Deep Neural Network Based Predistorter with ReLU Activation for Doherty Power Amplifiers

Reina Hongyo [#1], Yoshimasa Egashira [#2], Keiichi Yamaguchi [#3]

[#] *Wireless System Laboratory, Corporate Research & Development Center, Research & Development Div., Toshiba Corp.*
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki-shi, Japan
[1]reina1.hongyo@toshiba.co.jp
[2]yoshimasa1.egashira@toshiba.co.jp
[3]keiichi.yamaguchi@toshiba.co.jp

***Abstract* — In this paper, we propose a deep neural network based DPD (DNN-DPD) where ReLU activation is used instead of sigmoid activation to avoid the gradient vanishing problem. The ReLU function is implemented using a half-wave rectifier that acts as the activation function for the hidden layers in the DNN-DPD. Experimental results using a Doherty PA will show that the proposed DNN-DPD can achieve an additional 3 to 4 dB of linearity compensation compared to a conventional Volterra-DPD or shallow NN. Furthermore, to achieve comparable performance to sigmoid activation, the number of multipliers can be reduced by a factor of 100. This confirms that the DNN-DPD approach has a large potential to overcome the complicated memory effects of Doherty PAs with reduced DPD complexity.**

Index Terms — digital predistorter, deep neural network, ReLU, Doherty power amplifiers

I. Introduction

Doherty PAs [1] are subjected to intensive research in recent years because of their high power efficiency. Due to its non-isolated nature, Doherty PAs usually exhibit memory effects which are a function of the history of the input signal. Traditionally, Volterra series [2] based digital predistorter (Volterra-DPD) [3] have been used to compensate for the memory effects in Doherty PAs. However, since Volterra series have a high correlation between polynomial bases, it is hard to improve the compensation performance even if the number of polynomial terms were to be increased.

Recently, deep neural networks (DNN) have attracted the attention of researchers in many fields; in particular, state-of-the-art performance and its flexibility have been reported in areas such as image recognition and natural language processing [4], [5]. In addition, the effects of NN applied to DPD have been investigated in [6]-[8]. However, these approaches are based on relatively shallow network architectures using only 3 or 4 layers with sigmoid activation. DPD performance using deeper neural networks have not been reported thus far.

In this paper, we propose a deep neural network based DPD (DNN-DPD) with ReLU activation for the hidden layers therefore overcoming the gradient vanishing problem which is typically found in sigmoid based activation. Results will show that the proposed DPD can achieve superior compensation performance compared to a conventional Volterra-DPD or shallow NN when applied to a Doherty PA. In addition, the results will show that the number of multipliers used in the ReLU activated DNN-DPD can be reduced by up to a factor of 100 compared to sigmoid activated DNN-DPD without sacrificing performance.

II. DNN-DPD

DNN is an artificial neural network (ANN) with multiple hidden layers between the input and output layers. Since the generalization capability of a neural network grows exponentially with the depth of layers [4], it is fair to assume that an increase in the number of layers will improve the degree of compensation that can be achieved when using a DNN based DPD.

Fig. 1 shows a simple block diagram of the proposed DNN-DPD. The inputs $s_{\mathrm{I}}(n)$ and $s_{\mathrm{Q}}(n)$ are the in-phase and quadrature-phase baseband component of the transmitted signal at sampling time n with the parameters $y_{\mathrm{I}}(n)$ and $y_{\mathrm{Q}}(n)$ being the corresponding pre-distorted baseband output signals. The operation of each layer is given by:

1) Input layer ($l = 1$, Tap $= M$, Neuron $= 2M$)

$$\left[x_1^{(1)}, x_2^{(1)}, \cdots, x_M^{(1)}\right] = \left[s_{\mathrm{I}}(n), s_{\mathrm{I}}(n-1), \cdots, s_{\mathrm{I}}(n-M-1)\right] \quad (1)$$

$$\left[x_{M+1}^{(1)}, x_{M+2}^{(1)}, \cdots, x_{2M}^{(1)}\right] = \left[s_{\mathrm{Q}}(n), s_{\mathrm{Q}}(n-1), \cdots, s_{\mathrm{Q}}(n-M-1)\right] \quad (2)$$

2) Hidden layer ($l = 2, 3, \cdots, L-1$, Neuron$= D$)

$$x_i^{(l)} = f\left(\sum_{j=1}^{D} w_{i,j}^{(l)}\, x_j^{(l-1)} + b_i^{(l)}\right) \quad (3)$$

3) Output layer ($l = L$, Neuron $= 2$)

$$y_{\mathrm{I}}(n) = x_1^{(L)} = \sum_{j=1}^{D} w_{1,j}^{(L)} x_j^{(L-1)} + b_1^{(l)} \quad (4)$$

$$y_{\mathrm{Q}}(n) = x_2^{(L)} = \sum_{j=1}^{D} w_{2,j}^{(L)} x_j^{(L-1)} + b_2^{(l)} \quad (5)$$

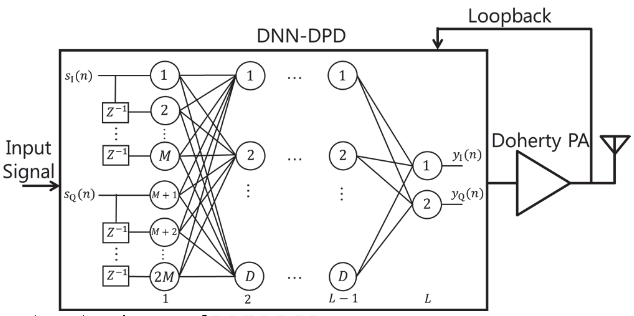

Fig. 1. A schema of DNN-DPD.

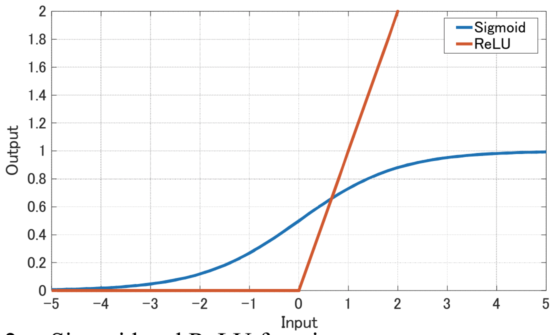

Fig. 2. Sigmoid and ReLU function.

where $x_i^{(l)}$ is the output of i-th neuron in l-th layer, $w_{i,j}^{(l)}$ is the synaptic weight connecting the j-th neuron in the (l-1)-th layer to the i-th neuron in the l-th layer, $b_i^{(l)}$ is the bias term used as an additional input for the i-th neuron in the l-th layer and $f(\cdot)$ is the activation function of the hidden layer. The input layer includes tapped delay lines (M taps) to cope with the memory distortion of Doherty PAs. The synaptic weights and biases were optimized by back propagation [9] and Adam [10] which is similar to a gradient descent method.

In previous works [6]-[8], a traditional sigmoid function (or hyperbolic tangent version) has been adopted as the activation function f of the hidden layers. However, a sigmoid function can cause the gradient vanishing problem [4] which makes it difficult to train a deep network. In order to avoid the gradient vanishing problem, a ReLU function [11] was adopted using the half-wave rectifier $f(z) = \max(z, 0)$ as the activation function of the hidden layers (see Fig. 2). The sigmoid function saturates when its argument is extremely positive or negative which then causes the back-propagated gradients to converge onto zero. On the other hand, the ReLU is a non-saturated function for positive arguments which in turn reduces the probability of the vanishing gradient problem when compared to sigmoid activation.

III. Experimental Results & Discussion

The proposed DNN-DPD is compared to three conventional DPD approaches. They are memoryless-DPD (ML-DPD), Volterra-DPD and neural network based DPD with one hidden layer (NN-DPD). The ML-DPD is equivalent to a Volterra-DPD without any delay taps. The Volterra-DPD consist of R

TABLE I
Measured ACLR for Different DPD Configuration

Kind of DPD	Number of kernels (R) for Volterra-DPD	Activation (f) for NN-DPD and DNN-DPD	ACLR at -3.3MHz [dBc]	ACLR at +3.3MHz [dBc]
w/o DPD	—	—	-20.664	-19.942
ML-DPD	9	—	-39.666	-37.712
Volterra-DPD	10	—	-49.412	-47.818
Volterra-DPD	50	—	-52.732	-52.106
Volterra-DPD	100	—	-52.003	-52.574
Voltera-DPD	200	—	-52.437	-52.049
NN-DPD	—	Sigmoid	-50.59	-49.466
DNN-DPD	—	Sigmoid	-51.773	-51.298
NN-DPD	—	ReLU	-51.377	-52.593
DNN-DPD	—	ReLU	-55.451	-55.632

optimal kernels that are pruned using an order range from 0 to 8 and a delay tap of up to 50. The DNN-DPD uses $L = 10$ layers, which means that there are 8 hidden layers. The input layer has I and Q components that use $M = 50$ taps. Each hidden layer has $D = 100$ neurons with a ReLU function. The output layer has two neurons with a linear activation function. The output spectrum for all DPD approaches applied to a Doherty PA are shown in Fig. 3.

For ease of comparison, the adjacent channel leakage ratio (ACLR) at $\pm$3.3 MHz offset is shown in Table I for the conventional DPD techniques. It is apparent that the ML-DPD approach is unable to compensate for the memory effects in a Doherty PA as it is unable to improve the ACLR beyond -40 dBc. The Volterra-DPD approach with delay taps improves the ACLR to about -48 to -53 dBc. However, even when the number of kernels are increased, an ACLR better than -53 dBc was not obtained. This confirms the practical limitations of a Volterra-DPD with delay taps. Table I also shows the comparison between NN-DPD and DNN-DPD both for a sigmoid and ReLU activation. As discussed previously, the sigmoid function suffers from the gradient vanishing problem which limits the ACLR performance to about -52 dBc at $\pm$3.3 MHz offset. When the activation function is changed to be a ReLU type, the ACLR performance does not greatly change for the NN-DPD approach. However, a 3 to 4 dB improvement is measured when using the DNN-DPD approach. This corresponds to an ACLR better than -55 dBc thus confirming the practical feasibility of the approach.

The relationship between ACLR and the number of layers used for both sigmoid and ReLU activation is shown in Fig. 4. In regards to ReLU, an ACLR improvement of 3 dB was measured when the number of layers was increased from 3 to 5. However for the sigmoid activated DNN-DPD, even with 9 layers, an ACLR improvement of only 1 dB was measured.

Fig. 5 shows the relationship between ACLR and the number of multipliers required for a variety of L (3, 4, 5, 6, 8, 10) and

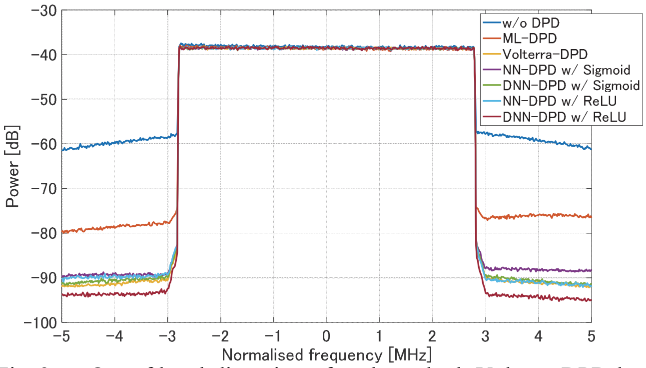

Fig. 3. Out-of-band distortion of each method. Volterra-DPD has 50 kernels, DNN-DPD with sigmoid and ReLU activation has 10 layers.

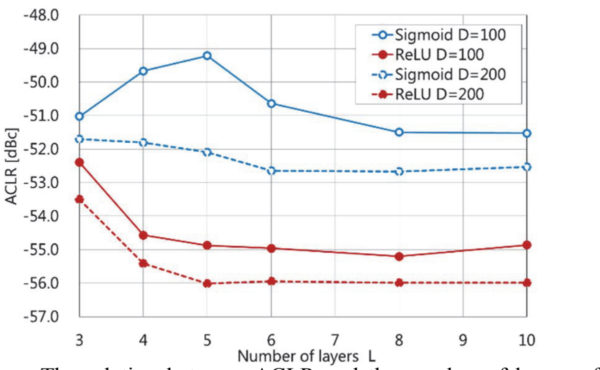

Fig. 4. The relation between ACLR and the number of layers of DNN-DPD with Sigmoid and ReLU activation.

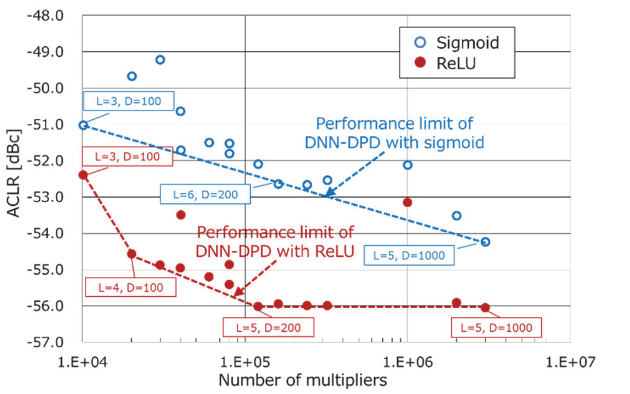

Fig. 5. The relation between ACLR and the computational complexity of DNN-DPD with sigmoid and ReLU activation.

D (100, 200, 1000) configurations. A dashed line is plotted to highlight the performance limit observed during the experiments for both the sigmoid and ReLU activated DNN-DPD. For clarity, to show the relationship between ACLR, L and D, the operating parameters for 7 unique DNN-DPD configurations are also shown in Fig. 5. To achieve an ACLR of -53 dBc, ReLU requires a factor of 10 fewer multipliers than the sigmoid activated DNN-DPD. This decreases to a factor of 100 fewer multipliers for an ACLR of -54 dBc. Taking into consideration that the number of multipliers is given by $D^2(L-2)+2D$, an increase in the number of neurons D causes an exponential increase in the number of multipliers. On the other hand, increasing the number of layers L results in a linear increase in the number of multipliers. Looking at Fig. 5, increasing the number of multipliers results in a 3 dB improvement in ACLR for sigmoid activation. However, although there is an initial improvement for ReLU activated DNN-DPD, increasing the number of multipliers beyond 10^5 yields no additional ACLR improvement. It can therefore be concluded that performance of sigmoid activated DNN-DPD is directly correlated to the number of neurons while the performance of ReLU activation is heavily influenced by the number of layers. As a result, ReLU activated DNN-DPD can achieve identical ACLR performance to sigmoid activation at a fraction of the number of multipliers.

IV. Conclusion

This paper, we proposed a DNN-DPD with ReLU activation that avoids the gradient vanishing problem. Experimental results using a Doherty PA confirms the proposed DNN-DPD approach can achieve superior compensation with an additional 3 to 4 dB of ACLR suppression reaching out-of-band distortion levels better than -55 dBc. In addition, the ACLR suppression limitations of a conventional Volterra-DPD with delay taps was measured to be about -52 dBc. Moreover, the results showed that the number of multiplier using ReLU activation is about 1/100 times of using sigmoid activation.

References

[1] B. Kim et al., "The Doherty power amplifier," *IEEE Microwave Magazine*, vol. 7, no. 5, pp. 42-50, October 2006.

[2] M. Schetzen, "The Volterra and Wiener theories of nonlinear systems," New York: Wiley, 1980.

[3] C. Eun et al., "A new Volterra predistorter based on the indirect learning architecture," *IEEE Trans. Signal Process.*, vol. 45, no. 1, pp. 223-227, January 1997.

[4] Y. Le Cun et al., "Deep learning," *Nature*, vol. 521, pp. 436-44, May 2015.

[5] G. E. Hinton et al., "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504-507, July 2006.

[6] B. E. Watkins et al., "Neural network based adaptive predistortion for the linearization of nonlinear RF amplifiers," *Military Commun. Conf. 1995*, pp. 145-149, San Diego, USA, November 1995.

[7] T. Liu et al., "Dynamic behavioral modelling of 3G power amplifiers using real-valued time-delay neural networks," *IEEE Trans. Microwave Theory & Tech.*, vol. 52, no.3, pp. 1025-1033, March 2004.

[8] M. Rawat et al., "Adaptive digital predistortion of wireless power amplifiers/transmitters using dynamic real-valued focused time-delay line neural networks," *IEEE Trans. Microwave Theory & Tech.*, vol. 58, no. 1, pp. 95-104, January 2010.

[9] D. E. Rumelhart et al., "Learnings representations by back-propagating errors," *Nature*, vol. 323, pp. 533-536, October 1986.

[10] D. P. Kingma et al., "Adam: A method for stochastic optimization," *Int. Conf. on Learning Representations (ICLR) 2015*, San Diego, USA, May 2015.

[11] X. Glorot et al., "Deep sparse rectifier neural networks," *The 14th Int. Conf. on Artificial Intelligence and Statistics (AISTATS) 2011*, vol. 15, JMLR: W&CP 15, pp. 315-323, FL, USA, 2011.