Hybrid Training for 5G Linearization Systems Based on Machine Learning

Guilherme C. Pereira*, Carmelo J. A. Bastos-Filho†
University of Pernambuco, Recife, Brazil
Email: *gcp.@ecomp.poli.br, †carmelofilho@ecomp.poli.br

Luiz A. M. Pereira‡, Arismar Cerqueira S. Jr§
National Telecommunications Institute (Inatel), Santa Rita do Sapucaí, Brazil
Email: ‡luiz.melo@inatel.br, ¶arismar@inatel.br

***Abstract*—Radio over Fiber systems blend the efficiency of fiber optic transmission with the flexibility of wireless communications but face significant challenges due to inherent nonlinearities in optical signal transmission. This paper proposes a hybrid training technique using simulated and real data to train machine learning models deployed for linearization, aiming to enhance performance and reliability. The proposal involves pre-training the model with a large synthetic dataset and fine-tuning it with real data collected from experiments. This method leverages synthetic data to establish initial weights, refined using real-world data to capture practical complexities. Results show that the hybrid training technique significantly outperforms models trained solely on synthetic data, with nearly a twofold improvement in performance as evidenced by lower Mean Squared Error (MSE) and Mean Absolute Error (MAE) values. Compared to models trained exclusively on real data, the hybrid method achieves comparable performance, proving effective in scenarios where real data is scarce or hard to obtain.**

***Index Terms*—5G, Machine Learning, Radio-over-Fiber, Linearization, Hybrid Training**

I. Introduction

The evolution of communication systems is driven by the growing demand for fast and efficient data transmission in personal and business contexts. Fifth Generation of Mobile Communications (5G) technology stands at the forefront, offering wideband, low latency, and support for applications such as Internet of Things (IoT), autonomous vehicles, and real-time industrial process control [1]. However, enhanced communication in remote areas (Enhanced Remote Area Communications (eRAC)) remains a critical challenge [2]. Projects like 5G-RANGE and 5G Rural First leverage technologies such as TV White Space (TVWS) to deliver high-quality connectivity to hard-to-reach regions [3]. Constant advancements in telecommunications require efficient solutions for managing signals in optical communication systems, such as fiber-wireless (FiWi). These systems combine the efficiency of Radio-over-Fiber (RoF) with the flexibility of wireless communications but face significant challenges due to nonlinearities in optical signal transmission. Overcoming these nonlinearities is essential to improve the quality and capacity of transmission, making RoF systems viable for future high-speed, low-latency networks.

Integrating radio with optical communication systems presents a promising solution for eRAC demands. By merging the flexibility of wireless networks with RoF high capacity, it is possible extending the reach of 5G signals and ensure high-quality communication in remote areas. However, using analog RoF systems in long-distance and high-power scenarios, face linearity challenges. Advanced signal processing techniques, including the use of machine learning models, such as Artificial Neural Network (ANN), are crucial to overcome these obstacles and improve RoF system efficiency [4]–[6].

Artificial Neural Networks (ANNs) are statistical models composed of artificial neurons organized in layers. These models have been widely used for regression, classification, and forecasting tasks.

979-8-3503-8817-6/24/$31.00 ©2024 IEEE

Previous works have demonstrated that ANNs can equalize signals [5], [6]. The shape of the dynamic equalizer is built based on previous data used to train the model. However, linearization process heavily relies on data quality and variability. Ensuring high-quality data can be challenging, especially with complex and specific signals, such as the RoF signals. Former solutions based on ANN deployed data acquired on simulations, which can be used to validate proposals but do not represent some real-world aspects of the RoF signals. On the other hand, real data are often scarce or difficult to measure. It is challenging since ANN models need considerable data to correctly train the model.

This paper proposes a technique that utilizes simulated and real data to train a linearization system to enhance the overall performance and reliability of RoF systems. Besides, applying Machine Learning (ML) techniques to RoF signals can be problematic as models are typically trained on specific databases and lose generalization capacity when system conditions change. One technique to address this issue is transfer learning, as demonstrated by [7]. In this work, we explore a similar approach, using a large synthetic database for pre-training and a smaller real database for fine-tuning the model.

II. Proposal

In our proposal, we first pre-train the model using the synthetic dataset for a pre-determined number of epochs. This initial step is crucial as it allows the model to familiarize itself with a large volume of data, albeit synthetic, which helps establish the foundational weights for the model. These weights serve as a starting point for further refinement. By exposing the model to synthetic data, we ensure that it has a broad understanding of the general patterns and characteristics of the signals it will later process more accurately using real-world data. This step leverages the abundance of synthetic data, which is often easier to generate and manipulate, providing a comprehensive initial learning phase for the model.

After the pre-training phase, we fine-tune the model using the real training dataset. This fine-tuning is essential because real-world data often contains nuances and complexities that synthetic data cannot fully replicate. We can adjust the weights and biases more precisely to reflect actual operating conditions by training the model on real data. This step significantly improves the model's performance and accuracy, making it more robust and reliable in practical applications. Combining synthetic pre-training followed by real-data fine-tuning ensures that the model is well-generalized and effectively tuned to handle real-world scenarios.

III. Methodology

This section presents the methodology deployed in this paper. First, we briefly describe the database used to train the model. After, we describe the mechanisms proposed for training ANN models using synthetic and real-world data.

A. Database

We used two types of datasets to train the linearization systems. First, we generated 29,491 examples of Orthogonal Frequency Division Multiplexing (OFDM) signals for the synthetic database, each with a non-linearity degree of 5, 2048 subcarriers, and an Signal-to-Noise Ratio (SNR) of 50dB, used for training. We also used real data available from [8], which provides both transmitted and received waveforms from the RoF system. Fig.1 shows a photography of the analog RoF setup used for generating the real dataset. A tunable laser produces a 10-dBm optical carrier at 1550.12 nm. This carrier is applied to the Mach-Zehnder Modulator (MZM) (model FTM7920FBA from Fujitsu), which modulates the optical carrier with the Radio Frequency (RF) signal, centered at 707 MHz. A polarization controller (PC) is employed to adjust the polarization of the light before it enters the MZM. The RF signal is generated by an software-defined radio (SDR)-based 5G transceiver (Universal Software Radio Peripheral (USRP) 2954R). The MZM is biased to its quadrature point by using a DC power supply.

The signals vary in bandwidth from 3, 6, 12, to 24 MHz, and the RF transmission power ranges from 0 to 11 dBm. The real data is split into training/validation and hold-out test sets based on RF transmission power, with powers 0 to 5 and 8 to 10 for training/validation and the rest for the blind test. We evaluate the performance of these models using two metrics: Mean Squared Error (MSE) and Mean Absolute Error (MAE).

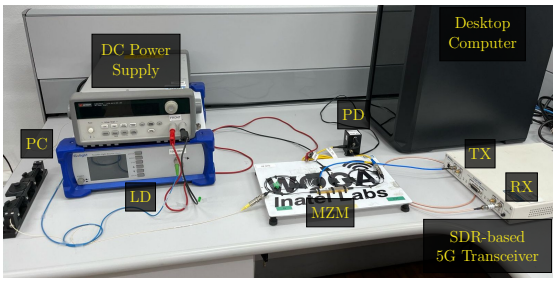

Fig. 1. The proposed setup to generate the dataset.

TABLE I
BEST MODEL SUMMARY

Layer (type)	Output Shape	Param #
input (InputLayer)	(None, 2)	0
dense (Dense)	(None, 64)	192
dense (Dense)	(None, 16)	1040
dense (Dense)	(None, 64)	1088
dense (Dense)	(None, 2)	130
Total params: 2450 (9.57 KB)		
Trainable params: 2450 (9.57 KB)		
Non-trainable params: 0 (0.00 Byte)		

B. Hybrid Training

The hybrid training process begins with an initial pre-training phase where the model is exposed to a synthetic dataset for 10 epochs. This step is crucial as it allows the model to learn general patterns and establish robust initial weights from a large volume of synthetic data. Once the model has grasped these general patterns, we move to the fine-tuning phase, where the model is trained with real data. This fine-tuning is essential for the model to learn the specific and intricate patterns unique to real-world data. By incorporating real data, the model adjusts and refines its weights to better handle the complexities and nuances found in practical applications. This two-step training process ensures that the model leverages the advantages of both synthetic and real data, resulting in improved performance and reliability.

We conducted a grid search to further optimize our linearization scheme model architecture. We experimented with three key hyperparameters: the number of hidden layers, neurons per layer, and the activation function. Specifically, we tested hidden layers ranging from 1 to 3, varied the neurons per layer among 16, 32, 64, 128, and 512, and included activation functions such as sigmoid, tanh, and relu. This comprehensive search covered 1269 different model configurations. This enable to identify the most effective architecture for our linearization scheme, which is shown in Table I.

IV. Results

In this section, we present the results of our experiments and compare the proposed hybrid training strategy with three other approaches. The comparison includes a model trained exclusively on synthetic data, a state-of-the-art RoF signal linearization scheme [5] (Baseline), the same model topology trained with real data, and our proposed hybrid training, which fine-tunes the synthetic model using a standard technique.

Table II shows the MSE values for all approaches, while Table III presents the MAE values. Our proposed ANN model, which employs a hybrid training technique incorporating both synthetic and real data, outperformed the baseline approach by nearly a factor of two. The Baseline model, trained only on synthetic data, exhibited higher error values, indicating that while synthetic data is useful for initial training, it lacks the complexity and variability of real-world data. This limitation affects the generalization of the model when real-world scenarios is considered.

The model trained exclusively on real data showed better performance than the Baseline, underscoring the importance of real-world data in achieving higher accuracy. However, acquiring sufficient real data can be challenging and resource-intensive. Our proposed method, which involves pre-training the model on synthetic data followed by fine-tuning with real data, demonstrated improved performance over the Baseline but did not surpass the model trained exclusively on real data.

By combining the strengths of both synthetic and real data, our comprehensive hybrid training process bridges the gap between synthetic and real-world data. This method results in a robust and reliable model for RoF signal linearization, achieving superior performance with lower MSE and MAE values. This approach effectively enhances the model's ability to generalize and perform well in practical applications.

TABLE II
MSE RESULTS FOR DIFFERENT MODELS

Models	Avg	Std	Max	Min
Baseline	2.88×10^{-2}	3.03×10^{-2}	9.61×10^{-2}	6.51×10^{-3}
Only Real Data	1.24×10^{-2}	1.08×10^{-2}	3.61×10^{-2}	2.17×10^{-3}
Proposed Method	1.24×10^{-2}	9.51×10^{-3}	3.33×10^{-2}	3.21×10^{-3}

TABLE III
MAE RESULTS FOR DIFFERENT MODELS

Models	Avg	Std	Max	Min
Baseline	1.22×10^{-1}	6.36×10^{-2}	2.51×10^{-1}	6.33×10^{-2}
Only Real Data	8.00×10^{-2}	3.82×10^{-2}	1.52×10^{-1}	3.55×10^{-2}
Proposed Method	8.11×10^{-2}	3.79×10^{-2}	1.50×10^{-1}	3.52×10^{-2}

V. DISCUSSION

The results indicate that the model with transfer learning outperformed the model trained solely on synthetic data, with a significance level greater than 0.05. It demonstrates that incorporating real data into the training process, even after an extensive pre-training phase with synthetic data, substantially improves the model's ability to generalize and adapt to real-world conditions. It is particularly noteworthy because synthetic data, while useful for initial training, often needs more complexity and variability found in real-world data. Thus, the transfer learning approach helps bridge this gap by refining the model with actual data, making it more accurate and reliable.

However, it is essential to note that these transfer learning models do not show a significant performance advantage over models trained exclusively on real data, with the significance level remaining at 0.05. It suggests that while transfer learning offers benefits when real data is limited or difficult to obtain, it may not necessarily surpass the effectiveness of a model trained entirely on real data. Despite this, the utility of transfer learning remains evident, particularly when acquiring a large volume of high-quality real data is challenging or impractical. By leveraging synthetic data for initial training and fine-tuning with available real data, transfer learning provides a viable solution for improving model performance.

VI. CONCLUSIONS

We conclude that the fine-tuning technique, which initially uses synthetic data for pre-training followed by training with real data collected from experiments, yields better results than models trained solely with synthetic data. This hybrid approach leverages the extensive availability of synthetic data to establish a robust initial model. The subsequent fine-tuning with real data allows the model to adjust to real-world complexities and nuances, enhancing its accuracy and performance. Our results demonstrate that this method significantly outperforms models trained exclusively on synthetic data without showing significant degradation compared to models trained entirely on real data. The hybrid training technique is particularly advantageous when acquiring high-quality real RoF signal data is challenging or impractical. By combining the strengths of both synthetic and real data, this approach provides a practical and efficient solution for developing high-performance models. It ensures the model can generalize well to real-world conditions while presenting a resource-efficiency. Future research could explore further optimization of this hybrid training process, including using more sophisticated synthetic data generation techniques and advanced fine-tuning strategies to continue improving the performance and reliability of RoF systems.

ACKNOWLEDGMENTS

This work received partial funding from Project XGM - AFCCT - 2024 - 2 - 15 - 1, supported by xG-Mobile – EMBRAPII - Inatel Competence Center on 5G and 6G Networks, with financial resources from the PPI IoT/Manufacturing 4.0 program of MCTI (grant number 052/2023), signed with EMBRAPII. The authors also thank the financial support from CNPq, CAPES, FINEP, FAPEMIG (Contracts # PPE - 00124 - 23, RED - 00194 - 23 and APQ - 02746 - 21).

REFERENCES

[1] W. Dias *et al.*, "Performance analysis of a 5G transceiver implementation for remote areas scenarios," in *2018 European Conference on Networks and Communications (EuCNC)*, pp. 363–367, IEEE, 2018.

[2] L. L. Mendes *et al.*, "Enhanced remote areas communications: The missing scenario for 5G and beyond 5G networks," *IEEE Access*, vol. 8, pp. 219859–219880, 2020.

[3] "Departure for Digital, Culture, Media and Sports, Technical Report 5G RuralFirst: New Thinking Applied to Rural Connectivity." https://www.5gruralfirst.org/wp-content/uploads/2019/10/5G-RuralFirst-New-Thinking-Applied-to-Rural-Connectivity-1.pdf. Accessed: 2023-03-20.

[4] J. He *et al.*, "Machine learning techniques in radio-over-fiber systems and networks," in *Photonics*, vol. 7, p. 105, MDPI, 2020.
[5] L. A. M. Pereira *et al.*, "Linearization schemes for radio over fiber systems based on machine learning algorithms," *IEEE Photonics Technology Letters*, vol. 34, no. 5, pp. 279–282, 2022.
[6] G. C. Pereira *et al.*, "Analyzing Machine Learning Paradigms for the Linearization of 5G Radio over Fiber Systems," in *2023 SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference (IMOC)*, pp. 19–21, IEEE, 2023.
[7] J. Zhang *et al.*, "Fast Remodeling for Nonlinear Distortion Mitigation Based on Transfer Learning," *Optics letters*, vol. 44, no. 17, pp. 4243–4246, 2019.
[8] I. CRR, "A-rof-linearization-dataset." https://github.com/inatelcrr/A-RoF-Linearization-Dataset, 2024.