Explainable Asymmetric Auto-Encoder for End-to-End Learning of IoBNT Communications

Roya Khanzadeh*[1], Stefan Angerbauer*[1], Jorge Torres Gomez[3], Pit Hofmann[4], Falko Dressler[3], Frank H.P. Fitzek[4,5], Andreas Springer[2], and Werner Haselmayr[2]

[1]Johannes Kepler University Linz, Institute for Communications Engineering and RF-Systems, JKU LIT SAL eSPML Lab, Linz, Austria
[2]Johannes Kepler University Linz, Institute for Communications Engineering and RF-Systems, Linz, Austria
[3]TU Berlin, School of Electrical Engineering and Computer Science, Berlin, Germany
[4]Technische Universität Dresden, Deutsche Telekom Chair of Communication Networks, Dresden, Germany
[5]Centre for Tactile Internet with Human-in-the-Loop (CeTI), Dresden, Germany

***Abstract*—The Internet of Bio-Nano Things (IoBNT) is envisioned to be a heterogeneous network of artificial and natural units that are connected to the Internet. Hence, it extends the connectivity and control to unconventional domains, such as the human body. A potential use case for IoBNT is the communication from the outside to the inside of the human body. In this scenario, typically the Receiver (RX) inside the human body has limited computational complexity, while the Transmitter (TX) outside has large computational resources. In this paper, we address this scenario and propose a novel Asymmetric Auto-Encoder (AAEC) architecture for end-to-end learning of a Molecular Communication (MC) system. It applies a Neural Network (NN) at the TX and a low-complexity slope detector at the RX. We discuss the different layers of the NN-based TX and the corresponding training approach. Moreover, we investigate the explainability of the NN-based TX and show through the use of meta modeling that it can be approximated by a linear model. In addition, we demonstrate that the proposed AAEC resembles an MC system with Zero Forcing (ZF) precoding for low and moderate Inter Symbol Interference (ISI). Finally, through numerical results, we confirmed the aforementioned findings and showed that the proposed AAEC outperforms MC systems with and without ZF precoding, especially in high ISI scenarios.**

***Index Terms*—Internet of Bio-Nano Things, Machine Learning, Molecular Communications, Auto-Encoder, Explainable Artificial Intelligence.**

I. Introduction

The Internet of Bio-Nano Things (IoBNT) represents a significant paradigm shift in the fields of nanotechnology and communication engineering and has the potential to enable transformative applications in healthcare and nanomedicine. It enables the development of intra-body sensing, communication, and actuation through clusters of nano-scale bio-compatible artificial or biological embedded computing devices, so-called Bio-Nano Things (BNT) [1], [2]. These nano-devices are expected not only to interact with each other, but also to communicate and exchange information with nearby external electronic devices outside the biological environment as gateways to the Internet. Molecular Communications (MC) has emerged as a promising communication method among BNTs, using molecules for information transmission [3], [4]. However, MC channels often suffer from significant Inter Symbol Interference (ISI). Several modulation (e.g., [5]) and detection (e.g., [6], [7]) techniques have been proposed to mitigate ISI effects in MC channels. Recently, also Neural Networks (NN) have been applied for detection, such as Recurrent NNs (RNN) [8] and Convolutional NNs (CNN) [9]. Moreover, in [10] and [11] both Transmitter (TX) and Receiver (RX) are replaced by NNs, a so-called Auto-Encoder (AEC), which are jointly optimized to maximize the information rate.

Despite the promising performance results of the AEC approach, it suffers from high computation complexity. This is especially problematic in the context of IoBNT, when applied to the human body. Thus, in this work, we present an Asymmetric AEC (AAEC), which applies a NN at the TX, while having a low-complexity slope detector at the RX. This approach is very practical for information transmission from the outside to the inside of the human body. In particular, the TX consists of an external electronic device and an electronic-biological interface. The signal to be transmitted is derived in the external device using a NN and is then converted into a molecule signal using the biological-electronic interface. The RX (e.g., BNT) is located in the human body and makes a simple decision based on the received molecule concentration. Hence, the proposed AAEC on the one hand reduces the computational capabilities required by the BNTs, but exploits the resources available outside the human body.

In addition to the computation complexity, NNs typically suffer from the lack of insights into the decision-making process, which limits the transparency and trustworthiness of NN-based systems. This is addressed through the emerging research field of Explainable Artificial Intelligence (XAI) [12], which aims to understand the logic behind the results of AI-based algorithms. Due to sensitive applications envisioned for IoBNT, XAI has recently gained attention also in MC. For the first time, the explainability of a NN-based detection

*These authors contributed equally to this work.

for an MC system has been studied in [13], based on the visualization approach to evaluate the neuron's features (local interpretability). The results show the analogy between the NN and standard peak and slope detectors. In this work, we investigate the explainability of the proposed AAEC, which we refer to as explainable AAEC (XAAEC). We use symbolic meta modeling to find a mathematical mapping function from the inputs to the outputs (global interpretability) [14] and studied the similarity to existing precoding methods.

The main contributions of this work can be summarized as follows:

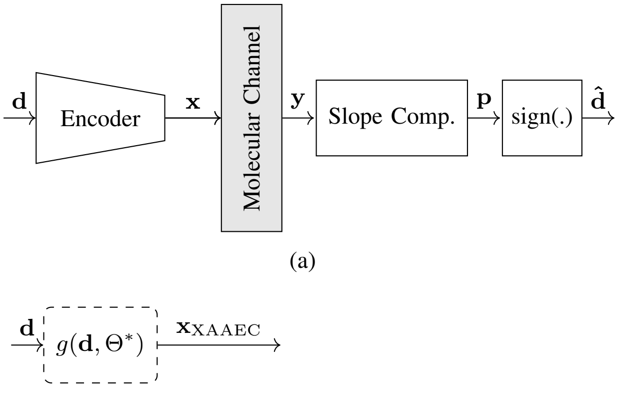
- We propose an Asymmetric Auto-Encoder (AAEC), which applies a NN at the TX and a low-complexity slope detector at the RX. This approach significantly improves the error performance compared to conventional MC systems.
- We study the explainability of the proposed NN-based TX and show that it can be interpreted as a linear precoder. Moreover, we show an analogy to a Zero Forcing (ZF) precoder for low and moderate ISI.

Notation: Vectors and matrices are denoted in bold face lower case $\mathbf{a}$ and upper case letters $\mathbf{A}$, respectively. The kth element of a vector $\mathbf{a}$ is named $a[k]$ and $[\mathbf{A}]_{k,l}$ addresses the element in row k and column l. The transpose operation and the Moore-Penrose inverse are expressed as $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\dagger}$, respectively.

II. Asymmetric Auto-Encoder

In this section, we introduce a novel Asymmetric Auto-Encoder (AAEC) concept, which allows us to find an end-to-end optimized TX for a low-complexity RX. Moreover, we discuss the corresponding training phase of the AAEC.

A. System Model



(a)

(b)

Fig. 1: (a) Proposed AAEC architecture; (b) meta-model of the encoder.

We consider the system shown in Fig. 1(a), which represents a potential IoBNT scenario for communication from the outside to the inside of the human body. The TX consists of an external computing unit (e.g., mobile phone), which has sufficient resources to run a NN, along with an electronic-biological interface (for simplicity, only the NN part of the TX is shown in Fig. 1(a)). The NN, shown as encoder, receives a block of M bits as input $\mathbf{d} \in \{0,1\}^M$ and generates the output $f(\mathbf{d}) = \mathbf{x} \in \{\mathbb{R}^+\}^M$, where $\mathbb{R}^+$ is the set of real positive values and $f(\mathbf{d})$ describes the mapping of the NN-based encoder. The hyperparameters of the applied NN are summarized in Table I. The NN-based encoder consists of a block-based CNN, where one-dimensional convolutional layers are used to only connect a few nearby neurons, which is inspired by the precoding principle in classical digital transmitters [15]. It includes three convolutional layers each coming with a batch normalization and a non-linear activation function, specifically a rectified linear unit (ReLu) function. The electronic-biological interface converts the output $\mathbf{x}$ of the NN into a molecular signal $x(t)$, i.e., it releases molecules into the human body. The relation between the output of the NN $\mathbf{x}$ and the molecular signal $x(t)$ is given by

$$x(t) = \sum_{n=0}^{\infty} x[n] \mathrm{rect}\left(\frac{t}{T_s} - \frac{n}{2}\right), \tag{1}$$

with $\mathrm{rect}(t) = 1$ for $-1/2 < t < 1/2$ and 0 otherwise, and T_s is the symbol duration. The released molecules propagate in the human body through blood vessels. The main propagation mechanisms are diffusion and advection and, thus, the Channel Impulse Response (CIR) can be expressed as [3]

$$h(t) = \frac{1}{\sqrt{4\pi D t}} e^{\frac{-(d-vt)^2}{4Dt}}, \tag{2}$$

with diffusion coefficient D, average flow velocity v, and the distance between TX and RX d. As an example, Fig. 2 shows the CIR for the parameters used throughout this work. Next, we derive an equivalent discrete representation of the system, which is required for the following discussions. The received molecular signal $y[n]$ (molecule concentration) in the nth interval is obtained as follows

$$y[n] = h_l * x[n] + w[n] = \sum_{l=0}^{L} h_l x[n-l] + w[n], \tag{3}$$

with additive Gaussian noise w, channel length L and h_l denotes the lth channel coefficient defined by

$$h_l = \int_{lT_s}^{(l+1)T_s} \frac{1}{\sqrt{4\pi D \tau}} e^{\frac{-(d-v\tau)^2}{4D\tau}} d\tau. \tag{4}$$

It is important to note that the channel length L needs to be chosen sufficiently large to cover the significant parts of the CIR (cf. Fig. 2), which also depends on the considered symbol duration.

We assume that the RX (e.g., BNT) has very limited computational resources and, thus, only applies a simple slope detection. The applied decision rule for the estimated bit in the nth interval can be expressed as

$$\hat{d}[n] = \begin{cases} 1, & \text{if } p[n] > 0 \\ 0, & \text{if } p[n] \leq 0 \end{cases}, \tag{5}$$

TABLE I: Layers of the proposed NN-based TX.

Encoder (TX)	
Type of layer	**Output size**
Input layer	M
Conv1d+BatchNorm1d+ReLu	$M \times 16$
Conv1d+BatchNorm1d+ReLu	$M \times 32$
Conv1d+ReLu+Max-pooling+Normalization	M

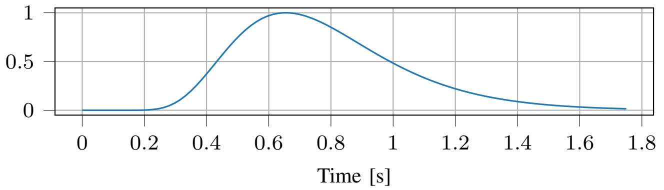


Fig. 2: Normalized Advection-diffusion CIR with $D = 1.24 \times 10^{-4}\,\mathrm{m}^2\mathrm{s}^{-1}$, $v = 0.055\,\mathrm{ms}^{-1}$, and $d = 38\,\mathrm{mm}$; the parameters are adopted from [18].

where $p[n]$ is the output of the slope computation and denotes the molecule concentration difference at the beginning of the current and the previous symbol interval

$$p[n] = y[n] - y[n-1]. \tag{6}$$

Based on the aforementioned results, we derive the following end-to-end model of the system depicted in Fig. 1(a), using matrix notation. Considering a noise-free transmission, the mapping from the input $\mathbf{x}$ to the output of the slope computation $\mathbf{p}$ can be expressed as

$$\mathbf{p} = \mathbf{W}\mathbf{x}, \tag{7}$$

where the matrix $\mathbf{W}$ is given by

$$\mathbf{W} = \begin{bmatrix} h_0 & 0 & 0 & 0 & \cdots & 0 \\ h_1 - h_0 & h_0 & 0 & 0 & \cdots & 0 \\ h_2 - h_1 & h_1 - h_0 & h_0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & -h_{L-2} & h_{L-1} - h_{L-2} & \cdots & h_1 - h_0 & h_0 \end{bmatrix}. \tag{8}$$

To illustrate, how this matrix can be obtained, we consider the first three channel coefficients h_l. Assuming zero-padding of $\mathbf{x}$, the first three values of $y[n]$ can be obtained using (3)

$$y[0] = h_0 x[0] \tag{9a}$$
$$y[1] = h_0 x[1] + h_1 x[0] \tag{9b}$$
$$y[2] = h_0 x[2] + h_1 x[1] + h_2 x[0]. \tag{9c}$$

By applying (6), the first three values of $p[n]$ can be derived as follows

$$p[0] = h_0 x[0] \tag{10a}$$
$$p[1] = (h_1 - h_0)x[0] + h_0 x[1] \tag{10b}$$
$$p[2] = (h_2 - h_1)x_0 + (h_1 - h_0)x[1] + h_0 x[2], \tag{10c}$$

Through rewriting these equations as the multiplication of a vector by a matrix ($\mathbf{p} = \mathbf{W}\mathbf{x}$), the structure of the matrix $\mathbf{W}$ can be clearly seen. Finally, the end-to-end model can be expressed as

$$\hat{\mathbf{d}} = \mathrm{sign}(\mathbf{W} f(\mathbf{d})), \tag{11}$$

where $\mathrm{sign}(\cdot)$ denotes the sign function. We refer to the system described above as Asymmetric Auto-Encoder (AAEC), which is an adaption of the classical AEC concept [16]. The principle of an AEC is that for a given channel, the NN at TX and RX are optimized with respect to the end-to-end performance [17], instead of optimizing their individual performance. Typically, the TX-NN and RX-NN have approximately the same complexity and, thus, AEC can be considered symmetric. However, especially in IoBNT applications it might be the case that the RX has limited computational complexity, while the TX has enough resources. This use case is addressed by the proposed AAEC, which applies a NN at the TX and a low-complexity slope detector at the RX. Nevertheless, the approach maintains the property, that the TX is optimized with respect to end-to-end performance.

TABLE II: Simulation parameters

Parameter	**Value**
Optimizer	Stochastic Gradient Decent
Learning Rate (η)	0.0008
Learning Rate Decay	0.999
Number of Epochs	4000
Batch Size	40
Input Size (M)	20

B. Model Training

The goal of the training of the NN-based TX is to minimize the error probability of the entire system (cf. Fig. 1), i.e., $\mathbf{d}$ and $\hat{\mathbf{d}}$ should differ as little as possible for any possible choices of $\mathbf{d}$ and channel conditions. However, the $\mathrm{sign}(\cdot)$ function used for the slope detection (cf. (11)) is not differentiable and, thus, does not allow for the optimization (gradient descent) of the end-to-end performance as discussed above for the AEC approach. To resolve this issue, during training, we replace the $\mathrm{sign}(\cdot)$ function by the differentiable sigmoid function $\sigma(\cdot)$, which leads to

$$\hat{\mathbf{d}}_t = \sigma(\mathbf{W} f(\mathbf{d})). \tag{12}$$

This enables the interpretation of the individual entries of $\hat{\mathbf{d}}_t$ as the probability of the bit at the nth interval belonging to the class labeled '1'. Hence, we can use the Binary Cross Entropy loss (BCE)

$$\begin{aligned} J(\theta_{\mathrm{TX}}) :=& \mathbb{E}_{\mathbf{y},\mathbf{d}}\big[-\log\big(p_{\theta_{\mathrm{TX}}}(\hat{\mathbf{d}}_t|\mathbf{y})\big)\big] \\ =& -\sum_{\forall \mathbf{d}} \big((d[n])\log(\hat{d}_t[n]) \\ &+ (1 - d[n])\log(1 - \hat{d}_t[n])\big), \end{aligned} \tag{13}$$

as the optimization loss function. The BCE can be interpreted as a measure for the average error probability using the AAEC and, thus, minimizing the BCE also minimizes the error probability. By the end-to-end-model in (12) and the BCE in (13) we now have a differentiable model and a loss function

TABLE III: Linearity score ε defined in (16) between AAEC and XAAEC.

T_s [s]	2	1.8	1.6	1.4	1.2	1
ε	0.9729	0.9750	0.9728	0.9734	0.9746	0.9638

TABLE IV: Metric for similarity between XAAEC and ZF precoder defined in (19) for $T_s = 1\,\mathrm{s}$.

M	20	50
ρ	0.9985	0.9990

that allows to train the NN-based TX using a gradient descent algorithm. The parameters used for training are summarized in Section II-A. In particular, we use Stochastic Gradient Decent (SGD) with a learning rate of $\eta = 0.0008$ and the model converges after around 4000 epochs of training.

III. Explainable AAEC

In this section, we study the interpretability of the proposed AAEC architecture, especially the NN-based TX. A NN is usually a black box, where we are not able to say why a certain input leads to a certain output. Shedding light onto this is the subject of explainable AI, which aims to understand the logic behind NNs. In the following, we apply this approach to the proposed AAEC. We take a two-step approach, which is described in the following subsections.

1) First Step - Finding a Meta Model: The NN-based encoder of the AAEC in Fig. 1(a) can be represented by the function $f(\mathbf{d})$. The aim of the meta model $g(\mathbf{d}, \mathbf{\Theta})$ is to find a less complex and interpretable approximation of $f(\mathbf{d})$ (cf. Fig. 1(b)). The model is parametrized by the matrix $\mathbf{\Theta}$, which is obtained through solving the following optimization problem using a gradient descent algorithm

$$\mathbf{\Theta}^* = \underset{\mathbf{\Theta}}{\mathrm{argmin}} \| f(\mathbf{d}) - g(\mathbf{d}, \mathbf{\Theta}) \|_{\mathrm{F}}, \tag{14}$$

where $\|.\|_{\mathrm{F}}$ denotes the Frobenius norm. We assume a linear model and, thus, the meta model can be expressed as

$$g(\mathbf{d}, \mathbf{\Theta}^*) = \mathbf{x}_{\mathrm{XAAEC}} = \mathbf{\Theta}^* \mathbf{d}, \tag{15}$$

with the $M \times M$ matrix $\mathbf{\Theta}^*$. Next, we evaluate how well the meta model approximates $f(\mathbf{d})$. As a metric, we use the linearity score defined by

$$\varepsilon = r(\mathbf{x}_{\mathrm{XAAEC}}, \mathbf{x}), \tag{16}$$

where $r(\cdot, \cdot)$ denotes Pearson's correlation coefficient [19]. Table III shows the linearity score for six different symbol durations. i.e., T_s. Moreover, Fig. 3 illustrates the transmitted signal generated by the meta model $\mathbf{x}_{\mathrm{XAAEC}} = g(\mathbf{d}, \mathbf{\Theta}^*)$ and the proposed NN $\mathbf{x} = f(\mathbf{d})$, along with their corresponding received signal $\mathbf{y}_{\mathrm{XAAEC}}$ and $\mathbf{y}$ after the MC channel, i.e. Since the linearity score is close to one and the signals in Fig. 3 match well, we conclude that the proposed AAEC can be very well approximated through the linear model in (15) and, thus, interpreted as linear precoder. In the following we refer to (15) as explainable AAEC (XAAEC).

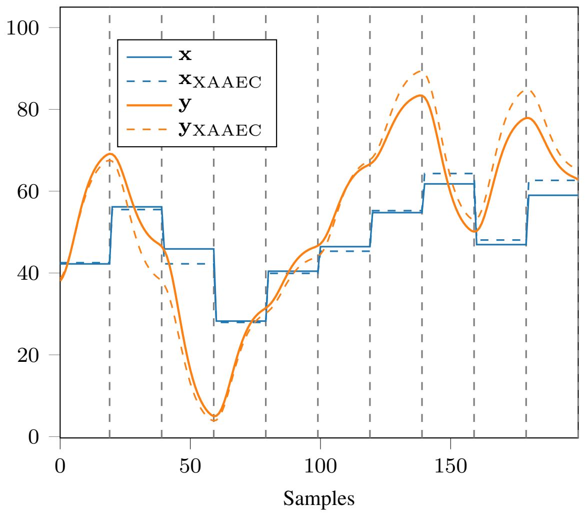


Fig. 3: Transmitted signals from the linear model $\mathbf{x}_{\mathrm{XAAEC}}$ and the actual encoder $\mathbf{x}$ and their corresponding received signals $\mathbf{y}_{\mathrm{XAAEC}}$ and $\mathbf{y}$ for the input bit sequence $\mathbf{d} = [1, 0, 0, 1, 1, 1, 1, 0, 1, 0]$ and $T_s = 1\,\mathrm{s}$. Vertical gray lines indicate symbol intervals.

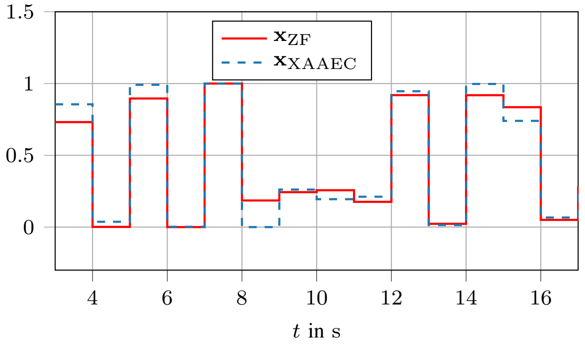


Fig. 4: Normalized transmitted signals of the XAAEC (linear model) $\mathbf{x}_{\mathrm{XAAEC}}$ and ZF precoder $\mathbf{x}_{\mathrm{ZF}}$ for the input bit sequence $\mathbf{d} = [1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0]$ and $T_s = 1\,\mathrm{s}$.

2) Second Step - Meta Model vs. Classical Approaches: In the previous step, we have shown that the linear model in (15), i.e., XAAEC, approximates the actual AAEC reasonably well. Now, we want to get further insights into this model. Therefore, we compare the XAAEC to well-know linear precoder designs [20]. We found that the XAAEC performs similarly to a Zero Forcing (ZF) precoder for low to moderate ISI (i.e., $T_s \geq 1\,\mathrm{s}$), which we will elaborate in the following. The output of the ZF precoder $\mathbf{x}_{\mathrm{ZF}}$ can be expressed as

$$\mathbf{x}_{\mathrm{ZF}} = \mathbf{W}_{\mathrm{ZF}} \mathbf{d}, \tag{17}$$

with the input bit sequence $\mathbf{d}$ (see Sec. II) and the precoding matrix given by [20]

$$\mathbf{W}_{\mathrm{ZF}} = \mathbf{W}^{\dagger}, \tag{18}$$

with the channel matrix $\mathbf{W}$ as defined in (8). Finally, we show that the XAAEC (15) and the ZF precoder (17) perform similar on a given input $\mathbf{d}$. We measure the similarity in two ways, namely through comparing the transmitted signals and the matrices, respectively. Fig. 4 illustrates the signals $\mathbf{x}_{\text{ZF}}$ and $\mathbf{x}_{\text{XAAEC}}$ for $T_s = 1\,\text{s}$ (moderate ISI) and we observe that they match very well. Next, we compare the matrices $\mathbf{W}_{\text{ZF}}$ and $\mathbf{\Theta}^*$ and define the following metric for it

$$\rho = \frac{1}{2}\left(\rho_{-1} + \rho_{+1}\right), \tag{19}$$

with

$$\rho_{\pm 1} = r(\Delta^{\pm}(\mathbf{\Theta}^*), \Delta^{\pm}(\mathbf{W}_{\text{ZF}})), \tag{20}$$

and $\Delta^{\pm}(\mathbf{Z}) = [\mathbf{Z}]_{i\pm 1,i} - [\mathbf{Z}]_{i,i}$ for a matrix $\mathbf{Z}$. Hence, (20) computes the Pearson's correlation coefficient $r(\cdot,\cdot)$ of the difference between the main and the first off-diagonal elements of the matrices $\mathbf{W}_{\text{ZF}}$ and $\mathbf{\Theta}^*$. More details about this metric can be found in the Appendix. Table IV shows the metric for two different numbers of transmitted bits M and we observe that for both cases the metrics are close to one. Hence, since the signals in Fig. 4 match well and the metric is close to one, we conclude that the XAAEC and the ZF perform similarly for low and moderate ISI scenarios (i.e. symbol duration above $1\,\text{s}$).[1]

IV. Simulation Results

In this section, we study the error performance (i.e., Bit Error Ratio (BER)) of the proposed AAEC architecture in comparison with a conventional MC system and ZF precoding. Moreover, we compare the robustness of the AAEC and ZF precoding against channel variations. Fig. 5 compares the BER of the proposed AAEC architecture with a conventional MC system using Concentration Shift Keying (CSK) at the TX and an advanced slope detector [18] at the RX. We observe that the AAEC significantly outperforms the conventional system for a wide range of symbol intervals T_{s}, which is expected due to the increased complexity at the TX. Nevertheless, it demonstrates the possible performance gain when increasing the TX complexity, which is especially interesting for IoBNT applications. A BER comparison of the AAEC and an MC system with a ZF precoder at the TX (cf. Sec. IV) and the basic slope detector defined in (5) at the RX is shown in Fig. 6. For low ISI scenario ($T_s = 1\,\text{s}$), AAEC and ZF precoder perform closely, while for high ISI scenario ($T_s = 0.15\,\text{s}$) the AAEC clearly outperforms the ZF precoder. The findings underscore the close parallel in performance between the AAEC and ZF precoders, while also highlighting that in scenarios with high ISI, the AAEC demonstrates a notably superior performance compared to the ZF precoder.

Finally, in Fig. 7 we compare the AAEC and ZF precoder with respect to their robustness to variations in the MC channel. Thus, we illustrate the BER for different relative channel mismatches given by $\delta = (D_{\text{real}} - D)/D$, with the real diffusion coefficient D_{real} and its mismatched value D. The results reveal that for a high ISI scenario, the proposed AAEC is clearly more stable to a model mismatch.

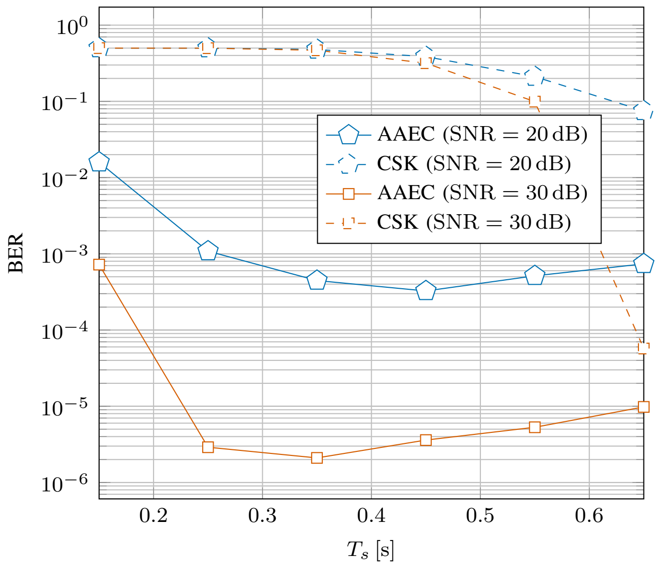


Fig. 5: BER performance of the proposed AAEC and an MC system using CSK.

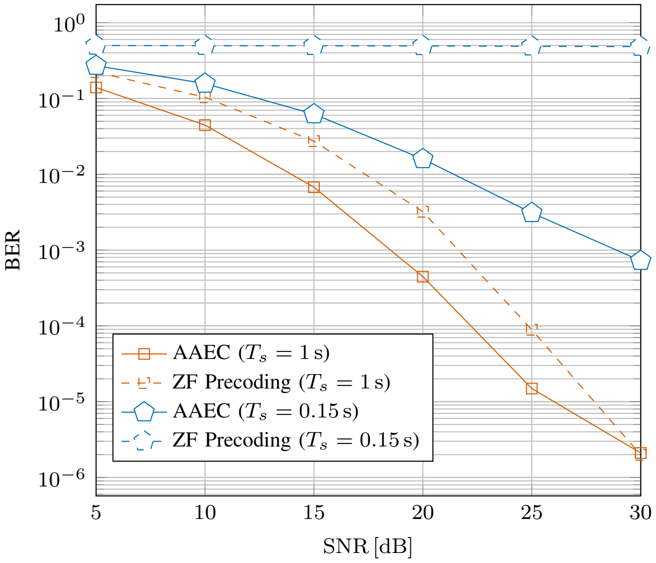


Fig. 6: BER performance of the proposed AAEC and an MC system using ZF precoding for low and high ISI.

V. Conclusion and Outlook

In this paper, we presented a novel AAEC architecture, with a NN at the TX and a low-complexity slope detector at the RX. The proposed AAEC overcomes computational constraints, which may occur in IoBNT communication scenarios. We presented a NN-based TX architecture and then an approach for training it. Furthermore, we investigated the explainability of the AAEC and showed that the NN-based TX can be approximated by a linear model and resembles a ZF precoder for low and moderate ISI regimes. In our numerical evaluations, we confirmed the aforementioned interpretation and showed that the proposed AAEC has superior error performance compared to MC systems with and without precoding

[1] Unfortunately, no precoder scheme was found that resembles the proposed AAEC for the high ISI case.

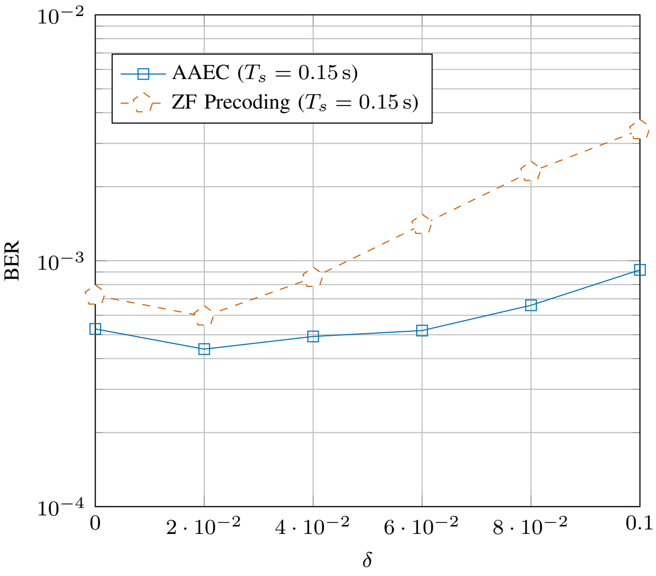


Fig. 7: BER performance of AAEC and an MC system using ZF precoding versus channel variations δ at SNR= 60 dB.

for high ISI scenarios. Interesting topics for future research include a detailed investigation of the explainability for high ISI and the design and comparison of AAEC with different low-complexity detectors (e.g., threshold detection).

Acknowledgment

This work has been in part supported by the "University SAL Labs" initiative of Silicon Austria Labs (SAL) and its Austrian partner universities for applied fundamental research for electronic-based systems. P. Hofmann and F. H.P. Fitzek acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the program of "Souveran. Digital. Vernetzt.". Joint project 6G-life, project identification number: 16KISK001K. Furthermore, F. H.P. Fitzek is funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy - EXC 2050/1 - Project ID 390696704 - Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden.

References

[1] F. Dressler and S. Fischer, "Connecting In-Body Nano Communication with Body Area Networks: Challenges and Opportunities of the Internet of Nano Things," *Elsevier Nano Communication Networks*, vol. 6, pp. 29–38, 6 2015.

[2] I. Akyildiz *et al.*, "The Internet of Bio-Nano Things," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 32–40, Mar. 2015.

[3] N. Farsad *et al.*, "A comprehensive survey of recent advancements in molecular communication," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1887–1919, Feb. 2016.

[4] W. Haselmayr *et al.*, "Integration of molecular communications into future generation wireless networks," *Proc. 6G Wireless Summit*, pp. 1–2, Mar. 2019.

[5] M. Å. Kuran *et al.*, "A survey on modulation techniques in molecular communication via diffusion," *Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 7–28, 2021.

[6] B. Li *et al.*, "Csi-independent non-linear signal detection in molecular communications," *IEEE Trans. Signal Process.*, vol. 68, pp. 97–112, 2020.

[7] M. Kuscu *et al.*, "Transmitter and receiver architectures for molecular communications: A survey on physical design with modulation, coding, and detection techniques," *Proc. IEEE*, vol. 107, no. 7, pp. 1302–1341, 2019.

[8] N. Farsad *et al.*, "Neural network detection of data sequences in communication systems," *IEEE Trans. Signal Process.*, vol. 66, no. 21, pp. 5663–5678, Sept. 2018.

[9] M. Bartunik *et al.*, "Using deep learning to demodulate transmissions in molecular communication," in *Proc. IEEE 16th International Symposium on Medical Information and Communication Technology (ISMICT)*, Sept. 2022, pp. 1–6.

[10] S. Mohamed *et al.*, "Model-based: End-to-end molecular communication system through deep reinforcement learning auto encoder," *IEEE Access*, vol. 7, pp. 70 279–70 286, May 2019.

[11] R. Khanzadeh, S. Angerbauer, F. Enzenhofer, A. Springer, and W. Haselmayr, "Towards end-to-end learning for salinity-based molecular communication," in *The 7th Workshop on Molecular Communications (2023)*, Apr. 2023.

[12] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, Sept. 2018.

[13] J. T. Gómez *et al.*, "Explainability of neural networks for symbol detection in molecular communication channels," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, July 2023.

[14] A. M. Alaa and M. van der Schaar, "Demystifying black-box models with symbolic metamodels," *Proc. 33rd International Conference on Neural Information Processing Systems*, vol. 32, Dec. 2019.

[15] Proakis, *Digital Communications 5th Edition*. McGraw Hill, 2007.

[16] S. Dörner *et al.*, "Deep learning based communication over the air," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 132–143, Dec. 2017.

[17] S. Cammerer *et al.*, "Trainable communication systems: Concepts and prototype," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5489–5503, June 2020.

[18] S. Angerbauer *et al.*, "Salinity-based molecular communication in microfluidic channels," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 9, no. 2, pp. 191–206, June 2023.

[19] K. Pearson, "Note on regression and inheritance in the case of two parents," *proceedings of the royal society of London*, vol. 58, no. 347-352, pp. 240–242, 1895.

[20] N. Fatema *et al.*, "Massive MIMO linear precoding: A survey," *IEEE systems journal*, vol. 12, no. 4, pp. 3920–3931, Dec. 2017.

Appendix A
Justification for the Metric in (19)

To obtain the metric in (19), we reconsider the matrix $\mathbf{W}$ defined in (8). For $T_s = 1\,\mathrm{s}$, we observe from the CIR shown in Fig. 2 that it is sufficient to set the channel length to $L = 2$, i.e., we consider the channel coefficients h_0, h_1, and h_2 (see (4)). Neglecting the initial and final phase of the transmission, the matrix $\mathbf{W}$ for $L = 2$ reads as

$$\mathbf{W} = \begin{bmatrix} h_2 - h_1 & h_1 - h_0 & h_0 & 0 & \cdots & 0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & h_2 - h_1 & h_1 - h_0 & h_0 & 0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & 0 & h_2 - h_1 & h_1 - h_0 & h_0 \end{bmatrix}. \tag{21}$$

According to the system model $\mathbf{p} = \mathbf{W}\mathbf{x}$, the output $p[n]$ is only affected by three samples of $x[n]$. Hence, we expect that a ZF precoder, which generates $x[n]$ will also only consider three symbols $d[n]$, when calculating $x[n]$. Thus, we conclude (and this tendency is also observed in simulation), that for the ZF precoder, the elements that have the greatest influence on the generated signal are located around the main diagonal of $\mathbf{W}_{\mathrm{ZF}}$. If, on the other hand, $\boldsymbol{\Theta}^*$ performs a similar operation to the ZF precoder, we expect that also in $\boldsymbol{\Theta}^*$ the most significant entries are located in and around the main diagonal. Inspired by the structure of (21), we decide to compare the three elements along the diagonal of $\boldsymbol{\Theta}^*$ and $\mathbf{W}_{\mathrm{ZF}}$ according to (19). Since our detector is a slope detector, it does not respond to a constant offset in the signal $y[n]$. Hence, our metric should neglect the influence of a constant offset in the generated channel input $x[n]$. Furthermore, it should measure, if the two signals $x_{\mathrm{ZF}}[n]$ and $x_{\mathrm{XAAEC}}[n]$ are positively correlated. The metric that fulfills both demands is the Pearson correlation coefficient, which was used (19).