Mixture of Experts Neural Network for Modeling of Power Amplifiers

Arne Fischer-Bühner[#$], Alberto Brihuega[*$], Lauri Anttila[$], Manil Dev Gomony[#], Mikko Valkama[$]
[#]Nokia Bell Labs, Antwerpen, Belgium, [*]Nokia Mobile Networks, Oulu, Finland,
[$]Tampere University, Finland

***Abstract*— A new Mixture of Experts Neural Network (ME-NN) approach is described and proposed for modeling of nonlinear RF power amplifiers (PAs). The proposed ME-NN is compared with various piece-wise polynomial models and the time-delay neural network (TDNN) regarding their ability to scale in terms of modeling accuracy and parameter count. To this end, measurements with GaN Doherty PA at 1.8 GHz and a load modulated balanced (LMBA) PA operating at 2.1 GHz with strong nonlinear behavior and dynamics are employed, assessing the potential benefits of ME-NN over the existing models. Implementation-related advantages of the proposed ME-NN over TDNNs at increasing network sizes are furthermore discussed. The measurement results show that the ME-NN approach offers increased modeling accuracy, particularly in the LMBA PA case, compared to the existing reference methods.**

***Keywords*— power amplifier, nonlinear distortion, behavioral modeling, neural network, digital predistortion, 5G and beyond**

I. Introduction

The evolving target capacities of modern wireless communication systems are coupled with stringent efficiency and linearity requirements. The waveforms employed, e.g., in 5G systems are challenging to amplify, due to their wide bandwidth and high peak-to-average power ratio [1]. Highly optimized components in the radio frequency (RF) front-end, most notably the RF power amplifier (PA), add strong nonlinear and dynamic distortion to the transmitted signals when operated in an efficient mode, while digital compensation techniques allow us to correct for the impairments and meet the linearity requirements. Digital pre-distortion (DPD) is the most established compensation approach, where appropriate inverse nonlinear distortion is induced prior to the RF transmitter thus yielding linearized transmission [2], [3].

Accurate models describing the nonlinear behavior of PAs are essential for DPD, with Volterra-based polynomial models (PM), such as the generalized memory polynomial (GMP), being commonly used. However, these models prove unable to model strong nonlinearities over a wide dynamic range [3], [4]. To overcome these limitations, piece-wise modeling approaches, such as the decomposed piece-wise (DPW-GMP) model [2] or the vector switched (VS-GMP) model [3], have been proposed, which utilize multiple sub-models for different input level regions. However, these models may lack the ability to properly model dynamics between the different sub-models [4]. To this end, the Mixture of Experts framework was introduced for GMP modeling (ME-GMP) in [4], which employs a soft partitioning of the input level range and combines several GMPs based on a probabilistic scheme that allows the sub-models to be overlapped. Thus, modeling of the dynamic behavior across regions was improved. However, as we show in this paper, the GMP-based modeling approaches may have limited capability to scale, i.e. they have a bounded accuracy as the number of parameters increases.

More recently, time-delay neural networks (TDNN) have gained attraction due to their superior modeling capabilities [5], [6]. While their complexity in terms of parameter count and training effort is usually large, they make it possible to scale the modeling capabilities further, when compared to the PMs – an aspect that is shown in this paper. However, increasing the NN size and depth results in many sequentially dependent computations which causes an increased processing latency and need for on-chip buffering of intermediate results. Therefore, implementation-related limitations arise when scaling TDNNs.

In this paper, we translate the successful ME framework to TDNNs by proposing a novel ME-NN structure for PA modeling. The structure consists of neural network (NN) experts, which are combined in a soft-switching manner by a gating NN. The ME-NN is trained as a single entity, so that the partitioning and gating of the experts are obtained during the training process. Based on true RF measurement data of a GaN Doherty PA at 1.8 GHz and a load modulated balanced (LMBA) GaN PA at 2.1 GHz, we compare the ME-NN with the PMs and the TDNN with regard to their ability to scale. We demonstrate that the ME-NN model is more capable than large TDNNs, while giving additional implementation benefits due to its parallel structure. Additionally, the results also show that the ME-NN approach can facilitate increased modeling accuracy, particularly in the LMBA PA case, compared to all the existing reference methods.

The remainder of the paper is organized as follows. In Section II, we describe TDNN modeling in general and extend it by introducing the ME-NN framework. Section III provides the RF measurement results, compares the performance of the various models regarding their scalability, and discusses implementation-related aspects. Finally, conclusions are drawn in Section IV.

II. Mixture of Experts Neural Network

With PMs, the PA input-to-output behavior is described as a linear mapping of nonlinear regressors to the output. Let $x(k)$ and $y(k)$ denote the complex-valued baseband input and output signals of the PA. The general model is given by

$$\hat{\boldsymbol{y}} = \Phi_{\mathrm{x}} \boldsymbol{a}\,, \tag{1}$$

where $\hat{\boldsymbol{y}} = [\hat{y}(1), ..., \hat{y}(K)]^{\mathrm{T}}$ and the model output $\hat{y}(k)$ is an approximation to $y(k)$. $\Phi_{\mathrm{x}} \in \mathbb{C}^{K \times W}$ is a matrix containing W input regressors, or input features, typically consisting of the complex-valued instantaneous and delayed

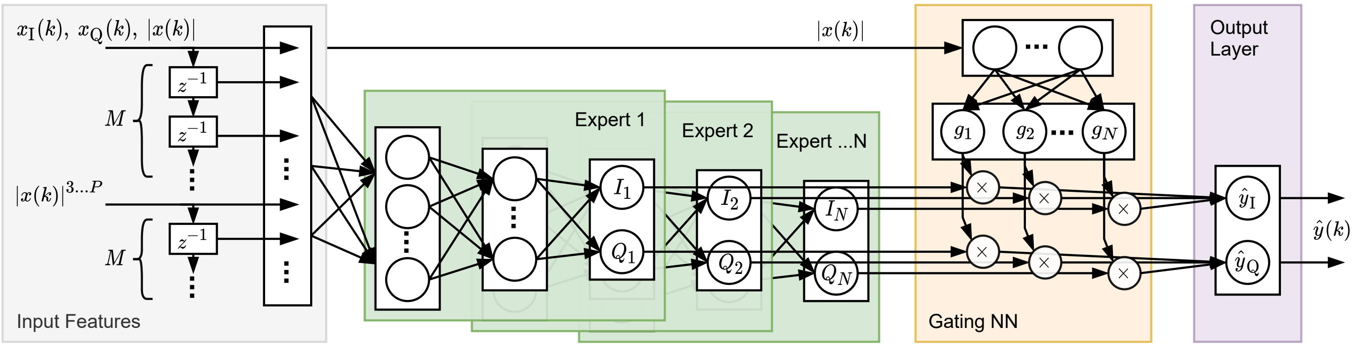

Fig. 1. Block diagram illustrating the ME-NN structure. N independent NN experts are combined by the gating NN based on the input envelope $|x(k)|$.

input samples of x(k), their envelope powers, and their partition into sub-regions. The input features are weighted by a set of parameters $\boldsymbol{a} \in \mathbb{C}^{W\times 1}$ chosen such that the difference between $\hat{y}(k)$ and $y(k)$ is minimized in the least-squares sense.

A. NN-based Modeling

In NN-based modeling, the linear weighting of the input features is replaced by a TDNN [5] such that

$$\hat{\boldsymbol{y}} = \Xi_{\chi}\left[\Phi_{\mathrm{x}}|\boldsymbol{c}\right] \ , \qquad (2)$$

where Ξ_{χ} denotes the mapping by the TDNN with structure χ given parameters $\boldsymbol{c}$. The TDNN is a forward-oriented network that consists of a number of neurons which are arranged in L consecutive, fully connected layers of width B_l. The output of a neuron b in layer l is given by

$$\upsilon_{b,l} = h_l\left[\sum_{i=1}^{B_{l-1}} \omega_{b,l,i}\xi_{l,i}\right] + \psi_{b,l} \ , \qquad (3)$$

with h_l being a nonlinear activation function, ξ_i representing the inputs to the neuron, $\omega_{b,l,i}$ are the respective input weights, and $\psi_{b,l}$ is an offset applied at the neuron's output. All $\omega_{b,l,i}$ and and $\psi_{b,l}$ form the set of adjustable parameters $\boldsymbol{c}$ and are learned during a training phase. We use the *Sigmoid* function as an activation function h_l for the layers $l = 1, ..., L-1$, also referred to as hidden layers (HL), while the final output layer $l = L$ uses a linear activation.

Since complex-valued NNs are cumbersome to operate and train, real-valued networks are employed to operate the parallel I and Q parts of the signal. Consequently, the output layer B_L has two neurons which provide the model output as real-valued I and Q components, interpreted as $\hat{y} = \hat{y}_{\mathrm{I}}(k) + j\hat{y}_{\mathrm{Q}}(k)$. The input features Φ_{x} serve as an input to the first layer. For the NN, Φ_{x} is composed from the instantaneous and previous I and Q samples of $x(k-m) = x_{\mathrm{I}}(k-m) + jx_{\mathrm{Q}}(k-m)$, as well as the envelope signal $|x(k-m)|$, $m = 1, 2, ..., M$. Although the NN is inherently nonlinear, it has been shown in [5] to be advantageous to also include the so called augmented terms and their respective delayed versions $|x(k-m)|^p$, $p = 3, 5, ..., P$. All input features are individually normalized to unit power.

B. Proposed Mixture of Experts Neural Network (ME-NN)

We now describe and construct the proposed ME-NN structure, with the end-to-end network structure shown in Figure 1. The underlying concept is to have multiple parallel TDNNs, $n = 1, ..., N$, that each specialize to a part of the PA's amplitude range. These so called expert NNs, Ξ_{e}, share an identical structure, but use their own set of specialized parameters $\boldsymbol{c}_n$. Each of the experts receives the full set of input features and has two real-valued outputs I_n and Q_n.

The joint output of the aggregate network is then composed from the experts' outputs through weighting and summation. An additional gating NN, Ξ_{g}, with parameters c_{g} is added to the network, to provide N weights in relation to the instantaneous envelope input to combine the experts in a probabilistic and soft manner. *Softmax* activation is used for the gating NN's output layer since *Softmax* layers are especially well-suited for modeling of probabilities and constrain the outputs of the gating NN to always sum up to one. The output of the aggregate ME-NN is thus given by

$$\hat{\boldsymbol{y}} = \sum_{n=1}^{N}\left(\Xi_{\mathrm{e}}\left[\Phi_{\mathrm{x}}|\boldsymbol{c}_n\right]\ \Xi_{\mathrm{g}}^{(n)}\left[\boldsymbol{A}_{\mathrm{x}}|c_{\mathrm{g}}\right]\right) \ , \qquad (4)$$

where $\boldsymbol{A}_{\mathrm{x}} = \left[\,|x(1)|, ..., |x(K)|\,\right]$, and $\Xi_{\mathrm{g}}^{(n)} = g_n$ is the n^{th} output of the gating NN.

III. RF Measurement Results and Analysis

A. Data and Training

In order to evaluate the modeling performance, the models are applied to real-world data measured from two different RF PAs. The first dataset is taken from a GaN Doherty PA (model RTH18008S-30) measured at 1.8425 GHz center frequency with an average output power of +39 dBm. The operated waveform consists of three 20 MHz wide OFDM carriers which together span a total bandwidth of 60 MHz. The sampling frequency is 360 Msamples/s. The second dataset is taken from a load modulated balanced (LMBA) GaN PA [7], operated at 2.1 GHz carrier frequency, and with a mean output power of +37 dBm under stimulus of a carrier aggregated signal comprising a total bandwidth of 320 MHz.

The NNs are implemented and processed using the *Keras* [8] framework and *Python*. We let the ME-NN train

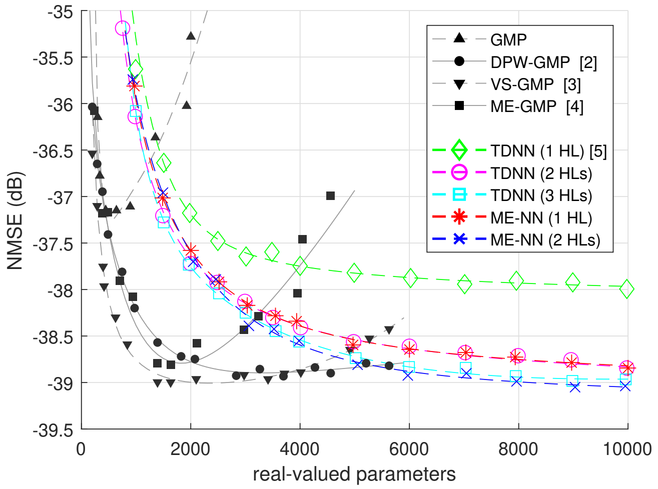

Fig. 2. Comparison of the modeling performance for the GaN Doherty PA at 1.8 GHz. The modeling NMSE of different PMs (gray/black) and NN models (colored) is shown with respect to the number of parameters. The solid lines illustrate the trends in the point clouds.

in its entirety, so that the experts and gating NN specialize during the training. For training of the NN parameters, we used the Adaptive Moment (*Adam*) optimizer and 300 epochs on a data set of $T=40$k samples. The training optimizes the NN towards minimizing the mean square error Γ given by

$$\Gamma = \frac{1}{2T}\sum_{k=1}^{T}\left((\hat{y}_{\mathrm{I}}(k) - y_{\mathrm{I}}(k))^2 + (\hat{y}_{\mathrm{Q}}(k) - y_{\mathrm{Q}}(k))^2\right) \qquad (5)$$

The trained models are evaluated in terms of normalized Γ (NMSE), normalized to the root-mean square power of the PA output, using separate validation data with 25k samples. Furthermore, we repeat the training 10 times and average the validation results to mitigate the impact of randomly initialized parameters during training. The trained TDNNs and ME-NNs use input features with $P = 3$ and $M = 9$, resulting in $W = 40$ input features for the Doherty PA case and $P = 3$ and $M = 11$, $W = 48$, in the LMBA PA case. In both cases, the size B_1 of the first HL is varied to generate TDNNs and experts of different sizes. The width of the remaining HLs is $B_2 = 10$ in the two layer case and $B_2 = 20$ and $B_3 = 5$ for a NN with three HLs. Below 3k parameters, the sizes of the respective second HLs were halved, to also allow for NNs of smaller size. For the ME-NNs, we use $N = 3$ experts and a gating NN with one HL $B_{\mathrm{g},1} = 10$. Note that the individual experts of the ME-NN are significantly smaller compared to the TDNN. The below comparision refers to the total size of the NNs, including the gate.

The different PMs use varied parametrization for the basis functions of the GMP. The provided instances are the best-performing ones selected from a large number of tested parameterizations. The piece-wise PMs use three regions or sub-models, as do the ME-NNs. The complex-valued nature of PM parameters is properly mapped to the corresponding real-valued parameter count for fair comparison.

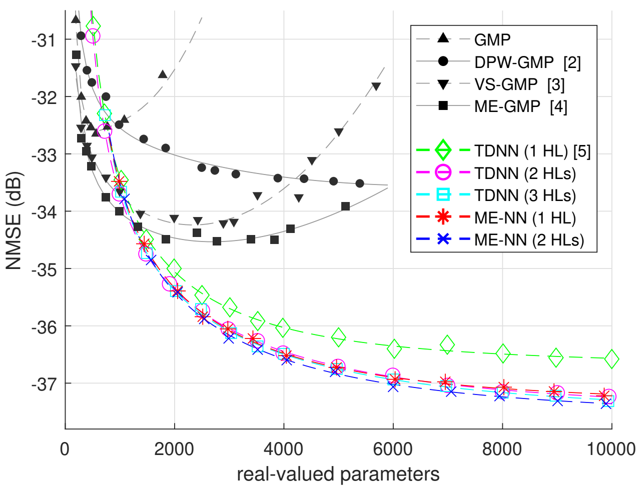

Fig. 3. Modeling NMSE for the LMBA PA behavior at 2.1 GHz. Comparison of various PMs (gray/black) and NN models (colored) versus their parameter count. The solid lines illustrate the trends in the point clouds.

B. Results

Fig. 2 and Fig. 3 show performance comparisons in terms of modeling NMSE of the various PMs (trained and assessed with the same data as the NNs), TDNNs, and ME-NNs with regard to their real-valued parameter count. We find that increasing the parameter count in PMs soon reaches a optimum for the NMSE. With increased model size, we observe the robustness of the GMP based models to decrease. A logical reasoning is that the only way to scale these models is to provide more basis functions through an increase of nonlinear order and memory depth, which can lead to ill-conditioned parameter estimation and rank deficiency. Therefore, *Ridge* regression [9] is applied to mitigate numerical impairments and to stabilize the coefficient estimate at large parameter counts. Piece-wise models benefit from employing multiple smaller sub-models allowing them to scale their accuracy further, but our experiments show limited margin.

In the LMBA PA case shown in Fig. 3, the NN-based models significantly outperform the PMs when the number of parameters is increased. The respective depths of the neural networks are thereby limiting the achievable modeling accuracy of the NN. We observe that ME-NNs significantly outperform TDNNs given an identical number of parameters and HLs. The ME structure allows each of the experts to specialize for a part of the nonlinear behavior, thus enabling the ME-NN to also map strong nonlinearities.

Examining the output of the gating NN, illustrated in Fig. 4 for the case of the Doherty PA, reveals how the experts are combined. The weights for the experts are determined by the gating NN as a function of the instantaneous input envelope $|x(k)|$ and can range from 0 to 1. These weights are then applied to the experts' outputs to form the joint output of the ME-NN. The distribution of the weights in Fig. 4 shows that in the Doherty PA case, none of the experts is fully suppressed or solely selected at any given input magnitude. Rather, the composition of the output varies as the input signal changes.

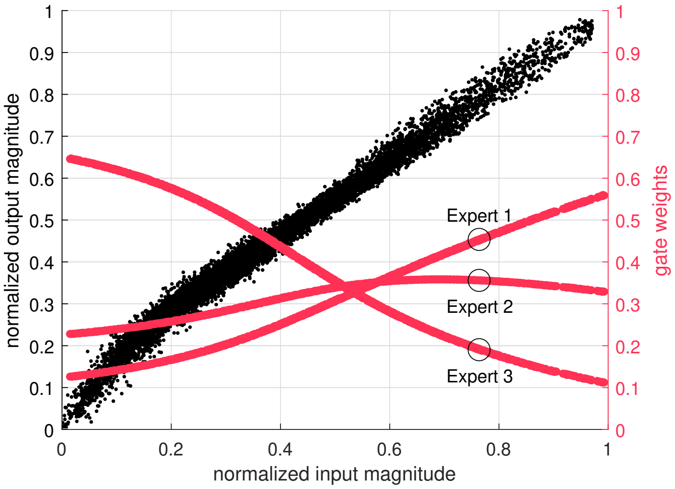

Fig. 4. Weights provided at the output of the gating NN to showcase the soft selection of experts as a function of the input signal's envelope $|x(k)|$ alongside the output magnitude of the GaN Doherty PA.

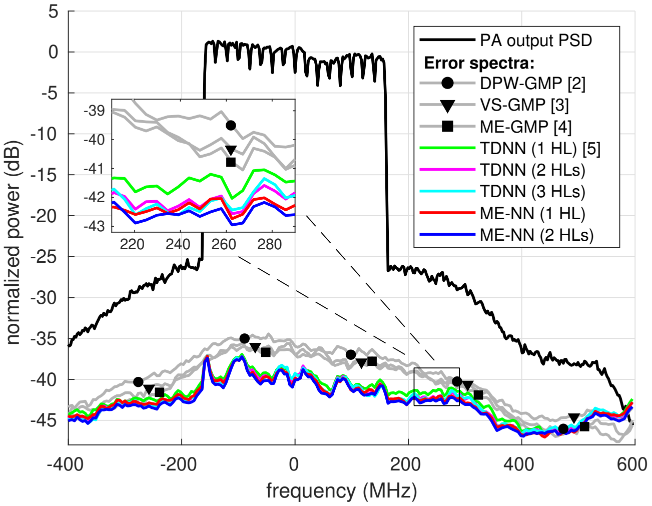

Fig. 5. Power spectral density (PSD) graph of the LMBA PA output and error spectra for selected PMs and NN models with approximately 2.5k parameters.

We observe that adding further HLs enables the TDNN to keep up with ME-NN, as more HLs improve the TDNN's nonlinear modeling abilities. The example error spectra of different models with approximately 2.5k parameters provided in Fig. 5 show the TDNN and ME-NN modeling errors to behave mostly similarly, however, a slight performance advantage can be credited towards the ME-NN.

C. Discussion

Although it is possible to scale the TDNN by adding more HLs, there are implementation-related aspects that favor the ME-NN over a regular TDNN. Since the ME-NN essentially employs several independent expert NNs, those can be processed in parallel before combining their results. A deeper NN structure through more HLs causes sequential dependencies while processing, increasing the number of dependent computations and thus processing latency. Consequently, a NN with fewer layers is favored, which benefits the ME-NN over the TDNN. In conventional TDNNs, every neuron in a layer is dependent on every other neuron's output in the previous layer. This will be limiting in terms of building a hardware implementation as on-chip buffering of data will grow with the network size for storing intermediate computations. Consequently, the latency of processing the NN with limited compute and memory resources will grow exponentially. By employing several smaller expert NNs, we reduce the average count of input connections per neuron yielding fewer dependencies. This relaxes the needs for local buffering (i.e. on-chip memory) and the amount of data transfers between the logic and memory, thus improving the overall power consumption and latency.

IV. Conclusion

In this paper, we studied the scaling ability of polynomial and TDNN approaches in RF PA behavioral modeling. Moreover, we proposed and investigated the new ME-NN model, which realizes the ME framework as an aggregate NN. All investigated polynomial models show bounded modeling abilities, while the NN models show excellent scaling abilities by adding more neurons and layers. We identified several implementation related aspects which favor the ME-NN over a conventional TDNN. In our future work, we will consider the ME-NN also in the context of digital pre-distortion and utilize the identified implementation advantages.

Acknowledgment

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860921. The work was also supported by the Academy of Finland under the grants #319994, #338224, and #332361.

References

[1] Z. Popovic, "Amping up the PA for 5G: Efficient GaN power amplifiers with dynamic supplies," *IEEE Microwave Magazine*, vol. 18, no. 3, pp. 137–149, 2017.

[2] 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, 2008.

[3] 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, 2012.

[4] A. Brihuega, M. Abdelaziz, L. Anttila, Y. Li, A. Zhu, and M. Valkama, "Mixture of experts approach for piecewise modeling and linearization of RF power amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 1, pp. 380–391, 2022.

[5] 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 Trans. Neural Netw.*, vol. 30, no. 1, pp. 242–254, 2019.

[6] Y. Zhang, Y. Li, F. Liu, and A. Zhu, "Vector decomposition based time-delay neural network behavioral model for digital predistortion of RF power amplifiers," *IEEE Access*, vol. 7, pp. 91 559–91 568, 2019.

[7] 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, 2020.

[8] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[9] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.