Autoencoder with Fitting Network for Terahertz Wireless Communications: A Deep Learning Approach

Zhaohui Huang[1], Dongxuan He[1,*], Jiaxuan Chen[1], Zhaocheng Wang[1,2,*], Sheng Chen[3]

[1] Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

[2] Shenzhen International Graduate School, Tsinghua Univeristy, Shenzhen 518055, China

[3] School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K.

[*] The corresponding author, email: dongxuan_he@mail.tsinghua.edu.cn

Abstract: Terahertz wireless communication has been regarded as an emerging technology to satisfy the ever-increasing demand of ultra-high-speed wireless communications. However, affected by the imperfections of cheap and energy-efficient Terahertz devices, Terahertz signals suffer from serve hybrid distortions, including in-phase/quadrature imbalance, phase noise and nonlinearity, which degrade the demodulation performance significantly. To improve the robustness against these hybrid distortions, an improved autoencoder is proposed, which includes coding the transmitted symbols at the transmitter and decoding the corresponding signals at the receiver. Moreover, due to the lack of information of Terahertz channel during the training of the autoencoder, a fitting network is proposed to approximate the characteristics of Terahertz channel, which provides an approximation of the gradients of loss. Simulation results show that our proposed autoencoder with fitting network can recover the transmitted symbols under serious hybrid distortions, and improves the demodulation performance significantly.

Keywords: Terahertz wireless communication; hybrid distortion; signal demodulation; autoencoder

Received: Sep. 01, 2021
Revised: Nov. 02, 2021
Editor: Jintao Wang

I. INTRODUCTION

With the exponential growth of data throughout demand, new spectral bands attract much attention in future wireless communications. Terahertz (THz) band (0.1 to 10 THz) has been regarded as one of the promising spectral bands to facilitate ultra-high-speed communications, which is capable of improving the throughput to Terabit-per-second (Tbps) [1–3]. Therefore, THz communication is envisioned as a key technology for the upcoming sixth-generation (6G) mobile communications and beyond [4, 5].

However, THz signals experience much more severe path loss than their counterparts in the lower frequency bands, which is induced by the spreading effect during propagation, the atmospheric attenuation effect caused by molecular absorption, and so on [3, 6]. In addition, ultra-high data rate leads to huge energy consumption and hardware costs. The imperfections of THz devices, including the in-phase/quadrature imbalance of ratio-frequency (RF) branches, the nonlinearity of power amplifier (PA), and the phase noise of local oscillator (LO), cause severe hybrid distortions to THz signals [7–9]. Furthermore, such distortions in THz communications cannot be effectively handled by the existing state-of-the-art techniques. For example, the widely-adopted minimum mean squared error (MMSE) equalization focuses on eliminating the influence of multi-path effects and inter-symbol interference by linear operation, which could not handle with the nonlinear imperfection in THz commu-

nication effectively. Besides, the powerful digital pre-distortion (DPD) methods require a high-rate and high-resolution RF device to obtain the exact distortion characteristics of the PA. These techniques may not be suitable to THz communication systems [10].

To improve the performance of the end-to-end communication system, deep learning based autoencoder has been used to learn the implementation details from both transmitter and receiver [11–13]. In [11], an adaptive transmission scheme was proposed to increase the transmission rate over an additive white Gaussian noise (AWGN) channel, where the noisy channel is treated as a noise layer and the transceiver is modeled as an autoencoder. Nevertheless, this scheme cannot handle the nonlinear distortions of hardware efficiently. In [12], an alternating algorithm was utilized to train the autoencoder-based system, which iterates between the training of the receiver by the actual gradients of loss and the training of the transmitter with an approximation of the gradients of the loss function. In [13], conditional generative adversarial net (GAN) was applied to model the effects of the unknown channel, which enables the gradients of the transmitter to be back-propagated from the receiver. With the aid of deep learning based autoencoder, reliable demodulation can be realized, revealing the potential of autoencoder-based transmission in handling the communications over complicated channels. However, for the THz channel, the channel model becomes much more complicated, where the scheme in [11] fails to design a proper autoencoder. Moreover, the performance of the schemes in [12] and [13] are usually unstable during the training process, due to the noise and complicated architectures.

In this paper, an autoencoder with fitting network for the THz channel is investigated, whereby the autoencoder is utilized to encode the transmitted symbols at the transmitter and recover these symbols at the receiver, respectively. Different from the existing state-of-the-art autoencoder-based transmission schemes, a fitting network, which is constructed from a deep feed-forward neural network (DFNN), is utilized to approximate the characteristics of the THz channel and the hybrid distortions of THz devices, which provides the gradients of loss during the training of autoencoder. Simulation results demonstrate the superior performance of our proposed autoencoder with fitting network, compared to the conventional schemes and the existing autoencoder-based counterparts.

II. SYSTEM MODEL

We consider a single-input single-output (SISO) THz communication system, where both the transmitter and the receiver are equipped with a single Cassegrain antenna. Without loss of generality, an equivalent baseband model is utilized to model the THz system, and the complex transmitted symbol is expressed as $y = y_I + \mathrm{j}\, y_Q$, where y_I and y_Q are the corresponding signals of the in-phase (I) and quadrature (Q) branches, respectively.

2.1 Hybrid Distortions in THz Transmission

At the transmitter, due to the imperfections of the quadrature modulator, the actual modulated signal can be expressed as

$$s = \mu_T y + v_T y^*, \tag{1}$$

where $(\cdot)^*$ denotes the complex conjugation operation, μ_T and v_T are the I/Q imbalance-related parameters at transmitter, given by [7]

$$\mu_T = \cos\left(\phi_T\right) - \mathrm{j}\,\epsilon_T \sin\left(\phi_T\right), \tag{2}$$
$$v_T = \epsilon_T \cos\left(\phi_T\right) - \mathrm{j}\,\sin\left(\phi_T\right), \tag{3}$$

in which ϵ_T and ϕ_T are the amplitude and phase imbalances between the I and Q branches, respectively.

Due to the severe nonlinearity of THz PA, the modulated signal suffers from both amplitude compression and phase rotation. In this paper, the odd order memoryless polynomial model, which is a general model of PA, is adopted to model the nonlinear distortion of the PA. The transmitted signal $\widetilde{s}$ is therefore related to the modulated signal s by [9]

$$\widetilde{s} = \sum_{k=1}^{K} \alpha_{2k-1} s|s|^{2(k-1)}, \tag{4}$$

where $2K-1$ is the order of nonlinearity, α_{2k-1} are the complex model parameters, and $|\cdot|$ represents the absolute value of a complex scalar.

After propagated through the THz channel, the received signal at the receiver can be expressed as $r =$

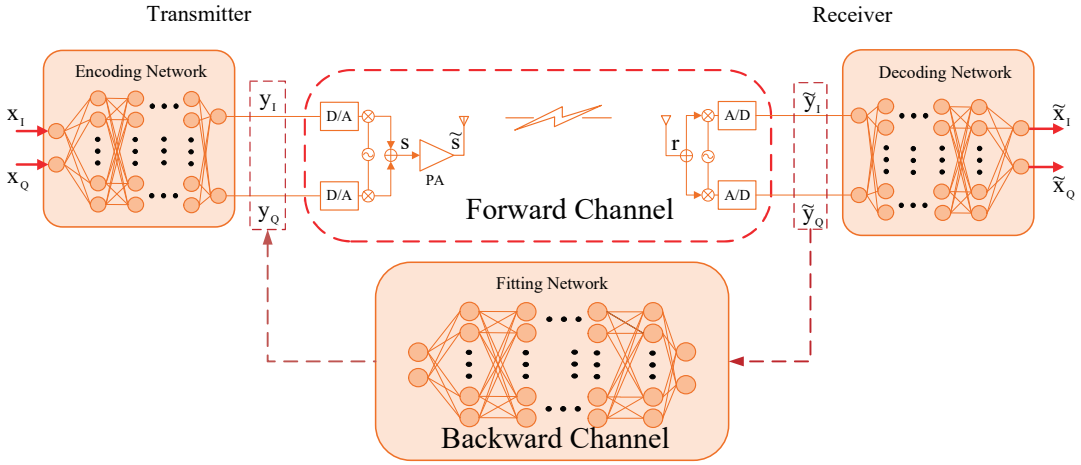

Figure 1. *The structure of our proposed autoencoder with fitting network.*

$H\widetilde{s}+\omega$, where $H = P_{\text{loss}} \cdot e^{\mathrm{j}\kappa}$ denotes the channel response, and ω denotes the baseband-equivalent AWGN with variance σ_ω^2. Here, P_{loss} is the variation of amplitude, and κ is phase shift [14].

Similar to the modulator, the imperfections of the quadrature demodulator also distort the received signal to

$$\widetilde{r} = \mu_R r + \upsilon_R r^*, \tag{5}$$

where μ_R and υ_R are the I/Q imbalance-related parameters at the receiver, given by

$$\mu_R = \cos\left(\phi_R\right) - \mathrm{j}\,\epsilon_R \sin\left(\phi_R\right), \tag{6}$$

$$\upsilon_R = \epsilon_R \cos\left(\phi_R\right) - \mathrm{j}\,\sin\left(\phi_R\right), \tag{7}$$

in which ϵ_R and ϕ_R are the amplitude and phase imbalances between the I and Q branches at the receiver, respectively.

Due to the instability of voltage-controlled oscillator (VCO), the phase noise, which increases with the square of the center frequency, affects the received symbols significantly, and the signal to be demodulated can be expressed as

$$\widetilde{y} = \widetilde{r} e^{\mathrm{j}\Delta\theta}, \tag{8}$$

where $\Delta\theta$ denotes the phase noise.

2.2 Problem Statement

To acquire the information from the received signal accurately, the distorted received signal $\widetilde{y}$ needs to be recovered to cancel the adverse effect of the THz channel before demodulation. However, due to the effects of hybrid distortions and noise, the distorted received signal cannot be recovered accurately when the received symbols corresponding to different constellation points overlap under the low signal noise ratio (SNR) region. In addition, it is worth noting that the gradients of the channel are unknown. Because the hybrid distortions of the THz communication system are highly complex, it is very challenging to directly estimate these distortions' parameters. Therefore, receiver and/or transmitter are blind to the specific parameters which they need to operate successfully.

To tackle this problem, as shown in Figure 1, an autoencoder-based transmission and reception with fitting network is proposed, which is capable of conveying the information successfully under severe THz hybrid distortions.

III. PROPOSED AUTOENCODER WITH FITTING NETWORK

We now detail our proposed autoencoder with fitting network shown in Figure 1. The proposed system consists of three networks, encoding network, fitting network, and decoding network, as well as two channels, forward channel and backward channel. The forward channel is the true THz channel that includes all the effects of the nonlinearity and distortions.

In our proposed scheme, the encoding network is added ahead the digital-to-analog converter (D/A) and the decoding network is added after the analog-to-

digital converter (A/D). The encoding network encodes the modulated symbol $x = x_I + \mathrm{j}\, x_Q$ into the transmitted symbol y, where x_I and x_Q are the corresponding symbols of the I and Q branches, respectively. The decoding network decodes the received symbol $\widetilde{y}$ into $\widetilde{x} = \widetilde{x}_I + \mathrm{j}\,\widetilde{x}_Q$, which is the estimate of x with $\widetilde{x}_I$ and $\widetilde{x}_Q$ as the estimates of x_I and x_Q, respectively. Due to the nonlinearity of THz devices and the complexity of the THz channel, the gradients of $\widetilde{y}$ with respect to y, which are also called the gradients of the forward channel, are hard to be obtained, and this makes the direct training of autoencoder impossible. Therefore, a fitting network, which is regarded as a backward channel, is introduced to facilitate the training of autoencoder by fitting the gradients of the forward channel.

3.1 System Architecture

As shown in Figure 1, all the networks in our proposed system are feedforward networks. Let the indexes e, f and d represent the encoding network, fitting network and decoding network, respectively. For each network, the number of hidden layers and the neurons in each hidden layer are denoted as L_i and $N_{i,j}$, for $i \in \{e, f, d\}$ and $j \in \{1, 2, \cdots, L_i\}$. In particular, improper hyper-parameters of network, including the number of hidden layers, the number of neurons and so on, will lead to underfitting or overfitting, which will deteriorate the performance of our proposed scheme. Therefore, numerical experiments are utilized to find the optimal hyper-parameters. Since all the inputs of the neural network are real-valued, the real and imaginary parts of the transmitted symbol are the input of the encoding network, and the real and imaginary parts of the received symbol are the input of the decoding network [15].

Each network has multiple hidden layers. The output of one layer is the input of the subsequent layer. Let the output vector of the j-th hidden layer in the network i be expressed as

$$\boldsymbol{o}_j^i = f\left(\boldsymbol{U}_j^i \boldsymbol{o}_{j-1}^i + \boldsymbol{b}_j^i\right), \tag{9}$$

where $\boldsymbol{U}_j^i$ and $\boldsymbol{b}_j^i$ are the weight matrix and the bias vector of the network's j-th layer, respectively, while $f(\cdot)$ denotes the activation function, which introduces nonlinearity to the network. We denote $\boldsymbol{W}_j^i = \left[\boldsymbol{U}_j^i\ \boldsymbol{b}_j^i\right]$ as the parameter matrix in the j-th layer of the network i. Since the output of each network should be normalized to within the interval $[-1,\ 1]$ to limit the power of the symbol, the *hardtanh* function is selected as the activation function, which can be expressed as

$$f(x) = \begin{cases} 1, & x > 1, \\ x, & -1 \le x \le 1, \\ -1, & x < -1. \end{cases} \tag{10}$$

It is evident that *hardtanh* function can provide nonlinearity and regularization to the network and it is beneficial to training because its derivative is easy to obtain.

The training process of our autoencoder with fitting network can be divided into two phases, training the fitting network and training the autoencoder network, where the mean square error (MSE) is used as the performance metric during these two training phases. Because the hybrid distortions of THz devices at receiver and the THz channel response H are unavailable to transmitter, the gradients of the loss function during the back-propagation of the encoder network are unavailable. Therefore, we train the fitting network (backward channel) to imitates the actual channel, so that we can use the gradients of the trained fitting network to assist the training of the encoder. As a result, the training of the fitting network must be prior to the training of the autoencoder network.

3.2 Training Fitting Network

In the system of Figure 1, $\boldsymbol{y} = [y_I\ y_Q]^{\mathrm{T}}$ is the output of the encoding network and $\widetilde{\boldsymbol{y}} = [\widetilde{y}_I\ \widetilde{y}_Q]^{\mathrm{T}}$ is the input of the decoding network. We use $\boldsymbol{y}$ and $\widetilde{\boldsymbol{y}}$ as the input and the desired output, respectively, to train the fitting network.

Let $h(\boldsymbol{y}, \omega)$ denote the response of the forward channel, which represents the actual response of THz devices and THz channel with the AWGN ω. That is, $\widetilde{\boldsymbol{y}} = h(\boldsymbol{y}, \omega)$. Further denote $g(\boldsymbol{y}; \boldsymbol{W}_f)$ as the response of the backward channel, which represents the mapping of the fitting network with $\boldsymbol{W}_f = \left[\boldsymbol{W}_1^f\ \boldsymbol{W}_2^f \cdots \boldsymbol{W}_{L_f}^f\right]$ being the weight matrix of the fitting network. In order to provide the response and gradients of the forward channel during the training of the autoencoder network, the loss function for training

the fitting network training can be expressed as

$$J_c(\boldsymbol{W}_f) = \mathbb{E}_{\boldsymbol{y},\omega}\Big[\big|h(\boldsymbol{y},\omega) - g(\boldsymbol{y};\boldsymbol{W}_f)\big|^2 + \big|\nabla_{\boldsymbol{y}} h(\boldsymbol{y},\omega) - \nabla_{\boldsymbol{y}} g(\boldsymbol{y};\boldsymbol{W}_f)\big|^2\Big], \quad (11)$$

where $\mathbb{E}_{\boldsymbol{y},\omega}[\cdot]$ denotes the expectation operator with respect to (w.r.t.) $\boldsymbol{y}$ and ω, while $\nabla_{\boldsymbol{y}}$ denotes the derivative w.r.t. $\boldsymbol{y}$.

Adam method [16] is adopted to train the fitting network by updating $\boldsymbol{W}_f$ until $J_c(\boldsymbol{W}_f)$ is lower than a predefined target. Moreover, batch algorithm is adopted to train the network to avoid overfitting [15]. For notational simplification, we use $\widetilde{g}(\boldsymbol{y})$ to represent the well-trained fitting network $g(\boldsymbol{y};\boldsymbol{W}_f)$ in the sequel.

The following proposition proves the effectiveness of training the autoencoder network based on the gradients generated by the backward channel.

Proposition 1. *The gradients generated by the backward channel can achieve nearly the same training performance as the actual gradients of the forward channel.*

Proof. See Appendix V.

To estimate the parameters of fitting network, the training complexity is about $\mathcal{O}(N_{\mathrm{f}}^2)$ with N_{f} denoting the total number of training neurons in fitting network [17]. Once the fitting network is well trained, it can be utilized to aid the training of the autoencoder.

3.3 Training Autoencoder Network

The input and output of the encoding network are the symbols $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, while the input and output of the decoding network are the symbols $\widetilde{\boldsymbol{y}}$ and $\widetilde{\boldsymbol{x}}$, respectively, where $\boldsymbol{x} = [x_I\ x_Q]^{\mathrm{T}}$ and $\widetilde{\boldsymbol{x}} = [\widetilde{x}_I\ \widetilde{x}_Q]^{\mathrm{T}}$. Let $\boldsymbol{y} = \Psi_e(\boldsymbol{x};\boldsymbol{W}_e)$ and $\widetilde{\boldsymbol{x}} = \Psi_d(\widetilde{\boldsymbol{y}};\boldsymbol{W}_d)$ be the mappings of the encoding network and decoding network, respectively, where $\boldsymbol{W}_e = \left[\boldsymbol{W}_1^e\ \boldsymbol{W}_2^e \cdots \boldsymbol{W}_{L_e}^e\right]$ denotes the weight matrix of the encoding network and $\boldsymbol{W}_d = \left[\boldsymbol{W}_1^d\ \boldsymbol{W}_2^d \cdots \boldsymbol{W}_{L_d}^d\right]$ denotes the weight matrix of the decoding network. It is evident that the mapping relationship between $\widetilde{\boldsymbol{x}}$ and $\boldsymbol{x}$ can be expressed as $\widetilde{\boldsymbol{x}} = \Psi_d\left(h\left(\Psi_e(\boldsymbol{x};\boldsymbol{W}_e),\omega\right);\boldsymbol{W}_d\right)$. To minimize the difference between $\widetilde{\boldsymbol{x}}$ and $\boldsymbol{x}$, the encoding network and decoding network should be trained simultaneously.

Specifically, the MSE between $\boldsymbol{x}$ and $\widetilde{\boldsymbol{x}}$, defined as

$$J(\boldsymbol{W}_e,\boldsymbol{W}_d) = \mathbb{E}_{\boldsymbol{x}}\left[\Psi_d(h(\Psi_e(\boldsymbol{x};\boldsymbol{W}_e),\omega);\boldsymbol{W}_d) - \boldsymbol{x}\right]^2, \quad (12)$$

is used as the loss function to train the autoencoder. As the gradient of the forward channel is unknown, during the training process, the fitting network is utilized to generate the approximate gradient of the forward channel. Adam method [16] and batch algorithm are also utilized to train the encoding network and decoding network simultaneously by updating $\boldsymbol{W}_e$ and $\boldsymbol{W}_d$ until $J(\boldsymbol{W}_e,\boldsymbol{W}_d)$ is lower than a predefined target, which are helpful to avoid underfitting or overfitting.

To estimate the parameters of encoding and decoding network, the training complexity is about $\mathcal{O}((N_{\mathrm{e}}+N_{\mathrm{d}})^2)$ with N_{e} and N_{d} denoting the total number of training neurons in encoding and decoding network, respectively. In particular, once the encoding and decoding networks are well-trained, only finite calculations are required to obtain the output.

IV. SIMULATION RESULTS AND DISCUSSION

We present the simulation results to verify the effectiveness of our proposed autoencoder with fitting network. Due to the directionality of Gassegrain antennas, a single path THz channel is considered. The parameters of I/Q imbalance are $\epsilon_T = \epsilon_R = 0.2$ and $\phi_T = \phi_R = 2°$, while the order of nonlinearity of PA is 5 with $\alpha_1 = 1.0108 + j0.0858$, $\alpha_3 = 0.0879 - j0.1583$, and $\alpha_5 = -1.099 - j0.8991$[1]. The phase noise is generated according to $\Delta\theta_{k+1} = \Delta\theta_k + \delta\theta_k$, where $\Delta\theta_k$ is the phase noise of the k-th block and $\delta\theta_k$ is the change of phase noise between adjacent blocks, which is a Gaussian random variable with $\delta\theta_k \sim \mathcal{N}\left(0,(5°)^2\right)$ [18].

Three schemes compared in the simulation study are:

1. MMSE: The transmitted signal is not processed at the transmitter and the frequency-domain MMSE equalization is executed to recover $\widetilde{x}$ from $\widetilde{y}$.
2. AE-AL: The autoencoder with alternating algorithm [12], where an approximate loss function's gradient is fed to the transmitter in each iteration of training the encoding network.

Table 1. *Structures of encoding, decoding and fitting networks.*

Network	Structures of hidden layers
Encoding Network	(16, 128, 256, 256, 128, 16)
Decoding Network	(16, 128, 256, 256, 128, 16)
Fitting Network	(40, 40, 40)

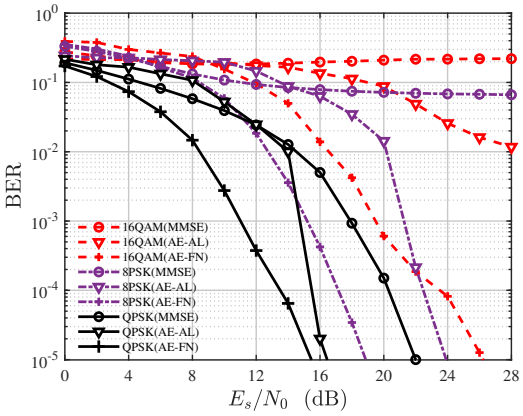

Figure 2. *BER performance comparison of three schemes with various modulations.*

3. AE-FN: Our proposed autoencoder with fitting network, as described in Section III.

Table 1 lists structures of the three neural networks adopted in our scheme, where for example, the structures of hidden layers (40,40,40) for the fitting network indicate that it has three hidden layers and each hidden layer has 40 neurons. In addition, 10^4 training examples are generated to train the neural networks, and the same number of training examples are utilized to train the MMSE estimator. In Figure 2, we plot the bit error rate (BER) versus E_s/N_0 for different schemes with different modulations, including QPSK, 8PSK and 16QAM, where E_s is the average power of the transmitted symbol, and N_0 is the power of the channel AWGN ω.

It can be seen from Figure 2 that the MMSE scheme has the worst BER performance, and for the 16-QAM and 8PSK modulations, it exhibits very high BER floors owing to failing to handle the hybrid distortions caused by the imperfections of THz devices. The AE-AL scheme [12] is much better but it still exhibits a high BER floor for the 16QAM modulation. Our proposed scheme significantly outperforms both the MMSE and AE-AL schemes, in terms of BER performance. In particular, for the QPSK modulation and at the BER level of 10^{-3}, our proposed scheme attains

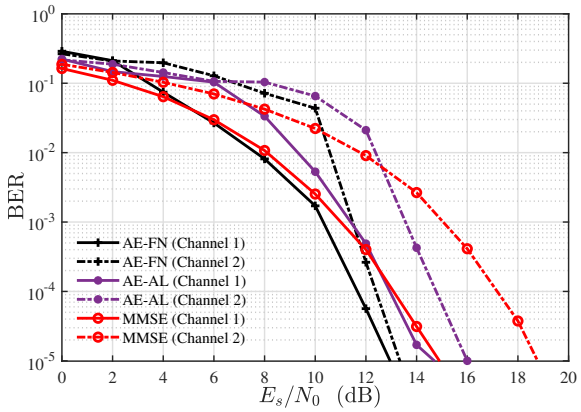

Figure 3. *BER performance comparison of three schemes experiencing various channels with QPSK modulation.*

Table 2. *I/Q imbalance parameters in different channels.*

Channels	ϵ_T	ϵ_R	ϕ_T	ϕ_R
Channel 1	0.15	0.15	1°	1°
Channel 2	0.30	0.30	5°	5°

around the 4 dB and 7 dB gains in the SNR, compared to the AE-AL [12] and MMSE, respectively. Moreover, the performance gain of our proposed scheme over the benchmark schemes increases with the modulation order. For example, at the BER level of 10^{-3}, the SNR gain of our AE-FN over the AE-AL increases to around 6 dB for the 8PSK modulation. For the 16QAM modulation, the AE-AL has the BER floor of around 10^{-2}, and the SNR gain of our proposed scheme over the AE-AL scheme is infinitely large, at the BER level of 10^{-3}.

In order to illustrate the applicability of our proposed AE-FN scheme, we study the performance of all three schemes with different I/Q imbalance. In Figure 3, the BER performance of AE-FN, AE-AL and MMSE experiencing different channels with different I/Q imbalance are compared, where I/Q imbalance parameters have been shown table 2.

Figure 3 shows that our proposed AE-FN scheme outperforms the other two schemes when the E_s/N_0 is high enough. With more serious distortions, the AE-FN scheme shows a more obvious performance gain compared to AE-AL and MMSE. For example, at the BER level of 10^{-3}, the SNR gain of AE-FN over MMSE is around 1 dB for channel 1 but around 3 dB for channel 2.

To examine the effectiveness of our proposed scheme in multi-antenna system, the BER perfor-

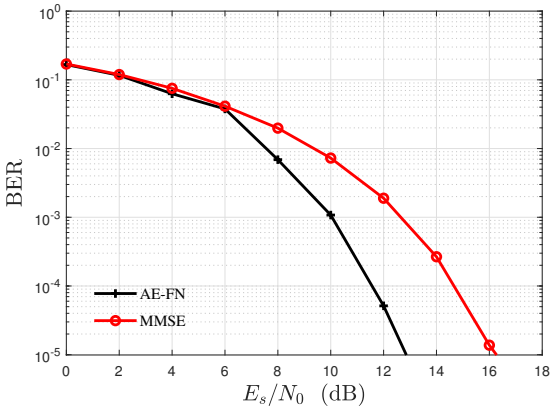

Figure 4. *Performance comparison of AE-FN and MMSE in MISO system with QPSK modulation.*

mance of multiple inputs and single output (MISO) scenario is presented in Figure 4. In detail, we consider a MISO system with four transmit antennas and QPSK modulation, the I/Q imbalance parameters of the four RF chains are $\epsilon_{T1} = 0.15, \epsilon_{T2} = 0.25, \epsilon_{T3} = 0.2, \epsilon_{T4} = 0.1, \phi_{T1} = 3°, \phi_{T2} = 2°, \phi_{T3} = 1°, \phi_{T4} = 2.5°$, respectively. It can be seen from the Figure 4 that the performance of our proposed AE-FN is significantly better than MMSE scheme. For example, at the BER level of 10^{-3}, the SNR gain of our AE-FN compared to MMSE is about 3 dB, which shows the superiority of our proposed scheme in MISO scenario.

V. CONCLUSIONS

In this paper, an improved autoencoder with fitting network has been proposed, which utilizes a deep learning based autoencoder to design transmitter and receiver for overcoming the hybrid distortions caused by the THz devices and the THz channel. To enable the training of the autoencoder, a deep feedforward neural network based fitting network has been introduced to approximate the characteristics of the THz devices and THz channel, which can generate the required gradient of the loss function for the back propagation of training the autoencoder. Simulation results have demonstrated that our proposed autoencoder with fitting network improves the BER performance considerably, compared with an existing state-of-the-art autoencoder counterpart.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (Grant 62101306), in part by the National Key R&D Program of China (Grant 2018YFB1801501), in part by Shenzhen Special Projects for the Development of Strategic Emerging Industries (20180608143929064O), in part by Shenzhen Wireless over VLC Technology Engineering Lab Promotion, and in part by Postdoctoral Science Foundation of China (Grant 2020M670332).

NOTES

[1]Here, more moderate I/Q imbalance is considered compared to [19].

APPENDIX

Proof of Proposition 1

We assume that the actual channel has the second-order derivative and its first-order derivative is bounded. When the fitting network is well trained, the loss function (11) is minimized. In particular, when the training set is sufficiently large, the following conditions are satisfied

$$\left|\widetilde{g}(\boldsymbol{y}) - \mathbb{E}_{\omega}\left[h(\boldsymbol{y},\omega)\right]\right| < \epsilon_1, \tag{13}$$

$$\left|\frac{\partial \widetilde{g}(\boldsymbol{y})}{\partial \boldsymbol{y}} - \mathbb{E}_{\omega}\left[\frac{\partial h(\boldsymbol{y},\omega)}{\partial \boldsymbol{y}}\right]\right| < \epsilon_2, \tag{14}$$

where ϵ_1 and ϵ_2 are any positive values.

For the decoding network, its training does not depend on the gradients of the forward channel, and the weight of the decoder $\boldsymbol{W}_d$ can be updated directly.

For the encoding network, the actual gradient of the loss function $J(\boldsymbol{W}_e, \boldsymbol{W}_d)$, given in (12), w.r.t. $\boldsymbol{W}_e$, denoted as $G(\boldsymbol{W}_e)$, can be expressed as

$$\begin{aligned} G(\boldsymbol{W}_e) =& \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\omega}\left[\nabla_{\boldsymbol{W}_e} J(\boldsymbol{W}_e, \boldsymbol{W}_d)\right]\right] \\ =& \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\omega}\left[\frac{\partial(\boldsymbol{x}-\widetilde{\boldsymbol{x}})^2}{\partial \widetilde{\boldsymbol{x}}}\frac{\partial \widetilde{\boldsymbol{x}}}{\partial \widetilde{\boldsymbol{y}}}\frac{\partial \widetilde{\boldsymbol{y}}}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{W}_e}\right]\right]. \end{aligned} \tag{15}$$

But $\frac{\partial \widetilde{\boldsymbol{y}}}{\partial \boldsymbol{y}}$ is unavailable, and we use $\frac{\partial \widetilde{g}(\boldsymbol{y})}{\partial \boldsymbol{y}} \approx \mathbb{E}_{\omega}\left[\frac{\partial \widetilde{\boldsymbol{y}}}{\partial \boldsymbol{y}}\right]$ to substitute for $\frac{\partial \widetilde{\boldsymbol{y}}}{\partial \boldsymbol{y}}$ in the back propagation.

Hence, the approximated gradient of the loss function $J(\boldsymbol{W}_e, \boldsymbol{W}_d)$ w.r.t. $\boldsymbol{W}_e$, denoted as $\widehat{G}(\boldsymbol{W}_e)$, can be expressed as

$$\widehat{G}(\boldsymbol{W}_e) = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\omega}\left[\frac{\partial(\boldsymbol{x}-\widetilde{\boldsymbol{x}})^2}{\partial\widetilde{\boldsymbol{x}}}\frac{\partial\widetilde{\boldsymbol{x}}}{\partial\widetilde{\boldsymbol{y}}}\mathbb{E}_{\omega}\left[\frac{\partial\widetilde{\boldsymbol{y}}}{\partial\boldsymbol{y}}\right]\frac{\partial\boldsymbol{y}}{\partial\boldsymbol{W}_e}\right]\right]. \tag{16}$$

The difference between $G(\boldsymbol{W}_e)$ and $\widehat{G}(\boldsymbol{W}_e)$ is given by

$$\begin{aligned}&\left|G(\boldsymbol{W}_e)-\widehat{G}(\boldsymbol{W}_e)\right| \\ &= \left|\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\omega}\left[\frac{\partial(\boldsymbol{x}-\widetilde{\boldsymbol{x}})^2}{\partial\widetilde{\boldsymbol{x}}}\frac{\partial\widetilde{\boldsymbol{x}}}{\partial\widetilde{\boldsymbol{y}}}\left(\frac{\partial\widetilde{\boldsymbol{y}}}{\partial\boldsymbol{y}}-\mathbb{E}_{\omega}\left[\frac{\partial\widetilde{\boldsymbol{y}}}{\partial\boldsymbol{y}}\right]\right)\frac{\partial\boldsymbol{y}}{\partial\boldsymbol{W}_e}\right]\right]\right| \\ &\leq \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\omega}\left[\left|\frac{\partial(\boldsymbol{x}-\widetilde{\boldsymbol{x}})^2}{\partial\widetilde{\boldsymbol{x}}}\frac{\partial\widetilde{\boldsymbol{x}}}{\partial\widetilde{\boldsymbol{y}}}\left(\frac{\partial\widetilde{\boldsymbol{y}}}{\partial\boldsymbol{y}}-\mathbb{E}_{\omega}\left[\frac{\partial\widetilde{\boldsymbol{y}}}{\partial\boldsymbol{y}}\right]\right)\frac{\partial\boldsymbol{y}}{\partial\boldsymbol{W}_e}\right|\right]\right].\end{aligned} \tag{17}$$

Since all the partial derivatives are bounded, the estimated gradient $\widehat{G}(\boldsymbol{W}_e)$ and actual gradient $G(\boldsymbol{W}_e)$ are sufficiently close once the difference $\left|\frac{\partial\widetilde{\boldsymbol{y}}}{\partial\boldsymbol{y}}-\mathbb{E}_{\omega}\left[\frac{\partial\widetilde{\boldsymbol{y}}}{\partial\boldsymbol{y}}\right]\right|$ is sufficiently small. Therefore, to prove **Proposition 1**, we need to show that the difference $\left|\frac{\partial\widetilde{\boldsymbol{y}}}{\partial\boldsymbol{y}}-\mathbb{E}_{\omega}\left[\frac{\partial\widetilde{\boldsymbol{y}}}{\partial\boldsymbol{y}}\right]\right|$ is sufficiently small when the SNR is sufficiently large or the noise ω is very small. First we have the following obvious lemma.

Lemma 1. *Let X be a Gaussian random variable following the distribution $X \sim \mathcal{N}(0, \sigma^2)$, and $Y = y(X)$ be a continue and bounded function of X. Clearly, X converges to 0 in probability, as $\sigma \to 0$. Due to the continuity of $y(\cdot)$, $Y = y(X)$ converges to $y(0)$ in probability, and $\mathbb{E}[Y]$ also converges to $y(0)$ because of the continuity and boundedness of $y(\cdot)$. As a result, $Y - \mathbb{E}[Y]$ converge to 0 in probability.*

Based on **Lemma 1**, $\left|\frac{\partial\widetilde{\boldsymbol{y}}}{\partial\boldsymbol{y}}-\mathbb{E}_{\omega}\left[\frac{\partial\widetilde{\boldsymbol{y}}}{\partial\boldsymbol{y}}\right]\right|$ is sufficiently small when the SNR is sufficiently large. This completes the proof.

References

[1] I. F. Akyildiz, J. M. Jornet, *et al.*, "Teranets: Ultra-broadband communication networks in the terahertz band," *IEEE Wireless Communications*, vol. 21, no. 4, 2014, pp. 130–135.

[2] K.-C. Huang and Z. Wang, "Terahertz terabit wireless communication," *IEEE Microwave Magazine*, vol. 12, no. 4, 2011, pp. 108–116.

[3] C. Han and Y. Chen, "Propagation modeling for wireless communications in the terahertz band," *IEEE Communications Magazine*, vol. 56, no. 6, 2018, pp. 96–101.

[4] M. Latva-aho, K. Leppänen, *et al.*, "Key drivers and research challenges for 6g ubiquitous wireless intelligence," 2020.

[5] I. F. Akyildiz, A. Kak, *et al.*, "6g and beyond: The future of wireless communications systems," *IEEE access*, vol. 8, 2020, pp. 133 995–134 030.

[6] G. A. Siles, J. M. Riera, *et al.*, "Atmospheric attenuation in wireless communication systems at millimeter and thz frequencies [wireless corner]," *IEEE Antennas and Propagation Magazine*, vol. 57, no. 1, 2015, pp. 48–61.

[7] Y. R. Ramadan, H. Minn, *et al.*, "Precompensation and system parameters estimation for low-cost nonlinear tera-hertz transmitters in the presence of i/q imbalance," *IEEE Access*, vol. 6, 2018, pp. 51 814–51 833.

[8] T. Mao, Q. Wang, *et al.*, "Spatial modulation for terahertz communication systems with hardware impairments," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, 2020, pp. 4553–4557.

[9] T. Schenk, *RF imperfections in high-rate wireless systems: impact and digital compensation*. Springer Science & Business Media, 2008.

[10] D. R. Morgan, Z. Ma, *et al.*, "A generalized memory polynomial model for digital predistortion of rf power amplifiers," *IEEE Transactions on signal processing*, vol. 54, no. 10, 2006, pp. 3852–3860.

[11] X. Chen, J. Cheng, *et al.*, "Data-rate driven transmission strategies for deep learning-based communication systems," *IEEE Transactions on Communications*, vol. 68, no. 4, 2020, pp. 2129–2142.

[12] F. A. Aoudia and J. Hoydis, "Model-free training of end-to-end communication systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 11, 2019, pp. 2503–2516.

[13] H. Ye, L. Liang, *et al.*, "Deep learning-based end-to-end wireless communication systems with conditional gans as unknown channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, 2020, pp. 3133–3143.

[14] A. R. Ekti, A. Boyaci, *et al.*, "Statistical modeling of propagation channels for terahertz band," in *2017 IEEE Conference on Standards for Communications and Networking (CSCN)*. IEEE, 2017, pp. 275–280.

[15] I. Goodfellow, Y. Bengio, *et al.*, *Deep learning*. MIT press, 2016.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] C. M. Bishop, *et al.*, *Neural networks for pattern recognition*. Oxford university press, 1995.

[18] G. Colavolpe, A. Barbieri, *et al.*, "Algorithms for iterative decoding in the presence of strong phase noise," *IEEE Journal on selected areas in communications*, vol. 23, no. 9, 2005, pp. 1748–1757.

[19] J. Antes and I. Kallfass, "Performance estimation for broadband multi-gigabit millimeter-and sub-millimeter-wave wireless communication links," *IEEE Transactions on*

Microwave Theory and Techniques, vol. 63, no. 10, 2015, pp. 3288–3299.

Biographies

Zhaohui Huang received the B.S. degree in electronic engineering from Tsinghua University in 2021. He is currently working towards the doctor's degree in Tsinghua University. His current interests include Terahertz communication and deep learning.

Dongxuan He received the B.S. degree in Automation from Beijing institute of technology (BIT) in 2013, and received Ph.D. degree in Information and communication systems from BIT in 2019. From 2017 to 2018, he was a visiting student in Singapore university of technology and design (SUTD). He is currently a postdoctoral researcher in department of electronic engineering, Tsinghua University. His current interests include Terahertz communication, deep leaning, physical layer security.

Jiaxuan Chen received the B.S. and Ph.D. degree from Tsinghua University, in 2016 and 2021 with the Department of Electronic Engineering, Tsinghua University. Her current research interests include wireless communications, signal processing, optical wireless communications, and vehicular ad-hoc networks.

Zhaocheng Wang received the B.S., M.S., and Ph.D. degrees from Tsinghua University in 1991, 1993, and 1996, respectively. From 1996 to 1997, he was a PostDoctoral Fellow with Nanyang Technological University, Singapore. From 1997 to 1999, he was a Research Engineer/Senior Engineer with OKI Techno Centre (Singapore) Pte., Ltd., Singapore. From 1999 to 2009, he was a Senior Engineer/Principal Engineer with Sony Deutschland GmbH, Germany. Since 2009, he has been a Professor with the Department of Electronic Engineering, Tsinghua University, where he is currently the Director of the Broadband Communication Key Laboratory, Beijing National Research Center for Information Science and Technology (BNRist).. His research interests include wireless communications, millimeter wave communications, and optical wireless communications.

Sheng Chen received the B.Eng. degree in control engineering from the East China Petroleum Institute, Dongying, China, in 1982, and the Ph.D. degree in control engineering from the City, University of London, London, U.K., in 1986. In 2005, he was awarded the higher doctoral degree, Doctor of Sciences (D.Sc.), from the University of Southampton, Southampton, U.K. From 1986 to 1999, he held research and academic appointments with the Universities of Sheffield, Edinburgh and Portsmouth, all in U.K. Since 1999, he has been with the School of Electronics and Computer Science, University of Southampton, where he holds the post of a Professor in intelligent systems and signal processing. His research interests include neural network and machine learning, wireless communications, and adaptive signal processing.