**

***Index Terms*— Behavioral modeling, digital predistortion (DPD), 5G New Radio (NR), mixture of experts (ME), nonlinear distortion, piecewise models, power amplifiers (PAs).**

I. Introduction

OVER the years, multiple power amplifier (PA) technologies have been developed with the goal of delivering enhanced power efficiency at different power back-off levels and over wide bandwidths [1]–[4]. Good examples are the Doherty PA (DPA) [5], [6] and the load-modulated balanced (LMBA) PA [2], [7], which leverages the concept of load modulation that allows the power efficiency to be optimized dynamically at a specific power back-off, by tuning the load impedance. In order to further increase the power efficiency, digital predistortion (DPD) solutions are commonly deployed to compensate for the strong nonlinear distortion that originates inside the PAs while being operated with high efficiency [4], [8]–[11]. However, due to the operation principle of DPAs and LMBA PAs, their nonlinear distortion characteristics become strongly amplitude-dependent. This makes their modeling and linearization through classical global polynomials (GPs) very challenging due to the global dependence on local effects [12].

Piecewise models, on the other hand, utilize separate submodels that operate over specific subregions of the overall PA response [13], [14]. Thus, they are capable of conveniently modeling such distinct amplitude-dependent behavior. Another important feature of piecewise models is the fact that global dependence on local effects can be largely avoided [12]. Consequently, piecewise models are well suited to model more complicated nonlinearities, and a number of piecewise models have been proposed in the literature [13]–[16]. To this end, a vector-switched (VS) model was proposed in [13] and is based on hard partitions of the PA input signal space, which defines the range of operation of each of the submodels. Zhu *et. al.* [15] proposed a decomposed piecewise (DPW) Volterra model, where each transmit sample is decomposed into several subsamples that are then processed by the different submodels before the final sample is reconstructed. A piecewise behavioral model based on a vector rotation decomposition of the canonical piecewise linear (CPWL) basis functions (BFs), referred to as DVR model, was proposed in [16] and was shown to require less amount of coefficients when modeling systems with non-Volterra-like behavior, e.g., as that exhibited by DPA or LMBA PAs.

Despite the noted piecewise models provide significantly better modeling accuracy than GPs, they have some inherent limitations. Specifically, the VS model does not impose any continuity constraint between the submodels, potentially compromising its performance [14], [16]. The DVR model in [16] considers an approximation of the original CPWL BFs

Manuscript received March 31, 2021; revised June 14, 2021; accepted June 17, 2021. Date of publication July 30, 2021; date of current version January 5, 2022. This work was supported in part by the Academy of Finland under Project 332361, Project 304147, Project 301820, and Project 319994, in part by Nokia Bell Labs, and in part by the Tampere University Doctoral School. This article is an expanded version from the RWW2021, January 2021 [22]. *(Corresponding author: Alberto Brihuega.)*

Alberto Brihuega, Lauri Anttila, and Mikko Valkama are with the Department of Electrical Engineering, Tampere University, 33100 Tampere, Finland (e-mail: alberto.brihuegagarcia@tuni.fi).

Mahmoud Abdelaziz is with the Zewail City of Science and Technology, Giza 12578, Egypt.

Yue Li and Anding Zhu are with the School of Electrical and Electronic Engineering, University College Dublin, D04 V1W8 Dublin 4, Ireland.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TMTT.2021.3098867.

Digital Object Identifier 10.1109/TMTT.2021.3098867

so that the model is linear in parameters, which may limit its performance. In addition, in general, memory modeling capabilities may be compromised in piecewise models as the different submodels operate independently, whereas memory effects may involve samples belonging to different subregions. As the signal BW increases, which is a general trend in wireless communication systems, such as 5G New Radio (NR) [17], complicated nonlinearities and memory effects are likely to appear, and hence, more robust models are needed.

In this article, we propose a new piecewise behavioral/DPD model for RF PAs based on the so-called mixture of experts (ME) framework, originally proposed in the context of learning theories in [18] with some modifications introduced in [19] and [20]. Furthermore, a good review of the ME theory and its applications can be found in [21]. The ME model is a probabilistic framework that allows to combine multiple regression functions, the so-called experts, and make them cooperate with a gating function to solve a regression or classification problem. Since its introduction, different experts based on support vector machine, Gaussian processes, or hidden Markov models, among others, have been considered and shown to provide systematically better performance when combined with ME [21]. ME is of special interest in the context of regression with piecewise data, or with data containing different patterns, where a given expert can focus on a specific pattern, generally providing better accuracy than the individual experts.

Motivated by the above, the ME approach can be of special interest also in the field of PA modeling and linearization, as shown in our early work in [22]. One distinct feature of the ME model compared to other piecewise models is the fact that ME utilizes soft partitions of the data. This implies that the submodels work across overlapping regions. This is a very important feature, as it avoids potential nonsmooth transitions between submodels, and facilitates the modeling of memory effects between regions. Furthermore, the soft partitions or gating networks are themselves nonlinear, which can enhance the overall nonlinear modeling capabilities. In addition, more sophisticated decision boundaries and regions can be defined by implementing a tree-structured regression model, referred to as the hierarchical ME (HME) model [20]. Hence, the ME framework stands as a very flexible and capable solution for modeling and linearization of RF PAs.

In this article, we extended our preliminary work in [22], where the basic single-layer ME model was considered for direct modeling of RF PAs. The main contributions of this article can be summarized and described as follows.

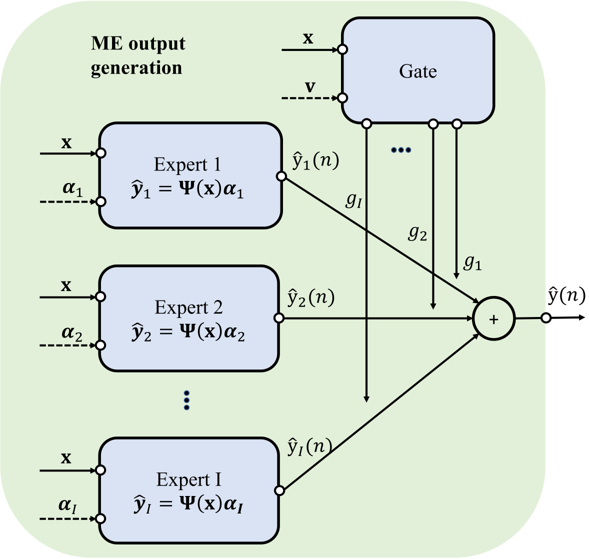

ME output generation
x
v
Gate
x
$\boldsymbol{\alpha}_1$
Expert 1
$\hat{\mathbf{y}}_1 = \boldsymbol{\Psi}(\mathbf{x})\boldsymbol{\alpha}_1$
$\hat{y}_1(n)$
g_I
g_2
g_1
x
$\boldsymbol{\alpha}_2$
Expert 2
$\hat{\mathbf{y}}_2 = \boldsymbol{\Psi}(\mathbf{x})\boldsymbol{\alpha}_2$
$\hat{y}_2(n)$
+
$\hat{y}(n)$
x
$\boldsymbol{\alpha}_I$
Expert I
$\hat{\mathbf{y}}_I = \boldsymbol{\Psi}(\mathbf{x})\boldsymbol{\alpha}_I$
$\hat{y}_I(n)$

Fig. 1. Block diagram illustrating the ME principle for estimating or approximating $y(n)$.

1) The ME framework is proposed for modeling and linearization of RF PAs. The classical ME theory that is commonly applied to model real-valued data is extended so that it can be applied to model complex baseband signals and nonlinearities. The ME fundamentals are carefully revisited, and the training algorithm to learn the parameters of the model is detailed.
2) The extension of the ME model to a multilevel regression tree is introduced and shown to provide better nonlinear modeling capabilities and linearization performance than more ordinary single-layer ME due to the stronger nonlinear behavior of the composite gating network.
3) Extensive sets of measurement results on a number of different PA technologies are reported to validate and showcase the capabilities of the ME framework in the context of PA direct modeling and DPD-based linearization. The ME model is compared against several state-of-the-art piecewise models in terms of complexity, modeling accuracy, and linearization performance.

The rest of this article is organized as follows. The ME theory and its extension to model complex baseband signals and nonlinearities are introduced in Section II, covering also the tree-structure HME model. The algorithm to train the parameters of the ME and HME model is described in III. The complexity analysis of the proposed models and corresponding comparisons against selected state-of-the-art piecewise models are provided in Section IV, whereas RF measurement results and their analysis are reported in Section V. Finally, Section VI provides the main concluding remarks.

II. Mixture of Experts Framework for PA Modeling and Linearization

A. Basic ME Model

In general, linear-in-the-parameters models are preferred in PA modeling and linearization as they can be trained by utilizing simple linear regression techniques, e.g., the least-squares (LS) fit or gradient-based methods, such as the least-mean-squares algorithm [23]. For instance, polynomial-based models from the Volterra-series family [8], [14], [15] or the modified CPWL BFs in [16] are good examples of such models and are very widely adopted in the literature. In this work, in the context of the ME framework, we, thus, also consider linear-in-the-parameters experts, more specifically, polynomial-based experts, while note that any other expert model can basically be adopted.

Let now $x(n)$ and $y(n)$ denote the I/Q samples of the input and target signals, respectively, in the context of the ME framework illustrated in Fig. 1. Considering a linear-in-the-parameters model, the target signal can be estimated or approximated as

$$\hat{\mathbf{y}} = \mathbf{\Psi}(\mathbf{x})\boldsymbol{\alpha} \tag{1}$$

where $\hat{\mathbf{y}} = [\hat{y}(1), \hat{y}(2), \ldots, \hat{y}(N)]^T$, with $\hat{y}(n)$ being an estimate of $y(n)$, and $\mathbf{x} = [x(1), x(2), \ldots, x(N)]^T$. Furthermore, $\mathbf{\Psi}(\mathbf{x}) \in \mathbb{C}^{N\times B}$ is the matrix containing the regressors of the model, B is the total number of regressors, and $\boldsymbol{\alpha} \in \mathbb{C}^{B\times 1}$ are the model coefficients. In the context of PA direct modeling, $\mathbf{x}$ is the PA input signal, whereas $\mathbf{y}$ corresponds to the PA output signal. On the other hand, when adopting the ME model as a postdistorter in the context of the indirect learning architecture (ILA) [24], $\mathbf{x}$ and $\mathbf{y}$ correspond to the PA output (divided by the target gain) and PA input data, respectively.

Assuming N statistically independent data points and I experts, the ME model can be formulated as the following decomposition of the input/output data [19], [21]:

$$P(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{N}\sum_{i=1}^{I} P(z_i(n)=1|x(n), \mathbf{v}_i)P(y(n)|x(n), \mathbf{w}_i) \tag{2}$$

where $z_i(n)$ is a hidden/latent variable and $P(z_i(n) = 1|x(n), \mathbf{v}_i)$ is the gating function of parameters $\mathbf{v}_i$, measuring the probability of the ith expert given the input. $P(y(n)|x(n), \mathbf{w}_i)$, in turn, denotes the probability of the ith expert, with parameters $\mathbf{w}_i$, for generating $y(n)$.

In general, the gating function can adopt multiple forms, the so-called mixture model being the most commonly adopted one [19], [21]. Such mixture model is defined as a convex sum of different density functions $P(\mathbf{x}|\boldsymbol{\pi}, \mathbf{v}_i)$ [25], that is,

$$P(\mathbf{x}|\boldsymbol{\pi}, \mathbf{v}) = \sum_{i=1}^{I} \pi_i P(\mathbf{x}|\mathbf{v}_i) \tag{3}$$

where π_i are the so-called mixing probabilities that sum up to one. By invoking Bayes' rule and the total probability theorem, the gating functions $P(z_i(n) = 1|x(n), \mathbf{v}_i)$ can be expressed as

$$P(z_i(n) = 1|x(n), \mathbf{v}_i) = \frac{a_i P(x(n)|z_i(n) = 1, \mathbf{v}_i)}{\sum_{j=1}^{I} a_j P\big(x(n)|\mathbf{v}_j\big)} \tag{4}$$

where $a_i / \sum_{j=1}^{I} a_j P(x(n)|\mathbf{v}_j)$ is the effective mixing probability, $a_i = P(z_i(n) = 1)$ is the prior probability of the ith gate, and $\sum_i a_i = 1$. In this work, $P(x(n)|\mathbf{v}_i)$ is considered to be a density among the exponential family, and specifically a Gaussian density, which allows obtaining $\mathbf{v}_i$ in the closed form [19]. Furthermore, as the PA nonlinearities act on the envelope of the transmit signal, the gates are assumed to make soft partitions based on the amplitude of the input signal, denoted as $A(n) = |x(n)|$, similar to other piecewise models [13], [14]. Hence, in the following, $P(z_i(n) = 1|x(n), \mathbf{v}_i)$ will be expressed as a function of $A(n)$ rather than of $x(n)$.

As for the experts, they are also commonly chosen from the exponential family so that their parameters can be obtained in closed form too. In this work, the experts are assumed to be Gaussian distributed, i.e.,

$$P(y(n)|x(n), \mathbf{w}_i) = \mathcal{N}\big(y(n)|\mathbf{\Psi}_i(x(n))\boldsymbol{\alpha}_i, \sigma_{e_i}^2\big) \tag{5}$$

where $\mathbf{w}_i = \{\boldsymbol{\alpha}_i, \sigma_{e_i}^2\}$, $\hat{y}_i(n) = \mathbf{\Psi}_i(x(n))\boldsymbol{\alpha}_i$ is the mean, and $\sigma_{e_i}^2$ is the variance. As $y(n)$ is complex-valued, the probability density function reads

$$P(y(n)|x(n), \mathbf{w}_i) = \frac{1}{\pi\sigma_{e_i}^2}\exp\left(-\frac{|y(n) - \hat{y}(n)|^2}{\sigma_{e_i}^2}\right). \tag{6}$$

It is important to note that, for a perfect sample estimate, i.e., $\hat{y}(n) = y(n)$, $P(y(n)|x(n), \mathbf{w}_i)$ reaches its maximum value. This is one of the probabilities that the training algorithm will try to maximize, resulting in the ME model iteratively yielding better sample estimates.

In order for the ME model to make a single prediction, the expectation of (2) is used, given as [21]

$$\hat{y}(n) = \sum_{i=1}^{I} g_i(A(n), \mathbf{v}_i)\hat{y}_i(n) \tag{7}$$

which is a weighted sum of the outputs of the estimates of the individual experts, and where $g_i(A(n), \mathbf{v}_i) = P(z_i(n) = 1|A(n), \mathbf{v}_i)$.

B. Hierarchical ME Model

The nonlinear modeling capabilities of the ME model can, in general, be enhanced by making the experts more nonlinear, e.g., by increasing the nonlinear order of the polynomial-based regression functions. However, high-order polynomials commonly have poor extrapolation properties and may easily overfit the data [26]. Alternatively, the gating networks are nonlinear too; hence, it is possible to make them more nonlinear by considering that the experts themselves are ME models. This approach results in an HME model [20], [21], which can be thought of as a tree-structured regression system, which essentially adds additional nonlinear decision layers. The leaves of the tree model contain the experts, whereas the nonterminal nodes of the tree contain the gating functions.

An example two-level HME model is depicted in Fig. 2. The gating network consists now of two layers, and each of them must take into consideration the nodes beneath. Considering a two level decision tree, the probabilistic model in (2) can be rewritten as follows [20]:

$$\begin{aligned} P(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{N}\Bigg(&\sum_{j=1}^{J} P\big(z_j(n) = 1|A(n), \mathbf{v}_j\big) \\ &\times \sum_{i=1}^{I_j} P\big(z_j(n) = 1|z_i(n) = 1|A(n), \mathbf{v}_{ij}\big) \\ &\times P\big(y(n)|x(n), \mathbf{w}_{ij}\big)\Bigg) \end{aligned} \tag{8}$$

where $P(z_j(n) = 1|A(n), \mathbf{v}_j)$ is the probability of selecting the jth gating network in the top layer given the current input sample, whereas J is the number of nodes in the top layer. I_j stands for the number of experts connected to the jth gating network, and $P(z_j(n) = 1|z_i(n) = 1|A(n), \mathbf{v}_{ij})$ is the ith

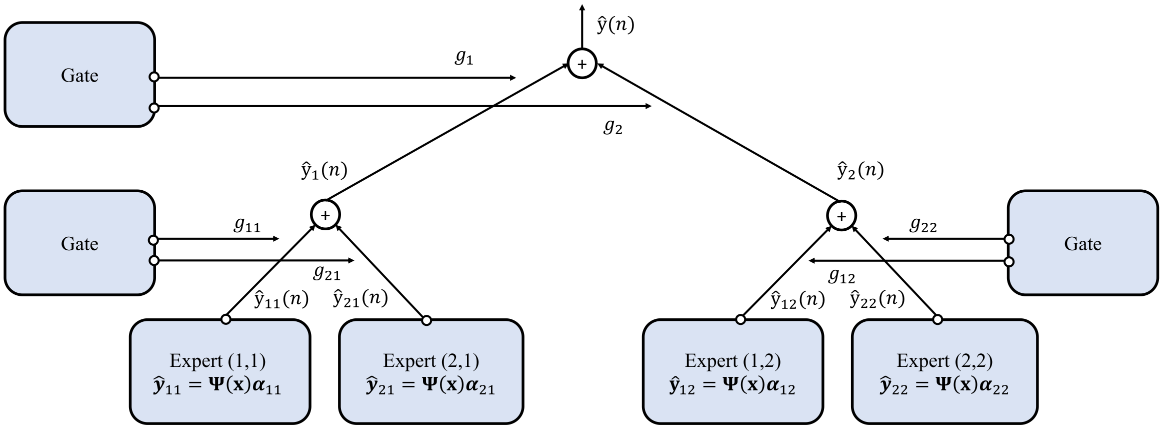
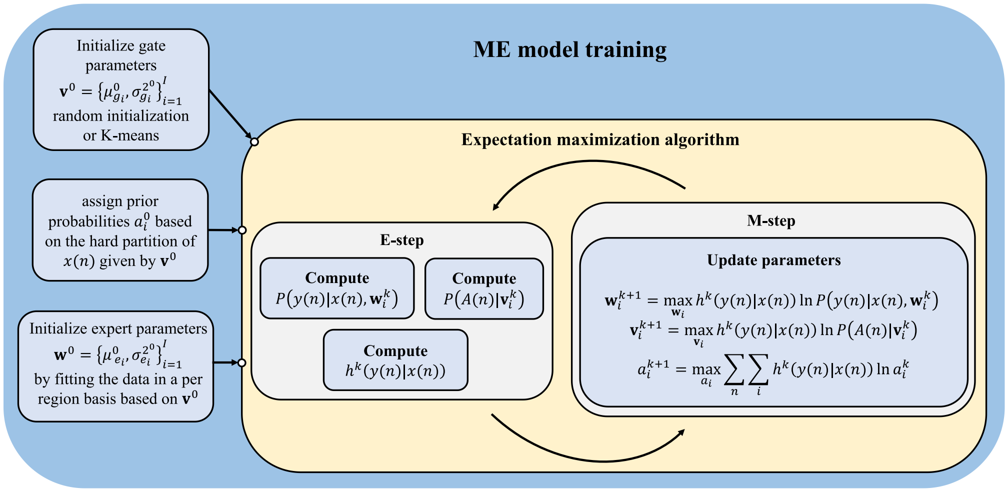

$\hat{y}(n)$
Gate
g_1
g_2
$\hat{y}_1(n)$
$\hat{y}_2(n)$
Gate
g_{11}
g_{21}
$\hat{y}_{11}(n)$
$\hat{y}_{21}(n)$
g_{22}
g_{12}
$\hat{y}_{12}(n)$
$\hat{y}_{22}(n)$
Gate
Expert (1,1)
$\hat{\mathbf{y}}_{11} = \mathbf{\Psi}(\mathbf{x})\boldsymbol{\alpha}_{11}$
Expert (2,1)
$\hat{\mathbf{y}}_{21} = \mathbf{\Psi}(\mathbf{x})\boldsymbol{\alpha}_{21}$
Expert (1,2)
$\hat{\mathbf{y}}_{12} = \mathbf{\Psi}(\mathbf{x})\boldsymbol{\alpha}_{12}$
Expert (2,2)
$\hat{\mathbf{y}}_{22} = \mathbf{\Psi}(\mathbf{x})\boldsymbol{\alpha}_{22}$

Fig. 2. High-level illustration of the HME principle with two layers.


ME model training

Initialize gate parameters $\mathbf{v}^0 = \{\mu^0_{g_i}, \sigma^{2^0}_{g_i}\}_{i=1}^{I}$ random initialization or K-means

assign prior probabilities a_i^0 based on the hard partition of $x(n)$ given by $\mathbf{v}^0$

Initialize expert parameters $\mathbf{w}^0 = \{\mu^0_{e_i}, \sigma^{2^0}_{e_i}\}_{i=1}^{I}$ by fitting the data in a per region basis based on $\mathbf{v}^0$

Expectation maximization algorithm

E-step

Compute $P\big(y(n)|x(n), \mathbf{w}_i^k\big)$

Compute $P\big(A(n)|\mathbf{v}_i^k\big)$

Compute $h^k(y(n)|x(n))$

M-step

Update parameters

$$\mathbf{w}_i^{k+1} = \max_{\mathbf{w}_i} h^k(y(n)|x(n)) \ln P\big(y(n)|x(n), \mathbf{w}_i^k\big)$$

$$\mathbf{v}_i^{k+1} = \max_{\mathbf{v}_i} h^k(y(n)|x(n)) \ln P\big(A(n)|\mathbf{v}_i^k\big)$$

$$a_i^{k+1} = \max_{a_i} \sum_n \sum_i h^k(y(n)|x(n)) \ln a_i^k$$

Fig. 3. Block diagram of the EM algorithm for ME model training.

gating function at the bottom layer in branch j and is also of the form of (4). The model formulation for an arbitrary tree depth is done in a similar fashion. Similarly, we may rewrite (7) as

$$\hat{y}(n) = \sum_{j=1}^{J} g_j\big(A(n), \mathbf{v}_j\big) \sum_{i=1}^{I_j} g_{i|j}\big(A(n), \mathbf{v}_{ij}\big)\hat{y}_{ij}(n) \tag{9}$$

where $g_{i|j}(A(n), \mathbf{v}_{ij})$ is used as shorthand for $P(z_j(n) = 1|z_i(n) = 1|A(n), \mathbf{v}_{ij})$.

In Section III, the algorithm to train the ME model parameters is described.

III. ME Parameter Learning: The Expectation–Maximization Algorithm

A. EM Algorithm for the Basic ME Model

To train the ME model, the expectation–maximization (EM) algorithm is usually considered [19], [20], [27]. EM is an iterative algorithm that calculates the maximum-likelihood (ML) parameters of a probabilistic model, in which some variables are observed and others are hidden/latent. For simplicity, the EM algorithm is formulated in the context of the basic ME, that is, a single layer model, whereas specific steps for training the HME model are detailed after the basic concepts are introduced.

The observable data are the input and target vectors $\mathbf{x}$ and $\mathbf{y}$, whereas it is unknown which expert generated each data point, formally expressed through the latent variable $\mathbf{z}$. As discussed above, in order for the gate and expert parameters to be analytically solvable, the densities must belong to the family of the exponential densities, and additionally, instead of the likelihood function in (2), one should consider the joint density $P(\mathbf{x}, \mathbf{y}) = P(\mathbf{y}|\mathbf{x})P(\mathbf{x})$ [21], which reads

$$P(\mathbf{x}, \mathbf{y}) = \prod_{n=1}^{N} \sum_{i=1}^{I} a_i g_i(A(n), \mathbf{v}_i) P(y(n)|x(n), \mathbf{w}_i). \tag{10}$$

The joint density essentially allows canceling out the denominator of the gating function in (4), which makes the optimization analytically solvable.

In order to train the ME parameters, the ML is calculated for $\ln P(\mathbf{x}, \mathbf{y}, |\mathbf{v}, \mathbf{w})$ and is done by iterating the EM algorithm [21], [27], which consists of the following two steps.

1) *E-Step:* In the kth iteration of the E-step, the expectation of the latent variables $h_i^k(y(n)|x(n)) = \mathbb{E}\{P(z(n)|y(n), x(n))\}$ is computed as

$$h_i^k(y(n)|x(n)) = \frac{a_i^k g_i\left(A(n), \mathbf{v}_i^k\right) P\left(y(n)|x(n), \mathbf{w}_i^k\right)}{\sum_j a_j^k g_j\left(A(n), \mathbf{v}_j^k\right) P\left(y(n)|x(n), \mathbf{w}_j^k\right)} \tag{11}$$

which measures the relative probability of $x(n)$ belonging to expert i, commonly referred to as membership probability or responsibility.

2) *M-Step:* Compute the maximum likelihood parameters weighted by the membership probabilities [19], expressed as

$$\begin{aligned} \mathbf{w}_i^{k+1} &= \arg\max_{\mathbf{w}_i} \sum_n h_i^k(y(n)|x(n)) \ln P\left(y(n)|x(n), \mathbf{w}_i^k\right) \\ \mathbf{v}_i^{k+1} &= \arg\max_{\mathbf{v}_i} \sum_n h_i^k(y(n)|x(n)) \ln P\left(A(n)|\mathbf{v}_i^k\right) \\ a_i^{k+1} &= \arg\max_{a_i} \sum_n \sum_i h_i^k(y(n)|x(n)) a_i^k. \end{aligned} \tag{12}$$

To compute the ML parameters in (12), one needs to differentiate with respect to the parameters and solve for them. As the densities are considered to be Gaussian, the parameters can be calculated in a straightforward manner. The gate parameters $\mathbf{v} = \{\mu_{g_i}, \sigma_{g_i}^2\}$ are given by

$$\begin{aligned} \mu_{g_i}^{k+1} &= \frac{\sum_n h_i^k(y(n)|x(n)) x(n)}{\sum_n h_i^k(y(n)|x(n))} \\ \sigma_{g_i}^{2^{k+1}} &= \frac{\sum_n h_i^k(y(n)|x(n)) \left(x(n) - \mu_{g_i}^{k+1}\right)^2}{\sum_n h_i^k(y(n)|x(n))} \end{aligned} \tag{13}$$

which are essentially the ML estimates of the mean and variance of a Gaussian distribution, i.e., the sample mean and the sample variance but weighted by the membership probabilities.

Similarly, the expert parameters $\mathbf{w}_i = \{\boldsymbol{\alpha}_i, \sigma_{e_i}^2\}$ are calculated as

$$\begin{aligned} \boldsymbol{\alpha}_i^{k+1} &= \left(\boldsymbol{\Psi}^H(\mathbf{x}) \mathbf{W}_i^k \boldsymbol{\Psi}(\mathbf{x})\right)^{-1} \boldsymbol{\Psi}^H(\mathbf{x}) \mathbf{W}_i^k \mathbf{y} \\ \sigma_{e_i}^{2^{k+1}} &= \frac{\sum_n h_i^k(y(n)|x(n)) |y(n) - \hat{y}(n)|^2}{\sum_n h_i^k(y(n)|x(n))} \end{aligned} \tag{14}$$

where $\mathbf{W}_i^k \in \mathbb{R}^{N \times N}$ is a diagonal matrix containing the responsibilities $h_i^k(y(n)|x(n))$, $n = 1, 2, \ldots, N$. The expression for $\boldsymbol{\alpha}_i^{k+1}$ is of the form of a weighted least-squares solution, where the responsibilities allow the expert parameters to be trained by giving more relevance to the samples that lie on the span of the corresponding expert. Alternatively, one could calculate the first- and second-order derivatives with respect to $\boldsymbol{\alpha}$ and approximate the closed-form solution with an iterative algorithm based on gradient-descent. This is the common approach when the model parameters are not analytically solvable, e.g., when soft-max gating networks are adopted [18].

Finally, the prior probabilities are updated as [19]

$$a_i^{k+1} = \frac{1}{N} \sum_n h_i^k(y(n)|x(n)) \tag{15}$$

which represents the average membership probability of the ith expert, or in other words, the proportion of the data that are assigned to the ith model.

Algorithm 1 EM Algorithm

```
**Inputs**: $\mathbf{x}$, $\mathbf{y}$, $\boldsymbol{\Psi}(\mathbf{x})$ and $I$
Initialize: $\mu_{g_i}^0$, $a_i^0$, $\sigma_{g_i}^{2^0}$, $\sigma_{e_i}^{2^0}$ and $\boldsymbol{\alpha}_i^0$
**while** learning **do**
    Calculate $P(A(n)|\mathbf{v}_i)$ and $P(y(n)|x(n), \mathbf{w}_i)$
    Update $h_i^k(y(n)|x(n))$ as per (11)
    Update $\mu_{g_i}^k$ and $\sigma_{g_i}^{2^k}$ as per (13)
    Update $\sigma_{e_i}^{2^k}$ and $\boldsymbol{\alpha}_i^k$ as per (14)
    Update $a_i^k$ as per (15)
**end while**
**return**: $\mu_{g_i}$, $\sigma_{g_i}^2$ and $\boldsymbol{\alpha}_i$
```

Prior to executing the EM algorithm, the gate and expert parameters need to be initialized. In the first place, the means of the gates $\mu_{g_i}^0$ are initialized, either randomly, or through K-means clustering [28]. Once the means are initialized, the membership probabilities are assigned in a *hard sense*, i.e., $h_i^k(y(n)|x(n)) = 1$ if the data point belongs to cluster i, and 0 otherwise. Then, the variances $\sigma_{g_i}^{2^0}$ can be calculated as per (13). Once the responsibilities are known, the expert parameters can be calculated as per (14), and the EM algorithm can be iterated until convergence. The EM algorithm is graphically illustrated in Fig. 3, and its pseudocode is provided in Algorithm 1. A stopping criterion can be set based on the maximum number of iterations or by checking the convergence of $\ln P(\mathbf{x}, \mathbf{y})$. In the measurement experiments reported in Section V, the convergence criterion is utilized, by comparing the increase in $\ln P(\mathbf{x}, \mathbf{y})$ in successive iterations against a threshold.

B. EM Algorithm for HME

The EM algorithm for the HME structure follows the same principle as the one discussed above, i.e., the maximization of the joint density is pursued. However, as there are now two gating layers that are mutually dependent, it is necessary to define the conditional posterior probability of the latent variables, which reads

$$h_{i|j}^k(y(n)|x(n)) = \frac{a_{i|j} g_{i|j}\left(A(n), \mathbf{v}_{ij}^k\right) P\left(y(n)|x(n), \mathbf{w}_{ij}^k\right)}{\sum_i a_{i|j} g_{i|j}\left(A(n), \mathbf{v}_{ij}^k\right) P\left(y(n)|x(n), \mathbf{w}_{ij}^k\right)} \tag{16}$$

and corresponds to the bottom layer responsibilities or membership probabilities, and where $a_{i|j}$ is the prior probability of the bottom layer gates and $\sum_i a_{i|j} = 1$, i.e., the prior probabilities of the gates within the same parent node sum up to one. On the other hand, the top layer responsibilities are of the same form as those for the single-layer case in (11), that is, $h_j^k(y(n)|x(n)) = \mathbb{E}\{P(z_j(n)|y(n), x(n))\}$. Finally, a joint posterior probability is also defined as $h_{ij}(y(n)|x(n)) = h_{i|j}(y(n)|x(n)) h_j(y(n)|x(n))$.

The learning rules for the parameters of the top layer gating function are of the same form as those in (13). On the other

TABLE I

DPD Main Path Processing Complexity per Linearized Sample

	VS model [13]	**DPW model [15]**	**DVR model [16]**	**ME model**	**HME model**
BF generation	$P+2+2G(P-1)$	$R\,(P+2+2G(P-1))+2R$	$11R+8RM_{\text{CPWL}}$	$P+2+2G(P-1)$	$P+2+2G(P-1)$
Filtering	$8B-2$	$R(8B-2)$	$8B_{\text{CPWL}}-2$	$I(8B-2)+4I-2$	$I(8B-2)+4(I+J-1)$

hand, the learning rules for the bottom layer parameters are given by the following expressions:

$$\begin{aligned}\mathbf{w}_{ij}^{k+1} &= \arg\max_{\mathbf{w}_{ij}} \sum_{n} h_{ij}^{k}(y(n)|x(n)) \ln P\big(y(n)|x(n),\mathbf{w}_{ij}^{k}\big)\\ \mathbf{v}_{ij}^{k+1} &= \arg\max_{\mathbf{v}_{ij}} \sum_{n} h_{ij}^{k}(y(n)|x(n)) \ln g_{i|j}\big(A(n),\mathbf{v}_{ij}^{k}\big)\\ a_{ij}^{k+1} &= \arg\max_{a_{ij}} \sum_{n}\sum_{i} h_{ij}^{k}(y(n)|x(n)) a_{ij}^{k} \qquad (17)\end{aligned}$$

where $a_{ij} = a_i a_{i|j}$.

It is noted that the original HME work [20] is formulated in the context of soft-max gating functions. Hence, the steps detailed in [19] to derive the simplified learning rules when Gaussian mixtures are adopted need to be considered.

Similarly, as in the single-layer model, the gate and expert parameters need to be initialized. To that end, the parameters of the top layer gating networks can be initialized following the same principle as that of the basic ME model. Then, we proceed with the initialization of the bottom layer gating networks. In this case, it is important that the means of the gates lie within the span of the gate in the parent node; otherwise, the conditional probabilities will be zero, and the algorithm would not be able to train the parameters. Then, the top and bottom layer responsibilities are calculated by assigning the data in the hard sense, and the expert parameters are initialized accordingly. For notational simplicity, it is assumed that all the submodels utilize the same parameterization, but these can be chosen freely in practice.

IV. ME Complexity Analysis and Comparison

In this section, we analyze the computational complexity of the ME and HME models and compare against the VS model in [13], the DPW model in [15], and the DVR in [16]. Here, we focus only on assessing the main path complexity, i.e., the complexity stemming from predistorting the transmit signal. The reason for this is that, in general, the main path complexity is far more critical than that of the learning path, as the predistortion process is to be executed in real time along with the data transmission, whereas the learning is executed at a much lower rate. In addition, it is noted that the iterative fashion in which the parameters of the ME model are trained seeks to find the optimal soft partition of the input data. Once the partition is known, the model parameters of the regression functions are learned with a single iteration of the weighted least-squares in (14). Consequently, assuming that the amplitude distribution of the transmit signal does not change significantly over time, the EM algorithm can be executed offline, while occasional parameter adaptation can be pursued to keep track of changes in the operating conditions of the transmitter system, e.g., due to device aging or temperature drifts through a single weighted least-squares fit.

It is assumed that the VS, the DPW, the ME, and the HME models build on polynomial-based regressors of the following form:

$$\begin{aligned}y(n) = &\sum_{\substack{p=1\\p\text{ odd}}}^{P}\sum_{m=0}^{M}\alpha_{p,m}x(n-m)|x(n-m)|^{p-1}\\ &+\sum_{\substack{p=1\\p\text{ odd}}}^{P}\sum_{m=0}^{M}\sum_{g=1}^{G}\beta_{p,m,g}x(n-m)|x(n-m-g)|^{p-1}\\ &+\sum_{\substack{p=1\\p\text{ odd}}}^{P}\sum_{m=0}^{M}\sum_{g=1}^{G}\gamma_{p,m,g}x(n-m)|x(n-m+g)|^{p-1} \qquad (18)\end{aligned}$$

where P is the maximum nonlinearity order, M is the memory depth, and G is the maximum envelope delay. On the other hand, the DVR model considers the BFs described in [16, eq. (17)].

The complexity analysis is done in terms of floating-point operations (FLOPs). It is assumed that a complex multiplication involves 6 FLOPs, whereas a complex addition and a real/complex multiplication both cost 2 FLOPs [29]. In order to generate the pth-order polynomial-based instantaneous BFs, it is considered that the process is done recursively, i.e., first the term $|x(n)|^2$ is calculated, and then, the pth-order instantaneous BFs denoted as $\Upsilon_p(n)$ are built as $\Upsilon_p(n) = \Upsilon_{p-2}(n)|x(n)|^2$, with $\Upsilon_1(n) = x(n)$. It is further assumed that generating the time-aligned memory BFs, i.e., the BFs corresponding to $m \neq 0$ on the first line of (18), does not cost any FLOP, as they are delayed versions of the instantaneous BFs. As all the piecewise models rely on the envelope of the transmit signal, it is assumed that it is known by all models, and hence, the cost involved in its computation is excluded from the comparison.

The exact complexity expressions for each of the models have been gathered in Table I, whereas specific complexity numbers are reported in Section V along with the corresponding experimental results. For completeness, the execution time required to train the different PW models is also provided. In Table I, R is the number of submodels/regions of the reference solutions, whereas B_{CPWL} and M_{CPWL} stand for the total number of regressors and memory depth of the DVR model. B stands for the number of regressors per submodel. The main path processing involves generating the corresponding BFs and the actual predistortion or filtering of the transmit signal. The BF generation for the VS, the ME, and the HME

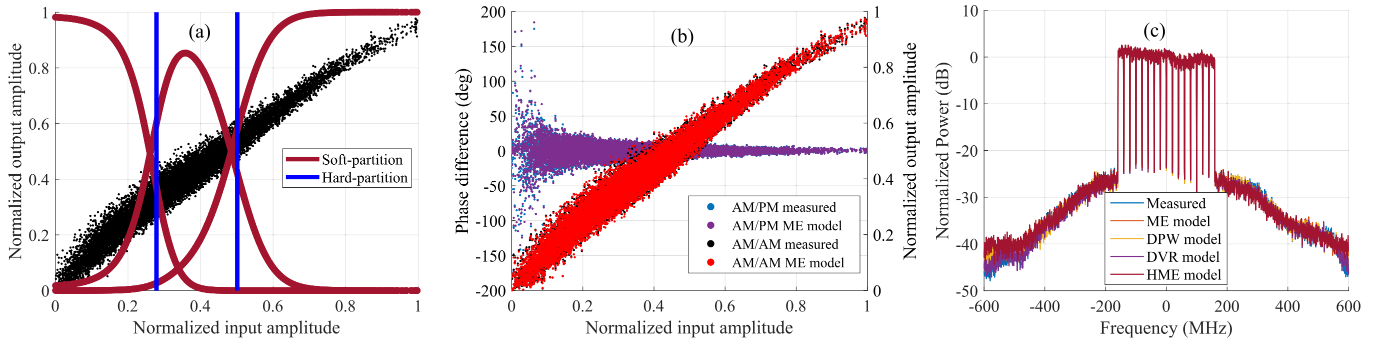


Fig. 4. (a) Hard and soft partitions provided by K-means and the ME model, respectively. (b) Measured and modeled AM/AM and AM/PM responses. (c) Measured and modeled spectra by different piecewise models.

model consists of generating the BFs given by (18). The DPW model is also assumed to utilize BFs of the form of (18); however, before doing so, it requires decomposing every transmit sample into subsamples by following the vector threshold decomposition approach in [15, eq. (3)]. Such decomposition is assumed to cost, on average, 2 FLOPs per region. On the other hand, the DVR is assumed to utilize the BFs in [16, eq. (17)]. The ME predistorting process basically builds on two steps: first, the linear transformation in (1) is computed for every expert, and then, their outputs are combined as per (7) and (9), for the ME and HME models, respectively. This implies that all regression functions are active at the same time. Similarly, for the DPW model, every transmit sample is decomposed into as many subsamples as submodels are defined, and each subsample is predistorted by its corresponding submodel prior to reconstructing the composite output signal, i.e., all submodels are active simultaneously. As for the DVR, the submodels are built in the actual CPWL BFs, which are all used to predistort every transmit sample. On the other hand, the VS model in [13] only computes the linear regression in (1) for the active submodel, as the gating network can be thought of as a binary decision.

V. RF Measurement Results

In order to evaluate the capabilities of the proposed ME and HME models, both in terms of direct modeling accuracy and linearization performance, several RF experiments, including different PA technologies, such as a gallium nitride (GaN) LMBA PA, a GaN DPA, and a GaN HEMT class AB PA, are conducted. As figures of merit, we consider the normalized mean squared error (NMSE), the adjacent channel error power ratio (ACEPR), and the adjacent channel leakage ratio (ACLR) [30]. The MATLAB implementation of the ME model and the EM algorithm is shared along with this article.

A. ME for Behavioral Modeling of RF PAs

The modeling accuracy of the different piecewise models is evaluated through RF measurements on an in-house designed LMBA GaN PA. The LMBA PA operates at 2.1-GHz carrier frequency with an average power of +37 dBm and 41% drain efficiency under the stimulus of a 320-MHz wide OFDM waveform composed of 16 20-MHz component carriers. Further details on the PA design and its characteristics can be found in [2]. The sampling frequency of the signal is 1.2 Gsamples/s, and its sample-level PAPR measured at 10^{-4} CCDF is ca. 8 dB.

TABLE II

Modeling Accuracy of Different Considered Models in Terms of NMSE, ACEPR, and Amount of Model Parameters

	# parameters	NMSE (dB)	ACEPR (dB)
GP model	640	-31.19	-32.40
VS model	672	-32.24	-32.42
DPW model	672	-31.55	-33.43
DVR model	663	-32.37	-33.89
ME model	672	-32.97	-34.42
HME model	484	-32.95	-34.32

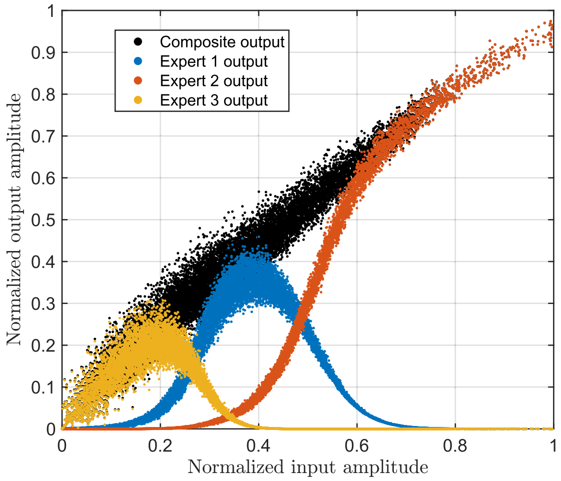


Fig. 5. Illustration of the outputs of the individual experts weighted by their corresponding gating functions and the composite model output. Direct PA modeling experiment with LMBA GaN PA at 2.1 GHz.

The ME model is assumed to utilize three experts, each of them utilizing the BFs given by (18) with $P = 7$, $M = 7$, and $G = 4$. The VS and DPW reference models are considered to utilize three regions, given by K-means, and have the same parameterization as the ME model, whereas the HME model considers $J = 2$ top layer nodes, each of them having $I_j = 2$

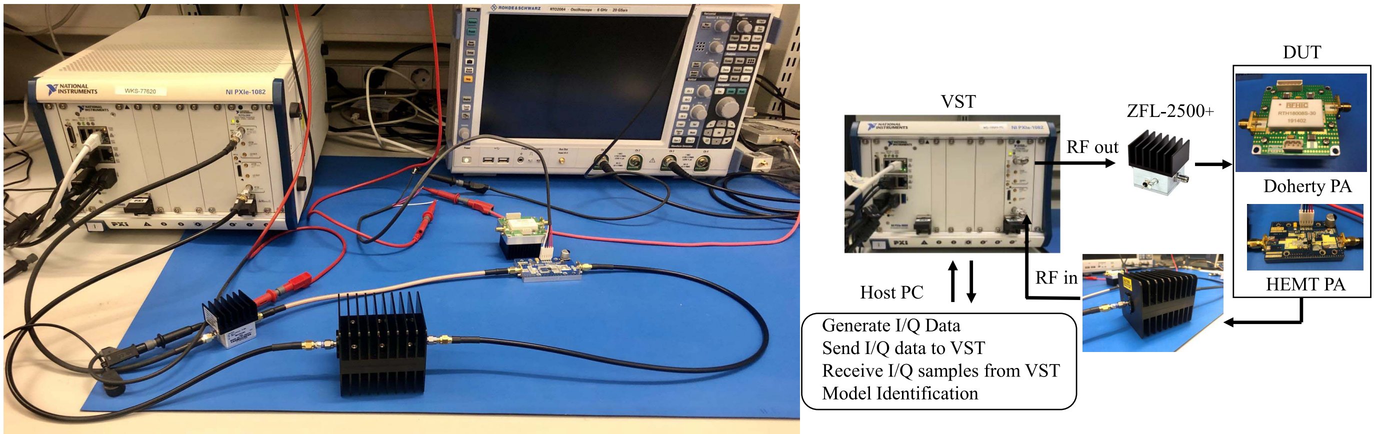


Fig. 6. RF measurement setup utilized in the DPD experiments containing the PXIe-5840 VST, the ZFL-2500VH+ driver amplifier, the DUTs, and a high-power attenuator.

experts with BFs parameters $P = 5$, $M = 10$, and $G = 2$. The DVR model utilizes 15 uniformly spaced regions and memory depth of $M_{\text{CPWL}} = 12$ [16]. The GP model utilizes $P = 7$, $M = 9$, and $G = 4$ and is assumed to also utilize even order nonlinear BFs on top of the odd order ones described in (18). The reason for this is that even order BFs can help in the modeling of complicated nonlinearities [31]. The PA output data are recorded, taken to baseband and synchronized to the digital waveform. Then, the reference models are fit by utilizing an LS approach, whereas the proposed ME and HME models are fit by iterating the EM algorithm until convergence.

The hard partitions provided by the K-means algorithm utilized in [13] and the soft partitions given by the ME model after the convergence of the EM algorithm are illustrated in Fig. 4(a). The vertical blue lines define the amplitude intervals over which the different submodels operate. On the other hand, the soft partitions should be interpreted as how much a given expert contributes—from zero to one—to generating a given output sample, whereas, with the hard partitions, these contributions are either one or zero. The different experts in the ME model can be interpreted as GPs that learn to specialize due to the responsibilities in the weighted least-squares fit. Fig. 5 illustrates the output of every expert weighted by its corresponding gating function, which corresponds to the terms $g_i(A(n), \mathbf{v}_\mathbf{i})\hat{y}_i(n)$, as well as the composite output of the ME model, which is the sum of the three submodels. This figure essentially shows the operation principle of the ME model, i.e., a set of experts cooperate to execute regression.

The performance of the different models is compared in terms of the NMSE and the ACEPR metrics and is given in Table II along with the number of coefficients of each model. As it can be seen, the modeling accuracy of the GP falls significantly behind the accuracy of the piecewise models due to the strong amplitude-dependent characteristics of the LMBA PA. The best modeling accuracy is provided by the proposed ME and HME models, both for the NMSE and the ACEPR metrics, as a result of its improved memory-modeling capabilities. The HME achieves a similar modeling accuracy with much fewer model parameters. This is due to the better nonlinear modeling capabilities provided by the two-layer gating network, which allows adopting lower polynomial orders for the experts.

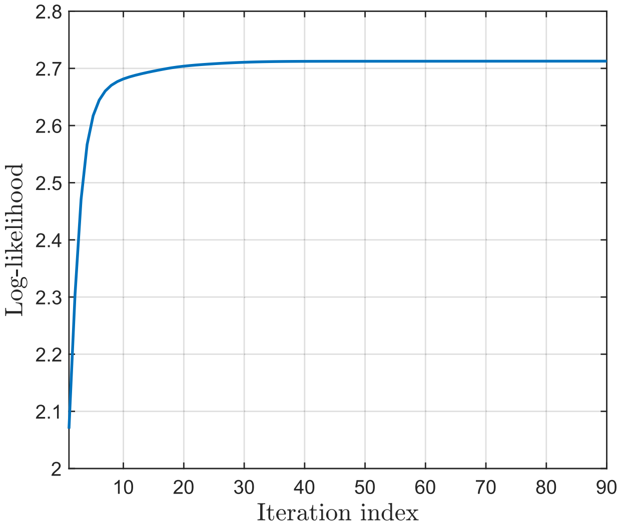


Fig. 7. Example convergence of the EM algorithm for ME-DPD parameter learning, within a single ILA iteration, with GaN DPA at NR band n3.

In the following, the linearization capabilities of the different PW models are assessed. As GPs are known to largely underperform in the linearization of wideband DPAs, as reported for instance in [13] and [14], they will not be considered in the following experiments.

B. ME for Linearization of RF PAs

The measurement setup for the DPD experiments is depicted in Fig. 6 and includes a National Instruments PXIe-5840 vector signal transceiver (VST), which serves both as a vector signal generator and as a vector signal analyzer. The baseband I/Q samples of the transmit waveform are generated with MATLAB in the VST environment, and the modulated waveform is upconverted to the desired carrier frequency utilizing the VST TX chain. The TX waveform is preamplified with a linear

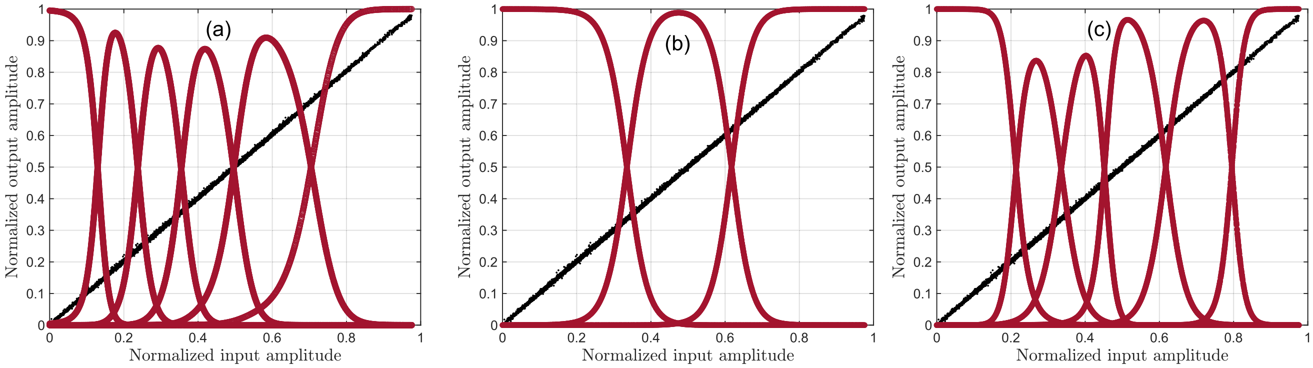


Fig. 8. (a) Soft partitions provided by the ME model. (b) Top layer soft partitions provided by the HME model. (c) Composite soft partitions provided by the HME model.

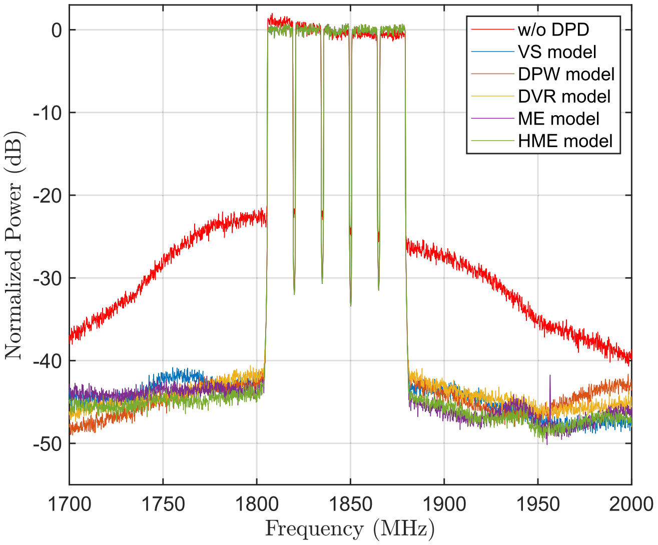


Fig. 9. Measured spectra at the GaN DPA output for a TX power of +39 dBm.

driver amplifier (ZFL-2500VH+) and then fed to the actual device under test (DUT). A GaN DPA and a class AB GaN HEMT PA are considered as DUTs. The DUT output signal is then attenuated with a high-power attenuator, whose output signal is observed via the VST receiver, where the signal is taken to the baseband and sampled. The received samples are processed in the VST environment utilizing MATLAB, where the different DPD solutions are trained and executed. All the models are learned through the ILA, and the reference models' parameters are fit through LS. The learning consists of three ILA iterations with a block size of $N = 10\,000$ samples. Within every ILA iteration, the ME and HME models run the iterative EM algorithm until convergence, an example of which is depicted in Fig. 7 for the ME model, where it can be seen that convergence is achieved within 40 EM iterations for the experiment considering the GaN DPA. It is, however, noted that the convergence speed heavily depends on the initialization of the gate parameters. For complexity assessment, we consider the expressions derived in Table I.

1) Measurement 1: GaN Doherty PA: The first DPD measurement experiment focuses on a GaN DPA operating at the 1.8425-GHz center frequency (NR band n3) at an average output power of +39 dBm, which corresponds to an output power back-off of ca. 7.5 dB with respect to saturation. The test waveform is composed of five 15-MHz component carriers, resulting in a total BW of 75 MHz that spans the whole NR band n3 and the PA BW. The PAPR of the test waveform, after iterative clipping and filtering (ICF), at 10^{-4} CCDF is ca. 7 dB, and the sampling frequency is 368.64 Msamples/s.

The ME and HME models are assumed to have $I = 6$ experts, each of them utilizing the polynomial-based BFs in (18) for $P = 5$, $M = 5$, and $G = 2$. The HME model considers $J = 3$ top layer nodes, each of them having $I_j = 2$ experts. The gating networks for the ME and HME models are shown in Fig. 8, together with the linearized amplitude response of the PA. Fig. 8(a) illustrates the soft partitions provided by the ME model, whereas Fig. 8(b) and (c) shows the top and composite, i.e., top times bottom layer, gates. The means of the gates of the ME model and the top layer gating network in the HME model are randomly initialized, whereas those of the bottom layer gating network are initialized so that they lie in the span of their corresponding top layer gating function.

The VS and DPW reference models are considered to utilize $R = 6$ regions, given by K-means, and have the same parameterization as the ME and HME experts. The DVR model utilizes $R = 9$ uniformly spaced regions and memory depth of $M_{\text{CPWL}} = 8$ [16].

The spectra at the output of the DUT when considering the different piecewise models are illustrated in Fig. 9, whereas their corresponding ACLR values are gathered in Table III. As it can be observed, the proposed ME models offer superior linearization capabilities compared to the state-of-the-art piecewise DPD models, with the HME providing the best performance due to its enhanced nonlinear modeling capabilities due to the two-layer gating network. To the best of our understanding, the overall improvement in the linearization/modeling performance is a result of the improved modeling between submodels due to the soft partitions and better modeling of the memory effects.

As for the complexity, all the models employ a similar number of model coefficients; however, the corresponding complexity in terms of FLOPs/sample depends heavily on their

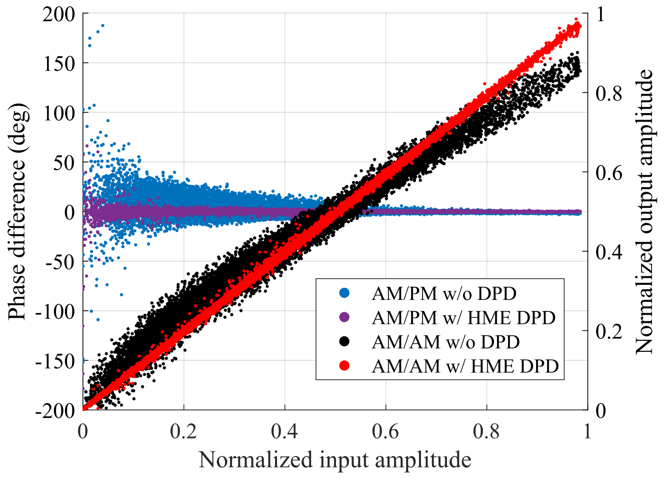


Fig. 10. Measured AM/AM and AM/PM responses for the GaN DPA w/o DPD and w/ the HME DPD for a TX power of +39 dBm.

TABLE III

LINEARIZATION PERFORMANCE IN TERMS OF ACLR, AMOUNT OF MODEL PARAMETERS, AND ASSOCIATED COMPLEXITY IN TERMS OF FLOPS/SAMPLE FOR THE GaN DPA EXPERIMENT

	# parameters	**BF gen.**	**Filtering**	**ACLR (dB)**
w/o DPD	–	–	–	24.52
VS model	264	23	350	42.23
DPW model	264	156	2100	43.36
DVR model	247	675	1974	42.83
ME model	264	23	2122	43.62
HME model	264	23	2132	44.25

operation principle. The VS, ME, and HME models require the lowest amount of operations to generate the required BFs because all the submodels employ the very same regressors, whereas the piecewise processing is embedded in their switching/gating functions. On the other hand, the DPW and DVR models incorporate the piecewise operations within the BFs; hence, the different submodels employ different regressors. The VS model presents the lowest filtering complexity, essentially because of its switching principle, in which only one of the submodels is active at a given time instant. On the other hand, the rest of the models requires filtering every transmit sample with all the submodels or the whole set of BFs. In addition, the ME and the HME models require applying the gating function to weigh the output of every submodel. However, the ME and HME models offer significantly better performance compared to the VS model, and it also outperforms the DPW and DVR at a similar overall computational cost. Fig. 10 illustrates the AM/AM and AM/PM responses of the PA without DPD and with the HME DPD model.

2) Measurement 2: GaN HEMT PA: The second DPD measurement experiment considers a GaN HEMT-based class AB PA (CGHV27030S-AMP1) operating at a 2.6-GHz center frequency at +35.5-dBm average output power. The test waveform is composed of eight 20-MHz component carriers, resulting in a total BW of 160 MHz. The PAPR of the test waveform, after ICF, at 10^{-4} CCDF is ca. 7 dB, and its sampling frequency is 645.12 Msamples/s. The ME and HME models are assumed to have $I = 6$ experts, each of them utilizing polynomial-based BFs with $P = 5$, $M = 5$, and $G = 1$. The HME model considers $J = 3$ top layer nodes, each of them having $I_j = 2$ experts. The VS and DPW reference models are considered to utilize $R = 6$ regions, given by K-means, and have the same parameterization as the ME and HME experts. The DVR model, on the other hand, utilizes $R = 7$ uniformly spaced regions and the memory depth of $M_{\text{CPWL}} = 7$ [16].

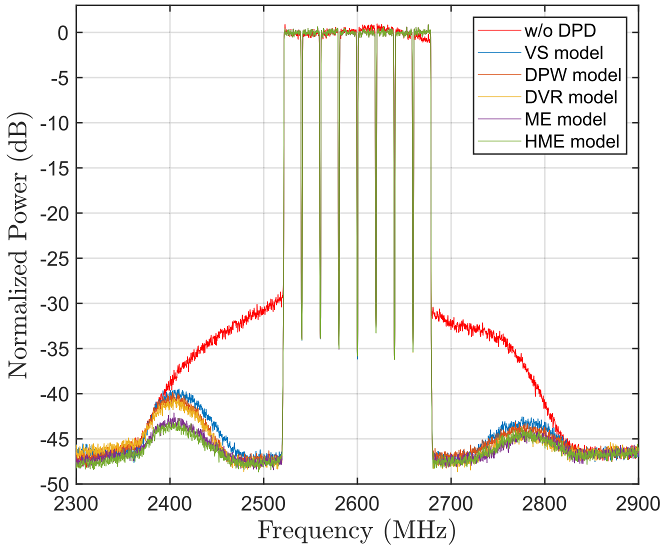


Fig. 11. Measured spectra at the HEMT-based PA output for a TX power of +35.5 dBm.

TABLE IV

LINEARIZATION PERFORMANCE IN TERMS OF ACLR, AMOUNT OF MODEL PARAMETERS, AND ASSOCIATED COMPLEXITY IN TERMS OF FLOPS/SAMPLE FOR THE HEMT-BASED PA EXPERIMENT

	# parameters	**BF gen.**	**Filtering**	**ACLR (dB)**
w/o DPD	–	–	–	33.59
VS	168	15	222	42.5
DPW	168	108	1332	43.13
DVR	158	469	1262	43.53
ME	168	15	1354	44.72
HME	168	15	1364	45.33

The linearized spectra at the PA output are illustrated in Fig. 11, and the associated complexity and specific ACLR values are gathered in Table IV. As it can be observed, the ME models again achieve the best linearization performance, giving up to 2 dB ACLR improvement while entailing similar complexity to that of the reference piecewise models. Based on the different conducted measurement experiments, it can be stated that the ME framework stands as a flexible and robust model for modeling and linearization of RF PAs. It allows working around some of the inherent limitations that state-of-the-art piecewise models commonly exhibit, which are mostly related to the way such models handle memory effects in the system.

C. Model Adaptation Runtime Comparison

In order to provide a more complete complexity comparison between the different PW models, their adaptation time is shortly discussed in the following.

The iterative principle of the EM algorithm seeks to find the optimal soft partition. Since the amplitude distribution of the transmit signal remains rather constant over time, the partition can be calculated offline and is then seldom updated. Once the partition is known, a single iteration of the weighted LS is utilized to calculate the DPD coefficients, which needs to be updated at a faster pace than the soft partitions (e.g., whenever the PA operating characteristics change). The following runtime numbers consider an Intel Core i7-10850H CPU @ 2.70-GHz machine running MATLAB 2021 and the parameterization of *Measurement 1*.

It is important to differentiate between the runtime required to obtain the region partitioning, which is 5.33 s for the ME model and 8.876 s for the HME model, and the runtime of the BF generation plus the weighted LS fit to estimate the DPD model parameters, which is 0.0787 s. Similarly, for the VS model, we differentiate between the runtime of 0.0216 s of the K-means algorithm to find the region partitioning, which is seldom executed, and the runtime of the BF generation plus LS fit, which totals 0.0667 s. On the other hand, the DPW model also requires to execute the K-means algorithm, whereas the BF generation plus the LS fit takes 0.1532 s. The DVR model considers equally spaced regions, which is assumed to require no computing time, whereas the BF generation plus the LS fit require 0.1359 s.

Overall, the adaptation complexity of the ME, HME, and VS models is rather similar, whereas the DPW and DVR models need approximately twice the time for adapting their coefficients.

VI. Conclusion

In this article, a new piecewise model for modeling and linearization of RF PAs based on the ME framework was proposed. The ME model utilizes soft partitions of the data, which implies that the different submodels overlap with one another. This ensures that the overall regression function is smooth and can thus facilitate accurate modeling of memory effects between regions. This feature is a notable improvement over the other existing piecewise models, wherein the partitions are commonly disjoint and the models operate and are also being learned independently. The ME model was also extended to a tree-structured regression model with multilayer nonlinear gating networks, which allows for further enhanced nonlinear modeling capabilities. The proposed ME approach was shown to provide the best modeling accuracy and linearization performance among the tested models in a large variety of RF measurement experiments on different PA technologies.

In the reported results, the gating network was considered to make partitions based on the envelope of the signal. However, more sophisticated decisions can be studied and incorporated in the model, e.g., by considering bivariate densities or by exploiting the multilayer gating network structure. Overall, the ME approach is a new framework for PA modeling and DPD research, with rich opportunities for further developments and tailoring to different linearization tasks.

References

[1] L. Guan and A. Zhu, “Green communications: Digital predistortion for wideband RF power amplifiers,” *IEEE Microw. Mag.*, vol. 15, no. 7, pp. 84–89, Nov. 2014.

[2] J. Pang, C. Chu, Y. Li, and A. Zhu, “Broadband RF-input continuous-mode load-modulated balanced power amplifier with input phase adjustment,” *IEEE Trans. Microw. Theory Techn.*, vol. 68, no. 10, pp. 4466–4478, Oct. 2020.

[3] R. S. Pengelly, S. M. Wood, J. W. Milligan, S. T. Sheppard, and W. L. Pribble, “A review of GaN on SiC high electron-mobility power transistors and MMICs,” *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 6, pp. 1764–1783, Jun. 2012.

[4] F. Mkadem and S. Boumaiza, “Physically inspired neural network model for RF power amplifier behavioral modeling and digital predistortion,” *IEEE Trans. Microw. Theory Techn.*, vol. 59, no. 4, pp. 913–923, Apr. 2011.

[5] W. H. Doherty, “A new high efficiency power amplifier for modulated waves,” *Proc. IRE*, vol. 24, no. 9, pp. 1163–1182, Sep. 1936.

[6] B. Kim, J. Kim, I. Kim, and J. Cha, “The Doherty power amplifier,” *IEEE Microw. Mag.*, vol. 7, no. 5, pp. 42–50, Oct. 2006.

[7] D. J. Shepphard, J. Powell, and S. C. Cripps, “An efficient broadband reconfigurable power amplifier using active load modulation,” *IEEE Microw. Wireless Compon. Lett.*, vol. 26, no. 6, pp. 443–445, Jun. 2016.

[8] 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. Signal Process.*, vol. 54, no. 10, pp. 3852–3860, Oct. 2006.

[9] F. M. Ghannouchi and O. Hammi, “Behavioral modeling and predistortion,” *IEEE Microw. Mag.*, vol. 10, no. 7, pp. 52–64, Dec. 2009.

[10] A. Brihuega, L. Anttila, M. Abdelaziz, T. Eriksson, F. Tufvesson, and M. Valkama, “Digital predistortion for multiuser hybrid MIMO at mmWaves,” *IEEE Trans. Signal Process.*, vol. 68, pp. 3603–3618, 2020.

[11] J. Reina-Tosina, M. Allegue-Martínez, C. Crespo-Cadenas, C. Yu, and S. Cruces, “Behavioral modeling and predistortion of power amplifiers under sparsity hypothesis,” *IEEE Trans. Microw. Theory Techn.*, vol. 63, no. 2, pp. 745–753, Feb. 2015.

[12] C. De Boor, *A Practical Guide to Splines*. New York, NY, USA: Springer, 1978.

[13] S. Afsardoost, T. Eriksson, and C. Fager, “Digital predistortion using a vector-switched model,” *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 4, pp. 1166–1174, Apr. 2012.

[14] A. Brihuega *et al.*, “Piecewise digital predistortion for mmwave active antenna arrays: Algorithms and measurements,” *IEEE Trans. Microw. Theory Techn.*, vol. 68, no. 9, pp. 4000–4017, Sep. 2020.

[15] A. Zhu, P. J. Draxler, C. Hsia, T. J. Brazil, D. F. Kimball, and P. M. Asbeck, “Digital predistortion for envelope-tracking power amplifiers using decomposed piecewise Volterra series,” *IEEE Trans. Microw. Theory Techn.*, vol. 56, no. 10, pp. 2237–2247, Oct. 2008.

[16] A. Zhu, “Decomposed vector rotation-based behavioral modeling for digital predistortion of RF power amplifiers,” *IEEE Trans. Microw. Theory Techn.*, vol. 63, no. 2, pp. 737–744, Feb. 2015.

[17] H. Holma, A. Toskala, and T. Nakamura, *5G Technology: 3GPP New Radio*, 1st ed. Hoboken, NJ, USA: Wiley, 2019.

[18] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive mixtures of local experts,” *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 2012.

[19] L. Xu *et al.*, “An alternative model for mixtures of experts,” in *Proc. Adv. Neural Inf. Process. Syst.*, 1995, pp. 633–640.

[20] M. I. Jordan, “Adaptive mixtures of local experts,” *Neural Comput.*, vol. 6, no. 2, pp. 181–214, 1994.

[21] S. E. Yuksel, J. N. Wilson, and P. D. Gader, “Twenty years of mixture of experts,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1177–1193, Aug. 2012.

[22] A. Brihuega, M. Abdelaziz, L. Anttila, Y. Li, A. Zhu, and M. Valkama, “Mixture of experts approach for behavioral modeling of RF power amplifiers,” in *Proc. IEEE Topical Conf. RF/Microw. Power Modeling Radio Wireless Appl. (PAWR)*, Jan. 2021, pp. 1–3.

[23] S. Haykin, *Adaptive Filter Theory*, 5th ed. London, U.K.: Pearson, 2014.

[24] L. Anttila, P. Händel, O. Mylläri, and M. Valkama, “Recursive learning-based joint digital predistorter for power amplifier and I/Q modulator impairments,” *Int. J. Microw. Wireless Technol.*, vol. 2, no. 2, pp. 173–182, Apr. 2010.

[25] R. J. Schilling and S. L. Harris, “Mixture models: Theory, geometry and applications,” in *Proc. NSF-CBMS Regional Conf. Ser. Probab. Statist.*, vol. 5, 1995, pp. 163–171.

[26] P. L. Gilabert *et al.*, “Order reduction of wideband digital predistorters using principal component analysis,” in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2013, pp. 1–7.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. Roy. Statist. Soc. B, Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.

[28] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient K-means clustering algorithm: Analysis and implementation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.

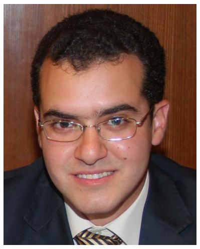
[29] R. J. Schilling and S. L. Harris, *Fundamentals of Digital Signal Processing Using MATLAB*. Boston, MA, USA: Cengage Learning, 2010.

[30] A. S. Tehrani, H. Cao, T. Eriksson, M. Isaksson, and C. Fager, “A comparative analysis of the complexity/accuracy tradeoff in power amplifier behavioral models,” *IEEE Trans. Microw. Theory Techn.*, vol. 58, no. 6, pp. 1510–1520, Jun. 2010.

[31] L. Ding and G. T. Zhou, “Effects of even-order nonlinear terms on power amplifier modeling and predistortion linearization,” *IEEE Trans. Veh. Technol.*, vol. 53, no. 1, pp. 156–162, Jan. 2004.

Alberto Brihuega (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in telecommunications engineering from the Universidad Politécnica de Madrid, Madrid, Spain, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree at Tampere University, Tampere, Finland.

He is currently an RF System Simulations Engineer with Nokia Mobile Networks, Oulu, Finland. His research interests include statistical and adaptive digital signal processing for compensation of hardware impairments in large-array antenna transceivers.

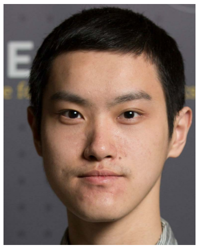


Mahmoud Abdelaziz (Member, IEEE) received the D.Sc. degree (Hons.) in electronics and communications engineering from the Tampere University of Technology, Tampere, Finland, in 2017.

From 2007 to 2012, he was a Communications and Signal Processing Engineer and an Embedded Systems Engineer with Newport Media Inc., Cairo, Egypt, (acquired by Atmel, San Jose, CA, USA), Etisalat Egypt, Cairo, and Axxcelera Broadband Wireless, Cairo. From 2018 to 2019, he was a Post-Doctoral Research Fellow with the Tampere University of Technology. Since 2019, he has been a Visiting Researcher with Tampere University, Tampere. He is currently an Assistant Professor with the Zewail City of Science and Technology, Giza, Egypt. His research interests include machine learning and statistical signal processing algorithms for flexible radio transceivers, in particular, behavioral modeling and digital predistortion of power amplifiers in single-antenna and multiple-antenna transmitters.

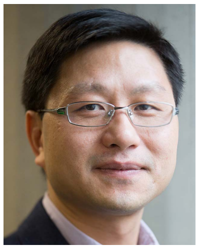
Lauri Anttila (Member, IEEE) received the M.Sc. and D.Sc. (Hons.) degrees in electrical engineering from the Tampere University of Technology (TUT), Tampere, Finland, in 2004 and 2011, respectively.

Since 2016, he has been a University Researcher with the Department of Electrical Engineering, Tampere University (formerly TUT), Tampere. From 2016 to 2017, he was a Visiting Research Fellow with the Department of Electronics and Nanoengineering, Aalto University, Helsinki, Finland. He has coauthored over 100 refereed articles and three book chapters. His current research interests include radio communications and signal processing, with a focus on the radio implementation challenges in systems, such as 5G, full-duplex radio, and large-scale antenna systems.



Yue Li (Member, IEEE) received the B.E. degree in information engineering from Southeast University, Nanjing, China, in 2016, and the Ph.D. degree in electronic engineering from University College Dublin (UCD), Dublin, Ireland, in 2020.

He is currently a Post-Doctoral Researcher with the RF and Microwave Research Group, UCD. His current research interests include behavioral modeling and digital predistortion for RF power amplifiers.



Anding Zhu (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from University College Dublin (UCD), Dublin, Ireland, in 2004.

He is currently a Professor with the School of Electrical and Electronic Engineering, UCD. He has published more than 150 peer-reviewed journal articles and conference papers. His research interests include high-frequency nonlinear system modeling and device characterization techniques, high-efficiency power amplifier design, wireless transmitter architectures, digital signal processing, and nonlinear system identification algorithms.

Prof. Zhu is an elected member of MTT-S AdCom, the Chair of the Electronic Information Committee, and the Vice-Chair of the Publications Committee. He is also the Chair of the MTT-S Microwave High-Power Techniques Committee. He has served as the Secretary of MTT-S AdCom in 2018. He was the General Chair of the 2018 IEEE MTT-S International Microwave Workshop Series on 5G Hardware and System Technologies (IMWS-5G) and a Guest Editor of the IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES on 5G Hardware and System Technologies. He is also an Associate Editor of the *IEEE Microwave Magazine* and a Track Editor of the IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES.

Mikko Valkama (Senior Member, IEEE) received the M.Sc. (Tech.) and D.Sc. (Tech.) degrees in electrical engineering (EE) from the Tampere University of Technology (TUT), Tampere, Finland, in 2000 and 2001, respectively.

In 2003, he was a Visiting Post-Doctoral Research Fellow with the Communications Systems and Signal Processing Institute, San Diego State University (SDSU), San Diego, CA, USA. He is currently a Full Professor and the Department Head of Electrical Engineering with the newly established Tampere University (TAU), Tampere. His current research interests include radio communications, radio localization, and radio-based sensing, with particular emphasis on 5G and 6G mobile radio networks.

Dr. Valkama was a recipient of the Best Ph.D. Thesis Award of the Finnish Academy of Science and Letters for his Ph.D. dissertation.