Applying Deep Learning for Phase-Array Antenna Design

Peng   Zhang

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Engineering in Electrical Engineering
Prince of Songkla University
2022

Applying Deep Learning for Phase-Array Antenna Design

Peng   Zhang

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Engineering in Electrical Engineering
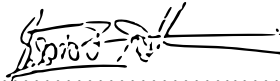**Prince of Songkla University**
2022

Thesis Title   Applying Deep Learning for Phase-Array Antenna Design

Author   Mr. Peng Zhang

Major Program   Electrical Engineering

Major Advisor   **Examining Committee:**

…………………………………   …….………………..……Chairperson

(Assoc. Prof. Dr. Mitchai Chongcheawchamnan)   (Assoc. Prof. Dr. Nutapong Somjit)

………………………………Committee

(Assoc. Prof. Dr. Mitchai Chongcheawchamnan)

………………………………Committee

(Dr. Kiattisak Wongsophanakul)

The Graduate School, Prince of Songkla University, has approved this thesis as partial fulfillment of the requirements for the Master of Engineering Degree in Electrical Engineering.

…………………………………………

(Prof. Dr.Damrongsak Faroongsarng)

Dean of Graduate School

This is to certify that the work here submitted is the result of the candidate's own investigations. Due acknowledgement has been made of any assistance received.

…………………….……….Signature
(Assoc. Prof. Dr. MitchaiChongcheawchamnan)
Major Advisor

…………………………. Signature
(Mr. Peng Zhang)
Candidate

I hereby certify that this work has not been accepted in substance for any degree, and is not being currently submitted in candidature for any degree.

…………………………………. Signature
(Mr. Peng Zhang)
Candidate

Thesis Title   Applying Deep Learning for Phase-Array Antenna Design

Author   Mr. Peng Zhang

Major Program   Electrical Engineering

Academic Year   2021

ABSTRACT

Hybrid beamforming (HBF) can provide rapid data transmission rates while reducing the complexity and cost of massive multiple-input multiple-output (MIMO) systems. However, channel state information (CSI) is imperfect in realistic downlink channels, introducing challenges to hybrid beamforming (HBF) design. For HBF designs, we had a hard time finding the proper labels. If we use the optimized output based on the traditional algorithm as the label, the neural network can only be trained to approximate the traditional algorithm, but not better than the traditional algorithm. This thesis proposes a hybrid beamforming neural network based on unsupervised deep learning (USDNN) to prevent the labeling overhead of supervised learning and improve the achievable sum rate based on imperfect CSI. Compared with the traditional HBF method, the unsupervised learning-based method can avoid the labeling overhead as well as obtain better performance than the traditional algorithm. The network consists of 5 dense layers, 4 batch normalization (BN) layers and 5 activation functions. After training, the optimized beamformer can be obtained, and the optimized beamforming vector can be directly output. The simulation results show that our proposed method is 74% better than manifold optimization (MO) and 120% better than orthogonal match pursuit (OMP) systems. Furthermore, our proposed USDNN can achieve near-optimal performance.

Keywords: Deep learning, Hybrid beamforming, massive MIMO, unsupervised deep learning.

ACKNOWLEDGEMENT

CONTENT

List of Tables

List of Figures

Chapter 1 Introduction

Hybrid beamforming (HBF) technology is a promising solution for mm-Wave multiple-input multiple-output (MIMO) systems to support ultra-high transmission capacity and low complexity. HBF provides a better solution than digital and analog beamformers (ABF and DBF) which struggles with non-convex optimization problem in multi-user scenarios. For the optimization of HBF, the hardest part is that the constant modulus constraint of the analog beamformer due to the phase shifter makes the problem highly non-convex and difficult to solve. Deep learning (DL) has proven to be an excellent tool for dealing with complex non-convex optimization problems. Its unique network structure can approximate any function under specific conditions. Using DL to solve the optimization problem of beamforming design can eliminate the complexity caused by too many iterations and realize real-time calculation. However, the method proposed by the researchers to train the neural network with the beamforming matrix obtained by the traditional scheme as the label, its performance is limited by the traditional scheme. It does not make full use of the powerful approximation ability of the neural network. For the existing one-to-one neural network beamforming system, in practical applications, when there are multiple users at the receiving end if a one-to-one communication system and method are adopted, it is necessary to establish multiple systems and separate training coefficients, which will result in wasted resources and increased costs. And for realistic downlink channels, the channel state information (CSI) is imperfect, which greatly increases the difficulty of HBF design. Therefore, we propose a multi-user hybrid beam neural network based on unsupervised deep learning (USDNN) to maximize the hardware limitations and imperfect CSI to improve the reachable sum rate.

1.1 Background

Antennas are an essential part of mobile communication systems. With the development of mobile communication technology, antenna structures are becoming more and more diverse, and the technology is becoming more and more complex [1]. Entering the 5G era and beyond, massive MIMO, beamforming (BF), etc., have become vital technologies, prompting the antenna to evolve towards an active and complex direction. The antenna design method also needs to keep pace with the times,

using advanced simulation methods to deal with complex design requirements and meet the ever-increasing performance requirements of antennas in the 5G era [2]. In addition, 5G networks need to adapt to scenarios such as large bandwidth, high reliability, low latency, and large connections. This requires 5G antennas to support more channels, flexible and real-time beam adjustment, and the ability to support high-frequency communication.

The critical evolution direction is the Active Phased Array Antenna (APAA) [3]. "Phased Array" is an abbreviation for "Phase Controlled Array". The phased array antenna (PAA) is an antenna that changes the shape of the pattern by controlling the feeding phase of the radiating element in the array antenna. Controlling the phase can change the direction of the maximum value of the antenna pattern for beam scanning purposes. In exceptional cases, it is also possible to control the sidelobe level, the minimum position, and the overall pattern's shape [4]. When the antenna is rotated mechanically, there will be disadvantages such as large inertia and slow speed. The PAA overcomes these disadvantages. The scanning speed of the beam is high, and its feed phase is generally controlled by an electronic computer, and the phase changes quickly (milliseconds). That is, the antenna pattern changes rapidly in the direction of the maximum value or other parameters [5]. This is the most prominent feature of the PAA.

Most PAAs designed over the past few years have used ABF technique, where phase adjustment is performed at radio frequency (RF) [6] or intermediate frequency (IF) [7] frequencies, and a set of data converters are employed throughout the antenna. However, recently there has been a growing interest in DBF technique. Each antenna element has a set of data converters, and the phase adjustment is made digitally in an Field Programmble Gate Array (FPGA) [8] or some data converter. DBF has many benefits, including transmitting multiple beams quickly and even changing the number of beams on the fly. This superior flexibility is attractive in many applications and is driving its popularity. In addition, continued improvements in data converters have reduced power consumption and extended to higher frequencies. RF sampling in the L-band and S-bands [9] has made this technology viable for radar systems.

There are many factors to consider when considering both ABF and DBF options, but the analysis usually depends on the number of beams required, power consumption, and cost goals. The DBF [10] method typically have high power consumption with one data converter per element but are incredibly flexible and convenient in forming multiple beams. However, data converters require a higher dynamic range since blocking-rejecting beamforming can only be done after digitization. On the other hand, ABF [11] can support numerous beams, but each beam requires an additional phase adjustment channel. For example, to form a 100-beam system, the number of RF phase shifters for a 1-beam system needs to be multiplied by 100, so cost considerations for data converters and phase adjustment integrated circuits (ICs) may change depending on the number of beams. Similarly, power consumption is typically lower for ABF methods that utilize passive phase shifters. However, as the number of beams increases, power consumption increases if additional gain stages are required to drive the distribution network.

A common trade-off is a hybrid analog and digital beamforming (HBF) approach [12] to achieve spectral efficiency maximum (SEM) [13] and increases the reachable sum rate. This is an increasingly hot industry area and will continue to grow in the years to come.

1.2 Research motivation

With the rapid rise of 5G millimeter-wave communications and broadband low-orbit satellite communications, millimeter-wave APAAs have begun an unprecedented development. The core of signal processing is BF [14]. Recently, HBF designs have received much attention for their ability to provide high BF gain to compensate for severe path loss at affordable hardware cost and power consumption [15]. However, CSI [16] is critical and practical data describing the channel in wireless communication. In wireless communication, the CSI represents the propagation characteristics of the communication link, which describes the joint influence of various effects such as scattering, fading, power attenuation, etc., in the channel. Hence, CSI is usually imperfect in practical applications. This makes HBF designs more challenging.

Machine learning (ML) [17] solves the problem of how to build computers that automatically improve through experience. It is one of the fastest-growing technology fields today, at the intersection of computer science and statistics

and the heart of artificial intelligence and data science. Entering the 5G era, 5G ultra-high throughput and low latency requirements can be met by ML technology. Furthermore, intelligent base stations and mobile terminals must mimic complex human-like learning and decision-making capabilities. Among them, DL [18] is a new research direction in ML, which is introduced into ML to make it closer to the original goal - artificial intelligence. Recent research on intelligent communication shows that data-based DL methods have great potential in solving traditionally challenging problems [19]–[23].

DL also known as deep neural network technology, consists of simple linear and nonlinear operations. Its unique network structure can approximate arbitrary functions under certain conditions. Therefore, using DL to solve optimization problems such as power allocation and BF design can eliminate the complexity caused by too many iterations and realize real-time calculation. Inspired by these works, this research aims to apply DL methods to solve complex HBF design problems with hardware limitations in imperfect CSI scenario.

1.3 Research purpose

It is well known that HBF design is a rather complex non-convex problem due to joint optimization of multiple variables and constant modulus constraints (CMC) (i.e. phase shifters) [24], and it is unlikely to find a closed-form optimal solution. However, since DL has been recognized as an efficient way to deal with intractable problems, DL can be used to solve HBF optimization problems. The issues to be solved are as follows:

1. Design a system model of a phased-array-based hybrid digital and analog beamforming transceiver.

   1.1 Active phased array design.

   1.2 Hybrid precoder design.

2. A neural network is designed to optimize the beamformer to maximize the achievable sum-rate with limited hardware and imperfect CSI.

   2.1 The structure of each layer of the DNN network is analyzed, and the function of each layer of the network is explained.

   2.2 Acquire the dataset through the simulated channel.

2.3 Set the parameters of the DNN network.

2.4 Train the DNN model with the dataset.

2.5 Verify the performance of the DNN model.

1.4 The importances of research

As an essential part of mobile communication, the research and design of antennas play a vital role in mobile communication. Entering the 5G era and beyond, applications will be significantly enriched. 5G networks need to adapt to scenarios such as large bandwidth, high reliability, low latency, and large connections, which require 5G antennas to support more channels, flexible real-time beam adjustment, and support for high-frequency communication. Currently, MIMO antenna configurations are mainly relied on to multiply the capacity of wireless base station antenna links [12]. These antennas can concentrate signal strength into a small spatial area, increasing overall efficiency and throughput by directing the signal exactly where it is needed. This BF capability can be increased by adding additional antennas.

In the 4G system, DBF is often used due to the small number of antennas. The advantage of this process is that the amplitude and phase of the signal can be adjusted arbitrarily, approaching infinite precision, and in the Orthogonal Frequency Division Multiplexing (OFDM) [19] system, different beamforming can be done for different subcarriers. However, this also means that each antenna needs to be equipped with an independent RF chain, including expensive digital-to-analog/analog-to-digital converters, etc., which also brings high energy and hardware expenses, which increase proportionally with the number of antennas. The advantage of the APAA is that it uses the microwave integration method to integrate chips such as phase shifters, filters, attenuators, power amplifiers, and low-noise amplifiers in the chip, which realizes the miniaturization and lightness of the equipment. High pointing accuracy and certain beam sidelobe suppression capability [3].

Moreover, the APAA has no moving parts and has excellent reliability. Therefore, even if a few antenna elements in the array fail, the overall performance of the antenna will not be affected [4]. In this experiment, the original DBF is divided into two parts. One part is realized by low-dimensional DBF. The other is realized by high-dimensional ABF, reducing the number of RF links and significantly reducing the Propagation power consumption and hardware cost. Moreover, most BF algorithms

invented before 5G era cannot be copied to the HBF structure due to hardware constraints. Therefore, this experiment will lay the foundation for the future HBF design of PAAs.

1.5 Research limitations

In HBF, in addition to the difficulty of joint optimization of the four BF variables (transmitting and receiving analog beamformers and digital beamformers), the constant modulus constraints of the ABFs due to the phase shifters make this problem highly non-convex and difficult to solve compared to traditional full DBF designs. Model-based design methods are used in the existing work to deal with this difficulty, and an algorithm based on orthogonal matching pursuit (OMP) is proposed [25]. However, ABFs are limited to predefined codebooks. To improve the performance of OMP, [26], [27] applied Manifold Optimization (MO) method to the optimization of simulated BF. Finally, a unitized iterative algorithm is proposed to optimize ABFs. However, these algorithms either require some approximation to simplify the original objective function or require a lot of time-consuming serial iterations to arrive at a solution. And in all these algorithms, it is assumed that there is perfect CSI, and there is a certain gap with reality.

1.6 Thesis structure

This study mainly introduces the optimization problem of applying DL to HBF. The first chapter introduces the study's background, motivation, purpose, importance, and limitations. Chapter 2 reviews some previous BF design methods. Chapter 3 introduces the HBF technique and the system model based on PAA. Chapter 4 introduces the channel model and gives a description of the neural network model based on unsupervised learning. Chapter 5 presents simulation results to evaluate the method's performance and compare it with existing methods. Finally, chapter 6 is mainly a summary and outlook.

Chapter 2 Literature Review

2.1 Deep learning for wireless communication

This paper discusses emerging DL techniques in future wireless communication networks. It will show that data-driven approaches should not replace but complement traditional design techniques based on mathematical models. A broad motivation for why artificial neural network-based DL will be an indispensable tool for future wireless communication network design and operation is given. Then, their vision for how artificial neural networks can be integrated into future wireless communication network architectures is presented. A comprehensive description of DL methods is provided, starting with the general ML paradigm, and then a more in-depth discussion of DL and artificial neural networks, covering the most widely used artificial neural network architectures and their training methods. DL will also be linked to other major learning frameworks such as reinforcement learning and transfer learning. Next, a comprehensive survey of the DL literatures for wireless communication networks is provided, followed by a detailed description of several novel case studies in which the use of DL proves to be very useful for network design. Each case study will show how the use of (even approximate) mathematical models can significantly reduce the amount of real-time data that needs to be acquired/measured to implement a data-driven approach. Finally, concluding remarks describe what they see as the main directions for future research in this field [28].

Wireless communication is expected to bring about huge changes in the future, and various emerging applications such as virtual reality and the Internet of Things (IoTs) will become a reality. However, these compelling applications bring many new challenges, including unknown channel models, low latency requirements in large-scale ultra-dense networks, etc. The astonishing success of DL in various fields, especially in wireless communications, has recently sparked interest in applying it to address these challenges. Therefore, in this review, some practical approaches to wireless communication using DL are investigated. The first is a DL-based architecture design that breaks the classic model-based block design rules for wireless communications over the past few decades. The second, DL-based algorithm design, will be illustrated by a few examples in a range of typical technologies envisioned for

5G and beyond. Their rationale, key features, and performance improvements will be discussed. Open questions and future research opportunities will also be pointed out, highlighting the interplay between DL and wireless communications. The author hopes that this review will inspire more novel ideas and exciting contributions to intelligent wireless communication.

DL, mainly implemented through Deep Neural Networks (DNN), has achieved remarkable success and excellent results in various fields such as image recognition, Chinese Go, and other complex games. These successful applications have stimulated increasing interest in the application of DL in wireless communications. Specifically, DL is a powerful tool capable of learning the intricate interrelationships between variables, especially those difficult to describe using mathematical models accurately. This enables one to design wireless communication systems without knowing precise mathematical models, which is not possible in the context of existing wireless design principles. Furthermore, for a family of lightweight DNNs of limited size, passing inputs through them requires only a limited number of operations, making DL methods computationally efficient. Even for the highly complex DNNs that may be necessary to solve large-scale communication problems, distributed parallel architectures and accelerated tools for DL promise high computational efficiency. Therefore, DNNs are attractive for solving large-scale wireless problems associated with numerous antennas/users/devices [29].

Inspired by the above literature, DL has been widely used in wireless communications. For example, state-of-the-art techniques for physical layer communication utilizing DL are summarized in [21]. Furthermore, [30] and [31] comprehensively investigate the application of DL in designing IoT and 5G cellular networks at various layers of the protocol stack, respectively. Compared with the conclusions of the above-related literature, we can apply DL technology to wireless communication more profoundly and comprehensively by introducing some design methods, namely DL-based architecture design and DL-based algorithm design.

2.2 Beamforming design based on deep learning

Although antenna selection has been extensively studied, most works provide sub-optimal solutions with different selection strategies. However, obtaining the best solution requires a high computational burden. In this paper, the author

considers the joint antenna selection and HBF design issues in a DL environment to reduce computational complexity while maintaining optimality. They designed a CNN in which antenna selection and HBF design can be performed together. To realize the antenna selection in DL, this problem is regarded as a classification problem, which considers the multilayer classification of CNN. They will also solve the HBF design problem in the classification framework. They evaluated the performance of the proposed method through numerical simulations and showed that their CNN framework provides significantly better performance than traditional techniques such as orthogonal matching pursuit [32].

BF is considered to be one of the essential technologies for designing advanced MIMO systems. In the existing design standards, sum-rate maximization (SRM) under a total power constraint is a challenge due to its nonconvexity. The current SRM problem technology can only obtain an optimal local solution, but many calculations are required due to its complex matrix operations and iterations. Unlike conventional methods, they proposed a fast BF design method based on DL, which does not require complicated operations and iterations. Specifically, the first derived heuristic solution structure of downlink BF based on the best minimum mean square error (MMSE) receiver through the virtual equivalent uplink channel, which divided the problem into power allocation and virtual uplink beamforming (VUBF) design. Next, the BF prediction network (BP-Net) is designed to perform joint optimization of power allocation and VUBF design. Besides, BP-Net uses a two-step training strategy for offline training. The simulation results show that their proposed method is fast and can obtain performance comparable to the latest method [33].

After reviewing several research documents, I summarized some current HBF design methods based on ML and DL, as shown in Table 1-1. They designed and optimized different neural networks, thereby reducing the number of iterations and propagation power consumption. However, most of the research is based on perfect CSI scenario. Because perfect CSI can reduce training overhead and the complexity of beam training, it is however well known that CSI is imperfect in practical applications. So we need to consider how to design a HBF system based on DL in the case of imperfect CSI scenario.

Different ML and DL algorithms are applied for beamforming optimization (CNN: Convolutional Neural Network; DNN: Deep Neural Network; BFNN: Beamforming Neural Network; BPNet: Beamforming Prediction Network; SVM: Support Vector Machine)

Table 2-1   HBF design method based on ML and DL.

Imperfect CSI	T. Lin, [15]	BFNN
	L. Sung, [34]	DNN
Perfect CSI	A. Alkhateeb, [32]	DL
	H. Huang, [33]	BPNet
	A. M. Elbir, [35]	CNN
	H. Huang, [36]	DNN
	J. Chen, [37]	ML
	Y. Long, [38]	SVM

DL-based algorithms can learn how to optimize the beamformer through training, thereby improving the performance of the HBF system. Besides, these algorithms have strong robustness and can handle imperfect CSI, and are expected to be applied in real situations. However, these algorithms also have shortcomings, because processing imperfect CSI requires sufficient beam training, which will cause a lot of training overhead. Also, some algorithms under the assumption of perfect CSI scenario can get rid of the complexity caused by excessive iteration and can improve system performance to a certain extent without incurring too much training overhead. However, we cannot ignore the impact of imperfect CSI in practical applications. Therefore, we can combine the advantages of these algorithms, continuously optimize the algorithm, and design new models to deal with imperfect CSI and reduce the complexity of beam training, maximize spectral efficiency, and achieve the optimal reachable sum rate.

2.3 Hybrid beamforming design based on imperfect CSI

Deploying multiple antennas at both the transmitter and receiver, MIMO, is undoubtedly a major advancement in wireless communication systems. However, an implementation of a millimeter-wave massive MIMO system faces many technical

difficulties, and it is still very challenging so far. In this paper, the design of hybrid beamformers for uplink connections in massive MIMO systems with imperfect CSI is investigated in both single-user (SU) and multi-user (MU) cases. The norm-limited channel error model is used to capture the characteristics of imperfect CSI in the existing system. The objective function is based on the worst-case robustness of the MMSE. Consider the SU and MU reception modes of the millimeter-wave massive-MIMO base station. A hierarchical BF optimization and joint precoder/combiner optimization for users with limited and extended computing power for the SU scenario was studied. These optimization techniques were subsequently extended to the MU case, in which a new hybrid robust combiner design was proposed. The proposed simulation results confirm the design superiority as compared with the latest robust hybrid design published in the literature [16].

To support two-way communication, a two-way relay method (two-way communication) between two users (transceivers) has received a lot of attention recently. As an extension of the one-way relay solution, the two-way relay solution can be designed to avoid simultaneous transmission from two users. However, this method is not bandwidth efficient because it requires four-time slots to complete the data exchange between two users. Most of the current work is limited to the ideal network assumption without interference. However, interference is the main limiting factor, especially it is complicated to obtain the perfect CSI of the interfering link, and these problems will seriously affect the performance of the wireless network.

A study of the optimal BF and power allocation problems of amplifying and forwarding-based two-way relay networks in the presence of interference and imperfect CSI. They also obtained the BF vector and the transmit power under the two assumptions of the CSI availability of the interfering link, namely the bounded uncertainty model and the second-order statistical scheme. To this end, the author has developed two design methods. The first method is based on the total transmit power minimization technique. They start with the uncertainty model with bounded norms and arrive at the optimal solution to the corresponding problem. They also developed a low-complexity algorithm that provides performance very close to the best algorithm to reduce computational complexity. The second method applied a signal-to-interference plus noise ratio (SINR) balancing technique. They proposed another low-complexity

algorithm based on the SINR balance standard. Next, the author considers the availability of second-order statistical information on CSI. Similarly, they start with minimizing total power consumption and derive the best and sub-optimal algorithms. Finally, they applied SINR balancing technology to this situation and developed another low-complexity algorithm suitable for practice [39].

CSI is a very important and practical data describing the channel in wireless communication. However, the CSI obtained in practical applications is usually not perfect, making the HBF design more challenging. Having investigated several literatures, I found that the most commonly used solutions for imperfect CSI problems in HBF designs are based on formulating MSE, MMSE, and SINR, as shown in Table 2-2.

Table 2-2   Different ways to solve imperfect CSI.

A. Morsali, [16]	MMSE
S. Salari, [39]	SINR
J. Wang, [40]	MMSE
M. B. Shenouda, [41]	MSE
N. Vucic, [42]	MSE

Note: MSE: mean square error; MMSE: minimum mean square error; SINR: signal to interference plus noise ratio.

Chapter 3 Research on Beamforming of Phase Array Antenna

This chapter presents three different BF techniques and system models based on PAA. 5G and beyong, low-earth orbit satellite commuication and modern radar systems highly demand low-cost and small antenna architectures which consume less power. In addition to these requirements, rapid retargeting against new threats or new users is required, multiple data streams are transmitted, and operating life is extended at an ultra-low-cost. Some applications need to counteract the effect of incoming blocking signals to reduce the probability of interception. PAA whichs offers beam steeering function without a mechanical motor enable many new applications. They have been gained much interest from the industry as the promising solutions to these challenges. The past shortcomings of PAAs are being addressed with advanced semiconductor technology to ultimately reduce the size, weight, and power of these solutions.

3.1 Development of phased array antennas

Radio subsystems that rely on antennas to send and receive signals have operated for over 100 years. These electronic systems will continue to improve and refine as accuracy, efficiency, and more advanced metrics become more important. Dish antennas have been widely used for transmitting (Tx) and receiving (Rx) signals over the past few years, where directivity is critical, and after years of optimization, many of these systems can perform well at a relatively low-cost run. These dishes have a robotic arm that rotates the radiation direction, and they do have some drawbacks, including slow steering, large physical size, poor long-term reliability, and only one radiation pattern or data stream that meets the requirements. As a result, people have turned to advanced PAA technology to improve these characteristics and add new capabilities. PAAs use an electronic steering mechanism, which has many advantages over traditional mechanical steering antennas, such as low height/small size, better long-term reliability, fast steering, and beam flexibility. With these advantages, PAAs are already widely used in military applications, satellite communications, and 5G telecommunications including the Internet of Vehicles.

Compared with traditional MIMO, massive MIMO can effectively improve the performance of the core is based on phased array technology. The so-called phased array refers to a type of array antenna that changes the direction of the directional beam by controlling the feeding phase of the radiating element in the array antenna. The main purpose of the phased array is to realize the spatial scanning of the array beam, that is, the so-called electrical scanning. In the early days, phased arrays were mainly used in military applications - phased array radars. Due to its fast scanning speed and strong multi-task capability, phased array radar has been widely used in the field of military radar and has become one of the symbols of military strength. In addition, phased array technology is also widely used in civil fields such as meteorological forecasting.

Looking back at the development history of mobile communications, it can also be seen from the evolution trend of base station antennas that phased array technology is an inevitable choice for improving system capacity and spectrum utilization, reducing interference and enhancing coverage in the 5G era. First of all, from passive antennas to active antenna systems, this means that antennas may be intelligent, miniaturized, and customized. In the future, the network will become more and more detailed, and it needs to be customized according to the surrounding scenes. For example, the deployment of stations in urban areas will be more refined, rather than simple coverage. 5G communication will use high-frequency bands, obstacles will have a great impact on communication, and customized antennas can provide better network quality. Second, the systematization and complexity of antenna design, such as beam arrays (to achieve space division multiplexing), multi-beam, and multi/high-frequency bands. All of these put forward high requirements for the antenna, which will involve the whole system and the problem of mutual compatibility. In this case, the antenna technology has surpassed the concept of components and gradually entered the system design.

A PAA is a collection of antenna elements assembled together, where the radiation pattern of each element is structurally combined with the radiation patterns of adjacent antennas to form a main lobe which emits radiated energy at the desired location, while the antenna, by design, is responsible for destructively interfering with signals in unwanted directions, creating unwanted signals and side lobes. The antenna

array is designed to maximize the energy radiated from the main lobe while reducing the energy radiated from the side lobes to acceptable levels. The radiation direction can be manipulated by changing the phase of the signal fed to each antenna element.

Figure 3-1 shows how the main lobe can be steered in the desired direction of the linear array by adjusting the phase of the signal in each antenna. As a result, each antenna in the array has independent phase and amplitude settings to form the desired radiation pattern. Since there are no mechanical moving parts, it is easy to understand the properties of fast beam steering in a phased array. Another advantage of PAAs over the mechanical antennas is that they can radiate multiple beams simultaneously, allowing them to track multiple targets or manage user data for multiple data streams. This is achieved by digital signal processing of multiple data streams at baseband frequencies.

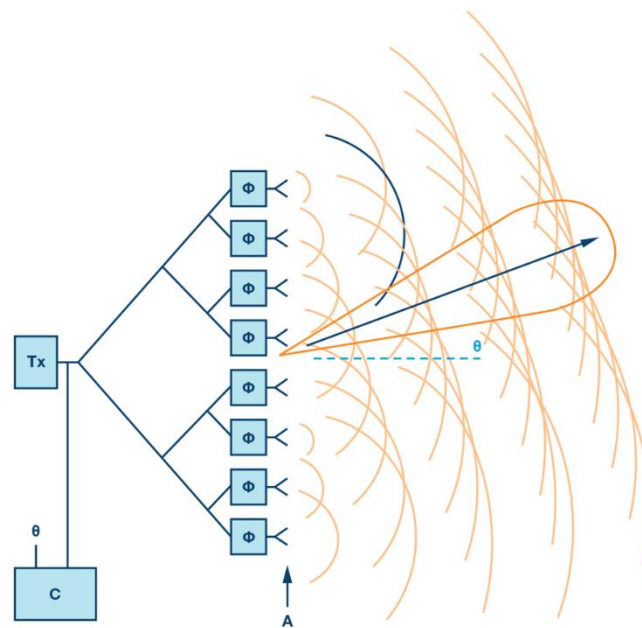Figure 3-1   The basic theoretical block diagram of the phased array unit.

3.2 Beamforming technology

Beamforming is the combination of antenna technology and digital signal processing for directional signal transmission or reception. As early as the 1960s, the array signal processing technology using antenna diversity reception has been highly valued in electronic countermeasures, phased array radar, sonar and other

communication equipment. The DBF-based adaptive array jamming zeroing technology can improve the anti-jamming capability of the radar system and is a key technology that must be used in the new generation of military radars. The positioning communication system obtains the sound field information through the microphone array, uses the principles of BF and power spectrum estimation, processes the signal, and determines the direction of the incoming wave of the signal, so as to accurately orient the signal source.

BF was derived from the adaptive antenna concept. The signal processing at the receiving end can form the desired ideal signal by weighting and synthesizing the signals received by the multi-antenna array elements. From the perspective of the antenna pattern, this is equivalent to forming a beam in a defined direction. For example, the previous omnidirectional receiving pattern is converted into a lobe pattern with zero and maximum pointing. The same principle applies to the transmitter. The amplitude and phase of the antenna element feed can be adjusted to form a pattern of the desired shape. If beamforming technology is to be used, the premise is that a multi-antenna system must be used. For example, MIMO, not only using multiple receive antennas, but also multiple transmit antennas. Since multiple groups of antennas are used, the wireless signals from the transmitter to the receiver correspond to the same spatial stream , including multiple paths. Using a certain algorithm at the receiving end to process the signals received by multiple antennas can significantly improve the signal-to-noise ratio of the receiving end. Even when the receiving end is far away, better signal quality can be obtained. MIMO can greatly improve network transmission rate, coverage and performance. When multiple independent spatial streams are simultaneously delivered based on MIMO, the throughput of the system can be multiplied.

In BF, we have many wave sources (ie antenna arrays). By carefully controlling the relative phases and amplitudes between the waves emitted/received by the wave sources, it is possible to control the electromagnetic wave radiation/reception gains to be strong in one direction (i.e. reception The location where the receiver/transmitter is located) and null in other (that is, the interference to other receivers is reduced/the chance of being interfered by other transmitters is reduced).

We take the receiving antenna array as an example, as shown in Figure 3-2. For electromagnetic waves traveling in the direction we want, the wavefront arrives at each antenna in the antenna array at a different time (phase). For each antenna, we adjust a specific phase delay to compensate the phase of the wavefront arriving at the antenna. Therefore, after the phase delay, we align the signals received by each antenna in phase, so that they are different from each other. The useful signal received by the antenna will become very large after summing. On the other hand, when the interfering signals propagating in other directions reach the antenna array, the delay corresponding to each antenna does not match the time difference between the signals arriving at the antennas, so the amplitude does not increase after the summation. In this way, the antenna array can equivalently implement a directional antenna using multiple common antennas with a specific delay. According to the principle of reciprocity of the antenna, the same architecture can also be used in the transmit antenna array to be equivalent to a highly directional antenna. In addition, the direction of the antenna radiation can be realized by changing the relative delay and amplitude between the wave sources, and the change of the relative position between the transmitting end and the receiving end can be easily tracked.
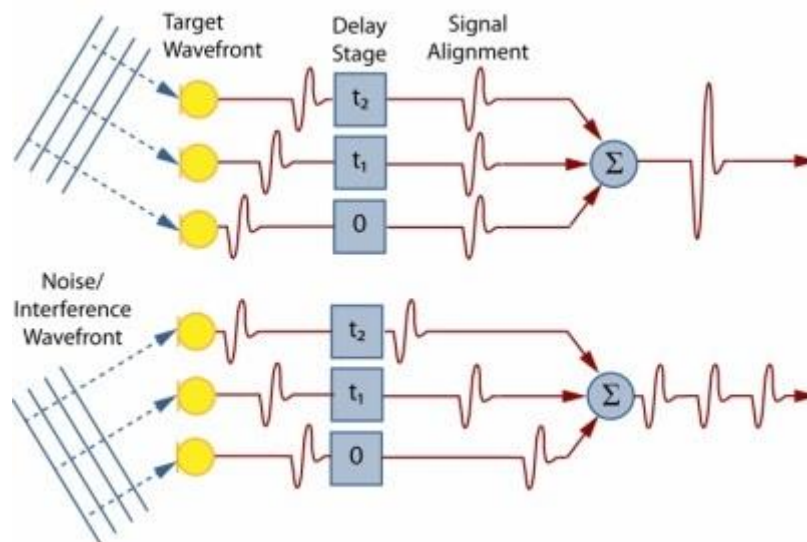
Figure 3-2   The phase delay of the signal.

BF is proposed to implement in mm-Wave applications such as 5G and beyond. The wavelength of the mm-Wave band is so small such that multiple antennas can be integrated into a mobile phone to realize BF in the mm-Wave band. BF and mm-

Wave technology are a match made in heaven. The use of mm-Wave can bring greater bandwidth to signal transmission, and BF can solve the problem of spectrum utilization, making 5G communication even more powerful.

## 3.3 Classification of beamforming techniques

3.3.1 Digital beamforming

DBF has multiple digital channels and realizes BF by adjusting the phase of each digital channel. Each antenna element is connected to an independent digital channel, and the DBF can generate multiple beams. Since the phase adjustment is directly achieved through the digital channel, this method is called DBF. DBF technology was originally used in the receiving system of phased array radar. It forms the main lobe in the desired direction to pass the useful signal and uses the side lobe to suppress the signal in the undesired direction, so as to achieve the purpose of spatial filtering. But because the array antenna has the reciprocity characteristic in the transmitting and receiving state, DBF can also be used in the transmitting system. The transmitting system of the digital array radar is also a multi-channel system. For one digital T/R channel, each antenna unit controls the array phase by changing the initial phase of the waveform generated by the direct digital synthesis (DDS) of each channel, so as to control the beam pointing purpose. The signal parameters (such as frequency, bandwidth, time width, frequency modulation form, etc.) of each channel are independently controllable, so the flexibility of BF is very good.

For antenna arrays, DBF is the ultimate means to achieve optimal performance, which can generate multiple independently steerable beams with flexibility and high quality. For a sufficiently large array with RF chains per element, any number of beams can be generated by assigning different complex weights (amplitude & phase) to each element in the digital domain. More advanced DBF strategies employ algorithms such as eigen-beamforming to obtain optimal SINR. In the future wireless communication network, channel capacity, delay, data rate, communication security, etc. are the most challenging indicators, and full DBF is an effective means for large-scale arrays to overcome these challenges.

Figure 3-3 shows a highly digitized transmitter BF network where each element in an antenna array is equipped with an RF transmit chain consisting of filters, local oscillator (LO), low noise amplifiers (LNA), power amplifier (PA), frequency

converter, analog-to-digital converter (ADC) and digital-to-analog converter (DAC). Based on this, each antenna element corresponds to a signal chain. However, this method is often very expensive in practical applications, and its cost is mainly reflected in computing resources and hardware devices such as RF devices, ADCs, DACs, and FPGAs. For most application scenarios, the cost of realizing all the functions required for ultra-high data rate communication systems through large all-digital arrays in the future is unacceptable. Not only that, under the existing device technology, this technology is still difficult to implement in many 5G base station antenna arrays.
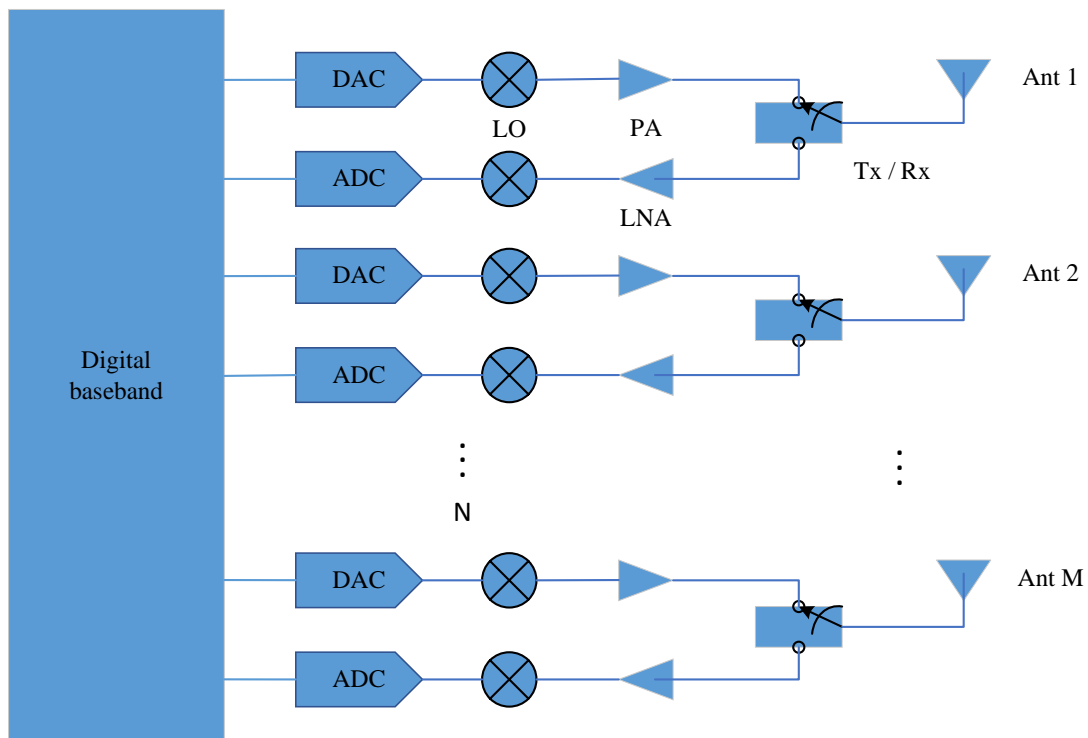
LO: Local Oscillator; PA: Power Amplifier; LNA: Low Noise Amplifier.
N: Number of ADC and DAC combinations; M: Number of antennas.
N = M

Figure 3-3   DBF system architecture.

3.3.2   Analog beamforming

There is only one digital channel in ABF, and a power divider is used to divide one digital channel into multiple channels, and each channel is connected to an independent antenna through an analog phase shifter. BF is achieved by adjusting the

phase of each analog phase shifter. Since the phase adjustment is achieved by the analog phase shifter at the back end of the digital channel, this method is called ABF.

In ABF technology, a single signal is amplified by analog phase shifters and directed to the desired receiver, and then fed to each antenna element in the antenna array. Among them, the amplitude/phase change is applied to the analog signal at the transmitting end, and the transmitting end sums the signals of different antennas and performs ADC. At present, ABF technology is the most cost-effective BF antenna array manufacturing technology, however, its disadvantage is that only one signal beam can be managed and generated.
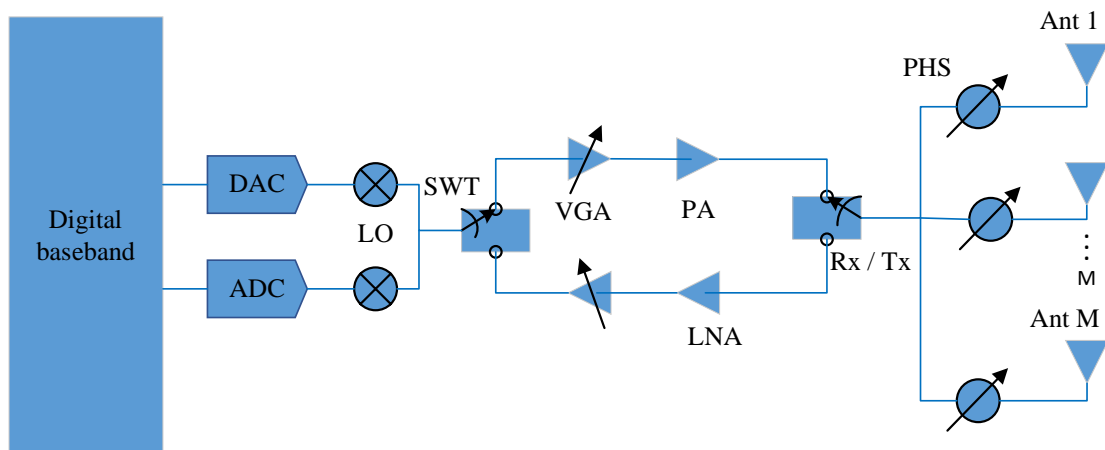
SWT: Switch; VGA: Variable Gain Amplifier; PHS: Phase Shifter.
N = 1, N < M

Figure 3-4   ABF system architecture.

Figure 3-4 shows that for ABF, one set of ADC\DACs is used for the entire array. This means that tuning the array with a simple software rewrite of all antenna elements is limited to what a single set of ADC\DACs can perform. This means that for the foreseeable future, HBF techniques combined with digital and analog may be the preferred strategy for implementing large multi-steerable beam antenna arrays.

3.3.3   Hybrid beamforming

A common trade-off these days that offers more flexibility than analog options but reduces cost compared to all-digital options is a HBF approach, which uses a subarray of ABF followed by a digital combination of the subarray signals. For

example, there may be a set of ADC/DACs for every 4 or 16 elements. This hybrid approach is popular in applications that require DBF but may not be practical due to size, power, or cost constraints.

HBF combines the advantages of ABF and DBF techniques, which can approach the optimal performance achievable by an all-digital system while reducing hardware cost and signal processing complexity. The core idea of HBF is to divide the entire large array into multiple small sub-arrays, that is, each antenna element is no longer completely independent. Among them, each subarray is an ABF array, and the number of divided subarrays determines the beamforming degree of freedom of the entire large array. In 5G terminology, this is also known as Array of Subarrays (AOSA).

Since ABF can be implemented with only analog phase shifters and the like, the cost of the system is greatly reduced due to the reduction in the number of complete RF chains required. However, the number of data streams or beams that can be supported by an HBF array is also reduced compared to a full DBF array. In practical applications, the design of this type of antenna array needs to comprehensively consider BF capability, system complexity, system budget, etc., and these issues are directly affected by the required number of controllable beams and the acceptable cost. In addition, although the reduction of the number of RF chains will reduce the number of data streams, the per-user performance of the HBF system can be approximated by a rational design of an all-digital system. HBF is a more practical option for current mm-Wave communication systems because radios in this band rely on line-of-sight propagation and the number of single-base-station users is reduced.

Figure 3-5 shows the basic structure of the transmitting and receiving system based on HBF. The antenna array is divided into N sub-arrays, each sub-array contains M antennas and RF units, and the devices in it can be further divided into multiple sub-arrays based on the actual design. The antenna unit is shared. For ease of understanding, a HBF array including N sub-arrays and M sub-array antenna elements is referred to as an N×M hybrid array herein.
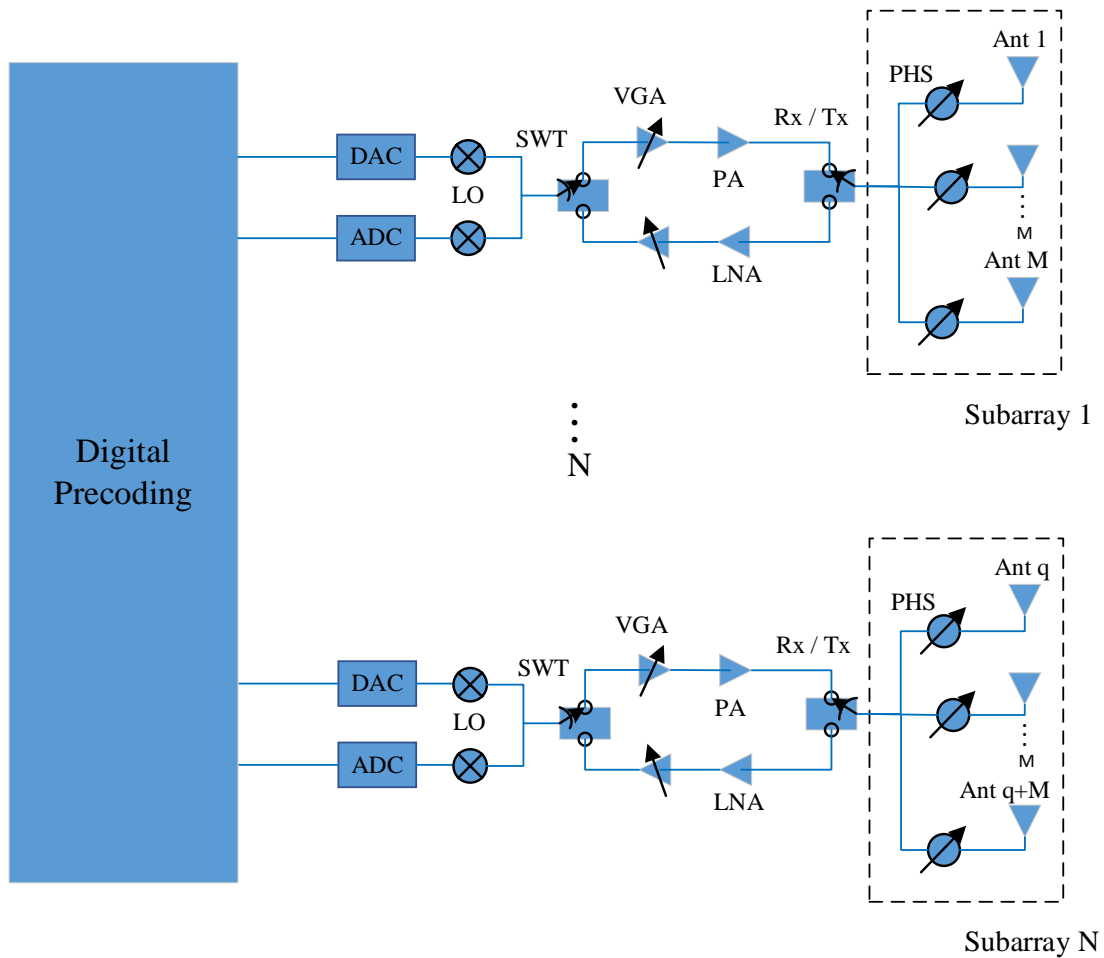
Figure 3-5　HBF system architecture.

　　　　Specifically, for an array of a given size, the size of the subarray (ie, the determination of M and N) represents a trade-off between system cost and performance. For example, when M is larger, the gain of the antenna is improved and the cost is lower, but the number of users that the array can support will be limited. The distance between corresponding elements in adjacent subarrays is called the subarray spacing, which is determined by the desired multibeam performance and the available physical area of the array. All sub-array signal processing will be agreed in the baseband processor. In the transmitting array and the receiving array, the sub-array and the unified processor are connected through ADCs and DACs respectively.

3.4 HBF system model based on phased array antenna

3.4.1 Active phased array

The potential to use mm-Wave frequencies in future wireless cellular communication systems has inspired research into PAAs for highly directional BF. However, due to the high cost and high power consumption of RF chain components in high frequencies, conventional all-DBF approaches requiring one RF chain for each antenna element are not feasible from a cost and power consumption perspective . Therefore, hybrid analog-digital beamforming with a reduced number of digitization chains and digital processing units is gradually applied to BS systems. In this HBF architecture, DBF is performed in a digital processing unit, such as a FPGA, while ABF is implemented by an analog phased array. Unlike DBF, which requires one digitizing chain per antenna element, the HBF architecture replaces all digitizing chains with phased-array subarrays. Therefore, the number of FPGAs, digital-to-analog converters (DACs), analog-to-digital converters (ADCs), mixers and amplifiers can be greatly reduced.

Electromagnetic waves at mm-Wave frequencies suffer from high attenuation caused by free-space path loss and shadowing [43], [44]. Fortunately, the shorter wavelengths of mm-Wave signals enable greater antenna gain using antenna arrays with a large number of antenna elements. It is well known that existing mm-Wave point-to-point communication systems with large antenna arrays can achieve multi-gigabit data rates over line-of-sight distances of several kilometers. However, fixed narrow beams have limited geographic coverage and cannot support mobile communication environments well. According to this situation, some advanced multi-beam or beam-steerable antenna array technologies have been recently adopted to realize 5G mm-Wave cellular communication, such as passive multi-beam antennas in [45]–[47], lens-based beam - Switched antenna systems in [48] and [49], and active phased arrays in [50] and [51]. In general, active BF systems can provide higher transmit power and better beamforming flexibility than passive multi-beam antenna arrays. Combined with MIMO technology, the performance of active BF systems can be further improved. With advanced BF precoding, MIMO communication systems can generate multiple beams to provide multiple data streams to support SU and MU MIMO communications [51].

3.4.2 Hybrid precoder design

Multi-user precoding involves assigning weight vectors to different mobile stations (MS) prior to transmission over multiple antennas of a base station (BS). Proper selection of weight vectors enables spatial separation between users, enabling multiplexing of multiple data streams. In lower frequency conventional MIMO systems, precoding is performed at the baseband by a digital signal processing unit. This design requires dedicated RF chains for each antenna element. Unfortunately, the high cost and high power consumption of current mm-Wave mixed-signal hardware technologies make all-digital transceiver architectures impractical. Therefore, mm-Wave systems require suitable MIMO architectures and signal processing algorithms. Recent work on precoding/combining design for mm-Wave systems advocates the use of hybrid analog/digital precoders/combiners. In this hybrid structure, the analog precoder/combiner allows BF gain, while the digital precoder/combiner provides multiplexing gain. ABF/combining can be accomplished using phase shifters, switches, or even lenses. Using a phase shifter, the relative phase of the RF signal is changed to steer the transmit/receive beam in the desired direction. The phase change may be digitally controlled, so there is only a quantized value.

Hybrid precoding for single-user mm-Wave systems is studied in [25]. The results show that the hybrid precoding/combining can achieve almost the same performance as the all-digital design. By exploiting the low scattering properties of mm-Wave channels, the angle of departure (AoD) and angle of arrival (AoA) response vectors assigned to the analog precoders and combiners to most major channel paths are near-optimal [25]. Using the obtained RF precoder/combiner, the baseband precoder/combiner can then be derived such that the resulting hybrid precoder/combiner is as close as possible to the digital precoder/combiner. In addition, hybrid precoding/combining for multi-user mmWave systems is investigated [52]–[54]. In [52], [53], the authors propose a two-stage hybrid precoding design. In the first stage, each MS and BS jointly select the "best" combination of RF combiners and RF BFs to maximize the channel gain for that particular MS. The baseband digital precoder is then derived as a zero-forcing (ZF) precoder by inverting the effective channel. In [55], an iterative hybrid precoding/combining algorithm that exploits channel reciprocity is proposed for multi-user systems with single-stream transmission per user. The work in

[56] established the required number of RF chains and phase shifters to enable hybrid precoding to achieve the same performance as digital precoding.

In this work, we study a multi-user mm-Wave system similar to those in [52], [53].The simplified model is shown in Figure 3-6. Specifically, we consider the practical limitations of the RF chain and design an RF precoder by extracting the phase of the conjugate transpose of the aggregated downlink channels to obtain large array gains in massive MIMO systems. The low-dimensional band ZF precoding is then performed based on the equivalent channel obtained by the product of the RF precoder and the actual channel matrix. This hybrid precoding scheme, called Phased-ZF (PZF), was shown to approach the performance of near-optimal but practically infeasible full-complexity ZF precoding in massive multi-user MIMO scenarios.
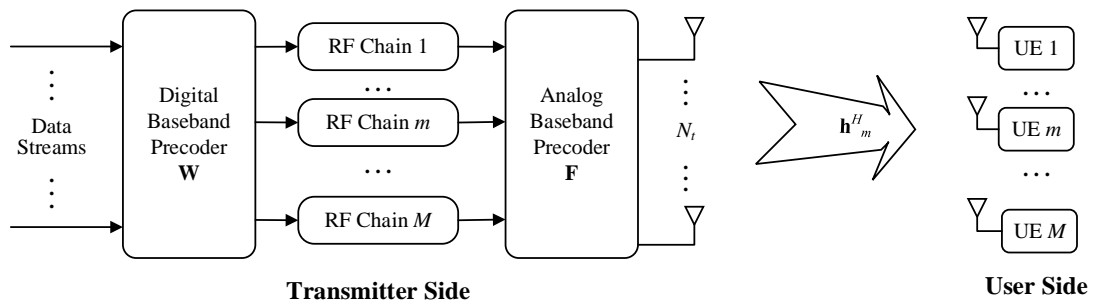
Figure 3-6　Simple model of hybrid mm-Wave precoding structure.

3.4.3　HBF system model

We consider a massive MIMO BS system model for narrowband scenarios. The BS uses a half-wavelength spacing uniform linear array (ULA) and the user terminal is equipped with a single antenna. In addition, it is assumed that all users have the same priority. The number of RF chains is twice the total number of data streams. The HBF structure can accurately achieve the same performance as any fully DBF regardless of the number of antennas. Therefore, we set the system to include N_t antennas and N_{RF} RF chains to transmit N_s data streams to M single-antenna users where each user receives d data streams. Let $N_s \leq N_{RF} \leq N_t$, where $N_s = Md$. As shown in Figure 3-7, the BS uses an HBF precoder. We consider a fully connected structure in which each RF chain is coupled to all antennas on the BS through 2 phase shifters. In the HBF structure, the BS first uses the $N_{RF} \times N_s$ digital precoder $\mathbf{V}_D$ to

digitally modify the baseband data stream. It then upconverts the processed signal to the carrier frequency through NRF RF chains. After that, the BS uses the $N_\text{t} \times N_\text{RF}$ RF precoder $\mathbf{V}_\text{RF}$, which is implemented using an analog phase shifter. Therefore, the transmitted signal is:

$$\mathbf{x} = \mathbf{V}_\text{RF}\mathbf{V}_\text{D}\mathbf{s} = \sum_m^M \mathbf{V}_\text{RF}\mathbf{V}_{\text{D}_m}\mathbf{s}_m \tag{1}$$

where $\mathbf{s} \in \mathbb{C}^{\text{Ns}\times 1}$ is a vector that represents the data symbol. It is a cascade of data flow vectors for each user. In addition, assume that $\mathbb{E}[\mathbf{ss}^\text{H}] = \mathbf{I}_{\text{N}_\text{s}}$. For the mth user, the received signal can be written as:

$$\mathbf{y}_m = \mathbf{h}_m^H \mathbf{V}_\text{RF}\mathbf{V}_{\text{D}_m}\mathbf{s}_m + \mathbf{h}_m^H \sum_{\ell \neq m} \mathbf{V}_\text{RF}\mathbf{V}_{D_\ell}\mathbf{s}_\ell + \mathbf{z}_m \tag{2}$$

where $\mathbf{h}_m \in \mathbb{C}^{\text{Nt}\times 1}$ represents the complex channel gain vector from the transmit antennas of the BS to the mth user, and $\mathbf{z}_m \sim \mathcal{CN}(0, \sigma^2)$ represents additive white Gaussian noise.
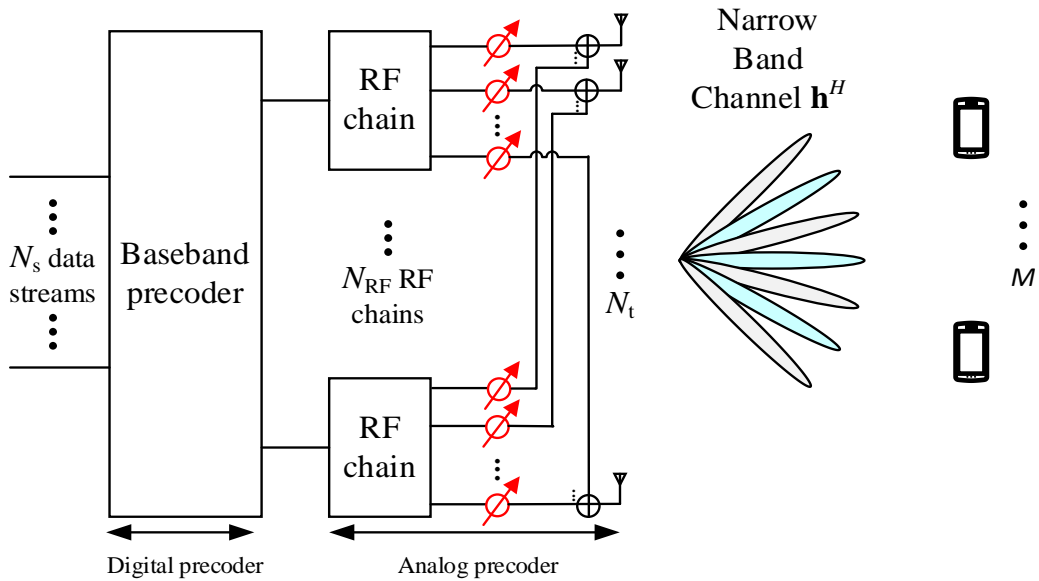

Figure 3-7   System structure of mm-Wave massive MIMO HBF.

Chapter 4 HBF Design Based on Deep Learning

This chapter mainly introduces the neural network model based on unsupervised learning and uses the classic Saleh-Valenzuela(SV) mm-Wave channel model to obtain the dataset through simulation and estimation. The simulation results are given and discussed, and finally, the complexity analysis and summary are described.

4.1 Concept

For the optimization of HBF, the hardest part is that the constant modulus constraint of the analog beamformer due to the phase shifter makes the problem very highly non-convex and difficult to solve. The latest research on intelligent communication shows that the DL technique has great potential in solving traditionally challenging problems [19]–[22]. The DL technique has proven to be an excellent tool for solving complex nonconvex optimization problems. Its unique network structure can approximate any function under specific conditions and use DL to solve the optimization problem of BF design, which subsequently leads to eliminate the complexity caused by too many iterations and realize real-time calculations. Although the training process is very time-consuming, the training is carried out in an offline mode. Therefore, the DL technology is an effective method to reduce cellular network delay. At the same time, to avoid the labeling overhead of labels, we intend to design HBF based on unsupervised learning. We develop an unsupervised deep neural network (USDNN) for the BF design problem that can be trained on how to optimize beamformers to maximize the reachable sum rate.

4.2 Channel model

Channel State Information (CSI) is very important and for describing characteristics of wireless communication channel. This information is used to define the known channel properties of the communication link, describing how a signal propagates through the channel from the transmitter to the receiver. It characterizes the synthesis of a series of effects, such as scattering, fading, and attenuation with distance. CSI, which can be obtained by channel estimation algorithms, enables us to adjust the transmission according to the current channel conditions, estimate the quality of the

wireless channel through the CSI matrix, and adjust the communication rate based on the CSI, thereby ensuring reliable and effective data transmission.

There are mainly two kinds of CSI, transient CSI and statistical CSI. Transient CSI require the priori knowledge of channel state which is decribed with impulse response of the channel. This makes it possible to adjust the transmitted signal according to the impulse response, thereby optimizing the received signal for spatial multiplexing or low error rates. On the other hand, statistical CSI means that the statistical characteristics of the current channel are known. These statistical features include the type of fading, average channel gain, line-of-sight components, and spatial correlation. As in the transient case, this information can also be used to optimize the transmission.

The acquisition of CSI is mainly limited by how fast the channel state changes. In a fast fading system, the transmission of each information symbol is accompanied by a rapid change of the channel state, so only statistical CSI can be obtained. In slow fading systems, on the other hand, it is also feasible to use transient CSI to adjust the transmission over a period of time. In practical applications, a compromise is generally chosen: transient CSI including quantization errors and statistical CSI. We can efficiently collect the spatial transfer function between each antenna and each user terminal, collecting the CSI information in the matrix **H**, as shown in Figure 4-1.

$$H = \begin{bmatrix} h_{1,1} & \cdots & h_{1,K} \\ \vdots & \ddots & \vdots \\ h_{M,1} & \cdots & h_{M,K} \end{bmatrix}$$
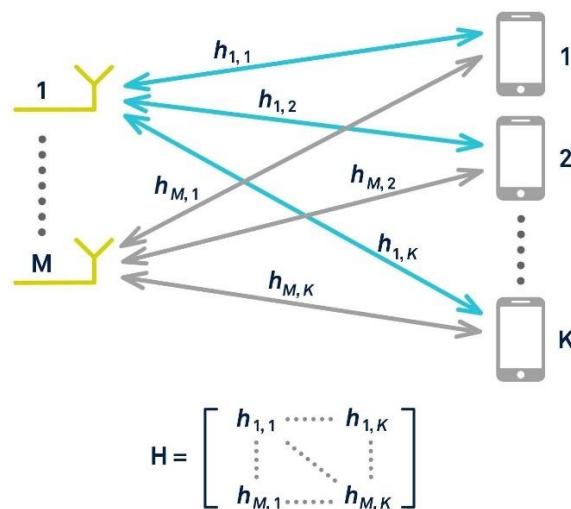
Figure 4-1   CSI used to characterize massive MIMO system.

In the mathematical model of the MIMO wireless channel, $h_{i,j}$ is the complex channel gain of the ith antenna at the transmitter and the jth antenna at the receiver. The gain comes from the superposition of multiple rays, and each ray reaches the receiver through multiple different paths.

$$h_{i,j} = \sum_i a^i e^{j\theta_t} \tag{3}$$

The modulus of the channel gain $|h_{i,j}|$ obeys the Rayleigh distribution, and if there is a strong direct line-of-sight (LoS) path in addition to a large number of scatterers, then the modulus of the channel gain $|h_{i,j}|$ obeys the Rice distribution. For the low-frequency case, it can be directly assumed that each sub-channel $h_{i,j}$ obeys Rayleigh or Rice distribution.

Different from the low-frequency channel, the mm-Wave channel basically propagates in a straight line and has poor diffraction ability. The channel has fewer scattering paths, which are often far less than the number of transmitting and receiving antennas. Therefore, its channel model has rich geometric characteristics. However, low-frequency channels are often modeled as random channels such as Rayleigh distribution due to the abundance of scattering paths, and therefore do not contain information about the communication environment. So we consider the SV channel model [6], where the channel contains a LoS path and an L_m non-LoS (NLoS) path. For the SV channel model, the channel vector $\mathbf{h}_m^H$ of the mth user can be expressed as

$$\mathbf{h}_m^H = \sqrt{\frac{N_{\text{t}}}{L_m+1}} \sum_{l=0}^{L_m} \beta_m^{(l)} \boldsymbol{a}^H\left(\phi_m^{(l)}\right) \tag{4}$$

where $\beta_m^{(0)} \boldsymbol{a}^H\left(\phi_m^{(0)}\right)$ is the LoS path of $\mathbf{h}_m$ with $\beta_m^{(0)}$ presenting the complex gain and $\phi_m^{(0)}$ denoting the spatial direction, $\beta_m^{(l)} \boldsymbol{a}^H\left(\phi_m^{(l)}\right)$ for $1 \le l \le L_m$ is the lth NLoS path of $\mathbf{h}_m$, and $\mathbf{a}(\phi)$ is the $N_{\text{t}} \times 1$ array steering vector at the BS.

Since millimeter wave signal propagates in a straight line and has poor diffraction ability, the number of paths cannot be too small. After the multipath signals with different amplitudes and phases are superimposed, the amplitude and phase of the synthesized signal $\mathbf{y}_m$ will fluctuate violently, that is, multipath fading occurs. Inter-symbol interference (ISI) caused by multipath effects affects the quality of signal

transmission. So for the mth user, the parameters of the spatial channel are set as follows:
1) the number of paths: the number of LoS paths is 1, the number of NLoS (L_m) is 2,
and the total number of paths is L = 3; 2) Path gain: $\beta_m^{(0)} \sim \mathcal{CN}(0,1)$ and
$\beta_m^{(l)} \sim \mathcal{CN}(0, 10^{-0.5})$ for l = 1, 2; 3) Spatial direction: $\phi_m^{(0)}$ and $\phi_m^{(l)}$ are uniformly
distributed within [-0.5, 0.5].

The signal-to-interference-noise ratio (SINR) of the signal received by
the mth user is expressed as:

$$\text{SINR}\left(\mathbf{V}_{\text{RF}}, \mathbf{v}_{\text{D}_m}\right) = \frac{\left|\mathbf{h}_m^H \mathbf{V}_{\text{RF}} \mathbf{v}_{\text{D}_m}\right|^2}{\sigma^2 + \sum_{\ell \neq m}\left|\mathbf{h}_m^H \mathbf{V}_{\text{RF}} \mathbf{v}_{\text{D}m}\right|^2} \tag{5}$$

There are three methods for channel estimation: the first is channel
estimation based on pilot symbols; the second category is channel estimation based on
decision feedback; the third category is blind channel estimation based on the limited
character characteristics and statistical characteristics of the transmitted information
symbols. In this thesis, the first method to estimate the SV channel is chosen.

4.3 The proposed DL model

As shown in Figure 4-2, the network consists of 5 dense layers, 4 batch
normalization (BN) layers, and 5 activation functions. The dense layer has 1024, 512,
256, 128, and $N_t N_{\text{RF}}$ neurons respectively. The number of neurons is determined by
empirical experiments to ensure sufficient learning ability. The BN layer is to enhance
convergence and is used to stabilize the training process. For the activation function,
the first four dense layers use a rectified linear unit (ReLu), and the last dense layer uses
a linear linear activation function to output phase-shift predictions.
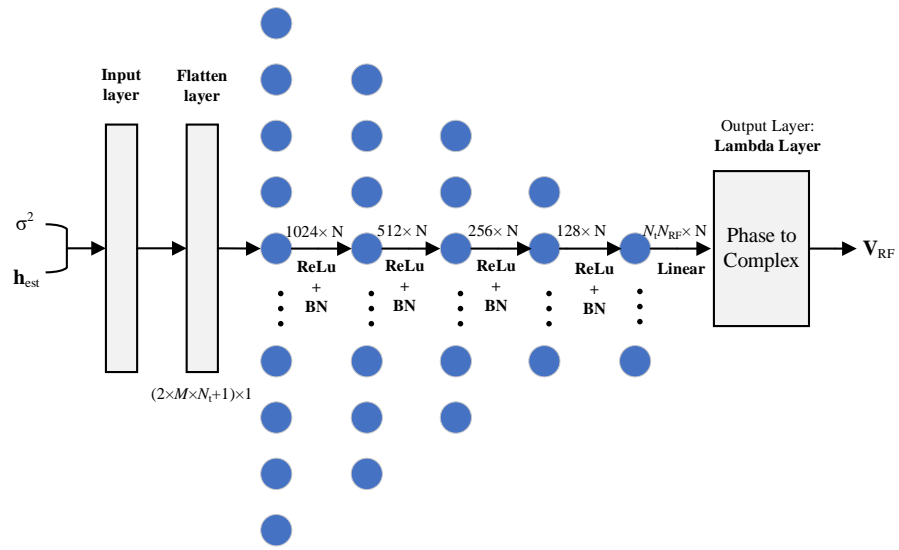
Figure 4-2   The proposed DNN architecture.

Training stage: As shown in Figure 4-3, we simulate the SV channel model to obtain the perfect CSI to calculate the loss function and update the calculated weight information to the USDNN model. Next, we apply the proposed classic mm-Wave channel estimator [57]. The BS sends pilot symbols with beamformers and receives user feedback signals $\mathbf{y}_p$ to estimate the channel estimate $\mathbf{h}_{est}$ (imperfect CSI) and noise power σ^2 which will be the input to the USDNN model. Using the loss function calculated to update the model's weights, the model then learns how to optimize $\mathbf{V}_{RF}$.

Testing stage: All model parameters have been trained and fixed. The imperfect CSI and σ^2 are used as input for testing, and the optimized beamformer can be obtained, which directly outputs the optimized beamforming vector $\mathbf{V}_{RF}$.

Figure 4-3   Two-stage design approach: training and testing.

In the previous work [33], the received user feedback signal $\mathbf{y}_p$ is used as input, and $\mathbf{y}_p$ itself is a function of $\mathbf{V}_{RF}$, which needs to be optimized. So we estimate the channel based on the $\mathbf{y}_p$ and obtain the $\mathbf{h}_{est}$ and σ^2 through the classical channel estimation algorithm [57] to form the input of the USDNN. The channel vector and the bea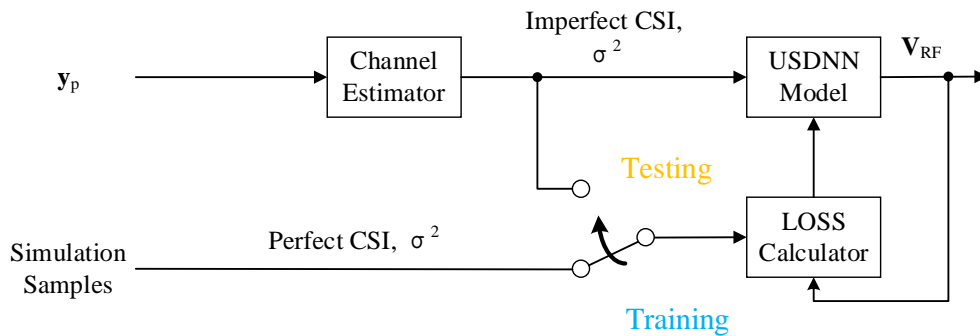mforming vector $\mathbf{V}_{RF}$ are both complex-valued vectors. Therefore, we consider the input layer and output layer as follows:

Input layer: For the real and imaginary parts and noise power σ^2 of each antenna's downlink channel CSI, the input data shape is a matrix of $2 \times M \times N_t + 1$ that is input through the flattened layer.

Output layer: Inspired by [15], the following lambda layer is used for the complex-valued output of USDNN. This layer converts the predicted real-valued phase into a complex-valued form $\mathbf{V}_{RF}$, where $\mathbf{P}_{pred}$ is the predicted real-valued phase.

$$\mathbf{V}_{RF} = e^{j \cdot \mathbf{P}_{pred}} = \cos(\mathbf{P}_{pred}) + j \cdot \sin(\mathbf{P}_{pred}) \tag{6}$$

We use the same USDNN architecture for the digital precoder $\mathbf{V}_D$ to set the number of neurons in the last dense layer to $2N_s N_{RF}$ and output the $N_s \times N_{RF}$ complex vector $\mathbf{V}_D$.

In this thesis, to compare performance, we calculated the achievable sum rate, expressed as:

$$R = \sum_{m=1}^{M} log_2 \left(1 + \text{SINR}(\mathbf{V}_{RF}, \mathbf{v}_{D_m})\right) \tag{7}$$

The optimization problem of $\mathbf{V}_{RF}$ and $\mathbf{V}_D$ is expressed as follows:

$$\max_{\mathbf{V}_{RF}, \mathbf{V}_D} R$$

$$\text{s.t.} \quad \left|[\mathbf{V}_{RF}]_{i,j}\right| = 1, \quad \forall\, i,j \tag{8}$$

$$\|\mathbf{V}_{RF}\mathbf{V}_D\|_F \leq 1,$$

The first expression is the constant modulus constraint of the phase shifter for ABF, and the second expression is the total transmission power constraint with normalized power.

To train the network, we use the Adam optimizer with an initial learning rate of 1×10^{-3}. The maximum number of epochs is set to 20,000 and tolerated early stop

20 is adopted to improve training efficiency. In addition, to accelerate the convergence speed, the learning rate is attenuated by a factor of 0.2 whenever the verification loss does not decrease for 20 consecutive periods. Table I shows the hyperparameters selected for USDNN training.

Table 4-1　Hyperparameters for USDNN training.

Parameter	Set Value
Epochs	20,000
Batch Size	256
Initial Learning Rate	1×10^{-3}
Minimum Learning Rate	5×10^{-5}
Reduce LR On Plateau (factor)	0.2
Reduce LP On Plateau (patience)	20

A computationally heavy optimization problem in supervised learning must solve each sample in the dataset and obtain data labels, which adds excessive difficulty to training. In unsupervised learning, our design does not require labels. It can be trained from the data obtained from the estimated channel, which can significantly reduce the computational complexity during training the network. In the training stage, the model learns under the perfect CSI as much as possible. In the testing stage, all model parameters have been trained and fixed, and imperfect CSI can be accepted as input for testing and directly output the optimized analog beamforming vector $\mathbf{V}_{\text{RF}}$. We train the network with the following new loss function directly related to the following performance goal:

$$Loss = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}\log_2\left(1 + \frac{\left|\mathbf{h}_{m,n}^{H}\mathbf{V}_{\text{RF},n}\mathbf{V}_{\text{D}m,n}\right|^2}{\sigma_n^2 + \sum_{\ell \neq m}\left|\mathbf{h}_{m,n}^{H}\mathbf{V}_{\text{RF},n}\mathbf{V}_{\text{D}m,n}\right|^2}\right) \tag{9}$$

where N represents the total number of the training samples, and $\sigma_n^2, \mathbf{h}_{m,n}^{H}, \mathbf{V}_{\text{RF},n}$ and $\mathbf{V}_{\text{D}m,n}$ represent the noise power, CSI, output BF, and DBF associated with the nth sample, respectively. The minus sign indicates that the sum rate is maximized when the USDNN is trained to minimize the loss function. We generate 10^5, 2×10^4 and 2×10^4 samples for training, verification, and testing, respectively,

while considering training efficiency, test performance, and stability. It is also worth emphasizing that labels are not required since unsupervised learning is used. Hence the cost of obtaining training samples is considerably reduced.

4.4 Results and discussion

We consider a system equipped with 64 antennas on the BS throughout the simulation process, and 8 RF chains transmit 4 data streams to 4 single-antenna users for communication. The signal-to-noise ratio (SNR) is defined as an indicator of the channel estimation level. Figure 4-4 shows that our proposed USDNN has excellent performance when the CSI is imperfect. When L-train = L-test = 3, N_t = 64, N_{RF} = 8, M = 4, and N_s = 4, we can easily observe that the achievable sum rate of our proposed method speeds up with the increase of SNR. Our proposed USDNN network can achieve near-optimal performance.

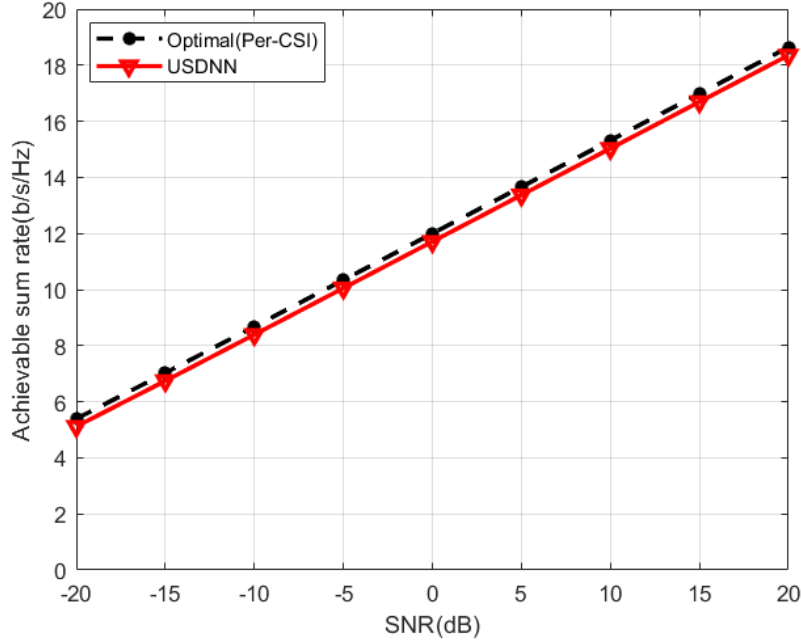Figure 4-4   Achievable sum rate under different SNRs (M = 4, N_t = 64, N_{RF} = 8, N_s = 4, L-train = L-test = 3).

Next, we evaluate how the system performance changes when the number of users $M \in \{2, 4, 6, 8\}$ and the number of antennas $N_t \in \{16, 32, 64, 128\}$ are changed. The SNR is set to 0 dB, and the number of channel paths is L-train = L-

test = 3. When increasing the number of antennas or users of the BS, a more complex DNN should be required. However, Figures 4-5 and 4-6 show that the proposed architecture is sufficiently complicated to have excellent performance with different antennas and users. As shown in Figure 4-5, when the number of users increases, the rate of change of the proposed method decreases, because when increasing the number of users and the number of fixed antennas will generate more inter-user interference. But the rate of our method still grows rapidly, always close to the optimal sum rate.
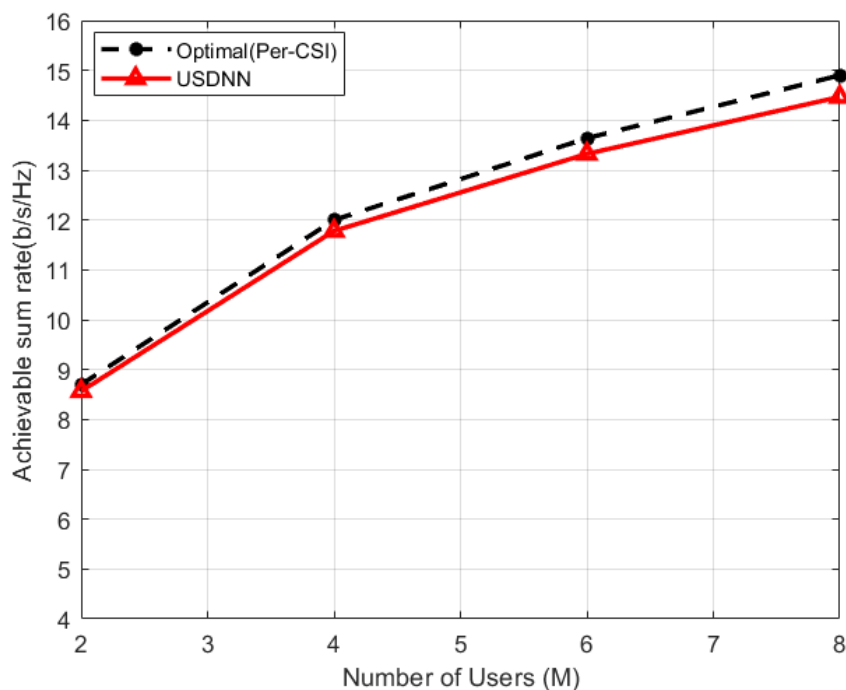
Figure 4-5   Achievable sum rate with different numbers of users ($N_t = 64$, $N_{RF} = 8$, d = 1, L-train = L-test = 3).

In Figure 4-6, we evaluate our proposed method with different numbers of antennas. When the number of antennas increases, the rate of change decreases with the increase in overhead signaling. Our proposed method is always close to optimal performance.
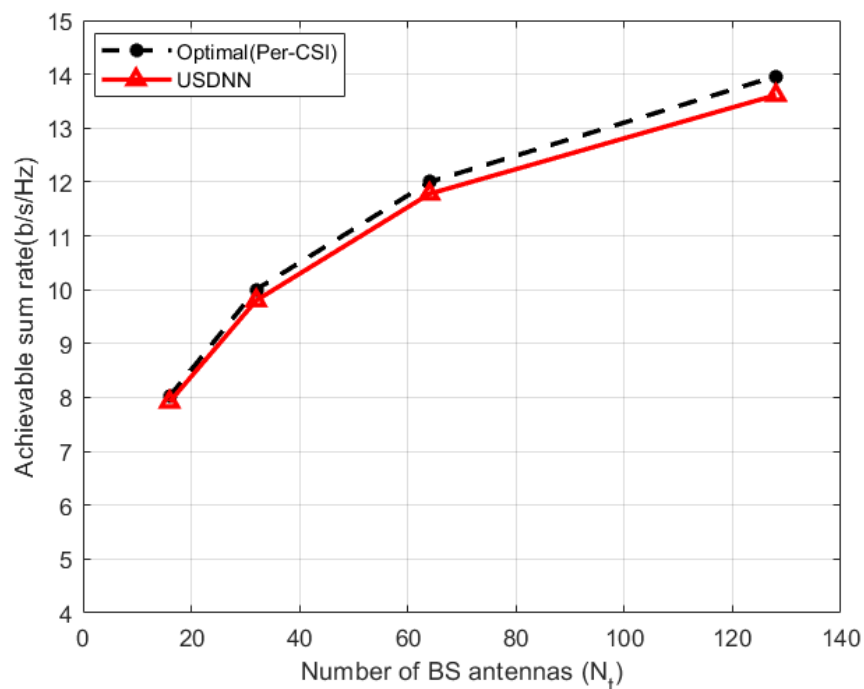
Figure 4-6 Achievable sum rate with different numbers of antennas ($M = 4$, $N_{RF} = 8$, $N_s = 4$, L-train = L-test = 3).

There usually are inconsistencies between the values of the parameters during testing and training in practical applications, resulting in unacceptable deviations between the results. Therefore, the robustness of the model should be investigated. Figure 4-7 shows the impact of the mismatch between the number of channel paths in the training stage and testing stage. It is assumed that the number of channel paths during training is L-train = 3, and L-test = 2, 3, 4, 5, and 6 are used for the testing stage. The best performance is obtained when the number of paths is consistent during training and testing. When the number of paths is inconsistent, there is a model mismatch, but the performance loss is limited, indicating that our model has a certain degree of robustness.
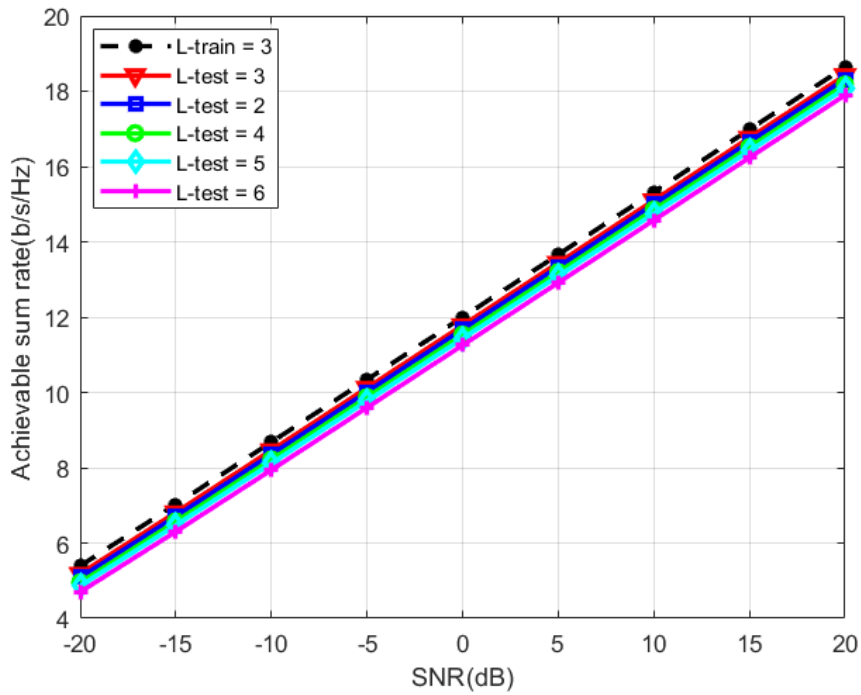
(a)
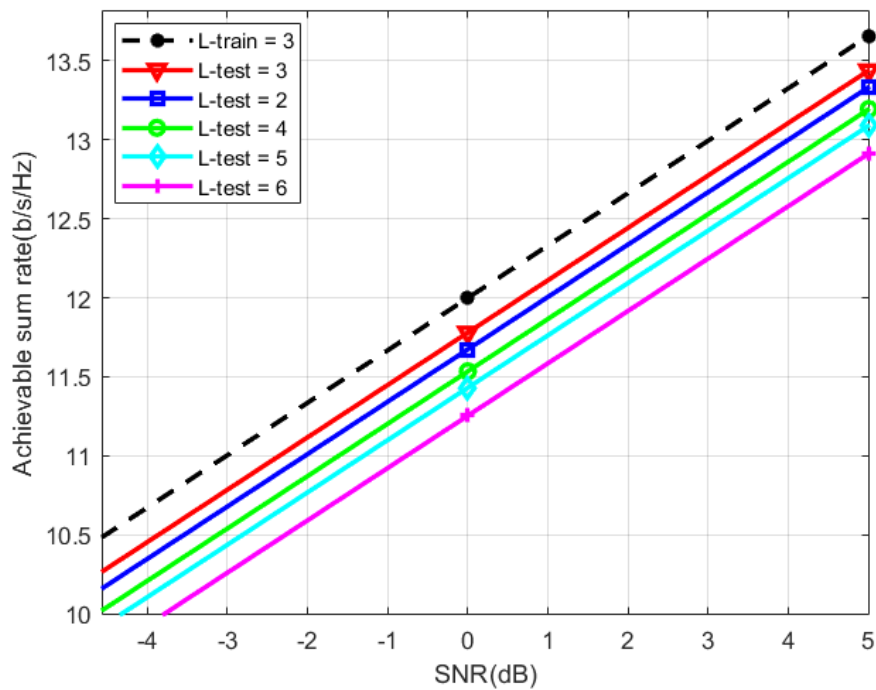
Figure 4-7　(a) Achievable sum rate when the channel path numbers of the model do not match ($M = 4$, $N_t = 64$, $N_{RF} = 8$, $N_s = 4$). (b) The contrast after magnification.

Finally, comparing the complexity of neural networks and traditional algorithms is a difficult task. For traditional model-based HBF designs [25], [26], the asymptotic computational complexity in complex multipliers is $O(N_t^3)$ because they involve operations such as singular value decomposition and matrix inversion. We analyze the computational complexity from the number of floating-point operations (FLOPs) proposed USDNN. Because the training is done offline, only the complexity of the online testing phase is calculated. The FLOP number of the Dense layer is given by $(2N_I - 1) N_O$, where N_I and N_O denote the input and output dimensions, respectively. But considering the number of FLOPs alone may not be sufficient to draw a fair conclusion, as how these operations are performed and the architecture used can also have profound effects.

Therefore, in this subsection, we also make an approximate comparison of the running time of all methods, as shown in Table II. These methods are run on the same Intel(R) Core (TM) i7-10700F CPU @ 2.90GHz and GeForce RTX 2080Super hardware configuration for a fair comparison. Unlike traditional algorithms, the large-scale matrix multiplication of USDNN can be effectively accelerated through the parallel computing of graphics processing units (GPU). It can be seen that the proposed USDNN runs dozens of times faster than the traditional HBF algorithm. Note that the computational time for training is not considered here since the training process is done offline.

Table 4-2　Comparison of running time (seconds) of different methods.

Algorithm / Parameters	USDNN	MO	OMP
$N_t = 16$	0.046	1.657	1.845
$N_t = 32$	0.052	1.693	1.962
$N_t = 64$	0.071	1.732	2.133
$N_t = 128$	0.102	1.834	2.368

($M = 4$, $N_{RF} = 8$, $N_s = 4$, L-train = L-test = 3).

4.5 Summary

BF design is a rather complex non-convex problem due to the joint optimization of multiple variables and constant modulus constraints. Our use of DL to

solve the optimization problem of BF is a very efficient method. Through training iterations with a large number of samples, it is demonstrated that the DL-based scheme has the ability to understand the complex characteristics of wireless channels. And the simulation results show that the DL-based method can be robust to imperfect CSI generation. In previous studies based on supervised learning, supervised learning needs to obtain a large number of training labels, thus requiring a lot of additional computing resources to find these labels using traditional optimization methods. For BF designs, we had a hard time finding a proper label. Therefore, we developed the USDNN model, and the proposed method based on unsupervised learning can not only avoid the labeling overhead of labels but also achieve very good performance.

Chapter 5 Performance Comparison with the Perfect CSI Scenario

This chapter starts from the principle of the algorithm, introduces two different HBF algorithms, and compares the results. Aiming at the non-convex problem of BF design, in previous studies, a model-based design approach was adopted to deal with this difficulty, and [25] proposed an OMP-based algorithm. However, analog beamformers are limited to predefined codebooks. In order to improve the performance of OMP and apply the manifold method to the optimization of analog BF, [26] proposed a MO-based hybrid beamforming algorithm that directly handles the constant modulus constraint of the analog components. All of the above algorithms assume a perfect CSI environment.

5.1 Orthogonal matching pursuit (OMP) algorithm

Compressed sensing (CS) is a new theory for signal sampling. Traditional information sampling is based on the Nyquist sampling theorem, which considers that the sampling rate of the signal is only not less than twice the highest frequency, the sampled digital signal retains all the information of the analog signal, and the signal can be restructured accurately. However, CS theory points out that if a signal is a sparse signal or can be represented as a sparse signal in a certain transform domain, then the signal can be projected onto a low-dimensional space through a matrix that is not related to the transform basis. This projection on the low-dimensional space contains all the information of the original signal and can accurately reconstruct the original signal. Under CS theory, signal sampling does not directly measure the signal itself, but the key is to sort out the structure and content of information in the signal.

CS theory was first proposed by Donoho and Candes et al. in 2004. CS focuses on how to use the sparsity of the signal itself to recover the original signal from some observed samples. At present, the main research directions of CS theory are divided into three parts: signal sparsity, measurement matrix, and reconstruction algorithm. The sampling process in CS is relatively simple, but the reconstruction is very complicated, so the research on the reconstruction algorithm is a very important aspect of CS. At present, the greedy iterative algorithm in the reconstruction algorithm is a very widely used class. This kind of algorithm selects the best support for the signal

by an iterative method, then selects the local optimal solution based on the greedy criterion, and gradually approaches the original signal. The most typical greedy algorithm is the matching pursuit algorithm (MP), and the improved OMP is based on it.

An overcomplete dictionary matrix $\mathbf{D} \in R^{n*k}$ is given, where each column of it represents an atom of a prototype signal. Given a signal y, it can be represented as a sparse linear combination of these atoms. The signal y can be expressed as y = $\mathbf{D}$x, where x is the original signal, y is the compressed signal, and $\mathbf{D}$ is the compressed matrix. The so-called overcompleteness in the dictionary matrix means that the number of atoms is much larger than the length of the signal y (its length is obviously n), that is, n<<k.

The basic idea of the MP algorithm: from the dictionary matrix D (also known as the overcomplete atom library), select an atom (that is, a column) that best matches the signal y, construct a sparse approximation, and find the signal residual, then Continue to select the atoms that best match the residual of the signal, and iteratively, the signal y can be linearly summed by these atoms, plus the final residual value to represent. Clearly, if the residual values are in a negligible range, the signal y is a linear combination of these atoms. The MP algorithm has a big disadvantage. If the vertical projection of the signal (residual value) on the selected atoms is non-orthogonal, this will make the result of each iteration not many optimal but sub-optimal, it takes many iterations to converge. So the OMP algorithm is introduced.

OMP algorithm: OMP algorithm is improved on the basis of the MP algorithm, as shown in Table 5-1. The same is that the standard for selecting atoms is the same as that of the MP algorithm, that is, the dictionary atoms that best match the test sample x are selected in the dictionary matrix D. The difference is that the OMP algorithm first performs the Schmidt orthogonalization operation on all the selected atoms in each iteration to ensure that the result of each loop is the optimal solution. Under the condition of the same precision, the performance of the OMP algorithm is better, and its convergence speed is also faster. At this point, there are two questions: Q1: How to ensure that the same atom will not be selected multiple times? A1: In OMP, the residuals are always orthogonal to the selected atoms. This means that an atom will not be selected twice and the result will converge in a finite number of steps. Q2: How

to ensure that the atoms in the space are orthogonal? A2: Atoms in space can be manipulated by Schmidt orthogonalization.

Table 5-1   OMP algorithm.

Initialization:
$f_0 = 0$, $\mathbf{R}_0 f = f$, $D_0 = \{ \}$ $x_0 = 0$, $a_0^0 = 0$, $k = 0$
(I) Compute $\{\langle \mathbf{R}_k f, \ x_n \rangle; \ x_n \in D \backslash D_k\}$.   (II) Find $x_{n_{k+1}} \in D \backslash D_k$ such that $\left\lvert\langle \mathbf{R}_k f, x_{n_{k+1}}\rangle\right\rvert \geq \alpha \, \overset{\sup}{\underset{j}{}} \left\lvert\langle \mathbf{R}_k f, x_j\rangle\right\rvert, \ 0 < \alpha \leq 1$.   (III) If $\left\lvert\langle \mathbf{R}_k f, x_{n_{k+1}}\rangle\right\rvert < \delta, (\delta > 0)$ then stop.   (IV) Recorder the directionary D, by applying the permutation $k + 1 \leftrightarrow n_{k+1}$.   (V) Compute $\{b_n^k\}_{n=1}^k$, such that, $x_{k+1} = \sum_{n=1}^k b_n^k x_n + \gamma_k$ and $\langle \gamma_k, x_n \rangle = 0, n = 1, \dots, k$.   (VI) Set, $a_{k+1}^{k+1} = \alpha_k = \|\gamma_k\|^{-2} \langle \mathbf{R}_k f, x_{k+1}\rangle$.

According to previous research, we use the sparse characteristics of the mm-Wave channel to express the precoding/combining problem as a sparse reconstruction problem. Using the principle of the OMP algorithm, we obtain an algorithm that can approximate the optimal unconstrained precoder and combiner, which can be implemented in low-cost RF hardware. $\mathbf{F}_{BB}$ is a digital precoding matrix of size $N_{RF} \times N_s$, and $\mathbf{F}_{RF}$ is an analog precoding matrix of size $N_t \times N_{RF}$. The optimal precoding problem of $\mathbf{F}_{RF}\mathbf{F}_{BB}$ can be described as:

$$\left(\mathbf{F}_{RF}^{opt}, \mathbf{F}_{BB}^{opt}\right) = \underset{\mathbf{F}_{RF}, \mathbf{F}_{BB}}{\mathrm{argmax}} \, I(\mathbf{F}_{RF}, \mathbf{F}_{BB}),$$

$$s.t. \quad \mathbf{F}_{RF} \in \mathbf{F}_{RF}, \tag{10}$$

$$\|\mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2 = N_s$$

Then, the above problem of jointly designing $\mathbf{F}_{RF}$ and $\mathbf{F}_{DB}$ is transformed into a univariate sparse constraint matrix reconstruction problem, and a spatial sparse precoding algorithm based on OMP is designed, as shown in Figure 5-2.

Table 5-2   Spatial Sparse Precoding Algorithm via OMP.

Require: $\mathbf{F}_{opt}$

1: $\mathbf{F}_{RF}$ = Empty Matrix

2: $\mathbf{F}_{res} = \mathbf{F}_{opt}$

3: **for** $i \leq N_t^{RF}$ **do**

4: $\mathbf{\Psi} = \mathbf{A}_t^* \mathbf{F}_{res}$

5: $k = \arg max_{\ell=1,\ldots,N_{cl}N_{ray}} (\mathbf{\Psi}\mathbf{\Psi}^*)_{\ell,\ell}$

6: $\mathbf{F}_{RF} = \left[\mathbf{F}_{RF} | \mathbf{A}_t^{(k)} \right]$

7: $\mathbf{F}_{BB} = (\mathbf{F}_{RF}^* \mathbf{F}_{RF})^{-1} \mathbf{F}_{RF}^* \mathbf{F}_{opt}$

8: $\mathbf{F}_{res} = \dfrac{\mathbf{F}_{opt} - \mathbf{F}_{RF}\mathbf{F}_{BB}}{\|\mathbf{F}_{opt} - \mathbf{F}_{RF}\mathbf{F}_{BB}\|_F}$

9: **end for**

10: $\mathbf{F}_{BB} = \sqrt{N_s} \dfrac{\mathbf{F}_{BB}}{\|\mathbf{F}_{RF}\mathbf{F}_{BB}\|_F}$

11: return $\mathbf{F}_{RF}$, $\mathbf{F}_{BB}$

It is worth noting that the compression matrix D is composed of a series of base signals or atoms, when two atoms in D are correlated, the OMP algorithm may get a wrong reconstructed signal.

5.2 HBF algorithm based on MO

Optimization on Riemannian manifolds, also known as Riemannian optimization, whose goal is to optimize a real-valued function defined on a manifold, has received increasing attention in recent years. Many practical problems can eventually be transformed into optimization problems defined on a manifold, such as signal separation, machine learning, recommender systems, network component analysis and classification, computer vision and graphics, etc. Many problems themselves are large in scale, but the data itself may fall on a low-dimensional manifold in a high-dimensional space, so it can be transformed into an optimization problem on the manifold.

Manifold constrained optimization problem refers to a special kind of constrained optimization problem whose constraints have manifold structure. Consider the general question:

$$\min_{x} f(x), \ s.t. \ x \in \mathcal{M}. \qquad (11)$$

where $\mathcal{M}$ represents the manifold constraint. This problem can be simply regarded as a constrained problem, and then solved by some traditional optimization methods for constrained optimization problems, such as the augmented Lagrangian method. There are three drawbacks of using the tradition methods. Firstly, the prevailing constraints are usually non-convex, the convergence is hard to guarantee. Secondly, these methods cannot guarantee that the iterative points always satisfy the constraints, and thridly, the manifold constraints are structured, and these methods do not explore this structural information. The three points mentioned above are exactly the three advantages of MO:

- All iteration points maintain constraint feasibility, ie always on the manifold.
- MO understands the previous point as an unconstrained optimization problem on a manifold, which facilitates analysis of convergence.
- Take advantage of the structural information of the manifold constraints themselves.

MO algorithm can be applied to beamforming, because MO is essentially a gradient descent method. However, the basic gradient descent method descends in the whole Euclidean space, so it cannot guarantee that the solution after descent still meets the constant mode constraint, so it cannot be directly used to solve $\mathbf{V}_{\mathrm{RF}}$. For MO, all feasible solutions satisfying the constant modulus constraint are first expressed as a manifold, and then the solution is mapped back to this manifold after each iteration, thus ensuring that the result will always satisfy the constant modulus constraint. The MO-HBF algorithm is shown in Figure 5-3.

Table 5-3   MO-HBF Algorithm.

Input: $\mathbf{H}_1, \sigma^2, \omega$ Output: $\mathbf{V}_{\mathrm{RF}}, \mathbf{V}_{\mathrm{U}}, \beta$
1: Initialize $\mathbf{V}_{\mathrm{RF},0}$ randomly and set $i = 0$;
2: **repeat**
3: Compute $\nabla J(\mathbf{V}_{\mathrm{RF},i})$ according to (11);
$\quad \nabla J(\mathbf{V}_{\mathrm{RF}}) = \frac{1}{\sigma^2 \omega}\left(\mathbf{V}_{\mathrm{RF}}(\mathbf{V}_{\mathrm{RF}}^H \mathbf{V}_{\mathrm{RF}})^{-1}\mathbf{V}_{\mathrm{RF}}^H - \mathbf{I}_{N_t}\right) \times \mathbf{H}_1 \mathbf{P}^{-2} \mathbf{H}_1^H \mathbf{V}_{\mathrm{RF}}(\mathbf{V}_{\mathrm{RF}}^H \mathbf{V}_{\mathrm{RF}})^{-1}.$ (11)
4: Use the MO method to compute $\mathbf{V}_{\mathrm{RF},(i+1)}$;

5: $i \leftarrow i + 1$;

6: **Until** a stopping condition is satisfied;

7: Compute β and $\mathbf{V}_U$ according to (7) and (8).

5.3 Results and discussion

Throughout the simulation, we still consider a system with 64 antennas on the BS and 8 RF chains delivering 4 data streams to 4 single-antenna user communications. We considered the SV channel model with the same parameters. Considering the comparison between MO and OMP algorithms, SNR is defined as an index of channel estimation level. Figure 5-1 shows that our proposed USDNN has excellent performance under imperfect CSI. When L-train = L-test = 3, $N_t = 64$, $N_{RF} = 8$, $M = 4$, $N_s = 4$, we easily observe that the reachability rates of all methods speed up with increasing SNR. Our proposed method is 74% better than Mo and 120% better than OMP, and our proposed USDNN can achieve near-optimal performance. Notably, both MO and OMP require perfect CSI knowledge and do not apply to real channels.
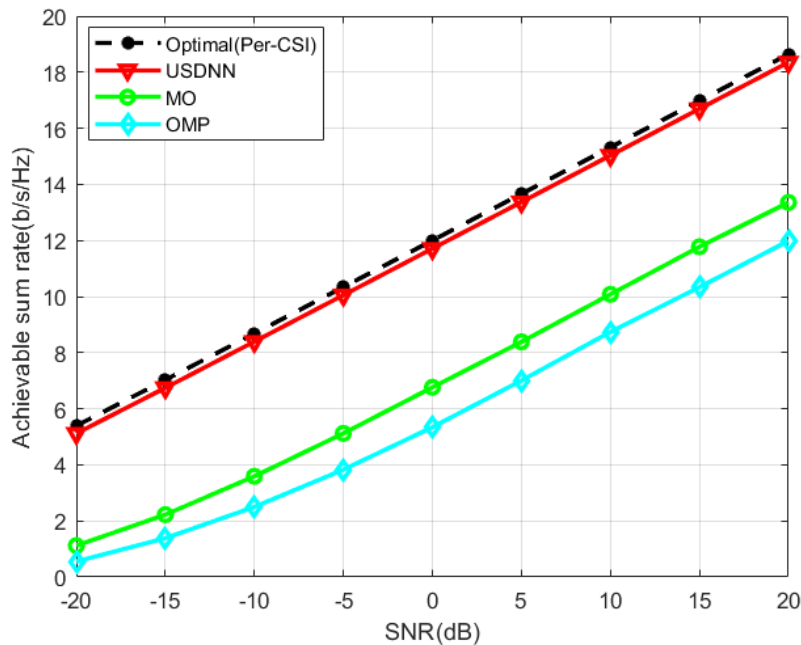

Figure 5-1　The achievable sum rate under different SNRs ($M = 4$, $N_t = 64$, $N_{RF} = 8$, $N_s = 4$, L-train = L-test = 3).

Next, we evaluate how the system performance changes when the number of users $M \in \{2,4,6,8\}$ and the number of antennas $N_t \in \{16,32,64,128\}$ are changed. The SNR is set to 0 dB, and the number of channel paths is L-train = L-test = 3. As shown in Figure 5-2, when the number of users increases, the rate of change of all methods decreases because increasing the number of users and the number of fixed antennas generates more inter-user interference. As a result, the MO and OMP grow slowly, and the rate of our method still grows rapidly, always close to the optimal sum rate.
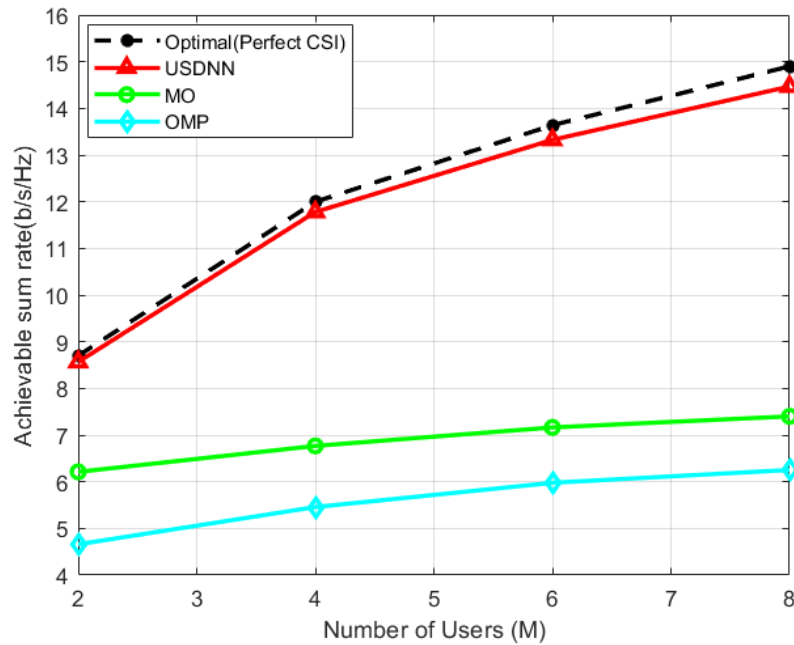
Figure 5-2   The achievable sum rate with different numbers of users ($N_t = 64$, $N_{RF} = 8$, d = 1, L-train = L-test = 3).

In Figure 5-3, we evaluate our proposed method with different numbers of antennas. When the number of antennas is low, the performance of the MO and OMP methods is poor. When the number of antennas increases, the performance gap becomes more significant, and the rate of change decreases with the overhead signaling increases. Our proposed method is consistently close to the optimal sum rate.
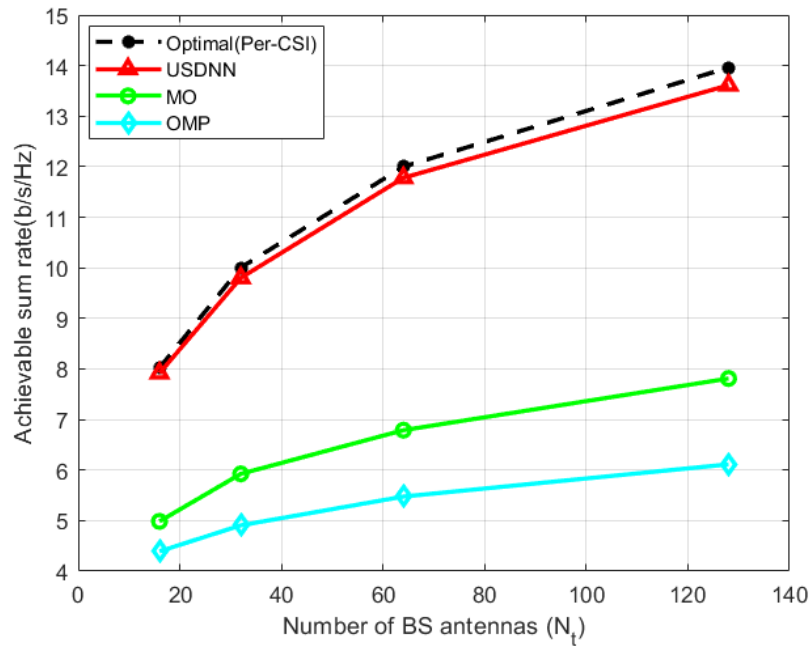
Figure 5-3 The achievable sum rate with different numbers of antennas ($M = 4$, N_{RF} = 8, N_s = 4, L-train = L-test = 3).

5.4 Summary

OMP is the most widely used algorithm and usually provides reasonably good performance. The algorithm requires selecting the columns of the analog precoding matrix from some candidate vectors, such as the channel's array response vector and the discrete Fourier transform beamformer. Therefore, the OMP-based hybrid precoder design can be viewed as a sparse constrained matrix reconstruction problem. While the design problem is greatly simplified in this way, limiting the space of feasible analog precoding solutions inevitably results in some performance penalty. In addition, getting the information of the array response vector ahead of time introduces additional overhead.

The MO is used to solve the constrained optimization problem, transforming the constrained optimization problem in the Euclidean space into an unconstrained optimization problem on the Riemannian manifold. Many things about optimization theory in Euclidean space can be extended, such as convexity of functions, smoothness, and so on. However, the MO method cannot solve all constrained optimizations, only those constraints that can be regarded as Riemannian manifolds,

such as orthogonal constraints, symmetric positive definite constraints, low-rank constraints, and so on. And the proximity class algorithm of Euclidean space is not easy to extend to the Riemannian manifold, at least in terms of calculation, because the property of "proximity friendly" in Euclidean space is not extended to the Riemannian manifold.

The two algorithms above either require some approximation to simplify the original objective function, or require a lot of time-consuming serial iterations to arrive at a solution. Moreover, in these algorithms, it is assumed that there is perfect channel state information, and there is a certain gap with the actual situation. However, the USDNN model proposed by us is not only suitable for the environment of imperfect CSI, but also can run the calculation quickly due to the acceleration of parallel computing, so it is more suitable for high-speed communication in the actual environment.

Chapter 6 Summary and Discussion

This chapter summarizes the use of DL techniques to solve non-convex optimization problems for BF. The competitive performance of USDNN is discussed, providing valuable insights for future BF designs.

6.1 Summary

This thesis studies the HBF optimization problem of narrowband mm-Wave MIMO communication systems. Instead of maximizing spectral efficiency as in most existing works, we use the reachable sum rate as a performance metric to characterize the performance of the transmission. The specific contents are summarized as follows:

1. We use DL technology to solve the non-convex optimization problem of analog beamformers, which can get rid of the complexity caused by too many iterations and realize real-time calculation. It can also reduce the computational complexity of the DNN model.

2. For BF designs, we had a hard time finding a proper label. If we use the optimized $\mathbf{V}_{RF}$ based on the traditional algorithm as the label, then the neural network can only be trained to be close to the traditional algorithm, but not better than the traditional algorithm. So we developed USDNN, a network that can be trained on how to optimize the beamformer to maximize the reachable sum rate. The simulation results show that, compared with the traditional BF algorithm, the method based on unsupervised learning can not only avoid the labeling overhead of labels, but also obtain better performance than the traditional algorithm.

3. Moreover, our research is based on imperfect CSI environment and is applicable to realistic downlink communication channels.

6.2 Discussion

HBF is a fundamental technology for massive MIMO systems that can reduce the number of RF chains and improve the achievable sum rate of the system. In this thesis, we propose a DL method based on unsupervised learning to design HBF for narrowband scenarios. In unsupervised learning, BS can directly use the estimated channel data for training without the need to obtain an optimal solution, thus significantly reducing training time and cost. The experimental results verify that this

method can perform better than traditional methods in massive MIMO systems with multiple RF chains. It can also manage hardware limitations and imperfect CSI challenges in mm-Wave systems. In the future, the proposed method can be extended to broadband scenarios and more complex beamforming problems.

References

[1]   S. Bellofiore, C. A. Balanis, J. Foutz, and A. S. Spanias, "Smart-antenna systems for mobile communication networks. Part 1. Overview and antenna design," *IEEE Antennas Propag. Mag.*, vol. 44, no. 3, pp. 145–154, Jun. 2002, doi: 10.1109/MAP.2002.1039395.

[2]   N. Ojaroudiparchin, M. Shen, S. Zhang, and G. F. Pedersen, "A Switchable 3-D-Coverage-Phased Array Antenna Package for 5G Mobile Terminals," *IEEE Antennas Wirel. Propag. Lett.*, vol. 15, pp. 1747–1750, 2016, doi: 10.1109/LAWP.2016.2532607.

[3]   J. Bang, Y. Hong, and J. Choi, "MM-wave phased array antenna for whole-metal-covered 5G mobile phone applications," in *2017 International Symposium on Antennas and Propagation (ISAP)*, 2017, pp. 1–2. doi: 10.1109/ISANP.2017.8228816.

[4]   M. Stangl, R. Werninghaus, and R. Zahn, "The TERRASAR-X active phased array antenna," in *IEEE International Symposium on Phased Array Systems and Technology, 2003.*, 2003, pp. 70–75. doi: 10.1109/PAST.2003.1256959.

[5]   *Phased Array Antenna Handbook, Third Edition*. Accessed: Apr. 24, 2022. [Online].   Available: https://books.google.com/books/about/Phased_Array_Antenna_Handbook_Thir d_Edit.html?hl=zh-CN&id=EKN5DwAAQBAJ

[6]   X. Gao, L. Dai, S. Han, C.-L. I, and X. Wang, "Reliable Beamspace Channel Estimation for Millimeter-Wave Massive MIMO Systems with Lens Antenna Array," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 9, pp. 6010–6021, Sep. 2017, doi: 10.1109/TWC.2017.2718502.

[7]   R. Zhang, J. Zhou, J. Lan, B. Yang, and Z. Yu, "A High-Precision Hybrid Analog and Digital Beamforming Transceiver System for 5G Millimeter-Wave Communication," *IEEE Access*, vol. 7, pp. 83012–83023, 2019, doi: 10.1109/ACCESS.2019.2923836.

[8] X. Xin, Y. Qinghong, S. Zhaolin, L. Qingjiang, W. Yinan, and L. Wenli, "The Realization of Digital Beamforming Based on FPGA and DSP," in *2010 International Conference on Intelligent System Design and Engineering Application*, 2010, vol. 2, pp. 713–716. doi: 10.1109/ISDEA.2010.418.

[9] L. Pettersson, M. Danestig, and U. Sjostrom, "An experimental S-band digital beamforming antenna," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 12, no. 11, pp. 19–29, 1997, doi: 10.1109/62.636800.

[10] U. R. O. Nickel, "Properties of Digital Beamforming with Subarrays," in *2006 CIE International Conference on Radar*, 2006, pp. 1–5. doi: 10.1109/ICR.2006.343380.

[11] V. Venkateswaran and A.-J. van der Veen, "Analog Beamforming in MIMO Communications With Phase Shift Networks and Online Channel Estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4131–4143, 2010, doi: 10.1109/TSP.2010.2048321.

[12] F. Sohrabi and W. Yu, "Hybrid Digital and Analog Beamforming Design for Large-Scale Antenna Arrays," *IEEE J. Sel. Top. Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016, doi: 10.1109/JSTSP.2016.2520912.

[13] X. Zhao, T. Lin, Y. Zhu, and J. Zhang, "Partially-Connected Hybrid Beamforming for Spectral Efficiency Maximization via a Weighted MMSE Equivalence," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 12, pp. 8218–8232, 2021, doi: 10.1109/TWC.2021.3091524.

[14] W. Roh *et al.*, "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, doi: 10.1109/MCOM.2014.6736750.

[15] T. Lin and Y. Zhu, "Beamforming Design for Large-Scale Antenna Arrays Using Deep Learning," *IEEE Wirel. Commun. Lett.*, vol. 9, no. 1, pp. 103–107, 2020, doi: 10.1109/LWC.2019.2943466.

[16] A. Morsali and B. Champagne, "Robust Hybrid Analog/Digital Beamforming for Uplink Massive-MIMO with Imperfect CSI," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2019, pp. 1–6. doi: 10.1109/WCNC.2019.8886103.

[17] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, Jul. 2015, doi: 10.1126/science.aaa8415.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, Art. no. 7553, May 2015, doi: 10.1038/nature14539.

[19] S. Dörner, S. Cammerer, J. Hoydis, and S. ten Brink, "Deep Learning Based Communication Over the Air," *IEEE J. Sel. Top. Signal Process.*, vol. 12, no. 1, pp. 132–143, 2018, doi: 10.1109/JSTSP.2017.2784180.

[20] H. Ye, G. Y. Li, and B.-H. Juang, "Power of Deep Learning for Channel Estimation and Signal Detection in OFDM Systems," *IEEE Wirel. Commun. Lett.*, vol. 7, no. 1, pp. 114–117, 2018, doi: 10.1109/LWC.2017.2757490.

[21] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep Learning in Physical Layer Communications," *IEEE Wirel. Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019, doi: 10.1109/MWC.2019.1800601.

[22] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep Learning for Massive MIMO CSI Feedback," *IEEE Wirel. Commun. Lett.*, vol. 7, no. 5, pp. 748–751, 2018, doi: 10.1109/LWC.2018.2818160.

[23] X. Gao, S. Jin, C.-K. Wen, and G. Y. Li, "ComNet: Combination of Deep Learning and Expert Knowledge in OFDM Receivers," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2627–2630, 2018, doi: 10.1109/LCOMM.2018.2877965.

[24] O. Aldayel, V. Monga, and M. Rangaswamy, "Tractable Transmit MIMO Beampattern Design Under a Constant Modulus Constraint," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 2588–2599, 2017, doi: 10.1109/TSP.2017.2664040.

[25] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially Sparse Precoding in Millimeter Wave MIMO Systems," *IEEE Trans. Wirel. Commun.*,

vol. 13, no. 3, pp. 1499–1513, Mar. 2014, doi: 10.1109/TWC.2014.011714.130846.

[26] T. Lin, J. Cong, Y. Zhu, J. Zhang, and K. Ben Letaief, "Hybrid Beamforming for Millimeter Wave Systems Using the MMSE Criterion," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3693–3708, 2019, doi: 10.1109/TCOMM.2019.2893632.

[27] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating Minimization Algorithms for Hybrid Precoding in Millimeter Wave MIMO Systems," *IEEE J. Sel. Top. Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016, doi: 10.1109/JSTSP.2016.2523903.

[28] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless Networks Design in the Era of Deep Learning: Model-Based, AI-Based, or Both?," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, 2019, doi: 10.1109/TCOMM.2019.2924010.

[29] L. Dai, R. Jiao, F. Adachi, H. V. Poor, and L. Hanzo, "Deep Learning for Wireless Communications: An Emerging Interdisciplinary Paradigm," *IEEE Wirel. Commun.*, vol. 27, no. 4, pp. 133–139, Aug. 2020, doi: 10.1109/MWC.001.1900491.

[30] "Machine learning for wireless communications in the Internet of Things: A comprehensive survey," *Ad Hoc Netw.*, vol. 93, p. 101913, Oct. 2019, doi: 10.1016/j.adhoc.2019.101913.

[31] R. Li *et al.*, "Intelligent 5G: When Cellular Networks Meet Artificial Intelligence," *IEEE Wirel. Commun.*, vol. 24, no. 5, pp. 175–183, Oct. 2017, doi: 10.1109/MWC.2017.1600304WC.

[32] 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, doi: 10.1109/ACCESS.2018.2850226.

[33] H. Huang, Y. Peng, J. Yang, W. Xia, and G. Gui, "Fast Beamforming Design via Deep Learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1065–1069, 2020, doi: 10.1109/TVT.2019.2949122.

[34] L. Sung and D.-H. Cho, "Multi-User Hybrid Beamforming System Based on Deep Neural Network in Millimeter-Wave Communication," *IEEE Access*, vol. 8, pp. 91616–91623, 2020, doi: 10.1109/ACCESS.2020.2990317.

[35] A. M. Elbir and K. V. Mishra, "Deep Learning Design for Joint Antenna Selection and Hybrid Beamforming in Massive MIMO," in *2019 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting*, Jul. 2019, pp. 1585–1586. doi: 10.1109/APUSNCURSINRSM.2019.8888753.

[36] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-Learning-Based Millimeter-Wave Massive MIMO for Hybrid Precoding," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3027–3032, Mar. 2019, doi: 10.1109/TVT.2019.2893928.

[37] J. Chen *et al.*, "Hybrid Beamforming/Combining for Millimeter Wave MIMO: A Machine Learning Approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11353–11368, 2020, doi: 10.1109/TVT.2020.3009746.

[38] Y. Long, Z. Chen, J. Fang, and C. Tellambura, "Data-Driven-Based Analog Beam Selection for Hybrid Beamforming Under mm-Wave Channels," *IEEE J. Sel. Top. Signal Process.*, vol. 12, no. 2, pp. 340–352, 2018, doi: 10.1109/JSTSP.2018.2818649.

[39] S. Salari, M. Z. Amirani, I.-M. Kim, D. I. Kim, and J. Yang, "Distributed Beamforming in Two-Way Relay Networks With Interference and Imperfect CSI," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 6, pp. 4455–4469, Jun. 2016, doi: 10.1109/TWC.2016.2542071.

[40] J. Wang and M. Bengtsson, "Joint Optimization of the Worst-Case Robust MMSE MIMO Transceiver," *IEEE Signal Process. Lett.*, vol. 18, no. 5, pp. 295–298, 2011, doi: 10.1109/LSP.2011.2123092.

[41] M. B. Shenouda and T. N. Davidson, "On the Design of Linear Transceivers for Multiuser Systems with Channel Uncertainty," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 6, pp. 1015–1024, Aug. 2008, doi: 10.1109/JSAC.2008.080817.

[42] N. Vucic, H. Boche, and S. Shi, "Robust Transceiver Optimization in Downlink Multiuser MIMO Systems," *IEEE Trans. Signal Process.*, vol. 57, no. 9, pp. 3576–3587, Sep. 2009, doi: 10.1109/TSP.2009.2020030.

[43] 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, doi: 10.1109/JPROC.2014.2299397.

[44] M. Xiao *et al.*, "Millimeter Wave Communications for Future Mobile Networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017, doi: 10.1109/JSAC.2017.2719924.

[45] P. Chen *et al.*, "A Multibeam Antenna Based on Substrate Integrated Waveguide Technology for MIMO Wireless Communications," *IEEE Trans. Antennas Propag.*, vol. 57, no. 6, pp. 1813–1821, Jun. 2009, doi: 10.1109/TAP.2009.2019868.

[46] Y. J. Cheng *et al.*, "Substrate Integrated Waveguide (SIW) Rotman Lens and Its Ka-Band Multibeam Array Antenna Applications," *IEEE Trans. Antennas Propag.*, vol. 56, no. 8, pp. 2504–2513, 2008, doi: 10.1109/TAP.2008.927567.

[47] M. Jiang, Z. N. Chen, Y. Zhang, W. Hong, and X. Xuan, "Metamaterial-Based Thin Planar Lens Antenna for Spatial Beamforming and Multibeam Massive MIMO," *IEEE Trans. Antennas Propag.*, vol. 65, no. 2, pp. 464–472, 2017, doi: 10.1109/TAP.2016.2631589.

[48] J. Ala-Laurinaho *et al.*, "2-D Beam-Steerable Integrated Lens Antenna System for 5G E -Band Access and Backhaul," *IEEE Trans. Microw. Theory Tech.*, vol. 64, no. 7, pp. 2244–2255, Jul. 2016, doi: 10.1109/TMTT.2016.2574317.

[49] J. Brady, J. Hogan, and A. Sayeed, "Multi-Beam MIMO Prototype for Real-Time Multiuser Communication at 28 GHz," in *2016 IEEE Globecom Workshops (GC Wkshps)*, 2016, pp. 1–6. doi: 10.1109/GLOCOMW.2016.7848964.

[50] Y. Kim *et al.*, "Feasibility of Mobile Cellular Communications at Millimeter Wave Frequency," *IEEE J. Sel. Top. Signal Process.*, vol. 10, no. 3, pp. 589–599, Apr. 2016, doi: 10.1109/JSTSP.2016.2520901.

[51] S. Sun, T. S. Rappaport, R. W. Heath, A. Nix, and S. Rangan, "Mimo for millimeter-wave wireless communications: beamforming, spatial multiplexing, or both?," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 110–121, Dec. 2014, doi: 10.1109/MCOM.2014.6979962.

[52] A. Alkhateeb, R. W. Heath, and G. Leus, "Achievable rates of multi-user millimeter wave systems with hybrid precoding," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, Jun. 2015, pp. 1232–1237. doi: 10.1109/ICCW.2015.7247346.

[53] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited Feedback Hybrid Precoding for Multi-User Millimeter Wave Systems," *IEEE Trans. Wirel. Commun.*, vol. 14, no. 11, pp. 6481–6494, 2015, doi: 10.1109/TWC.2015.2455980.

[54] T. E. Bogale and L. B. Le, "Beamforming for multiuser massive MIMO systems: Digital versus hybrid analog-digital," in *2014 IEEE Global Communications Conference*, 2014, pp. 4066–4071. doi: 10.1109/GLOCOM.2014.7037444.

[55] R. Méndez-Rial, N. González-Prelcic, and R. W. Heath, "Adaptive hybrid precoding and combining in MmWave multiuser MIMO systems based on compressed covariance estimation," in *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2015, pp. 213–216. doi: 10.1109/CAMSAP.2015.7383774.

[56] T. E. Bogale, L. B. Le, A. Haghighat, and L. Vandendorpe, "On the Number of RF Chains and Phase Shifters, and Scheduling Design With Hybrid Analog–Digital Beamforming," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 5, pp. 3311–3326, 2016, doi: 10.1109/TWC.2016.2519883.

[57] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems," *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 5, pp. 831–846, 2014, doi: 10.1109/JSTSP.2014.2334278.

Appendix

List of Publication and Proceeding

P. Zhang, L. Pan, T. Laohapensaeng and M. Chongcheawchamnan, "Hybrid Beamforming based on an Unsupervised Deep Learning Network for Downlink Channels with Imperfect CSI," in IEEE Wireless Communications Letters, doi: 10.1109/LWC.2022.3179362.

*

Hybrid Beamforming based on an Unsupervised Deep Learning Network for Downlink Channels with Imperfect CSI

Peng Zhang, Liangrui Pan, Teeravisit Laohapensaeng, and Mitchai Chongcheawchamnan*

Abstract—Hybrid beamforming can provide rapid data transmission rates while reducing the complexity and cost of massive multiple-input multiple-output (MIMO) systems. However, channel state information (CSI) is imperfect in realistic downlink channels, introducing challenges to hybrid beamforming (HBF) design. This paper proposes an unsupervised deep learning neural network (USDNN) for hybrid beamforming to prevent the labeling overhead of supervised learning and improve the achievable sum rate based on imperfect CSI. The simulation results show that our proposed method is 74% better than MO and 120% better than orthogonal match pursuit (OMP) systems; our proposed USDNN can achieve near-optimal performance.

Index Terms—Deep learning, Hybrid beamforming, massive MIMO, unsupervised deep learning.

I. INTRODUCTION

With the rapid development of 5G millimeter-wave (mm-Wave) communications and broadband low-earth-orbit satellite communications, there has been an unprecedented new development of mm-Wave active phased array antennas. The signal processing component is a core in the active phased array system for beamforming tasks. Currently, there are three types of beamformers, analog beamformer (ABF), digital beamformer (DBF), and hybrid beamformer (HBF). An HBF, which includes both analog and digital beamforming (BF) design, has received widespread attention because they provide high beamforming gains to compensate for severe path loss with affordable and low-power consumption hardware [1]. An HBF uses a small number of ABFs to drive multiple antennas to form beams. Each beam is connected to a digital precoder through a single radio frequency (RF) chain. This hybrid combination of phase-based analog and baseband digital BFs reduces the number of transmission chains while maintaining the data reachability rate at an acceptable level. However, channel state information (CSI) [2] is essential and practical data describing wireless communication channels. In wireless communication, CSI represents the propagation characteristics of the communication link, which describes the combined influence of multiple effects such as scattering, fading, and power attenuation in the channel. In a frequency division duplex (FDD) system, the downlink CSI is obtained by the user end through downlink pilot estimation and then transmitted back to the base station end through the feedback link. Therefore, CSI is usually imperfect in practical applications, making HBF design more challenging.

For the optimization of HBF, the most challenging aspect is that the constant modulus constraint of the ABF caused by the phase shifter makes the problem highly nonconvex and challenging to solve. In previous studies, a model-based design method was used to address this difficulty. An algorithm based on orthogonal match pursuit (OMP) was proposed in [3]. However, the ABF is limited to a predefined codebook. To improve the performance of OMP and apply the manifold optimization (MO) method to the optimization of simulated BFs, [4] proposed a hybrid beamforming algorithm based on MO to address the constant modulus constraint of the analog component directly. All of the above algorithms assume perfect CSI and do not apply to natural channels. For the processing of imperfect CSI, the design of an HBF for the uplink of a massive multiple-input multiple-output (MIMO) system under imperfect CSI is studied in [5], and the norm-bounded channel error model is used to capture the characteristics of imperfect CSI in the actual system. In [3], in the presence of interference and imperfect CSI, a two-way relay network's optimal beamforming and power distribution problems based on Amplify and Forward were designed for an HBF. In [6], a DL-based HBF design method was proposed, and a beamforming neural network was developed and proved robust under imperfect CSI conditions. However, this method requires a large number of serial iterations to obtain a solution.

The latest research on intelligent communication shows that the deep learning (DL) technique has great potential for solving problems that have traditionally presented challenges. The DL technique has proven to be an excellent tool for solving complex nonconvex optimization problems [7]. Its unique network structure can approximate any function under specific conditions and use DL to solve the optimization problem of BF design [8], which can eliminate the complexity caused by too many iterations and realize real-time calculations. Although the training process is time-consuming, the training is carried out in an offline mode. Therefore, DL technology is an effective method to reduce cellular network delay. In [9], a two-step training method was proposed based on supervised pretraining and unsupervised retraining, and a beamforming predictive network was trained in the offline training phase. In [10], a technique based on supervised learning was proposed to train a

Corresponding author
P. Zhang and M. Chongcheawchamnan are with Faculty of Engineering, Prince of Songkla University, Hat Yai, Songkhla, 90110 Thailand (e-mail: zpeng0213@gmail.com; mitchai.c@psu.ac.th).
L. Pan is with College of Computer Science and Electronic Engineering, Hunan University, Chang Sha, China (e-mail: lip141772@gmail.com).

T. Laohapensaeng is with School of Information Technology, Mae Fah Luang University, Chiang Rai, Thailand (e-mail: teeravisit.lao@mfu.ac.th).
Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org

deep neural network (DNN) in the offline training phase. In these studies, supervised learning required many training labels, so additional computing resources were needed to determine these labels using traditional optimization methods. For BF designs, we had a hard time finding the proper labels. If we use the optimized output based on the traditional algorithm as the label, the neural network can only be trained to approximate the traditional algorithm, but not better than the traditional algorithm.

To prevent the overhead tagging task mentioned above, we intend to design a BF based on unsupervised learning. Therefore, we have developed an unsupervised deep neural network (USDNN) for the BF design problem and training to optimize a BF to maximize the achievable sum rate. The simulation results show that, compared with the traditional BF algorithm, the method based on unsupervised learning can avoid the labeling overhead of labels and obtain better performance than the traditional algorithm.

The rest of this article is organized as follows. First, section II introduces the massive MIMO system and channel model. Then, in section III, the neural network model based on unsupervised learning is explained. Next, the simulation results are shown in section IV to evaluate the method's performance and compare it with existing methods. Finally, the conclusions are given in section V.

Symbols: Scalars, vectors, and matrices are represented by italic letters, boldface lowercase letters, and boldface uppercase letters, respectively. The superscript $*$ and H subscripts indicate complex conjugates and Hermitian transposes, respectively. $\mathbb{C}^{x \times y}$ represents an $x \times y$ complex vector or matrix space. I denotes the identity matrix with appropriate dimensions. $|.|$ is the absolute value, $\| . \|$ represents the Euclidean norm, and $\log(.)$ and $\mathbb{E}[.]$ represent the logarithm and expectation operator, respectively. Thus, $\mathcal{CN}(0, R)$ represents a zero-mean complex Gaussian distribution with covariance R.

II. SYSTEM MODEL

We consider a massive MIMO base station (BS) system model for narrowband scenarios. The BS uses a half-wavelength spacing uniform linear array (ULA), and the user terminal is equipped with a single antenna. In addition, it is assumed that all users have the same priority. The number of RF chains is twice the total number of data streams. The HBF structure can accurately achieve the same performance as any fully DBF regardless of the number of antennas [11]. Therefore, we set the system to include N_t antennas and N_{RF} RF chains to transmit N_s data streams to M single-antenna users, and each user receives d data streams. Let $N_s \leq N_{RF} \leq N_t$, where $N_s = Md$. As shown in Figure. 1, the BS uses an HBF precoder. We consider a fully connected structure in which each RF chain is coupled to all antennas on the BS through 2 phase shifters. In the HBF structure, the BS first uses the $N_{RF} \times N_s$ digital precoder V_D to digitally modify the baseband data stream. It then transfers the processed signal up to the carrier frequency through N_{RF} RF chains. After that, the BS uses the $N_t \times N_{RF}$ RF precoder V_{RF}, which is implemented using an analog phase shifter. Therefore, the sending signal is:

$$\mathbf{x} = \mathbf{V}_{RF}\mathbf{V}_D\mathbf{s} = \sum_m^M \mathbf{V}_{RF}\mathbf{V}_{D_m}\mathbf{s}_m \qquad (1)$$

where $\mathbf{s} \in \mathbb{C}^{N_s \times 1}$ is a vector that represents the data symbols. It is a cascade of data flow vectors for each user. In addition, assume that $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \mathbf{I}_{N_s}$. For the mth user, the received signal can be written as:

$$\mathbf{y}_m = \mathbf{h}_m^H\mathbf{V}_{RF}\mathbf{V}_{D_m}\mathbf{s}_m + \mathbf{h}_m^H \sum_{\ell \neq m} \mathbf{V}_{RF}\mathbf{V}_{D_\ell}\mathbf{s}_\ell + \mathbf{z}_m \qquad (2)$$

where $\mathbf{h}_m \in \mathbb{C}^{N_t \times 1}$ represents the complex channel gain vector from the transmit antennas of the BS to the mth user, and $\mathbf{z}_m \sim \mathcal{CN}(0, \sigma^2)$ represents additive white Gaussian noise.
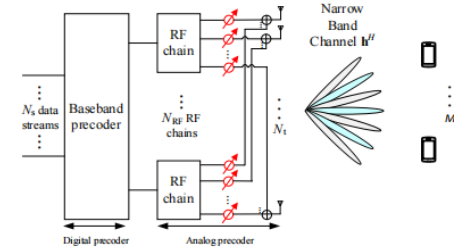

Fig. 1. System structure of mm-Wave massive MIMO HBF.

We considered the Saleh-Valenzuela (SV) channel model [12], where the channel contains one line-of-sight (LoS) path and L_m non-LoS (NLoS) paths. For the SV channel model, the channel vector $\mathbf{h}_m^H$ of the mth user can be expressed as

$$\mathbf{h}_m^H = \sqrt{\frac{N_t}{L_m+1}} \sum_{l=0}^{L_m} \beta_m^{(l)} \boldsymbol{a}^H\left(\phi_m^{(l)}\right) \qquad (3)$$

where $\beta_m^{(0)} \boldsymbol{a}^H(\phi_m^{(0)})$ is the LoS path of $\mathbf{h}_m$ with $\beta_m^{(0)}$ presenting the complex gain and $\phi_m^{(0)}$ denoting the spatial direction, $\beta_m^{(l)} \boldsymbol{a}^H(\phi_m^{(l)})$ for $1 \leq l \leq L_m$ is the lth NLoS path of $\mathbf{h}_m$, and $\mathbf{a}(\phi)$ is the $N_t \times 1$ array steering vector at the BS.

The signal-to-interference-noise ratio (SINR) of the signal received by the mth user is expressed as:

$$SINR(\mathbf{V}_{RF}, \mathbf{v}_{D_m}) = \frac{|\mathbf{h}_m^H \mathbf{v}_{RF}\mathbf{v}_{D_m}|^2}{\sigma^2 + \sum_{\ell \neq m} |\mathbf{h}_m^H \mathbf{V}_{RF}\mathbf{v}_{D_m}|^2} \qquad (4)$$

In this article, to compare performance, we calculated the achievable sum-rate, expressed as:

$$R = \sum_{m=1}^M log_2\left(1 + SINR(\mathbf{V}_{RF}, \mathbf{v}_{D_m})\right) \qquad (5)$$

The optimization problem of $\mathbf{V}_{RF}$ and $\mathbf{V}_D$ is expressed as follows:

$$\max_{\mathbf{V}_{RF}, \mathbf{V}_D} R$$
$$s.t. \quad |[\mathbf{V}_{RF}]_{i,j}| = 1, \quad \forall i, j \qquad (6)$$
$$\|\mathbf{V}_{RF}\mathbf{V}_D\|_F \leq 1,$$

The first expression is the constant modulus constraint of the phase shifter for ABF, and the second expression is the total transmission power constraint with normalized power.

III. DEEP NEURAL NETWORK HYBRID BEAMFORMING BASED ON UNSUPERVISED LEARNING

The performance of suboptimal solutions obtained by traditional optimization-based methods cannot be guaranteed, and they usually have high complexity. Therefore, this section introduces the DNN-based neural network architecture and unsupervised training.
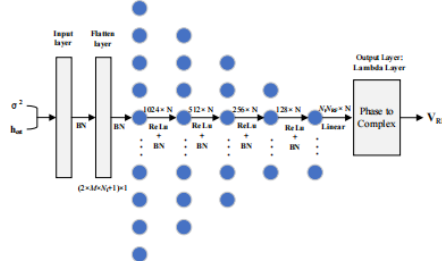

Fig. 2. The proposed USDNN architecture.

A. DNN Architecture

As shown in Figure. 2, the network consists of five dense layers, four batch normalization (BN) layers, and five activation functions. The five dense layers have 1024, 512, 256, 128, and N_tN_{RF} neurons, and empirical experiments determine the number of neurons to ensure sufficient learning ability. The BN layer enhances convergence and is used to stabilize the training process. The first four dense layers use the corrected linear unit (ReLu) for the activation function. Finally, the last dense layer uses the linear activation function to output the phase shift prediction.

B. Training and testing

Training stage: As shown in Figure. 3, we simulate the SV channel model to obtain the perfect CSI to calculate the loss function and update the calculated weight information to the USDNN model. Next, we apply the proposed classic mm-Wave channel estimator [13]. The BS sends pilot symbols with beamformers and receives user feedback signals y_p to estimate the channel, resulting in channel estimate $\mathbf{h}_{est}$ (imperfect CSI) and noise power σ^2 as the input to the USDNN model. By using the loss function calculated to update the model's weights, the model then learns how to optimize $\mathbf{V}_{RF}$.

Testing stage: All model parameters have been trained and fixed. The imperfect CSI and σ^2 are used as input for testing, and the optimized beamformer can be obtained, which directly outputs the optimized beamforming vector $\mathbf{V}_{RF}$.
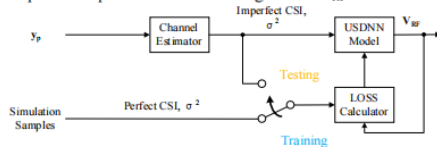

Fig. 3. Two-stage design approach: training and testing.

In the previous work [9], the received user feedback signal y_p is used as input, and y_p itself is a function of $\mathbf{V}_{RF}$, which needs to be optimized. So we estimate the channel based on the y_p and

obtain the $\mathbf{h}_{est}$ and σ^2 through the classical channel estimation algorithm in [13] to form the input of the USDNN. The channel vector and the beamforming vector $\mathbf{V}_{RF}$ are both complex-valued vectors. Therefore, we consider the input layer and output layer as follows:

Input layer: For the real and imaginary parts and noise power σ^2 of each antenna's downlink channel CSI, the input data shape is a matrix of $2 \times M \times N_t + 1$ that is input through the flattened layer.

Output layer: Inspired by [6], the following lambda layer is used for the complex-valued output of DNN. This layer converts the predicted real-valued phase into a complex-valued form $\mathbf{V}_{RF}$, where $\mathbf{P}_{pred}$ is the predicted real-valued phase.

$$\mathbf{V}_{RF} = e^{j\mathbf{P}_{pred}} = \cos(\mathbf{P}_{pred}) + j \cdot \sin(\mathbf{P}_{pred}) \qquad (7)$$

We use the same DNN architecture for the digital precoder $\mathbf{V}_D$ to set the number of neurons in the last dense layer to $2N_sN_{RF}$ and output the $N_s \times N_{RF}$ complex vector $\mathbf{V}_D$.

To train the network, we use the Adam optimizer with an initial learning rate of 1×10^{-3}. The maximum number of epochs is set to 20,000 and tolerated early stop 20 is adopted to improve training efficiency. In addition, to accelerate the convergence speed, the learning rate is attenuated by a factor of 0.2 whenever the verification loss does not decrease for 20 consecutive periods. Table I shows the hyperparameters selected for DNN training.

TABLE I
HYPERPARAMETERS FOR DNN TRAINING

Parameter	Set Value
Epochs	20,000
Batch Size	256
Initial Learning Rate	1×10^{-3}
Minimum Learning Rate	5×10^{-5}
Reduce LR On Plateau (factor)	0.2
Reduce LP On Plateau (patience)	20

B. Unsupervised Training

A computationally heavy optimization problem in supervised learning must solve each sample in the data set and obtain data labels, which adds excessive difficulty to training. In unsupervised learning, our design does not require labels. It can be trained from the data obtained from the estimated channel, which can significantly reduce the computational complexity of training the network. In the training stage, the model learns under the perfect CSI as much as possible. In the testing stage, all model parameters have been trained and fixed, and imperfect CSI can be accepted as input for testing and directly output the optimized analog beamforming vector $\mathbf{V}_{RF}$. We train the network with the following new loss function directly related to the performance goal:

$$Loss = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}\log_2\left(1 + \frac{\left|\mathbf{h}_{m,n}^H\mathbf{V}_{RF,n}\mathbf{V}_{D_{m,n}}\right|^2}{\sigma_n^2 + \sum_{\ell\neq m}\left|\mathbf{h}_{m,n}^H\mathbf{V}_{RF,n}\mathbf{V}_{D_{m,n}}\right|^2}\right) \qquad (8)$$

where N represents the total number of training samples, and σ_n^2, $\mathbf{h}_{m,n}^H$, $\mathbf{V}_{RF,n}$ and $\mathbf{V}_{D_{m,n}}$ represent the noise power, CSI, output BF, and DBF associated with the nth sample, respectively. The minus sign indicates that the sum rate is maximized when the DNN is trained to minimize the loss

function. We generate 10^5, 2×10^4 and 2×10^4 samples for training, verification, and testing, respectively, while considering training efficiency, test performance, and stability. It is also worth emphasizing that labels are not required since unsupervised learning is used. Hence the cost of obtaining training samples is considerably reduced.

IV. SIMULATION RESULTS

We consider a system equipped with 64 antennas on the BS throughout the simulation process, and 8 RF chains transmit 4 data streams to 4 single-antenna users for communication. We considered the SV channel model with the same parameters as [12]. For the mth user, the parameters of the spatial channel are set as follows: 1) the number of paths: the number of LoS paths is 1, the number of NLoS (L_m) is 2, and the total number of paths is L = 3; 2) Path gain: $\beta_m^{(0)} \sim \mathcal{CN}(0, 1)$ and $\beta_m^{(l)} \sim \mathcal{CN}(0, 10^{-0.5})$ for l = 1, 2; 3) Spatial direction: $\phi_m^{(0)}$ and $\phi_m^{(l)}$ are uniformly distributed within [-0.5, 0.5]. Considering the comparison between MO and OMP algorithms, the signal-to-noise ratio (SNR) is defined as an indicator of the channel estimation level. Figure. 4 shows that our proposed USDNN has excellent performance when the CSI is imperfect. When L-train = L-test = 3, N_t = 64, N_{RF} = 8, M = 4, and N_s = 4, we can easily observe that the achievable rates of all methods increase with increasing SNR. Our proposed method is 74% better than MO and 120% better than OMP, and our proposed USDNN can achieve near-optimal performance. Notably, both MO and OMP require perfect CSI knowledge and do not apply to real channels.
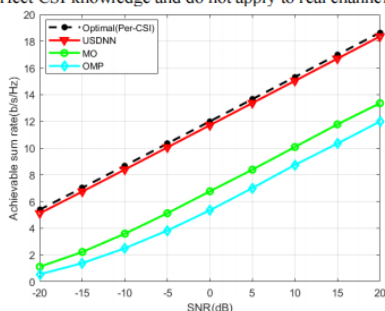

Fig. 4. The achievable sum rate under different SNRs (M = 4, N_t = 64, N_{RF} = 8, N_s = 4, L-train = L-test = 3).

A. Impact of Number of Users

This section evaluates how the system performance changes when the number of users $M \in \{2, 4, 6, 8\}$ and the number of antennas $N_t \in \{16, 32, 64, 128\}$ are changed. The SNR is set to 0 dB, and the number of channel paths is L-train = L-test = 3. When increasing the number of antennas or users of the BS, a more complex DNN should be required. However, Figures 5 and 6 show that the proposed architecture is sufficiently complex to have excellent performance with different antennas and users. As shown in Figure. 5, when the number of users increases, the rate of change of all methods decreases because increasing the number of users and the number of fixed antennas generates more inter-user interference. As a result, the

MO and OMP grow slowly, and the rate of our method still grows rapidly, always close to the optimal sum rate.
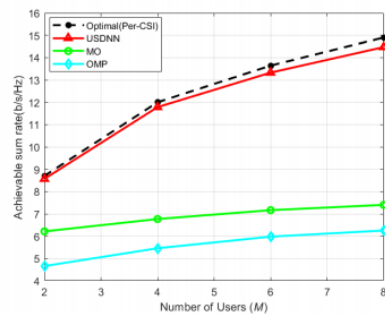

Fig. 5. The achievable sum rate with different numbers of users (N_t = 64, N_{RF} = 8, N_s = 8, d = 1, L-train = L-test = 3).

B. Impact of Number of Antennas

In Figure. 6, we evaluate our proposed method with different numbers of antennas. When the number of antennas is low, the performance of the MO and OMP methods is poor. When the number of antennas increases, the performance gap becomes more significant, and the rate of change decreases with the overhead signaling increases. Our proposed method is consistently close to the optimal sum rate.
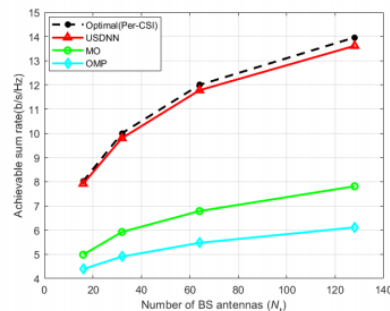

Fig. 6. The achievable sum rate with different numbers of antennas (M = 4, N_{RF} = 8, N_s = 4, L-train = L-test = 3).

C. Influence of Number of Channel Paths

There usually are inconsistencies between the values of the parameters during testing and training in practical applications, resulting in unacceptable deviations between the results. Therefore, the robustness of the model is an important consideration. Figure. 7 shows the impact of the mismatch between the number of channel paths in the offline training stage and online testing stage. It is assumed that the number of channel paths during offline training is L-train=3, and L-test=2, 3, 4, 5, and 6 are used for testing in online stage. The performance is best when the number of paths is consistent during training and testing. When the number of paths is inconsistent, there is a model mismatch, but the performance

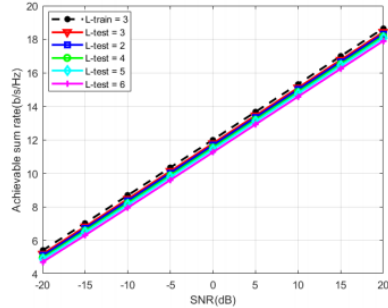loss is limited, indicating that our model has a certain degree of robustness.

Fig. 7. The achievable sum rate when the channel path numbers of the model do not match ($M = 4$, $N_t = 64$, $N_{RF} = 8$, $N_s = 4$).

D. Complexity Analysis

Comparing the complexity of neural networks and traditional algorithms is a difficult task. For traditional model-based HBF designs [3], [4], the asymptotic computational complexity in complex multipliers is $O(N_t^3)$ because they involve operations such as singular value decomposition and matrix inversion. We analyze the computational complexity from the number of floating-point operations (FLOPs) proposed USDNN. Because the training is done offline, only the complexity of the online testing phase is calculated. The FLOP number of the Dense layer is given by $(2N_I - 1)N_O$, where N_I and N_O denote the input and output dimensions, respectively. But considering the number of FLOPs alone may not be sufficient to draw a fair conclusion, as how these operations are performed and the architecture used can also have profound effects.

Therefore, in this subsection, we also make an approximate comparison of the running time of all methods, as shown in Table II. These methods are run on the same Intel(R) Core (TM) i7-10700F CPU @ 2.90GHz and GeForce RTX 2080Super hardware configuration for a fair comparison. Unlike traditional algorithms, the large-scale matrix multiplication of USDNN can be effectively accelerated through the parallel computing of graphics processing units (GPU). It can be seen that the proposed USDNN runs dozens of times faster than the traditional HBF algorithm. Note that the computational time for training is not considered here since the training process is done offline.

TABLE II COMPARISON OF RUNNING TIME (SECONDS) OF DIFFERENT METHODS

Algorithm Parameters	USDNN	MO	OMP
$N_t = 16$	0.046	1.657	1.845
$N_t = 32$	0.052	1.693	1.962
$N_t = 64$	0.071	1.732	2.133
$N_t = 128$	0.102	1.834	2.368

($M = 4$, $N_{RF} = 8$, $N_s = 4$, L-train = L-test = 3).

V. CONCLUSIONS AND FUTURE WORK

Hybrid beamforming is a fundamental technology for massive MIMO systems that can reduce the number of RF chains and improve the achievable sum rate of the system. In this letter, we propose a deep learning method based on unsupervised learning to design hybrid beamforming for narrowband scenarios. In unsupervised learning, BS can directly use the estimated channel data for training without the need to obtain an optimal solution, thus significantly reducing training time and cost. The simulation results verify that this method can perform better than traditional methods in massive MIMO systems with multiple RF chains. It can also manage hardware limitations and imperfect CSI challenges in mm-Wave systems. In the future, the proposed method can be extended to broadband scenarios and more complex beamforming problems.

REFERENCES

[1] W. Roh et al., "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, doi: 10.1109/MCOM.2014.6736750.

[2] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, "Channel State Information Prediction for 5G Wireless Communications: A Deep Learning Approach," IEEE Trans. Netw. Sci. Eng., vol. 7, no. 1, pp. 227–236, Jan. 2020, doi: 10.1109/TNSE.2018.2848960.

[3] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially Sparse Precoding in Millimeter Wave MIMO Systems," IEEE Trans. Wirel. Commun., vol. 13, no. 3, pp. 1499–1513, Mar. 2014, doi: 10.1109/TWC.2014.011714.130846.

[4] T. Lin, J. Cong, Y. Zhu, J. Zhang, and K. B. Letaief, "Hybrid Beamforming for Millimeter Wave Systems Using the MMSE Criterion," IEEE Trans. Commun., vol. 67, no. 5, pp. 3693–3708, May 2019, doi: 10.1109/TCOMM.2019.2893632.

[5] A. Morsali and B. Champagne, "Robust Hybrid Analog/Digital Beamforming for Uplink Massive-MIMO with Imperfect CSI," in 2019 IEEE Wireless Communications and Networking Conference (WCNC), Apr. 2019, pp. 1–6. doi: 10.1109/WCNC.2019.8886103.

[6] T. Lin and Y. Zhu, "Beamforming Design for Large-Scale Antenna Arrays Using Deep Learning," IEEE Wirel. Commun. Lett., vol. 9, no. 1, pp. 103–107, Jan. 2020, doi: 10.1109/LWC.2019.2943466.

[7] F. Liang, C. Shen, W. Yu, and F. Wu, "Towards Optimal Power Control via Ensembling Deep Neural Networks," IEEE Trans. Commun., vol. 68, no. 3, pp. 1760–1776, Mar. 2020, doi: 10.1109/TCOMM.2019.2957482.

[8] 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, doi: 10.1109/ACCESS.2018.2850226.

[9] H. Huang, Y. Peng, J. Yang, W. Xia, and G. Gui, "Fast Beamforming Design via Deep Learning," IEEE Trans. Veh. Technol., vol. 69, no. 1, pp. 1065–1069, Jan. 2020, doi: 10.1109/TVT.2019.2949122.

[10] C. Huang, G. C. Alexandropoulos, C. Yuen, and M. Debbah, "Indoor Signal Focusing with Deep Learning Designed Reconfigurable Intelligent Surfaces," in 2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Jul. 2019, pp. 1–5. doi: 10.1109/SPAWC.2019.8815412.

[11] F. Sohrabi and W. Yu, "Hybrid Digital and Analog Beamforming Design for Large-Scale Antenna Arrays," IEEE J. Sel. Top. Signal Process., vol. 10, no. 3, pp. 501–513, Apr. 2016, doi: 10.1109/JSTSP.2016.2520912.

[12] X. Gao, L. Dai, S. Han, C.-L. I, and X. Wang, "Reliable Beamspace Channel Estimation for Millimeter-Wave Massive MIMO Systems with Lens Antenna Array," IEEE Trans. Wirel. Commun., vol. 16, no. 9, pp. 6010–6021, Sep. 2017, doi: 10.1109/TWC.2017.2718502.

[13] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems," IEEE J. Sel. Top. Signal Process., vol. 8, no. 5, pp. 831–846, Oct. 2014, doi: 10.1109/JSTSP.2014.2334278.

VITAE

Name   Mr. Peng Zhang

Student ID   6310120035

Educational Attainment

Degree	Name of Institution	Year of Graduation
Bachelor of Engineering	Anhui Polytechnic University	2018
Electrical Engineering		

Scholarship Awards during Enrolment

Intelligent Automation Research Center From Engineering Faculty Scholarship

List of Publication and Proceeding

P. Zhang, L. Pan, T. Laohapensaeng and M. Chongcheawchamnan, "Hybrid Beamforming based on an Unsupervised Deep Learning Network for Downlink Channels with Imperfect CSI," in IEEE Wireless Communications Letters, doi: 10.1109/LWC.2022.3179362.