IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 22, 2021, accepted December 7, 2021, date of publication December 15, 2021,
date of current version December 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3135903

Reducing the System Overhead of Millimeter-Wave Beamforming With Neural Networks for 5G and Beyond

PENGFEI XUE[1], YUHONG HUANG[1], DONGZHI ZHU[1], YOUPING ZHAO[1], (Senior Member, IEEE), AND CHEN SUN[2], (Senior Member, IEEE)

[1]School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
[2]Sony (China) Ltd., Beijing 100028, China

Corresponding authors: Youping Zhao (yozhao@bjtu.edu.cn) and Chen Sun (Chen.Sun@sony.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1804901, and in part by the Research and Development Center China, Sony (China) Ltd.

ABSTRACT To accommodate the rapid change of radio propagation environment for mobile communication scenarios, millimeter-wave beamforming requires instantaneous channel state information (CSI) to update its operational parameters in real time, resulting in heavy system overhead. As the number of antennas increases, the system overhead associated with beam management will increase dramatically. To address this overarching problem, a neural network-aided millimeter-wave beamforming algorithm is proposed in this paper. A new parameter, referred to as ''beam adjustment interval'', is proposed to evaluate the beamforming performance. It is defined as the maximum time duration in which the signal-to-interference-plus-noise ratio (SINR) of the user equipment can be maintained above the predefined threshold. Besides, a predictive method of beam adjustment to maximize the beam adjustment interval is developed, which considers the SINR not only at the current location but also future possible locations. Simulation results show that the proposed algorithm can significantly increase beam adjustment interval and reduce the total number of beam adjustments for the moving user equipment, thus reducing the system overhead 41.4% on average over 10 randomly generated test traces.

INDEX TERMS Beamforming, beam adjustment interval, millimeter-wave, neural network.

I. INTRODUCTION

With the explosive growth of traffic demand in wireless communication networks, millimeter-wave frequency band (30∼300 GHz) has attracted increasing attention due to a large amount of underutilized spectrum. As a feasible solution to improve network capacity, millimeter-wave bands can be employed to achieve hundreds of times increase in network capacity for the current 5G cellular networks [1]. In comparison with the low frequency band, millimeter-wave suffers more severe path loss and is easily blocked, making it necessary to utilize high-gain directional transmission and reception. In addition, the short wavelength allows large-scale antennas fabricated in a small physical space, making it possible to combine the millimeter-wave and massive multiple-input multiple-output (MIMO) to obtain high gain [2], [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Olutayo O. Oyerinde.

Most millimeter wave beamforming algorithms claim both the transmitter and the receiver update beamforming parameters in real time based on perfect channel state information (CSI), especially in multiuser scenarios. On one hand, it is hard to obtain perfect CSI for the fast-changing channel when considering the highly-mobile millimeter-wave applications; on the other hand, real-time update causes heavy system overhead, including measurement, computation and processing overhead.

So far, many standards working groups have conducted research on beamforming and beam management. In third generation partnership project (3GPP) Release 16 for new radio (NR), beam management is defined as procedures to acquire and maintain a set of transmission reception beams that can be used for downlink and uplink transmission or reception. The procedure of beam management includes at least the following aspects: beam sweeping, beam measurement, beam determination and beam reporting. The overhead

of beam management and latency reduction are also considered in NR [4]. Beam management procedures should also flexibly support different antenna configurations [5]. In addition, a coexistence manager has been introduced in various standards to coordinate interference among neighboring cells of different networks [6]. However, when the user equipment (UE) is moving, the beam pattern has to be frequently updated so as to avoid harmful interference to the neighboring cells, also resulting in heavy system overhead. Some general guidelines for the selection of the beam management parameters were provided in [7]. The performance of the beam management schemes is assessed in terms of detection accuracy, reactiveness, and overhead. And the overhead is defined as the ratio between the amount of time and frequency resources allocated to beam management operations and all the available resources.

Recently, the application of artificial intelligence (AI) algorithms including machine learning (ML) algorithms in wireless communication has attracted a lot of attention in research community. AI can be applied in the following two beam management aspects: beam training and beam tracking [8]. Beam tracking is a widely-adopted approach to address the problem of device mobility. Generally speaking, factors like direction, velocity and location need to be taken into account for beam tracking. An intelligent millimeter-wave beam tracking method with reinforcement learning algorithm was proposed in [9]. The UE can always have an excellent service with the best millimeter-wave beam for each UE position. Besides, an adaptive beam management scheme based on deep learning for mobile high-speed 5G millimeter-wave networks was proposed in [10]. The proposed deep learning architecture can predict the matching beam through learning the mobility information and the matching beam of the UE. To enable highly-mobile millimeter-wave applications, a novel integrated machine learning with coordinated beamforming solution has been proposed. The solution can overcome the challenges such as the use of narrow beams, the sensitivity of millimeter-wave signals to blockage, the reliability of highly-mobile links, the frequent handoff between base stations in densely deployed networks, and the heavy system overhead as detailed in [11]. In addition to the above research, there are also researches using machine learning beam management for millimeter-wave intelligent reflecting surface (IRS) and unmanned aerial vehicles (UAV) [12], [13].

As a kind of machine learning algorithm, neural networks have been applied to beamforming due to their strong nonlinear fitting ability, adaptive learning ability, and classification capacity of complex data. In general, beamforming can be seen as an adaptation problem, which can be implemented by an artificial neural network. In [14], a phase transmittance radial basis function neural network (RBF-NN) beamforming for static and dynamic channels was proposed. A beamforming neural network was designed to deal with the power minimization problem in [15]. The functionality of neural networks relies on good training data. Each record of training data includes proper input and the corresponding desired output. By off-line training, a neural network model is established. Then it can operate online to give desired output corresponding to any proper input. The existing neural network-based beamforming algorithms usually aim to point the nulls to interference and the main beam to the desired signal. Input data is always the transformation of the correlation matrix of the received signal, for the reason that there exists sufficient direction information of the desired signal and interference that can be learned by neural network.

Most of the prior research has been focused on developing beamforming strategies and beam management. However, there are few studies dedicated to the overhead reduction of millimeter-wave beamforming [16]. To reduce the system overhead, a machine learning-based approach, more specifically, radial basis function neural network is employed in this paper to train the beam pattern and minimize the number of beam adjustments for the moving UEs. The main contributions of this paper can be summarized as follows.

- A neural network-aided millimeter-wave beamforming algorithm is proposed to reduce the system overhead. The algorithm is suitable for mobile communication scenarios, and we formulate our beam adjustment problem as a beam adjustment interval maximization problem.
- A new parameter, referred to as ''beam adjustment interval'', is defined to evaluate the beamforming performance and the system overhead. ''Beam adjustment interval'' is defined as the maximum time duration in which the signal-to-interference-plus-noise ratio (SINR) of the desired UE can be maintained above the predefined threshold.
- A predictive method of beam adjustment is presented to maximize the beam adjustment interval, which considers the SINR not only at the current location but also future possible locations.

The rest of this paper is organized as follows: the system model is presented in section II. Then in section III the proposed neural network-aided millimeter-wave beamforming algorithm is discussed in detail. After that, the performance of the proposed algorithm is simulated and analyzed in section IV. Finally, conclusion of this paper is given in section V.

II. SYSTEM MODEL

Consider a scenario as shown in Fig. 1, where each base station (BS) is equipped with a uniform linear array (ULA) of M antennas, and each BS can serve multiple users. In practical applications, the BS can flexibly select uplink and downlink beams according to needs. For downlink, it can adopt a narrow beam with high gain to transmit, and for uplink, it can adopt a wide beam with low gain to receive. For simplicity, assuming that each BS serves one UE within its coverage at a given time instance and the UE is equipped with an omnidirectional antenna. The UE moves randomly, and it is always connected with a beam from the serving BS.

In this paper, the neural network-based beamforming algorithm directs the main beam to the desired signal by default,

and the received signal y_k of the UE served by the k-th BS is represented as

$$y_k = \sqrt{P_k}\mathbf{h}_{kk}^H \mathbf{w}_k s_k + \sum_{i=1, i\neq k}^{K} \sqrt{P_i}\mathbf{h}_{ik}^H \mathbf{w}_i s_i + z_k \quad (1)$$

where P_k is the transmit power of the k-th BS, $\mathbf{h}_{ik} \in C^{M\times 1}$ is the channel response vector from the i-th BS to the UE served by the k-th BS, $(\,)^H$ denotes the conjugate transpose, $\mathbf{w}_i \in C^{M\times 1}$ is the transmit beamforming weight vector of the i-th BS, s_i is the modulated symbol with the unit norm for the UE served by the i-th BS, and z_k is the additive white Gaussian noise (AWGN) at the UE served by the k-th BS with the noise power of σ_k^2. According to the above signal model, the SINR of the UE served by the k-th BS is expressed as

$$SINR_k = \frac{P_k \mathbf{h}_{kk}^H \mathbf{w}_k \mathbf{w}_k^H \mathbf{h}_{kk}}{\sum_{i=1, i\neq k}^{K} P_i \mathbf{h}_{ik}^H \mathbf{w}_i \mathbf{w}_i^H \mathbf{h}_{ik} + \sigma_k^2} \quad (2)$$

By maximizing the beam adjustment interval (i.e., the maximum time interval during which the SINR of the desired UE can be maintained above the predefined threshold), the system overhead can be reduced while meeting the SINR requirement of the desired UE. The beam adjustment interval maximization problem can be further formulated as

$$\begin{aligned} \mathbf{w}_k &= \underset{\mathbf{w}_l}{arg\ \max}\ \Delta T_k \\ &s.t.\ SINR_k \geq SINR_k^{th} \\ &P_k \leq P_k^{\max} \end{aligned} \quad (3)$$

where ΔT_k is the beam adjustment interval of the UE served by the k-th BS, $\mathbf{w}_l$ is the l-th codeword in the selected codebook, $l = 0, 1, \cdots, L-1$, L is the number of codewords in the codebook. $SINR_k^{th}$ is the SINR threshold of the UE served by the k-th BS, and $P_k^{\max}$ is the maximum transmit power of the k-th BS. Assuming that codebook-based beamforming is employed, each column of the codebook is a codeword, which represents the beamforming weight vector that can form the corresponding beam pattern. In this paper, only the beamforming weight vectors of the UE's serving BS are optimized, while the transmit power and beamforming vectors of the other interfering BSs are fixed. Therefore, the goal of the maximization problem is to obtain the codeword $\mathbf{w}_k$ of the serving BS with the maximized beam adjustment interval.

III. NEURAL NETWORK-AIDED BEAMFORMING ALGORITHM

In this section, we first present a neural network-aided millimeter wave beamforming algorithm in detail. And then discuss a new approach to obtain the training data, which is used to construct the neural network for beamforming.

A. PROPOSED BEAMFORMING ALGORITHM

To maximize the beam adjustment interval for the desired UE, we need to choose the beam pattern that can provide satisfactory performance at the current location as well as future possible locations. Therefore, a neural network needs to be constructed to reach this goal based on a predictive method. The procedures of the proposed algorithm are comprised of the following four steps [17].

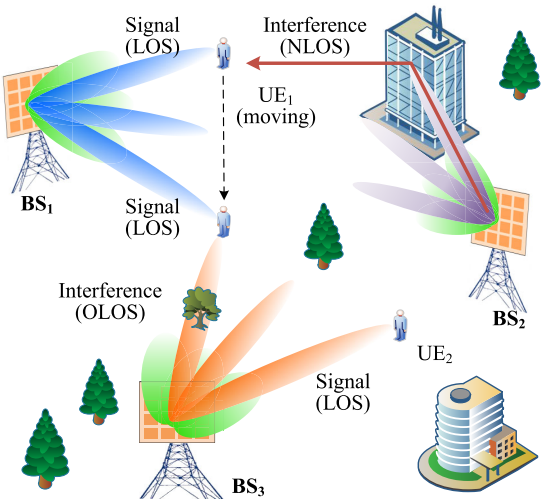


FIGURE 1. Illustration of millimeter-wave communication systems with mobile user equipment.

Step 1: an application scenario is chosen, which includes the desired signal and the co-channel interference. For both the desired signal and the co-channel interference, the channel condition can be divided into three types, namely, line of sight (LOS), none line of sight (NLOS) and obstructed line of sight (OLOS). For the LOS scenario, there is no obstruction between the transmitter and the receiver. For the NLOS scenario, there exist obstacles like buildings that will block the signal transmission between the transmitter and the receiver. For the OLOS scenario, there are obstacles like trees which may increase the path loss to some extent. In addition, the radio signal may experience rain attenuation and atmospheric attenuation during millimeter-wave radio propagation [18]. Thus, atmospheric attenuation is considered in this paper, although this is a negligible problem in microwave systems. An application scenario can be any combination of three types of the desired signal and the co-channel interference as shown in Fig. 1. In this paper, we only consider the desired signal and interference both under LOS scenario for simplicity.

Step 2: the training data is generated based on a predictive method to train the neural network. Firstly, the beamforming weight vectors and the corresponding array gain of the base station antenna will be generated. And then, a 360-degree judgment of the UE movement direction is performed according to the relative position between the UE and the serving BS, combined with the SINR threshold. Finally, the SINR and the beam adjustment interval can be calculated. This step will be detailed in Section III.B.

Step 3: a neural network is constructed according to the optimization problem as depicted by (3). The neural network is constructed as a multi-classification model of beam weight vector index. Any neural network algorithm that can solve the multi-classification problem can be employed to establish the

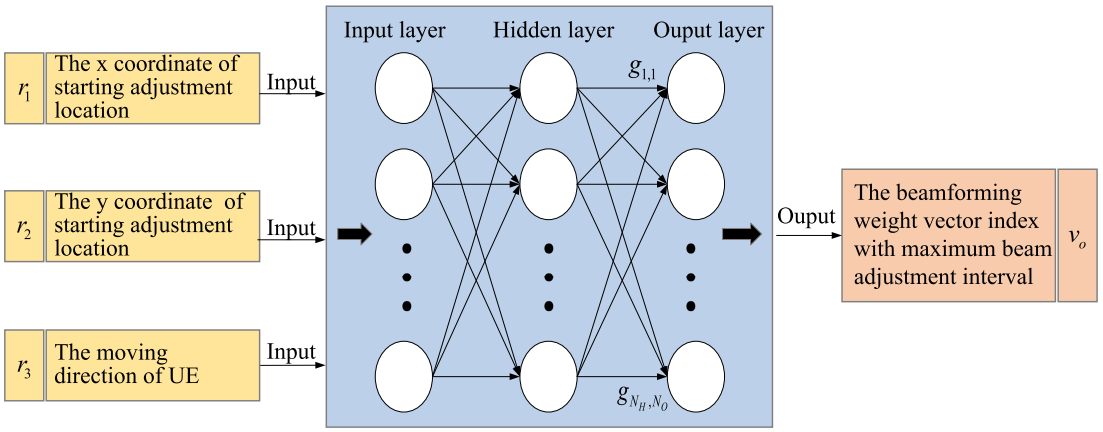


FIGURE 2. Block diagram of RBF-NN.

beam adjustment interval maximization model. RBF-NN is adopted to address this issue in this paper.

Step 4: the established neural network is applied into dynamic beam adjustment after training. When the SINR of the desired UE cannot be satisfied, meta-data such as the starting location for beam adjustment, moving direction of UE, beamwidth and direction of the interfering beam will be collected and input into the established neural network to obtain the beam pattern with the maximized beam adjustment interval.

A three-layer RBF-NN, as shown in Fig. 2, is adopted in the simulation owing to its prominent merits such as strong capacity in classification, fast learning convergence, and strong nonlinearity fitting ability. Certainly, the RBF-NN as a local approximation network can meet the real-time requirements, despite it is prone to the phenomenon of abnormal data when the data is insufficient. The possible problem can be solved by fully learning the historical source domain knowledge (i.e., the center vector of the radial basis function) to make up for the deficiency of the generalization ability of the current domain. The learning algorithm of RBF-NN based on self-organizing selection center has two stages: the first stage is a self-organizing learning stage in which the center and the variance of basis function need to be learned; the second stage is a supervised learning stage in which weights between the hidden layer and output layer need to be found.

Generally, RBF-NN employs Gaussian function as radial basis function. Assuming that there are N_I, N_H, N_O nodes in the input layer, hidden layer, and output layer, respectively. The n-th value v_n of the output vector $\mathbf{v} = [v_1, v_2, \cdots, v_{N_O}]$ can be expressed as

$$v_n = \sum_{j=1}^{N_H} g_{j,n} e^{-\frac{\|\mathbf{r}-c_j\|^2}{2\sigma_j^2}} \quad (n = 1, 2, \cdots, N_O) \tag{4}$$

The beamforming weight vector index with the maximum beam adjustment interval v_o is a single output in this paper. $\mathbf{r} = [r_1, r_2, \cdots, r_{N_I}]$ is the input vector, which includes starting adjustment location r_1, r_2, moving direction of UE r_3 in this paper. c_j and σ_j are the central vector and standard deviation of the j-th Gaussian function, respectively, and $g_{j,n}$ is the weight from the j-th node of the hidden layer to the n-th node in the output layer.

B. GENERATION OF TRAINING DATA

There are multiple ways to generate training traces, including uniformly-interval sampling the azimuth angle of UE, the distance from the serving BS, possible moving directions, and other ways like street information-based prediction or random mobility model-based prediction. In order to find the proper beamforming weight vector that can satisfy the SINR of the desired UE at most (if not at all) locations along the probable trajectories, a uniformly-interval sampling traversal search based predictive method is adopted to generate the training data. The starting positions of the UE in the serving area are uniformly sampled, and the possible movement directions at each starting position are judged. If the location of the next point is outside the serving area, no training will be conducted in that direction. And then training traces including all possible starting adjustment locations and possible moving directions are generated.

A codebook is a matrix, each column of which represents a beamforming weight vector that forms a certain beam pattern. The most commonly used codebooks are beamsteering codebook and phase shift-specific codebook [19], [20]. Beams generated by beamsteering codebook are directed to different directions for the reason that they are based on array steering vectors on quantified angle of arrivals (AOAs). While beams generated by phase shift-specific codebook are directed to a specific direction, since the phases of antenna elements are randomly chosen from a specific set of phases, such as $\{\pm 1, \pm j\}$. The generation procedures of beamsteering codebook are as follows: assuming that the angular range of serving sector is $(-\Phi/2, +\Phi/2)$, and the number of codewords in the codebook is L. Therefore mainlobe interval of adjacent beams is $\Phi/(L-1)$, and the l-th beam is directed to $\varphi_l = -\Phi/2 + l\Phi/(L-1)$. Thus the l-th codeword w_l is represented as

$$\mathbf{w}_l = \mathbf{W}(:, l) = \frac{1}{\sqrt{M}}[1, e^{j\frac{2\pi}{\lambda}d\sin\varphi_l}, \cdots, e^{j\frac{2\pi}{\lambda}(M-1)d\sin\varphi_l}]^T \tag{5}$$

where $\mathbf{W}$ is $M \times L$ matrix of the codebook that forms diversified beam patterns, $\mathbf{w}_l$ is l-th column of codebook, λ is wavelength of millimeter wave, d is spacing of the ULA, and the operator $()^T$ stands for transpose.

Considering the millimeter-wave propagation model derived from [21], the path loss could be obtained using

$$PL\,(d)\,[\mathrm{dB}] = \alpha + \beta 10\log_{10}(d) + \xi \tag{6}$$

where d is the distance from BS to UE in meters, the values of parameter α, β are the least square fits of floating intercept and slope over the measured distances, and $\xi \sim \mathcal{N}\left(0, \sigma^2\right)$, σ^2 is the lognormal shadowing variance. The values of α, β and σ^2 are 61.4, 2, 5.8, respectively in 28 GHz LOS scenario as shown in [21]. For each UE, let d be the distance between UE and BS where beam exists. The receive power P_r in dBm can be calculated by

$$P_r\,[\mathrm{dBm}] = P_t\,[\mathrm{dBm}] + G_t\,[\mathrm{dB}] - PL\,(d)\,[\mathrm{dB}] \tag{7}$$

where P_t is the transmit power of the BS in dBm, and G_t is the antenna array gain in dB.

For the beam pattern of each training trace, the SINR of the desired UE needs to be determined whether it can be maintained above the required SINR threshold, and the beam adjustment interval can be calculated. The calculation of UE's SINR needs to consider the interference from neighboring cells. For each training trace, the beam adjustment interval is calculated at the starting adjustment location by traversing each codeword, and then take the codeword corresponding to the maximum beam adjustment interval and record the beam index. By traversing each training trace, the beam index with the maximum adjustment interval of each training trace can be found. Due to the considerations of the SINR of the desired UE not only at the current location but also future possible locations, intelligent prediction can be achieved to a certain extent.

Then, a neural network model shown in Fig. 2 is established as a beamforming weight vector classification network. Input parameters of the neural network consist of starting adjustment location, moving direction of UE. The output of the neural network is the beamforming weight vector index with the corresponding maximized beam adjustment interval. The major procedures of the proposed training data generation algorithm are shown in Algorithm 1.

Algorithm 1 Training Data Generation

1: Adopt a uniformly-interval sampling traversal search based predictive method to generate training traces in the serving area;
2: Select a proper beam codebook $\mathbf{W}$;
3: Set the required SINR threshold $SINR_{th}$;
4: **for** each training trace **do**
5: Initialize transmit power P_t;
6: **for** each beam pattern $\mathbf{w}_l$ in the codebook **do**
7: Find out the corresponding beam adjustment interval ΔT as the UE moves along the training trace;
8: **end for**
9: Take the maximum value of ΔT as the maximum beam adjustment interval and record the beam index;
10: **end for**
11: Obtain the corresponding $\mathbf{w}$ for each training trace with the maximum ΔT;

IV. SIMULATION RESULTS

In this section, the performance of the proposed algorithm is evaluated. Fig. 3 shows the training scenario and part of the training traces. The cell radius of the serving BS is 30 m, the signal of the target UE (i.e., the desired signal) is under the LOS, and the signal of the UE that interferes with the BS (i.e., the interference signal) is also under the LOS.

Following the procedures described in Section III.B to generate training traces and training data. In order to ensure training efficiency and model accuracy, a compromised method is adopted to generate training data according to the following interval: the starting positions are taken every 1 degree in the sector arc direction and every 1 m in the radial direction. The possible moving directions are taken every 10 degrees within the range of 360 degrees. After excluding those starting points on the boundary of the serving cell, a total of 123868 training data are obtained. The starting adjustment location and moving direction of UE were taken as input, and the beamforming weight vector index with the corresponding maximized beam adjustment interval was taken as output to train the neural network. This training process can be implemented by the MATLAB Deep Learning Tool Box.

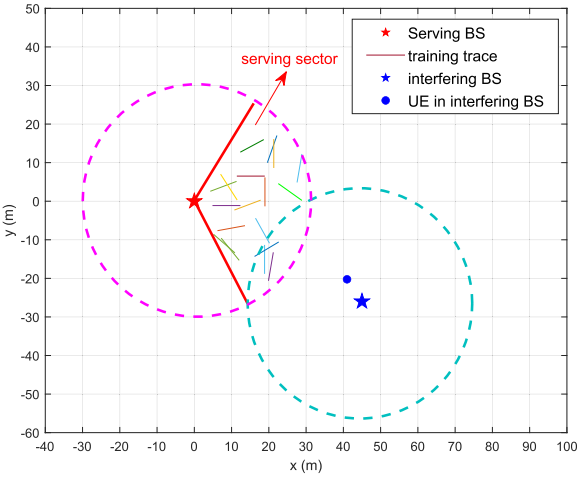
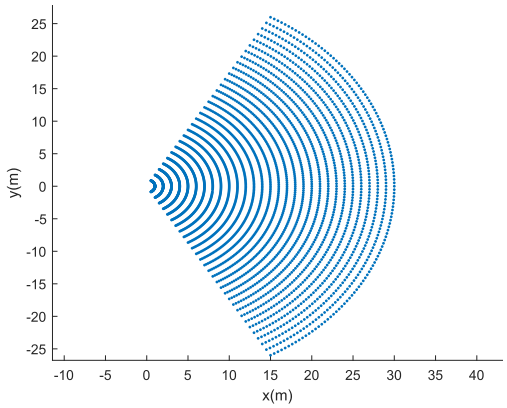
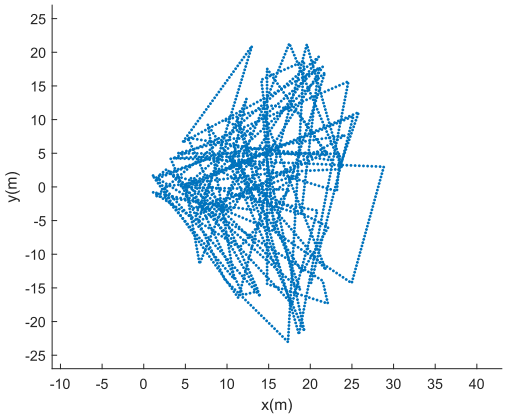


FIGURE 3. Training scenario.

In order to verify the accuracy and generalization ability of the trained neural network model, the test trace is generated by the random mobility model, which is different from the method used in training data generation. Fig. 4(a) and Fig. 4(b) show the difference between the two methods. Principles of random mobility model are as follows: randomly choose a starting point in the serving sector; then, move towards a randomly chosen ending point in the serving sector with a certain velocity, which is equivalent to randomly choosing a possible moving direction; repeat the last step until the desired UE stops moving.



(a) Uniform interval model used for training.



(b) Random mobility model used for testing.

FIGURE 4. Methods of training data generation and test trace generation.

TABLE 1. List of simulation parameters.

Parameter	Value
Operating frequency	28 GHz
Maximum transmit power of the serving BS	30 dBm
Equivalent isotropic radiated power of the interfering BS	5 dBm
Noise figure of receiver	5 dB
Bandwidth	40 MHz
Required SINR threshold of UE	20 dB
Number of transmit antennas at the serving BS	16
Number of receive antennas at the UE	1
Number of codewords (L)	25
Moving speed of the UE	0.5 m/s
Angular range of serving sector	[-60°,+60°]
Angular interval of trainings	1°
Distance interval of trainings	1 m
Number of test data	3182
Number of training data	123868

Table 1 lists the major simulation parameters. The transmit power of the serving BS is the total power provided by the BS transmitter. 30 dBm is the maximum transmit power of the serving BS, which can meet the basic communication requirements. For simplicity of simulation, the transmit power of the interfering BS is actually the equivalent isotropic radiated power (EIRP), and 5 dBm is the average of the radiated power in all directions. The trained neural network is applied to real time beam adjustment in the test trace. In the meantime, the total number of beam adjustments during the random moves is counted. And then the performance of beam adjustment is

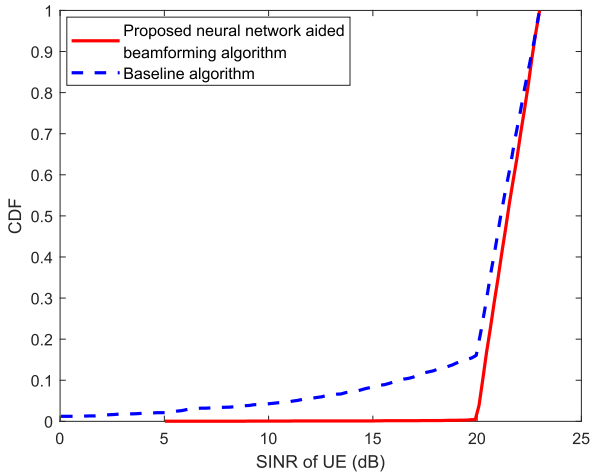


FIGURE 5. Comparison of the CDF of SINR.

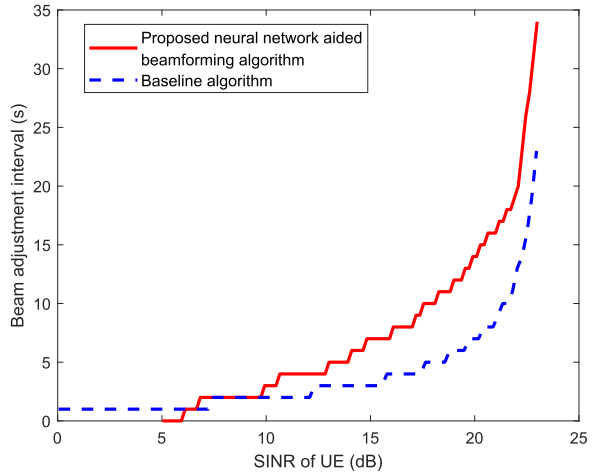


FIGURE 6. Comparison of the beam adjustment intervals.

compared between two beamforming algorithms, namely, the proposed neural network aided beamforming algorithm and a baseline beamforming algorithm. The baseline algorithm is similar to the exhaustive search algorithm proposed in [22]. If the SINR of the UE cannot be met, the BS starts to scan the beam through a predefined codebook. Once a beam higher than the SINR threshold is found, this beam will be selected to reach the UE.

Fig. 5 shows the cumulative distribution function (CDF) curve of SINR under the two algorithms. It can be seen from the figure that the beam adjustment algorithm proposed in this paper works better and the complexity is reduced. The probability of not reaching the SINR threshold is much smaller than the baseline algorithm, therefore the superiority of this algorithm is obvious. In the simulation, there are only 16 outages in 3182 test data, and the probability is only 0.5%. As shown in Table 2, the proposed algorithm helps to decrease the total number of beam adjustments from 608 to 316 when moving along the same trace, so the processing overhead can be reduced by 48.0%.

In Fig. 6, the beam adjustment interval presents an increasing trend with the increase of SINR of UE. However, the proposed algorithm has higher beam adjustment interval and

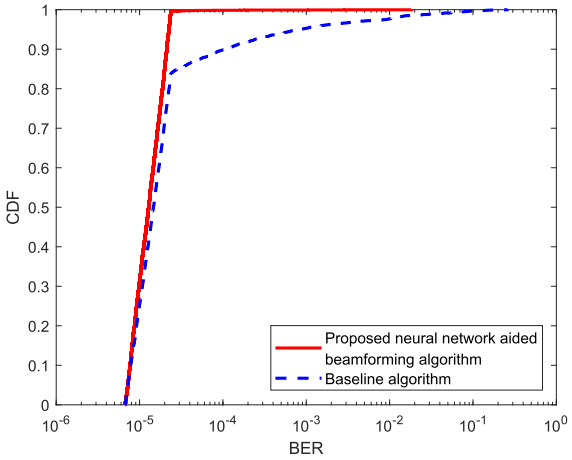
TABLE 2. System overhead comparison with different system settings.

Comparison parameters	Proposed	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed	Baseline
BS antenna configurations	16ULA		16ULA		36ULA		36ULA	
Bandwidth (MHz)	10		40		40		40	
Moving speed of UE (m/s)	0.5		0.5		0.5		3	
Number of beam adjustments	264	430	316	608	413	1113	110	184
Average beam adjustment interval (s)	11.98	7.40	10.02	4.49	7.54	2.86	4.73	2.88
Processing overhead reduction	38.6%		48.0%		62.9%		40.2%	

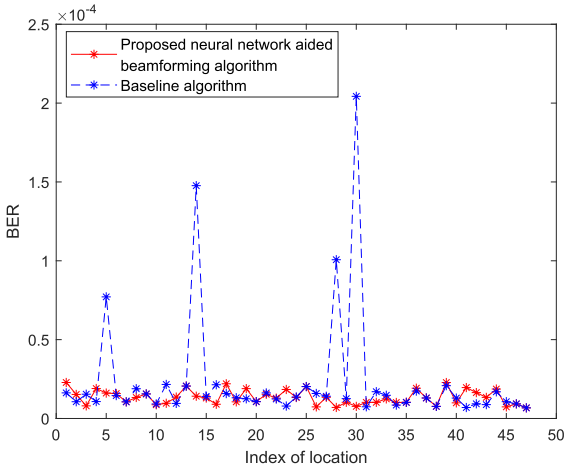
better SINR than the baseline algorithm. In the same searching time, the method proposed in the paper has fewer adjustments to the beam, and the frequency of beam adjustment is slow. While the comparison method requires constant search and adjustment of the beam, and scans the beam until a satisfactory beam is found once the SINR of the desired UE cannot be met, so the frequency of beam adjustment is relatively high.

Fig. 7 depicts the bit error rate (BER) of the UE's communication quality when moves randomly near the base station. Fig. 7(a) shows the CDF of BER along the test trace, and we take the logarithmic value of BER for comparison purposes. In Fig. 7(b), the 47 trajectory points are selected at equal intervals from the 3182 test data (e.g., a UE location point to select every 68s). From the average BER of the UE in the entire moving trace, the average BER of the method in this paper is 2.1×10^{-5}, while the average BER of the comparison algorithm is 1.3×10^{-3}. It can be seen that the BER of the proposed algorithm is 2 orders of magnitude less than that of the baseline algorithm. The possible reason is that the beam of the baseline algorithm is narrow. It is easier to move from an area with a high beam gain to an area with a low beam gain when the user is moving. Therefore, the research on the influence of beamwidth and beam type on user communication performance will be further research content.

Since the millimeter-wave frequency band is often accompanied by wider bandwidth and more antennas, the comparison of different antenna configurations and different bandwidths is also listed in Table 2. When we use the same antenna configuration, increasing the bandwidth will significantly reduce the processing overhead. The reason is that the noise will be increased when increasing the bandwidth, therefore the SINR will decrease. In order to meet the UE performance, the beam adjustment will be more frequent. However, the neural network-aided beam adjustment algorithm can choose the beam that keeps the user performance longer due to the existing learning foundation. Every adjustment will reduce a lot of overhead compared to the baseline algorithm. When the adjustment is more frequent, the cumulative effect makes the total overhead greatly reduced. Similarly, with an increase in the number of antennas, the beam will be narrower and more accurately aimed at the user, which also means more frequent beam adjustments. It can be seen that the total processing overhead can be reduced by 62.9%. In addition, we also compared the algorithm performance at different moving speeds of the UE. It is obvious that the overhead reduction is higher when the UE moves at a lower speed (e.g., 0.5 m/s).



(a) The CDF of BER along the test trace.



(b) Comparison of BER in the selected points.

FIGURE 7. Comparison of BER.

Next, let us take the first 50 beam adjustments as an example. As shown in Fig. 8, the vertical axis is the corresponding time instance of the current beam adjustment index during the random moves. The proposed algorithm significantly increases the beam adjustment interval, and in the meantime, greatly reduces the total number of beam adjustments (i.e., it is reduced from 608 beamforming adjustments to 316 adjustments).

For a small section of the moving trace of the desired UE from 174 seconds to 190 seconds, the beam is adjusted 9 times with the beam index of [1 2 3 1 2 3 1 2 3] and the corresponding adjustment interval of [2 1 3 1 1 3 1 1 3] seconds when using the baseline algorithm. In contrast, the beam is adjusted only 3 times with the beam index of [9 11 14] and the corresponding beam adjustment interval is as long as [8 2 6]

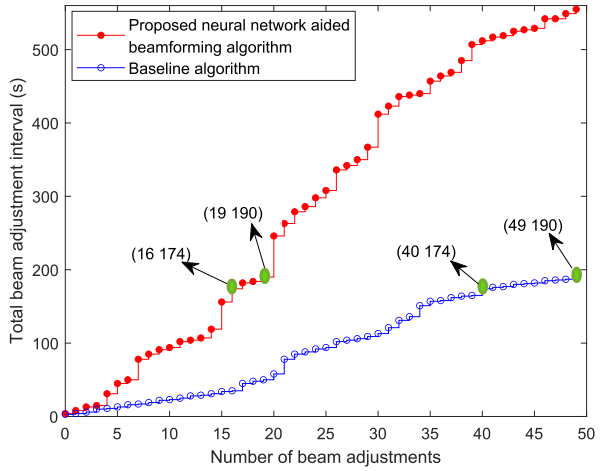


FIGURE 8. Comparison of the first 50 beam adjustments.

TABLE 3. Comparision of overhead in different test traces.

Index of test trace	Number of beam adjustments N_d^*	Number of beam adjustments N_d	Overhead reduction
1	299	497	39.8%
2	308	564	45.4%
3	315	548	42.5%
4	314	534	41.2%
5	393	572	31.3%
6	336	530	36.6%
7	354	661	46.4%
8	354	612	42.1%
9	257	477	46.1%
10	354	611	42.1%
Average	328	560	41.4%

seconds when using the proposed algorithm. The simulation results show that the proposed beamforming algorithm can significantly increase beam adjustment interval and reduce the number of beam adjustments, thus reducing the system overhead.

Table 3 shows the results of 10 simulations in different test traces, including the number of beam adjustments N_d^* when using the proposed algorithm, the number of beam adjustments N_d when using the baseline algorithm, the system overhead reduction, and the average system overhead reduction in 10 simulations. It is clear that the method proposed in this paper has fewer beam adjustments with an average of 328, while the baseline method has frequent beam adjustments and the average number of adjustments is 560. On the whole, the average reduction in system overhead of the method proposed in this paper is 41.4% in 10 simulations. Therefore, the method proposed in this paper can greatly reduce the system overhead.

The 3GPP provides frame structure and measurement signals for NR physical layer, both Frequency Division Duplexing (FDD) and Time Division Duplexing (TDD) will be supported [23], [24]. In millimeter-wave, TDD is generally used. In 5G cellular systems, beam management is required whether the UE is in initial access or in tracking stage. Note that the reference signals used for beam measurement are different. For the downlink, beam measurement is performed based on the synchronization signal (SS) block in

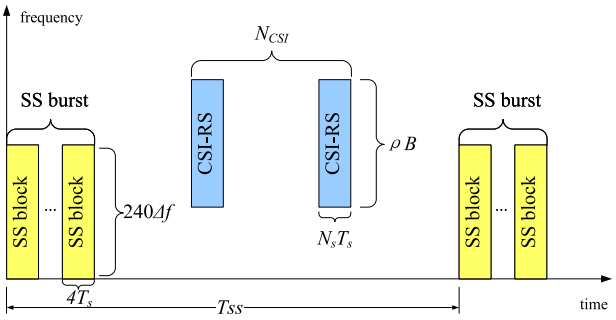


FIGURE 9. SS blocks and CSI-RSs time-frequency resources configurations.

TABLE 4. List of symbols and descriptions.

Symbols	Descriptions
N_{SS}	Number of SS blocks per SS burst
N_{CSI}	Number of CSI-RSs per SS burst periodicity
T_s	Duration of a symbol
Δf	Subcarrier spacing
N_s	Number of OFDM symbols for a CSI-RS
ρ	Portion of bandwidth B for CSI-RSs
B	Bandwidth
T_{SS}	SS burst periodicity
N_{RB}	Number of physical resource block
N_d	Number of beam adjustments when using the baseline algorithm
N_d^*	Number of beam adjustments when using the proposed algorithm
T_t	Duration of the test trace when the UE moves from the starting point to the ending point

initial access stage. Beam measurement is performed based on SS block and the channel state information-reference signal (CSI-RS) allocated to the UE in tracking stage. An SS block is a group of 4 orthogonal frequency division multiplexing (OFDM) symbols in time and 240 subcarriers in frequency. SS blocks are sent to the UE at a fixed period by beam sweeping. Multiple CSI-RSs can be configured after each SS burst, as shown in Fig. 9. In order to evaluate overhead reduction and the improvement in system capacity, the following analysis is carried out. All the symbols are shown in the Table 4.

As the synchronization signal for initial access, SS block needs to be sent periodically, while CSI-RS does not need to be sent periodically but only triggered when the beam is adjusted. In the entire beam management procedure shown in Fig. 10, the BS first sends the SS block during initial access. Then the UE determines the beam through beam measurement and reports it to the BS. During tracking, the BS sends SS block and CSI-RS to the UE to adjust the beam. When adopting the proposed algorithm, the number of beam adjustments can be reduced, and the transmission of CSI-RS can be reduced due to the existing learning foundation. The UE directly determines the beam and reports the beam information to the base station.

The total number of time-frequency resources R_{SS} scheduled for the transmission of N_{SS} SS blocks is given by

$$R_{SS} = N_{SS} 4 T_s 240 \Delta f = 960 N_{SS} T_s \Delta f \tag{8}$$

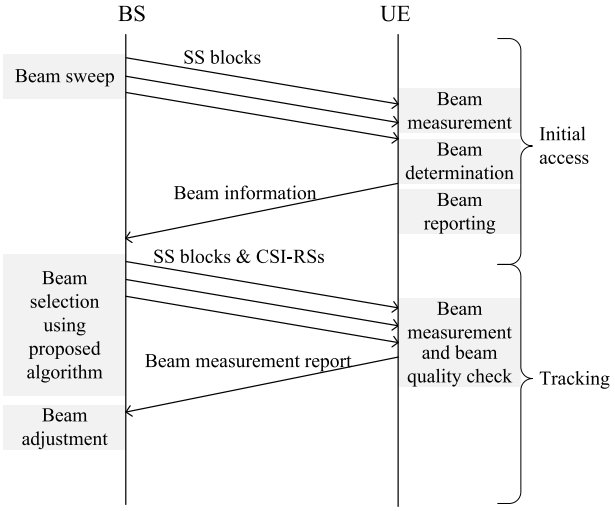


FIGURE 10. Beam management procedure.

Moreover, additional overhead is introduced by the transmission of CSI-RSs after the SS burst, so the total number of time-frequency resources R_{CSI} is given by

$$R_{CSI} = N_{CSI} N_s T_s \rho B \tag{9}$$

Besides, UE needs to report beam management information (e.g., cri-RSRP, ssb-Index-RSRP) to the base station according to the 3GPP standard [25]. This will also occupy time-frequency resources, which can be given by

$$R_{UE} = 4 T_s N_{RB} 12 \Delta f = 48 T_s N_{RB} \Delta f \tag{10}$$

Taking into account both the SS blocks, the CSI-RSs and the beam report, the total time-frequency resources occupancy ratio using the proposed algorithm and the baseline algorithm are Ω^* and Ω, respectively, which are given by

$$\Omega = \left(R_{SS} + \left(R_{SS} + R_{CSI} + R_{UE}\right) N_d\right) / \left(T_{SS} B N_d\right) \tag{11}$$

$$\Omega* = \left(R_{SS} + \left(R_{SS} + R^*_{CSI} + R_{UE}\right) N^*_d\right) / \left(T_{SS} B N_d\right) \tag{12}$$

Owing to the time-frequency resources saved by using the proposed algorithm, the system capacity for the serving BS can be improved. More specifically, the increase rate of system capacity for the serving BS can be defined as

$$I_C = \left(\Omega - \Omega^*\right) N_d / T_t \tag{13}$$

Substituting (8)-(12) into (13), the final formula of the increase rate of system capacity for the serving BS is

$$I_C = \frac{\left(R_{SS} + R_{UE}\right)\left(N_d - N^*_d\right) + R_{CSI} N_d - R^*_{CSI} N^*_d}{T_{SS} B T_t} \tag{14}$$

Supposing $N_{SS} = 64$, $T_s = 8.92\ \mu$s, $\Delta f = 120$ kHz, $N_s = 4$, $\rho = 0.72$, $B = 40$ MHz, $T_{SS} = 10$ ms, $N_{RB} = 4$, $N_{CSI} = 25$, $N_d = 560$, $N^*_{CSI} = 1$, $N^*_d = 328$, $T_t = 3182$ s, the result is $I_C = 2.3\%$. In other words, the system capacity for the serving BS can be increased by 2.3%. As indicated by (14), the increase rate of system capacity is related to the number of beam adjustments and total time-frequency resources occupied in the beam management procedure. And the number of beam adjustments is related to factors such as the antenna configurations, the moving speed of the UE, the bandwidth, the SINR threshold of the UE, the applicable channel model, etc. For example, different channel models will result in different calculated SINR, thereby affecting the number of beam adjustments. When the number of base station antennas increases, the beam pattern and beamwidth will change, thus affecting the number of beam adjustments. In sum, the system capacity for the serving BS can be affected by various system settings and system scenarios.

V. CONCLUSION

Nowadays, the system overhead associated with beam management in millimeter-wave massive MIMO is an increasingly prominent problem. In this paper, a radial basis function neural network-aided millimeter-wave beamforming algorithm is proposed to reduce the system overhead. By training the RBF-NN based beam adjustment model with the maximum beam adjustment interval as the optimization target, and applying the trained beam adjustment model to real-time adaptive beam adjustment, the number of beam adjustments can be greatly reduced for the moving user equipment. At the same time, it can improve the communication performance of the user equipment compared with the baseline algorithm. Simulation results show that the proposed algorithm can significantly increase beam adjustment interval and reduce the total number of beam adjustments, thus reducing system overhead remarkably, on average 41.4% in 10 different test traces. Our study shows many factors have impacts on the system overhead reduction, such as the bandwidth, the antenna configurations, the moving speed of user equipment, the applicable channel model, the beamwidth and so on. For future work, it is worthwhile to evaluate the performance in more practical scenarios.

REFERENCES

[1] W. Roh, J. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 106–113, Feb. 2014.

[2] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave cellular wireless networks: Potentials and challenges," *Proc. IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[3] L. Jiang and H. Jafarkhani, "Multi-user analog beamforming in millimeter wave MIMO systems based on path angle information," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 608–619, Jan. 2019.

[4] *Study on New Radio (NR) Access Technology*, document TR38.912 V16.0.0, 3GPP, 2020.

[5] *Study on New Radio Access Technology*, document TR38.802 V14.2.0, 3GPP, 2017.

[6] *IEEE Standard for Information Technology-Telecommunications and Information Exchange Between Systems-Local and Metropolitan Area Networks-Specific Requirements—Part 19: Wireless Network Coexistence Methods*, Standard 802.19.1-2018, IEEE, 2018, pp. 1–458.

[7] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "Standalone and non-standalone beam management for 3GPP NR at mmWaves," *IEEE Commun. Mag.*, vol. 57, no. 4, pp. 123–129, Apr. 2019.

[8] Y.-N.-R. Li, B. Gao, X. Zhang, and K. Huang, "Beam management in millimeter-wave communications for 5G and beyond," *IEEE Access*, vol. 8, pp. 13282–13293, 2020.

[9] R. Wang, O. Onireti, L. Zhang, M. A. Imran, G. Ren, J. Qiu, and T. Tian, "Reinforcement learning method for beam management in millimeter-wave network," in *Proc. U.K./China Emerg. Technol. (UCET)*, pp. 1–4, Aug. 2019.

[10] W. Na, B. Bae, S. Cho, and N. Kim, "Deep-learning based adaptive beam management technique for mobile high-speed 5G mmWave networks," in *Proc. IEEE 9th Int. Conf. Consum. Electron. (ICCE-Berlin)*, pp. 149–151, Sep. 2019.

[11] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 6, pp. 37328–37348, 2018.

[12] C. Jia, H. Gao, N. Chen, and Y. He, "Machine learning empowered beam management for intelligent reflecting surface assisted mmWave networks," *China Commun.*, vol. 17, no. 10, pp. 100–114, Oct. 2020.

[13] W. Xu, Y. Ke, C.-H. Lee, H. Gao, Z. Feng, and P. Zhang, "Data-driven beam management with angular domain information for mmWave UAV networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7040–7056, Nov. 2021.

[14] M. P. Enriconi, F. C. C. de Castro, C. Müller, and M. C. F. de Castro, "Phase transmittance RBF neural network beamforming for static and dynamic channels," *IEEE Antennas Wireless Propag. Lett.*, vol. 19, no. 2, pp. 243–247, Feb. 2020.

[15] W. Xia, G. Zheng, Y. Zhu, J. Zhang, J. Wang, and A. Petropulu, "Deep learning based beamforming neural networks in downlink MISO systems," in *Proc. IEEE Int. Conf. Commun. (ICC) Workshop*, May 2019, pp. 1–5.

[16] Y. Zhao, Y. Liu, G. Boudreau, A. B. Sediq, and X. Wang, "A low overhead angle-based beamforming using multi-pattern codebooks for mmWave massive MIMO systems," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2018, pp. 211–216.

[17] Y. Zhao, D. Zhu, C. Sun, and X. Guo, "Electronic apparatus and method in wireless communication system," U.S. Patent 10 327 155 B2, Jun. 18, 2019.

[18] *Attenuation by Atmospheric Gases and Related Effects*, document ITU-R P.676-12-2019, ITU-R, 2019.

[19] T. Kim, J. Park, J.-Y. Seol, S. Jeong, J. Cho, and W. Roh, "Tens of Gbps support with mmWave beamforming systems for next generation communications," in *Proc. IEEE GLOBECOM*, Dec. 2013, pp. 3685–3690.

[20] D. Ramasamy, S. Venkateswaran, and U. Madhow, "Compressive tracking with 1000-element arrays: A framework for multi-Gbps mmWave cellular downlinks," in *Proc. Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2012, pp. 690–697.

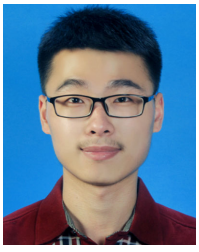
[21] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[22] M. Giordani, M. Mezzavilla, C. N. Barati, S. Rangan, and M. Zorzi, "Comparative analysis of initial access techniques in 5G mmWave cellular networks," in *Proc. Annu. Conf. Inf. Sci. Syst.*, Mar. 2016, pp. 268–273.

[23] *Physical Channels and Modulation*, document TS38.211 V16.7.0, 3GPP, 2021.

[24] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 173–196, Sep. 2018.

[25] *Physical Layer Procedures for Data*, document TS38.214 V16.6.0, 3GPP, 2021.



PENGFEI XUE received the B.S. degree in communication engineering from Qingdao University, Qingdao, China, in 2020. He is currently pursuing the M.S. degree with the School of Electronic and Information Engineering, Beijing Jiaotong University. His research interests include wireless channel modeling and intelligent communications for the next generation wireless communications.

YUHONG HUANG received the B.S. degree in communication engineering from Northeast Electric Power University, China, in 2019. She is currently pursuing the Ph.D. degree with the School of Electronic and Information Engineering, Beijing Jiaotong University. Her research interests include channel modeling and intelligent communications for the next generation wireless communications.

DONGZHI ZHU received the B.S. degree in communication engineering from Yanshan University, Qinhuangdao, China, in 2015, and the M.S. degree in communication and information systems from Beijing Jiaotong University, Beijing, China, in 2018. Her research interest includes machine learning-based radio resource management.

YOUPING ZHAO (Senior Member, IEEE) received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1994 and 1996, respectively, and the Ph.D. degree from Virginia Tech, Blacksburg, VA, USA, in 2007. He is currently a Professor with Beijing Jiaotong University, Beijing. His current research interests include cognitive radio and intelligent communications for the next generation wireless communications. He received the Best Editor Award of *China Communications* in 2018. Since 2017, he has been on the Editorial Board of *China Communications* cosponsored by the IEEE Communications Society.

CHEN SUN (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2005. From August 2004 to May 2008, he was a Researcher with ATR Wave Engineering Laboratories, Japan, working on adaptive beamforming and direction-finding algorithms of parasitic array antennas and a theoretical analysis of cooperative wireless networks. In June 2008, he joined the National Institute of Information and Communications Technology (NICT), Japan, as an Expert Researcher, working on distributed sensing and dynamic spectrum access in TVWS. He is currently the Director of the Sony Research and Development Center, Wireless Network Research Department, Beijing, China. He received the IEEE Standards Association Working Group Chair Award for Leadership in 2011, the IEEE 802.11af Outstanding Contributions Award in 2014, and the IEEE 802.19.1 Outstanding Contributions Award in 2018. From 2013 to 2015, he served as the Technical Editor for the IEEE 1900.6 Standard in 2011 and the Rapporteur for the European Telecommunications Standards Institute Reconfigurable Radio Systems EN 301 144. He currently serves as the Technical Editor for the IEEE 802.19.1a Working Group.