Optics Communications 590 (2025) 131949

Contents lists available at ScienceDirect

Optics Communications

journal homepage: www.elsevier.com/locate/optcom

Open dataset initiative for machine learning-based linearization in analog radio over fiber systems ☆

L.A.M. Pereira [a,*], L.L. Mendes [a], Mitchell A. Cox [b], Arismar Cerqueira S. Jr. [a]

[a] National Institute of Telecommunications, João de Camargo Ave, Santa Rita do Sapucaí, 37536-001, MG, Brazil
[b] University of the Witwatersrand, 1 Jan Smuts Avenue, Braamfontein, Johannesburg, South Africa

ARTICLE INFO

Dataset link: https://github.com/inatelcrr/A-RoF-Linearization-Dataset

Keywords:
Analog radio over fiber
Dataset
Digital pre-distortion
Linearization
Software-defined radio

ABSTRACT

In analog radio over fiber (A-RoF) systems, linearization is essential to mitigate nonlinear distortions. While many studies have explored both machine learning (ML)-based and traditional linearization schemes for A-RoF, they often do not provide the dataset used to train the ML models. We present a mathematical model for representing A-RoF systems and provide a comprehensive real dataset, which we make available via GitHub. We utilize a software-defined radio (SDR) approach to generate digital samples of the generalized frequency division multiplexing (GFDM) waveform, which are then applied to a real A-RoF system. This dataset, consisting of transmitted and received GFDM waveform samples, enables the training of both traditional and ML-based models, thereby supporting the development of robust linearization schemes. Furthermore, we conduct a performance evaluation of a ML-based digital pre-distortion (DPD) scheme, trained on both synthetic and real datasets. The linearization performance was assessed using key metrics, including the root mean square error vector magnitude (EVM_{RMS}), normalized mean square error (NMSE), and adjacent channel leakage ratio (ACLR). The results show that the DPD scheme successfully reduced the EVM_{RMS} from approximately 6% to 2%, along with improvements in NMSE and ACLR. Additionally, the dataset, in conjunction with the memoryless polynomial model presented in this paper, provides a robust framework for characterizing A-RoF systems. Researchers can leverage this publicly available dataset to model A-RoF systems and design novel linearization schemes, thereby minimizing nonlinear distortions and enhancing signal fidelity.

1. Introduction

Ubiquitous communications are an emergent feature of the sixth-generation of mobile network (6G) [1]. One of the main challenges is ensuring broadband communication coverage wherever connectivity is required, including remote and rural areas. The high costs associated with spectrum licensing, as well as the implementation and maintenance of complex communications infrastructure in remote areas, have hindered the fifth-generation of mobile network (5G) from adequately addressing this important communication scenario. Moreover, remote and rural areas exhibit lower potential for clientele due to the low (and often less affluent) population density [2]. This means that low cost and effective solutions are crucial to connecting the unconnected.

To address these challenges, the open radio access network (O-RAN) presents a compelling opportunity to bridge the digital divide and improve connectivity in remote areas by offering cost-effective, flexible, and innovative solutions [3]. The open interfaces of this architecture enables a multi-vendor system, which will boost innovative network solutions in a competitive market [4]. This promote enhanced operational efficiency and reducing both implementation and maintenance costs associated with communication infrastructure in remote areas. Enhancing network intelligence represents another pivotal feature envisioned in

☆ This work was partially funded by the project National Science Foundation (NSF) (1263236, 0968895, 1102301); The 863 Program (2013AA014402). Also XGM-AFCCT-2024-2-15-1, supported by xGMobile – EMBRAPII-Inatel Competence Center on 5G and 6G Networks, with financial resources from the PPI IoT/Manufacturing 4.0 program of MCTI, under grant number 052/2023, signed with EMBRAPII. Additional support was provided by FAPESP under grants Nos. 20/05127-2, 2021/06569-1, and 2022/09319-9, including the SAMURAI project. This work was also supported by FAPEMIG under contract No. PPE-00124-23, RED-00194-23, and APQ-02746-21, as well as CNPq-Brasil.

* Corresponding author.
E-mail addresses: luiz.melo@inatel.br (L.A.M. Pereira), lucianol@inatel.br (L.L. Mendes), Mitchell.Cox@wits.ac.za (M.A. Cox), arismar@inatel.br (A.C. S. Jr.).
URLs: https://inatel.br/docentes/ (L.A.M. Pereira), https://inatel.br/docentes/ (L.L. Mendes), https://www.wits.ac.za/oclab (M.A. Cox), https://inatel.br/docentes/ (A.C. S. Jr.).

https://doi.org/10.1016/j.optcom.2025.131949
Received 5 February 2025; Received in revised form 11 April 2025; Accepted 30 April 2025
Available online 16 May 2025

O-RAN. Thereupon, dedicated logical components will be available to facilitate the integration of artificial intelligence (AI)-driven solutions into the network [5]. These solutions can enhance network efficiency, optimize service delivery, and enable innovative applications tailored to the needs of remote communities.

Reducing the costs associated with deploying base stations (BSs) is necessary to provide effective connectivity in remote areas, which requires innovative and hybrid solutions. For instance, the centralized radio access network (C-RAN) architecture using analog radio over fiber (A-RoF) as the transport link [6], or perhaps even free-space optics [7]. A-RoF enables the long distance transport of pre-modulated passband signals in their analog format, allowing for simpler and thus more cost-effective BSs [8]. This means that analog-to-digital converters (ADCs), digital-to-analog converters (DACs) and the up/down-conversion electronics are not necessary at the remote BS. This means that the remote BS is only composed of an optical detector and radiofrequency (RF) front-end.

Another promising (and potentially complementary) approach is the use of TV white space (TVWS) communications, which leverage unoccupied digital television (TV) channels for mobile communications [9]. Although digital TV services operate within frequency ranges that are typically regulated, TVWS refers to the unused portions of these bands that can be utilized in an unlicensed manner for other communications without causing interference to primary services. Thus, using TVWS for mobile communication can help reduce the costs associated with extending connectivity to remote areas but it requires careful compliance with regulatory requirements to ensure that these communication signals do not interfere with digital TV services.

Although the adoption of C-RAN with A-RoF and the exploitation of TVWS potentially reduce the capital expenditures (CAPEX) and operational expenditure (OPEX), several challenges must be addressed and overcome. For instance, when extending the fronthaul link, it is necessary to increase both the optical and RF power to compensate for additional attenuation. Unfortunately, increasing the transmission power leads to distortions in the transmitted signal due to the nonlinear behavior of Mach–Zehnder modulators (MZMs), employed in A-RoF systems. If these distortions are not adequately addressed, the signal at remote BSs will exhibit excessive out-of-band emission (OOBE), leading to a scenario where using the TVWS becomes non-compliant with legal regulations. Linearization schemes, including digital pre-distortion (DPD) techniques, are crucial for addressing this challenge.

Conventional linearization approaches, such as those based on the Volterra series, have been proposed to mitigate nonlinear distortions in A-RoF systems [10,11]. Simultaneously, there is growing interest in machine learning (ML)-based linearization methods, with numerous supervised learning solutions based on artificial neural networks (ANNs) being proposed and evaluated, showing significant improvements [12–15]. In particular, our research group has been developing ML-based linearization schemes for A-RoF systems, including cascaded A-RoF with power amplifier (PA), as well as ML-based models capable of simultaneously addressing both linear and nonlinear effects, such as chromatic dispersion [16–19]. Classical models, such as the one proposed in [10,11] are sensitive to amplitude and require recalibration when the quiescent operating point changes. Additionally, system responses naturally drift over time due to aging, necessitating periodic retraining. In contrast, ML-based solutions offer superior adaptability, reducing the need for frequent recalibration. Their generalization capabilities allow them to maintain performance despite variations in system behavior. Moreover, while training ML models can be resource-intensive, the inference process is computationally efficient, making them suitable for real-time applications where performance and efficiency are critical. These advantages position ML-based methods as a promising alternative to traditional models in nonlinear system modeling. This approach requires a representative dataset, encompassing all potential effects in an A-RoF system. One method to generate datasets for training ANNs involves creating synthetic data using mathematical models that represent the nonlinearities of A-RoF systems. However, fully capturing all potential effects in A-RoF systems through modeling is a complex challenge. Preferably, datasets can be generated using experimental or deployed A-RoF systems, but most studies utilizing real data do not make their datasets available for training the models [12,13,15]. This limitation significantly hampers the research community in developing innovative linearization schemes, which could substantially propel advancements in this domain.

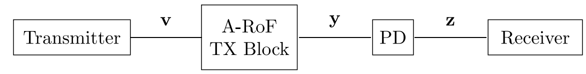


Fig. 1. Block diagram of the equivalent baseband model for an analog radio-over-fiber system.

The primary contribution of this paper is to provide access to a dataset derived from a real A-RoF system, which is described in detail later in this paper. We aim to facilitate research and development of innovative linearization schemes based on ML algorithms, among other possible applications, by making this dataset openly available to the research community. The second contribution is related to a detailed analysis of the memoryless polynomial model employed to represent the MZM nonlinearities. We demonstrate that this model can effectively map the input and output of A-RoF systems, enabling the construction of an accurate model to characterize A-RoF systems operating under various power and bandwidth configurations. This model can be utilized to generate new datasets for different protocols and waveforms, thereby reducing the necessity of costly experimental setups. Finally, we show that the synthetic data generated by the models presented in this paper closely aligns with the real system. We also evaluate the training process using synthetic data to train the neural network, which was subsequently tested on real data from our experimental setup. The results were promising, demonstrating that neural networks can be effectively trained with synthetic data for A-RoF system linearization. This approach facilitates the generation of large volumes of controlled data, which is crucial given that ML-based solutions are highly data-dependent. Moreover, we show that the trained model can be successfully applied to implement DPD in real systems, where it generalizes well to practical scenarios without the need for extensive data collection for each new system.

By providing this dataset and the accompanying analysis, we hope to enable researchers to develop effective linearization schemes for A-RoF systems. These schemes will support flexible and dynamic RF power allocation, which is crucial for addressing the connectivity needs of unserved or underserved remote regions. Ultimately, the work presented in this paper aims to contribute to the development of cost-effective and efficient solutions for bridging the digital divide and providing ubiquitous communication coverage in the 6G.

2. System description and models

Before exploring the experimental approach employed to compile our dataset, we present an overview of the system and the signals incorporated within it, alongside a baseband equivalent model that depicts the MZM used in A-RoF systems, illustrated in Fig. 1.

In this diagram, the transmitter block produces digital samples of a generalized frequency division multiplexing (GFDM), or other suitable waveform. The employed waveform for characterizing the nonlinearities of the optical modulator must be capable of driving the device into various operating conditions. This means that the waveform must have sufficiently large dynamic range, characterized by a high peak to average power ratio (PAPR). The GFDM waveform meet this criteria and was chosen for characterizing the MZM optical modulator employed in this work. Fig. 2 illustrates a set of digital samples of the GFDM waveform, indicating where the standard deviation (σ) and

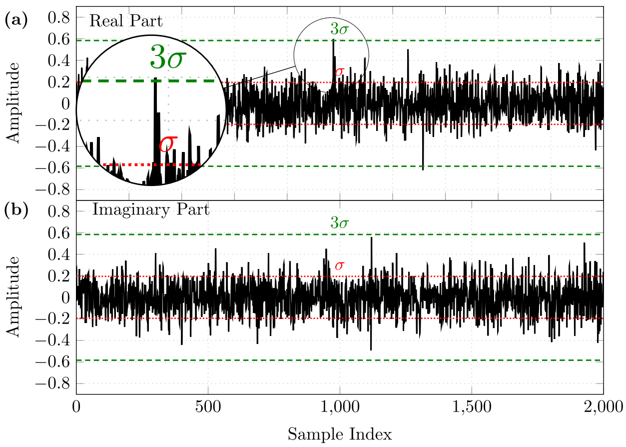


Fig. 2. GFDM waveform: (a) real part; (b) imaginary part.

(3σ) are situated. This demonstrates that the waveform includes values exceeding 3σ, thereby enabling the exploitation the full dynamic range of the optical modulator. In other words, it is an adequate waveform for characterizing the nonlinearities of the device. In this regard, GFDM modulation shares similar properties with the orthogonal frequency division multiplexing (OFDM) used in 5G [20]. GFDM is a multi-carrier waveform that enhances the OOBE control through the use of well designed pulse shaping filters [21]. This is particularly interesting for communications in remote areas utilizing TVWS technology.

GFDM is a non-orthogonal waveform with M sub-symbols in each of K sub-carriers. The vector containing the GFDM digital samples can be written as [22]

$$\mathbf{v} = \mathbf{A}\mathbf{d}, \tag{1}$$

where $\mathbf{A} \in \mathbb{C}^{KM\times KM}$, represents the transmit matrix, while $\mathbf{d} \in \mathbb{C}^{KM\times 1}$ denotes a data block comprising N elements. Hence, the total number of symbols is given by $N = KM$.

Other waveform alternatives, such as universal filtered OFDM (UF-OFDM), filtered OFDM (F-OFDM) and filterbank multicarrier (FBMC), may present varied trade-offs in complexity, spectral efficiency, channel condition robustness, and system compatibility. The choice of a waveform for a specific application depends on the requirements of the scenario, including the need for spectral efficiency, tolerance to latency and robustness to channel impairments [23].

The A-RoF TX Block consists of a laser diode (LD), a polarization controller (PC), an MZM modulator, and an adjustable direct current (DC) source for the bias voltage (V_{BIAS}). We selected the MZM for our experiment primarily due to its capability to handle high RF power levels before significant distortion occurs, while maintaining an acceptable insertion loss. It is worth noting that direct modulation could also be employed as a more cost-effective alternative. However, the nonlinear models typically used to characterize MZM nonlinearities may not accurately represent the behavior of directly modulated lasers, which necessitates further investigation. The LD emits light, which is then modulated by the MZM. Inside the MZM, the optical carrier is split into two separate paths. The RF electrical signal is applied to the electrodes along one of these paths, modulating the phase of the light in that arm of the interferometer. This configuration is known as a single-drive MZM. When the two paths recombine, the constructive and destructive interference between the phase-modulated and unmodulated light results in intensity modulation of the optical carrier, which directly corresponds to the RF signal. At the receiver, a direct detection mechanism is employed to recover the RF signal from the intensity-modulated optical carrier.

The MZM is typically biased at the quadrature point to mitigate nonlinearities, where the transfer function is most linear. This is achieved by setting the bias voltage (V_{BIAS}) to half of the voltage required for a phase shift of π radians (V_π) in the interferometer arms, i.e., $V_{\text{BIAS}} = V_{\pi/2}$. Operating at the quadrature point minimizes distortions introduced by the modulator. The MZM, polarized by a DC voltage supplied by the V_{BIAS} block, exhibits a memoryless nonlinear behavior that can be described by the following transfer function [10]:

$$y_n = \sum_{j=0}^{J-1} h_j |v_n|^j v_n, \tag{2}$$

where h_j represents the coefficients of the polynomial model, with J denoting the non-linear order of the model, v_n stands for the GFDM input signal samples and y_n the output signal. In matrix form, this model can be expressed as

$$\mathbf{y} = \mathbf{V}\mathbf{h}, \tag{3}$$

where $\mathbf{h} = [h_0, h_1, \dots, h_{J-1}]^{\mathrm{T}}, \mathbf{h} \in \mathbb{C}^{J\times 1}$ is the column vector containing the coefficients of the model, with $(\cdot)^{\mathrm{T}}$ denoting the matrix transposition operation, $\mathbf{y} = [y_0, y_1, \dots, y_{N-1}]^{\mathrm{T}}, \mathbf{y} \in \mathbb{C}^{N\times 1}$ represents the output signal, whereas $\mathbf{V} \in \mathbb{C}^{N\times J}$ is an input matrix described as follows

$$\mathbf{V} = [\mathbf{v}^0, \mathbf{v}^1, \dots, \mathbf{v}^{J-1}], \tag{4}$$

where

$$\mathbf{v}^j = \begin{bmatrix} |v_0|^j v_0 \\ |v_1|^j v_1 \\ \vdots \\ |v_{N-1}|^j v_{N-1} \end{bmatrix}. \tag{5}$$

At the receiver side, the photodetector (PD) may also potentially introduce nonlinear distortions, depending on the incident optical power. Considering remote and rural areas, PD will likely be situated far away from the transmitter. This allows us to simplify the model and disregard distortions produced by PD, due to the low optical power of the incident beam [24]. As such, the PD will only introduce noise:

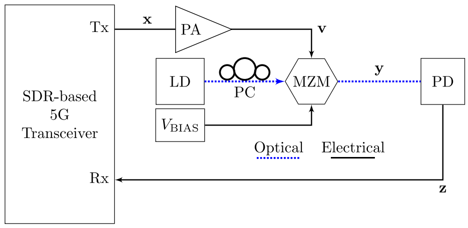
$$\mathbf{z} = \mathbf{y} + \mathbf{w}, \tag{6}$$

where $\mathbf{w}$ is the noises produced in the PD and associated electronics (see [7] for details) given by $\mathbf{w} \in \mathbb{C}^{N\times 1}$.

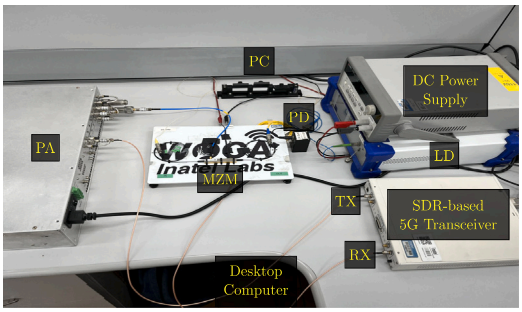
This model can be used to generate a simplified synthetic dataset for designing ML-based linearization schemes. The synthetic dataset consists of pairs of input signals ($\mathbf{v}_i$) and their corresponding output signals ($\mathbf{z}_i$) obtained by simulating the A-RoF system. While this dataset provides a valuable starting point for developing ML-based linearization techniques, fine-tuning and validation using experimental data from A-RoF system are essential to create robust solutions.

3. A-RoF dataset acquisition setup

While the simplified model and synthetic dataset provide a foundation for understanding the behavior of an A-RoF system, it is crucial to validate and refine the linearization techniques using experimental data from a real-world setup. In this section, we describe our specific experimental configuration designed to acquire a comprehensive dataset for an A-RoF system. Fig. 3(a) presents the block diagram of the A-RoF system, while Fig. 3(b) shows a photograph of the experimental setup. The process begins with a tunable laser from Golight, which generates a 1550.12 nm carrier with a optical power of 4 dBm. This carrier is subsequently introduced into a MZM, the FTM7920FBA model from Fujitsu, characterized by an insertion loss of approximately 5 dB. Despite being a specific type of modulator, it is highly representative of optical modulators in general due to its common usage. A PC is utilized to adjust the polarization of the light before feeding the MZM due to its sensitivity to polarization. An RF signal is fed into the PA before being introduced to the MZM. The class A PA is configured with sufficient back-off to ensure operation within its linear region, staying well below the 1 dB compression point. This setup is designed



(a) Block Diagram.



(b) Photography of the experimental setup.

Fig. 3. The proposed setup to generate the dataset.

to deliver high RF power (up to 15 dBm) at the RF input without introducing signal distortion. This approach allows for a high-power, undistorted signal to be provided at the MZM input, which cannot be achieved solely with the software-defined radio (SDR) RF front-end. Subsequently, the RF signals modulate the optical carrier through the MZM. This RF signal, centered at 707 MHz, is produced by a SDR-based 5G transceiver, specifically a National Instruments Universal Software Radio Peripheral (USRP) 2954R, which is connected to a desktop. The MZM is precisely biased to its quadrature point by adjusting its polarization voltage (V_{BIAS}) to 1.1 V, using a Keysight (E3640 A) power supply. For enhanced long-term stability in a production environment, the use of a dedicated MZM bias controller is recommended.

At the reception, a PD from EoT (model ET-5000F) receives the incoming −2 dBm optical signal. The modulation, signal processing, and coding were implemented using the GNU Radio framework for a real-time solution. The dataset is acquired by employing two Data Sync blocks from the GNU Radio Toolkit, which store the digital samples of the transmitted GFDM signal before the USRP RF front-end and after the received signal is sampled. It is important to note that the SDR front-end receiver was configured with an oversampling factor of four, ensuring that the captured signal includes any out-of-band distortions. A preamble is inserted at the beginning of the frame to enable time synchronization between the transmitted and received signals, as illustrated in Fig. 4. Synchronization is achieved by computing the cross-correlation between the transmitted and received datasets, identifying the index corresponding to the peak correlation. Once synchronized, the data becomes available for training ML-based model aimed at designing the linearization scheme.

Fig. 5 presents part of the GNU Radio flowgraph for the transmitter and receiver, illustrating the dataset acquisition process. The complete flowgraph diagram of the SDR-based 5G transceiver incorporates additional processing blocks. In Fig. 5(a), the File Sink Block stores the undistorted GFDM waveform samples that are transmitted to the UHD USRP Sink block, which sends the samples to a USRP 2954R device. These samples serve as dataset labels. The hardware device generates the analog RF GFDM signal, which is applied to the A-RoF system. The RF signal at PD output is sent to the receiver hardware, which samples the signal with a four-times oversampling factor to capture any out-of-band distortions necessary for linearizing the optical modulator

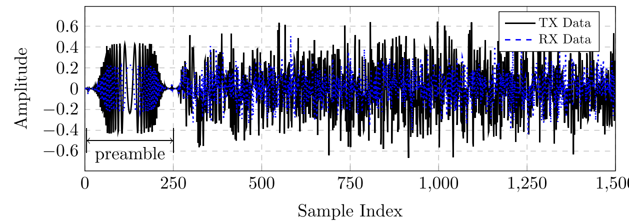
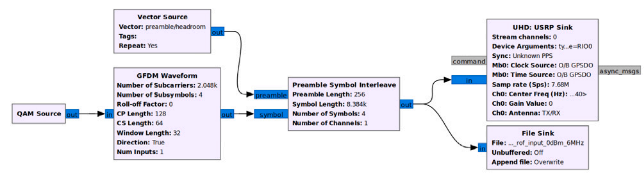
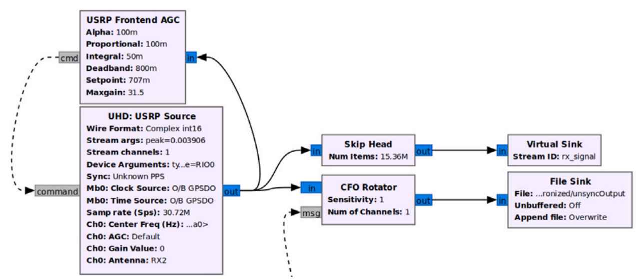


Fig. 4. Synchronization between transmitted (Tx) and received (Rx) data with preamble signaling.



(a) Simplified GNU Radio flowgraph diagram of the Transmitter.



(b) Simplified GNU Radio flowgraph of the Receiver.

Fig. 5. The simplified diagram of the GNU Radio flowgraph to illustrate the process of acquiring the dataset.

response. The UHD USRP Source Block, shown in Fig. 5(b), captures the analog signals connected to the USRP device, converts them into digital form, and feeds the digital samples into the GNU Radio flowgraph for further processing. Another File Sink captures the received samples, which have undergone the nonlinearities of the A-RoF system. After synchronization, these samples serve as training instances, enabling the learning of distortions characteristic of the A-RoF system.

4. Experimental validation, dataset description and utilization

This section reports the practical validation of the memoryless polynomial model described by (2) and the dataset description and guidelines for utilization.

4.1. Experimental validation of memoryless polynomial model

In the GNU Radio SDR-based 5G transceiver project, we store the digital samples of the transmitted and received GFDM waveform. After performing the synchronization process detailed in the previous section, we applied the normal equation to estimate the coefficients of the memoryless polynomial model for a given nonlinearity order J. The normal equations can be described as

$$\mathbf{h} = \left[\mathbf{V}^{\mathrm{H}}\mathbf{V}\right]^{-1}\mathbf{V}^{\mathrm{H}}\mathbf{z}, \tag{7}$$

here, $\mathbf{h} \in \mathbb{C}^{J\times 1}$ represents the coefficients of the memoryless polynomial model, $\mathbf{V}$ denotes the regression matrix as defined in (4), and $\mathbf{z}$ is specified by (6). For example, the coefficients of the polynomial

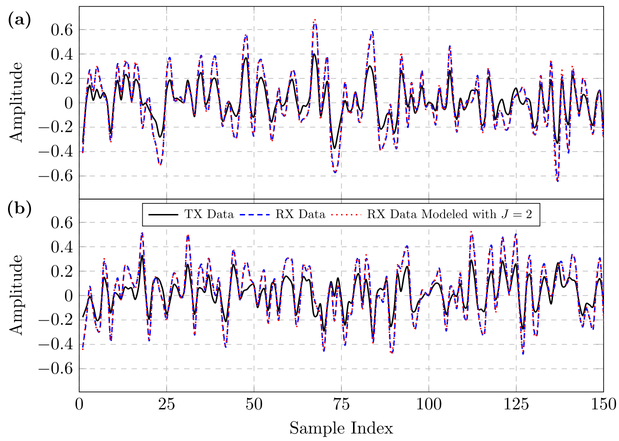


Fig. 6. Transmitted and received signals: (a) real part; (b) imaginary part.

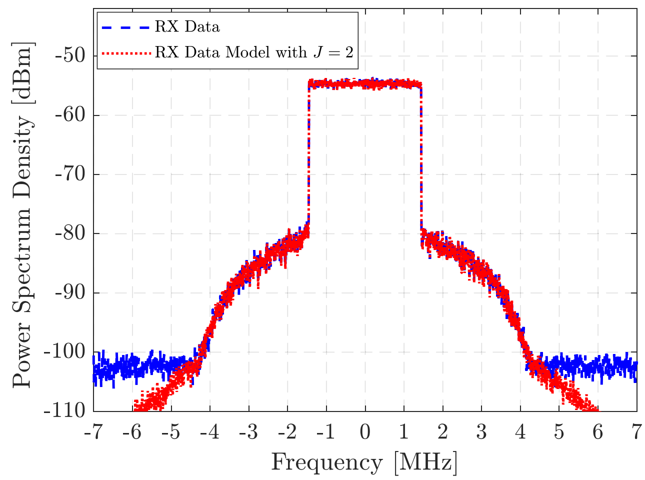


Fig. 7. Comparison between the measured and estimated spectra using a memoryless polynomial model for a 3 MHz bandwidth signal at 10 dBm RF power input to the MZM..

model for an input signal of 10 dBm at the MZM with a bandwidth of 3 MHz, using $J = 2$, are given by

$$\mathbf{h} = \begin{bmatrix} 1.1641 + j0.3312 \\ -1.5533 - j0.4325 \end{bmatrix}. \tag{8}$$

It is important to note that varying RF power levels and signal bandwidths will yield different coefficients for the model. The coefficients for all combinations of RF power and bandwidth can be estimated using the dataset provided in this paper and applying to Eq. (7).

The experimental validation of the polynomial model was conducted by comparing the actual output of the A-RoF system with the output generated by (3) using the coefficients derived from (8). Fig. 6 shows the real and imaginary components of the transmitted, received, and modeled GFDM signals. We observe that the memoryless polynomial model effectively captures the behavior of the A-RoF system, as evidenced by the comparison between the actual and modeled responses. This indicates that the model serves as an excellent tool for representing the A-RoF system under specific conditions. However, if the parameters of the A-RoF system change, the coefficients will also change, necessitating recalculation.

Fig. 7 presents a comparative spectral analysis between the measured and estimated data. The last one was estimated by the memoryless polynomial model. The measured data was collected using a 3-MHz bandwidth signal with 10 dBm of RF power at the MZM input. The estimated signal was derived by applying the transmitted signal samples

Table 1
Characterization of the MZM nonlinear model as a function of nonlinearity order with a 10 dBm RF input signal.

Nonlinearity Order	MSE	NMSE (dB)
$J = 2$	2.6584×10^{-5}	−25.7538
$J = 3$	2.4566×10^{-5}	−26.0966
$J = 4$	2.4541×10^{-5}	−26.1010
$J = 5$	2.4541×10^{-5}	−26.1011
$J = 6$	2.4539×10^{-5}	−26.1014
$J = 7$	2.4539×10^{-5}	−26.1014

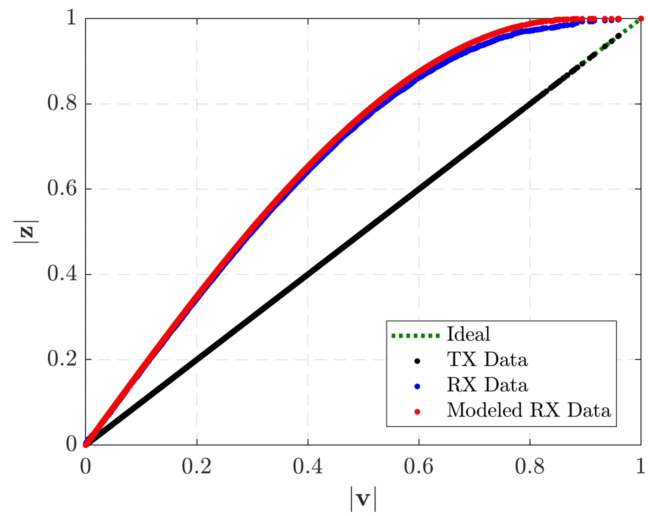


Fig. 8. Input magnitude as a function of output magnitude for 7 dBm RF power at MZM input.

to the memoryless model, with $J = 2$ as the memoryless polynomial model nonlinearity order and coefficients defined in (8). This result underscores the capability of the model to accurately approximate the actual data across the entire frequency spectrum.

We further validate the memoryless polynomial model by computing the absolute values of the signal input as a function of the signal output and comparing them with the actual data from the A-RoF system, as illustrated by Fig. 8. The memoryless polynomial model successfully captures the output of the A-RoF system, which endorses its suitability for representing such systems. Moreover, we explored various orders of nonlinearity and assessed the mean-squared error (MSE) and normalized mean square error (NMSE) between the actual output of the A-RoF system and the output predicted by the memoryless polynomial model, as demonstrated in Table 1. Alternative orders of nonlinearity (J) can effectively model the A-RoF system, yielding comparable levels of MSE and NMSE, which can reduce complexity by employing lower orders of nonlinearity.

4.2. ML-based pre-distortion

In this sub-section, we present the linearization results of a ML-based DPD scheme, previously described by us in [16]. The goal is to demonstrate the applicability of the proposed dataset for training a ML-based model and validate its linearization performance for the proposed real dataset. Initially, the proposed dataset was used to estimate the memoryless model coefficients for a desired bandwidth and RF power level. We selected a 3 MHz signal with an RF power of 10 dBm and varied the nonlinearity order (J) from 2 to 7. Each nonlinearity order resulted in distinct coefficient vectors, since $\mathbf{h} \in \mathbb{C}^{J \times 1}$.

Regarding the ML-based model, we employed a multi layer perceptron (MLP) neural network with the hyperparameters described in Table 2. The training was done throughout 750 epochs using a

Table 2
MLP Neural Network hyperparameters.

Hyperparameter	Input Layer	Hidden Layer 1	Hidden Layer 2	Hidden Layer 3	Output Layer
Number of Neurons	2	512	256	128	2
Activation Function	Linear	ReLu	ReLu	ReLu	Linear
Solver	Adam				
Loss Function	Mean Squared Error				
Epochs	750				
Batch Size	8192				

Table 3
DPD performance for distinct Nonlinearity order.

Nonlinearity	EVM_{RMS}(%)		NMSE (dB)		ACLR (dB)	
Order	Without DPD	With DPD	Without DPD	With DPD	Without DPD	With DPD
$J = 2$	6.31	1.94	−22.21	−28.31	−29.27	−34.24
$J = 3$	6.22	3.81	−22.29	−24.44	−29.00	−30.37
$J = 4$	6.74	4.48	−21.78	−23.02	−29.37	−30.62
$J = 5$	6.39	3.38	−22.06	−24.57	−28.61	−30.50
$J = 6$	6.55	3.89	−21.89	−23.86	−28.79	−30.47
$J = 7$	6.06	3.52	−22.49	−23.07	−28.97	−31.40

8192-samples batch. The number of epochs and neurons per layer were heuristically selected throughout observations during the training process. Hence, the linearization performance could be enhanced by conducting an exhaustive grid search to identify the optimal hyperparameters. During the training first epochs, MSE was about 10^{-3}. In addition, MSE rapidly decreases for the next epochs achieving $\approx 10^{-7}$. The first approach involved training the neural network using synthetic data, while the second approach was based on trained the MLP with real data from the provided dataset.

The performance evaluation was conducted with OFDM signals. This approach allows for validating the applicability of the dataset when different waveforms are employed, as the dataset includes the GFDM waveform. The generated OFDM signal has 2048 subcarriers, each one modulated by a 256-quadrature amplitude modulation (QAM) symbol, with a sampling rate of 3.84×10^6 samples per second. An oversample factor of 4 was used to enable visualize the out-of-band distortions due to the optical modulator. The OFDM samples energy was adjusted to ensure the same energy considered in the provided dataset. Table 3 shows the linearization performance as a function of MZM nonlinearity order (J) for the ML-based DPD trained with synthetic data. We varied the nonlinearity order parameter from 2 to 7 to identify the order that best characterizes the nonlinearities. The best results were achieved for $J = 2$, with improved root mean square error vector magnitude (EVM_{RMS}) from 6.31% to 1.94%. The NMSE and adjacent channel leakage ratio (ACLR) metrics were also enhanced, with improvements of approximately 6 dB and 5 dB, respectively.

We also utilized the provided dataset containing real data to train the ML-based DPD scheme. In this evaluation, the MLP-based neural network was retrained exclusively with real data, rather than synthetic data. The trained model was then tested on an OFDM signal using the same parameters as in the previous analysis. Similar results were achieved using this approach, with the EVM_{RMS} improving from 6.24% to 1.98%, while the NMSE and ACLR improved by 6.26 dB and 4.27 dB, respectively. It is important to note that the dataset was composed of samples of a GFDM signal, whereas the evaluation was performed for an OFDM signal. These results demonstrate that, regardless of the waveform used to generate the dataset, it is possible to design robust linearization schemes for different waveforms. As previously

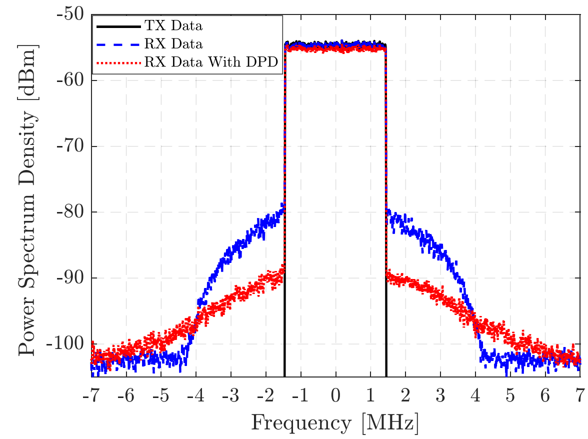


Fig. 9. Spectra from transmitted (TX) signal, the received (RX) signal, and the received signal after applying DPD trained on synthetic data and evaluated with real-world data..

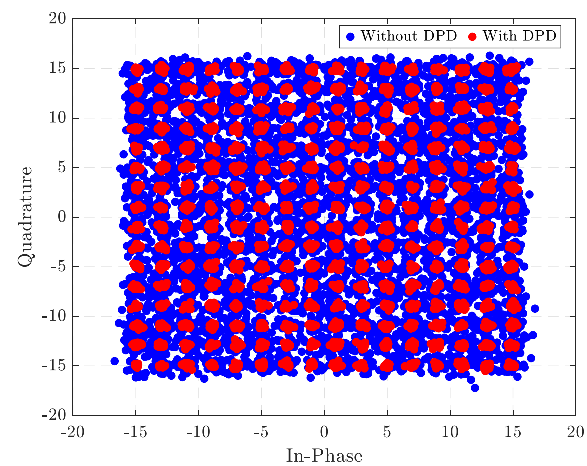


Fig. 10. Constellation diagram showing the In-Phase and Quadrature components of a signal with and without DPD..

mentioned, the waveform used to compose the dataset must have a high dynamic range, characterized by a high peak PAPR. This characteristic enables the capture of MZM nonlinearities, which are crucial for the proper design of the linearization scheme, whether it is based on ML solutions or not.

Finally, we conducted a new training process using synthetic data to train the neural network, which was subsequently evaluated on real data from our experimental setup. The objective was to assess the feasibility of using synthetic data for training ML-based linearization schemes in practical implementations. Fig. 9 illustrates the spectra for the received GFDM signal, with and without DPD technique. The approach yielded comparable results, with the EVM_{RMS} improving from 8.03% to 2.43%, and enhancements of 3 dB in NMSE and 3.8 dB in ACLR. These results demonstrate that synthetic data can be effectively employed to design linearization schemes for real-world datasets, reducing the need for extensive and costly acquisition of real data. It is important to emphasize that the models and coefficients used to train the DPD scheme must accurately represent the nonlinearities of A-RoF systems, which may require precise initial characterization, as highlighted in this paper. Fig. 10 illustrates the constellation of the received GFDM signal with and without ML-based DPD. The blue markers represent the signal constellation without DPD, while the red markers depict the improved constellation with DPD applied, indicating reduced distortion.

4.3. Dataset description and guidelines

The dataset was generated using 4-dBm optical power and RF signal centered at 707 MHz, with varying bandwidths, i.e. 3, 6, 12, 24 and 50 MHz. To compile a comprehensive dataset, we collected transmitted and received digital samples across various RF power levels. More specifically, we transmitted, received, and stored samples of RF signals across a range of RF power levels, from 0 to 15 dBm. For each configuration, we have stored 33536 samples, corresponding to one oversampled GFDM symbol. Each symbol comprises 2048 subcarriers, along with 128 samples for cyclic prefix and 64 for cyclic suffix, leading to 8384 samples for GFDM symbol before the oversampling process and 33536 after. This approach enables the capture of nonlinear behavior across a spectrum of intensities, ranging from mild to severe nonlinearities, as the nonlinearities inherent in A-RoF systems are power-sensitive.

We grouped each pair of datasets corresponding to the transmitted and received samples obtained at each tested RF power level. Therefore, eleven datasets pairs were produced, which can also be amalgamated to form a new dataset encompassing all the results, tailored to specific usage requirements. This can enhance the generalization capacity of the linearization scheme, resulting in robust solutions, which represents one of the most compelling aspects of employing ML in linearization schemes. The dataset has been pre-synchronized using the method detailed in Section 3, thereby ensuring that each training instance is appropriately matched with its corresponding label. The generated dataset is accessible on GitHub via the link provided in this paper.

Upon downloading the dataset from the repository, a DPD schemes can be designed. As an example, consider a dataset for 6-MHz and a 5-dBm powerRF signal. Algorithm 1 outlines the procedure to use the dataset to train a ML-based DPD scheme tailored for an A-RoF system. Selecting the appropriate ANN model is crucial for implementing ML-based DPD. Given that the memory effect of the A-RoF system can be disregarded, the MLP stands out as a compelling choice due to its trade-off between complexity and performance. [11,16]. After training the ANN, the weights must be saved in hierarchical data format version 5 (h5) and the ANN structure in a json file. These files can be loaded into an SDR-based transceiver to implement the ML-based linearization scheme for real A-RoF systems.

Algorithm 1 Training a neural network for DPD

```
Load transmission data (tx_data) from file as complex numbers
Load received data (rx_data) from file as complex numbers
Prepare the training dataset:
    Concatenate real and imaginary parts of rx_data to form train_matrix
    Concatenate real and imaginary parts of tx_data to form train_labels
Design the Neural Network:
    The input layer will have two neurons
    Select the number of hidden layers, determine the activation function type, and specify number of neurons for layer
    The output layer will have two neurons
Compile the model
Train the model using the selected configuration for a predefined number of epochs and a specified batch size
Save the trained model and model structure to files (.h5 and . json).
```

5. Conclusions

This paper introduced a comprehensive dataset generated from an experimental A-RoF system, available on GitHub. The current dataset encompasses GFDM signals with RF power ranges from 0 to 15 dBm, capturing distinct nonlinear behaviors as the nonlinearities within the A-RoF system intensify with increasing RF power. The dataset includes signals with bandwidths of 3, 6, 12, 24 and 50 MHz. By making this dataset open, we aimed to enhance the reproducibility of research, enabling researchers to compare results, validate models, and build upon existing work. Furthermore, this initiative could foster collaboration among researchers, leading to the development of more robust linearization techniques, advancing the field of ML-based linearization schemes.

Additionally, the paper presented the experimental validation of a memoryless polynomial model to represent A-RoF systems. This approach aids the generation of synthetic datasets to design new linearization schemes. When compared with the actual A-RoF output from practical experiments, the model yielded MSE and NMSE values close to 10^{-5} and -25 dB, respectively. This theoretical comparison underscored the proposed mode potential in simulating A-RoF systems, contributing and promoting technological advancement through open, collaborative research efforts.

Finally, the provided dataset offers valuable contributions to the design of linearization schemes for A-RoF systems. By utilizing this dataset, researchers can estimate the memoryless model coefficients for various RF power levels and bandwidths. This enables characterizing the memoryless nonlinear model to generate new sets of input and output signals tailored to distinct waveforms, protocols, and specific requirements. While we emphasized the potential of this dataset in supporting the design of ML-based linearization schemes, its utility is not confined to ML solutions. It can also be employed in the development of a wide range of linearization schemes, including those not based on machine learning. Moreover, the dataset can be used to test new models for optical modulators, including electro-optic modulators. Researchers can apply the input sample from our dataset as a reference or label when using ML-based solutions, generating a distorted version of the signal by passing it through their model. This allows training with labeled data and adapts the dataset to various system configurations, including changes to components such as optical power, dispersion, and modulation format.

CRediT authorship contribution statement

L.A.M. Pereira: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **L.L. Mendes:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Mitchell A. Cox:** Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation. **Arismar Cerqueira S.:** Writing – review & editing, Supervision, Resources, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Luiz Augusto reports writng assistance was provided by National Science Foundation. Luiz Augusto reports financial support was provided by Brazilian Company of Research and Industrial Innovation. Luiz Augusto reports financial support was provided by Minas Gerais tate Foundation of upport to the Research. Luiz Augusto reports financial support was provided by Federal Government of Brazil Ministry of Science Technology and Innovation. Luiz Augusto reports financial support was provided by National Council for cientfc and Technological Development. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially funded by the project National Science Foundation (NSF) (1263236, 0968895, 1102301); The 863 Program (2013AA014402). Also 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, under grant number 052/2023, signed with EMBRAPII. Additional support was provided by FAPESP under grants Nos. 20/05127-2, 2021/06569-1, and 2022/09319-9, including the SAMURAI project. This work was also supported by FAPEMIG under contract No. PPE-00124-23, RED-00194-23, and APQ-02746-21, as well as CNPq-Brasil.

Data availability

The datasets presented in this paper are available in https://github.com/inatelcrr/A-RoF-Linearization-Dataset.

References

[1] H. Tataria, M. Shafi, A.F. Molisch, et al., 6G wireless systems: Vision, requirements, challenges, insights, and opportunities, Proc. IEEE 109 (7) (2021) 1166–1199.

[2] L.L. Mendes, C.S. Moreno, M.V. Marquezini, et al., Enhanced Remote Areas communications: The missing scenario for 5G and beyond 5G networks, IEEE Access 8 (2020) 219859–219880.

[3] A. Giannopoulos, S. Spantideas, N. Kapsalis, P. Gkonis, L. Sarakis, C. Capsalis, M. Vecchio, P. Trakadas, Supporting intelligence in disaggregated open radio access networks: Architectural principles, AI/ML workflow, and use cases, IEEE Access 10 (2022) 39580–39595.

[4] S.K. Singh, R. Singh, B. Kumbhani, The evolution of radio access network towards open-RAN: Challenges and opportunities, in: 2020 IEEE Wireless Communications and Networking Conference Workshops, WCNCW, IEEE, 2020, pp. 1–6.

[5] A. Arnaz, J. Lipman, M. Abolhasan, M. Hiltunen, Toward integrating intelligence and programmability in open radio access networks: A comprehensive survey, Ieee Access 10 (2022) 67747–67770.

[6] Y.L. Lee, D. Qin, L.-C. Wang, et al., 6G massive radio access networks: Key applications, requirements and challenges, IEEE Open J. Veh. Technol. 2 (2020) 54–66.

[7] A. Trichili, M.A. Cox, B.S. Ooi, M.-S. Alouini, Roadmap to free space optics, J. Opt. Soc. Am. B 37 (11) (2020) A184, http://dx.doi.org/10.1364/JOSAB.399168, URL https://www.osapublishing.org/abstract.cfm?URI=josab-37-11-A184.

[8] Y.-J. Huang, G.-T. Lin, P.-H. Ting, et al., Analog mobile fronthaul supporting 128 RF chain with nonlinear compensator for B5G and 6G ma-MIMO beamforming, J. Lightwave Technol. 40 (20) (2022) 6867–6874.

[9] F. Mekuria, L. Mfupe, Spectrum Sharing for Unlicensed 5G Networks, in: 2019 IEEE Wireless Communications and Networking Conference, WCNC, IEEE, 2019, pp. 1–5.

[10] M. Noweir, Q. Zhou, A. Kwan, et al., Digitally linearized radio-over fiber transmitter architecture for cloud radio access network's downlink, IEEE Trans. Microw. Theory Tech. 66 (7) (2018) 3564–3574, http://dx.doi.org/10.1109/TMTT.2018.2819665.

[11] H. Chen, J. Li, K. Xu, et al., Experimental investigation on multi-dimensional digital predistortion for multi-band radio-over-fiber systems, Opt. Express 22 (4) (2014) 4649–4661.

[12] M.U. Hadi, K.U. Danyaro, A. Alqushaibi, et al., Digital predistortion based experimental evaluation of optimized recurrent neural network for 5G analog radio over fiber links, IEEE Access (2024).

[13] E. Liu, Z. Yu, Y. Song, et al., TL-ANN Based Nonlinear Equalization for Multi-Users Radio Over Fiber System, J. Lightwave Technol. 41 (5) (2022) 1399–1405.

[14] T. Xie, Q. Sheng, J. Yu, Deep learning equalizer connected with Viterbi-Viterbi algorithm for PAM D-band radio over fiber link, Sensors 23 (24) (2023) 9773.

[15] S. Liu, M. Xu, J. Wang, et al., A multilevel artificial neural network nonlinear equalizer for millimeter-wave mobile fronthaul systems, J. Lightwave Technol. 35 (20) (2017) 4406–4417.

[16] L.A.M. Pereira, L.L. Mendes, C.J.A. Bastos-Filho, Arismar Cerqueira S. Jr., Linearization schemes for radio over fiber systems based on machine learning algorithms, IEEE Photonics Technol. Lett. 34 (5) (2022) 279–282.

[17] L.A.M. Pereira, L.L. Mendes, C.J.A. Bastos-Filho, Arismar Cerqueira S. Jr., Machine learning-based linearization schemes for radio over fiber systems, IEEE Photonics J. 14 (6) (2022) 1–10.

[18] L.A.M. Pereira, L.L. Mendes, C.J.A. Bastos Filho, Arismar Cerqueira S. Jr., Amplified Radio-over-Fiber System Linearization Using Recurrent Neural Networks, J. Opt. Commun. Netw. 15 (3) (2023) 144–154.

[19] L.A.M. Pereira, L.L. Mendes, C.J.A. Bastos-Filho, Arismar Cerqueira S. Jr., Novel machine learning linearization scheme for 6G A-RoF systems, J. Lightwave Technol. 41 (23) (2023) 7245–7252.

[20] D. Zhang, M. Matthé, L.L. Mendes, G. Fettweis, A study on the link level performance of advanced multicarrier waveforms under MIMO wireless communication channels, IEEE Trans. Wirel. Commun. 16 (4) (2017) 2350–2365.

[21] G. Fettweis, M. Krondorf, S. Bittner, GFDM-generalized frequency division multiplexing, in: VTC Spring 2009-IEEE 69th Vehicular Technology Conference, IEEE, 2009, pp. 1–4.

[22] N. Michailow, M. Matthé, I.S. Gaspar, et al., Generalized frequency division multiplexing for 5th generation cellular networks, IEEE Trans. Commun. 62 (9) (2014) 3045–3061.

[23] M. Vaezi, Z. Ding, H.V. Poor, Multiple Access Techniques for 5G Wireless Networks and Beyond, Springer, 2019.

[24] K.J. Williams, R.D. Esman, M. Dagenais, Nonlinearities in p-i-n microwave photodetectors, J. Lightwave Technol. 14 (1) (1996) 84–96, http://dx.doi.org/10.1109/50.476141.