IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 14, 2017, accepted July 3, 2017, date of publication July 7, 2017, date of current version August 29, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2724037

Channel Estimation and Hybrid Precoding for Millimeter-Wave MIMO Systems: A Low-Complexity Overall Solution

ZHENYU XIAO[1,2,3], (Senior Member, IEEE), PENGFEI XIA[4], (Senior Member, IEEE), AND XIANG-GEN XIA[5,6], (Fellow, IEEE)

[1]School of Electronic and Information Engineering, Beihang University, Beijing 100191, China
[2]Beijing Key Laboratory for Network-Based Cooperative Air Traffic Management and Beijing Laboratory for General Aviation Technology, Beijing 100191, China
[3]Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China
[4]Key Laboratory of Embedded System and Service Computing, School of Electronics and Information Engineering, Tongji University, Shanghai 200092, China
[5]College of Information Engineering, Shenzhen University, Shenzhen 518060, China
[6]Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716, USA

Corresponding author: Zhenyu Xiao (xiaozy@buaa.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61571025 and Grant 91538204 and in part by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61521091.

ABSTRACT To enable multi-stream transmission and increase the achievable rate, a hybrid digital/analog precoding structure is usually adopted in millimeter-wave (mmWave) MIMO systems. However, it may require matrix operations with a scale of antenna size, which is generally large in mmWave communications. Moreover, the channel estimation is also rather time-consuming due to the large number of antennas at both Tx/Rx sides. In this paper, a low-complexity overall channel estimation and hybrid precoding approach is proposed. In the channel estimation phase, a hierarchical multi-beam search scheme is proposed to fast acquire N_S (the number of streams) multipath components (MPCs)/clusters with the highest powers. In the hybrid precoding phase, the analog and digital precodings are decoupled. The analog precoding is designed to steer along the N_S acquired MPCs/clusters at both Tx/Rx sides, shaping an $N_S \times N_S$ baseband effective channel, while the digital precoding is performed in the baseband with the reduced-scale effective channel. Performance evaluations show that, compared with the state-of-the-art scheme, while achieving a close or even better performance when the number of radio frequency chains or streams is small, both the time complexity of the channel estimation and the computational complexity of the hybrid precoding are reduced.

INDEX TERMS Hybrid precoding, millimeter-wave, mmWave, mmWave MIMO, beam search, hierarchical search.

I. INTRODUCTION

Millimeter-wave (mmWave) communication is a promising technology for future wireless communications owing to its large vacant spectrum resource, which enables a much higher capacity than the current alternatives. It has raised increasing attention as an important candidate technology in both the next-generation wireless local area network (WLAN) [2]–[6] and mobile cellular communication [7]–[13]. In general, mmWave communication faces the problem of high propagation loss due to the high carrier frequency. Thus, mmWave devices usually need large antenna arrays to compensate for the propagation loss.

Different from the conventional multiple-input multiple-output (MIMO) systems, where a fully-digital beamforming structure is usually exploited thanks to a small number of antenna branches, in mmWave communications the number of antennas is large; the high power consumption of mixed signal components, as well as radio-frequency (RF) chains, makes it impractical to realize a full-blown digital beamforming. In such a case, analog beamforming is usually adopted for mmWave communications, where all the antennas share a single RF chain and generally have constant-amplitude (CA) constraint on their weights [14], [15]. As entry-wise estimation of channel status information (CSI) is time costly due to large arrays and subspace observations of the channel [16], the training approach is generally adopted, including the power iteration method by exploiting the directional feature of mmWave channel [17], and the switching beamforming which probes on a pre-defined codebook and finds the best codeword within the codebook [14], [15]. For switching beamforming, multiple-stage hierarchical search algorithms were proposed to reduce the number of measurements [4], [14]–[16], [18]. These schemes first probe with low-resolution codewords, i.e. codewords with larger beam widths, and then probe with high-resolution codewords, i.e. codewords with thinner beam widths. Although these analog beamforming schemes reduce search complexity, they can shape only one communication beam, which limits the achievable rate.

In order to enable multi-stream transmission and increase the achievable rate, a hybrid analog/digital precoding structure was then proposed[1] [8]–[13]. Similar to the analog beamforming structure, in a hybrid analog/digital precoding structure a large number of antennas share a few RF chains; thus it enables parallel transmission and provides the potential to boost the system capacity. However, the large antenna size challenges the need of a low-complexity design of the channel estimation and hybrid precoding. In particular, the hybrid precoding may require matrix operations with a scale of antenna size, which is generally large in mmWave communication. Moreover, the channel estimation is also rather time consuming due to the large number of antennas at both Tx/Rx sides. Most of the existing literatures study either channel estimation only [8], [19]–[23] or hybrid precoding only [24]–[28]

In [29], an overall approach was proposed for channel estimation and hybrid precoding in mmWave MIMO systems. In the channel estimation phase, a hierarchical codebook was designed by exploiting the hybrid structure, which is different from the former ones with only analog combining [14], [15]. Based on the codebook, a hierarchical multi-beam search method was proposed to acquire at least N_{S} (the number of streams) multipath components (MPCs)/clusters with the highest powers. With these MPCs, the channel matrix was reconstructed. In the precoding phase [24], [29], the optimal precoding matrix is first obtained based on the estimated channel without considering the CA constraint, and then analog and digital precoding matrices are determined by minimizing the Frobenius distance between the product of the analog and digital precoding matrices and the unconstraint optimal one. By exploiting the sparse feature on the angle domain, the optimization problem is modeled to be a sparse reconstruction problem, and is solved by the orthogonal matching pursuit (OMP) approach. Although this overall approach [29] is theoretically feasible, its satisfactory performance requires a large number of RF chains, which makes it less attractive for the devices with only a few number of RF chains. Moreover, it has a high time complexity in the channel estimation[2] and high computational complexity in the hybrid precoding.

In this paper, we propose a low-complexity overall channel estimation and hybrid precoding approach. The differences between this approach and the one proposed in [29] are:

- In the channel estimation phase, we propose a new hierarchical search scheme, which uses a particularly pre-designed analog hierarchical codebook to search multiple beams. The pre-designed analog hierarchical codebook has an over-sampling layer to guarantee an accurate estimation of the beam direction, and is robust to the number of RF chains and N_{S}. Moreover, the proposed search scheme exploits the particular channel structure in mmWave communication and the hierarchical feature of the pre-designed codebook, which greatly reduces the required time slots.
- In the hybrid precoding phase, the analog and digital precodings are decoupled. The analog precoding is designed to steer along the N_{S} acquired MPCs/clusters at both Tx/Rx sides, shaping an $N_{\mathrm{S}} \times N_{\mathrm{S}}$ baseband effective channel, while the digital precoding operates on the $N_{\mathrm{S}} \times N_{\mathrm{S}}$ effective channel, which greatly lowers the operation size of the matrices.

Performance evaluations show that, compared with the approach proposed in [29], the newly proposed approach achieves a close performance to the alternative, or even a better one when the number of RF chains or streams is small. Moreover, the time complexity of the channel estimation and the computational complexity of the hybrid precoding are greatly reduced. In particular, $N_{\mathrm{S}}^3 M^2$-proportional time slots are reduced to $N_{\mathrm{S}} M$-proportional time slots for the channel estimation , while antenna-size matrix operations are reduced to stream-size matrix operations for the hybrid precoding.

The rest of this paper is organized as follows. In Section II, we introduce the system and channel models, and formulate the problem. In Sections III, we propose the hierarchical multi-beam search method for channel estimation. In Section IV, we present the hybrid precoding operation. In Section V, we show the performance evaluations. Lastly, we conclude the paper in Section VI.

Notation: a, $\mathbf{a}$, $\mathbf{A}$, and $\mathcal{A}$ denote a scalar variable, a vector, a matrix, and a set, respectively. $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$ denote conjugate, transpose and conjugate transpose, respectively. In addition, $[x_1, x_2, \ldots, x_M]$ denotes a row vector with its elements being x_i. Some other operations used in this paper are defined as follows.

$\mathbb{E}\{\cdot\}$	Expectation operation.
$\lvert x\rvert$	Absolute value of scalar variable x.
$\lVert\mathbf{x}\rVert$	2-norm of vector $\mathbf{x}$.
$\lVert\mathbf{x}\rVert_0$	0-norm of vector $\mathbf{x}$.
$\lVert\mathbf{X}\rVert_{\mathrm{F}}$	Frobenius norm of $\mathbf{X}$.
$[\mathbf{X}]_{i,j}$	The ith-row and jth-column element of $\mathbf{X}$.
$[\mathbf{X}]_{:,i:j}$	The ith to jth columns of $\mathbf{X}$.

[1] Beamforming in the case of multi-stream transmission is called precoding here.

[2] The time complexity of channel estimation refers to the time slots spent in the channel estimation.

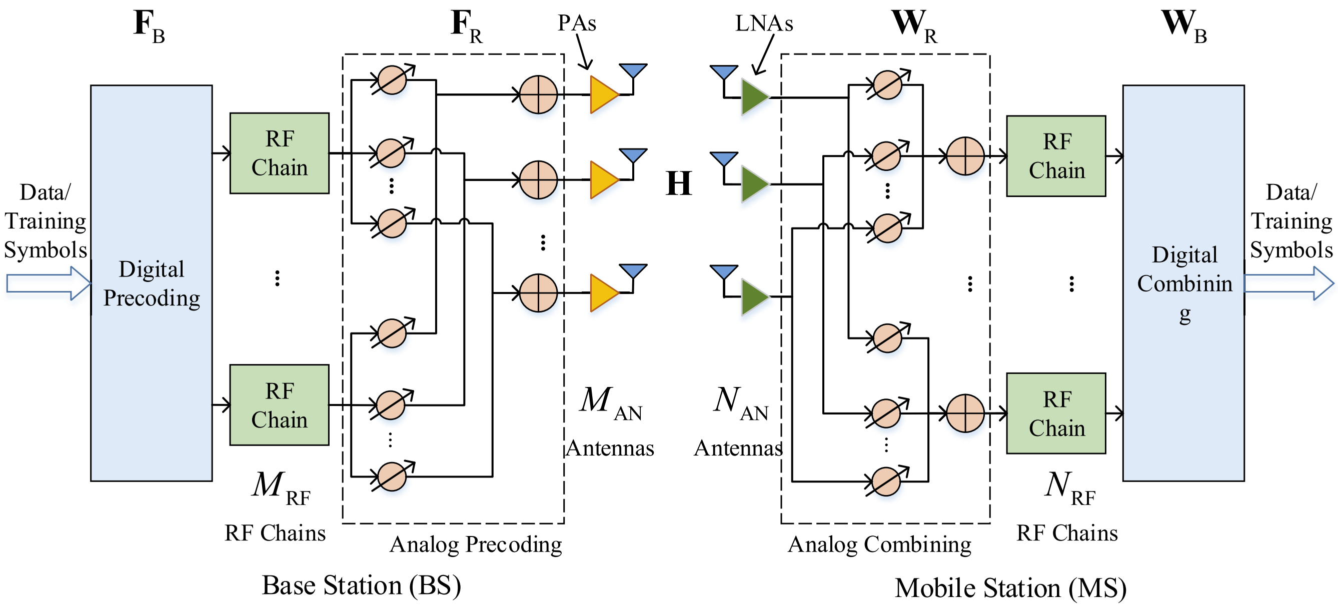

FIGURE 1. Illustration of the mmWave MIMO system with a hybrid analog/digital precoding and combing structure.

II. SYSTEM AND CHANNEL MODELS

A. SYSTEM MODEL

Without loss of generality, we consider a downlink point-to-point multiple-stream transmission from a base station (BS) to a mobile station (MS) in this paper, while the signal model and the proposed scheme are also applicable for an uplink transmission. An mmWave MIMO system with a hybrid digital/analog precoding structure is shown in Fig. 1, where relevant parameters are listed below.

- N_{S} Data streams transmitted from the BS to the MS.
- N_{R} The number of RF chains at the BS.
- N_{A} The number of antennas at the BS.
- M_{R} The number of RF chains at the MS.
- M_{A} The number of antennas at the MS.

Basically, we have $N_{\mathrm{R}} \leq N_{\mathrm{A}}$ and $M_{\mathrm{R}} \leq M_{\mathrm{A}}$, but in practical mmWave MIMO systems, N_{R} and M_{R} are far less than N_{A} and M_{A}, respectively. Moreover, it is noted that the supported number of data streams, i.e., N_{S}, is constrained by the number of RF chains, which means $N_{\mathrm{S}} \leq \min\{N_{\mathrm{R}}, M_{\mathrm{R}}\}$.

The BS performs digital precoding in the baseband and analog precoding in RF, respectively, while the MS performs analog combining in RF and digital combining in the baseband, respectively. Let $\mathbf{s}_{N_{\mathrm{S}}\times 1}$ denote the transmitted signal vector with normalized power, i.e., $\mathbb{E}(\mathbf{s}\mathbf{s}^{\mathrm{H}}) = \mathbf{I}_{N_{\mathrm{S}}}$, where $\mathbf{I}$ is an identity matrix. Considering a narrow-band block-fading propagation channel as in [24] and [29], the received signal vector at the MS writes

$$\mathbf{y} = \sqrt{P}\mathbf{W}_{\mathrm{B}}^{\mathrm{H}}\mathbf{W}_{\mathrm{R}}^{\mathrm{H}}\mathbf{H}\mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}}\mathbf{s} + \mathbf{W}_{\mathrm{B}}^{\mathrm{H}}\mathbf{W}_{\mathrm{R}}^{\mathrm{H}}\mathbf{n}, \tag{1}$$

where P is the transmission power per stream, $\mathbf{F}_{\mathrm{B}}$ and $\mathbf{F}_{\mathrm{R}}$ are the $N_{\mathrm{R}} \times N_{\mathrm{S}}$ digital and $N_{\mathrm{A}} \times N_{\mathrm{R}}$ analog precoding matrices at the BS, respectively, $\mathbf{W}_{\mathrm{B}}$ and $\mathbf{W}_{\mathrm{R}}$ are the $M_{\mathrm{R}} \times N_{\mathrm{S}}$ digital and $M_{\mathrm{A}} \times M_{\mathrm{R}}$ analog precoding matrices at the MS, respectively, $\mathbf{n}$ is a standard white Gaussian noise vector, i.e., $\mathbb{E}(\mathbf{n}\mathbf{n}^{\mathrm{H}}) = \mathbf{I}_{N_{\mathrm{S}}}$. In addition, we have the entry-wise CA constraint for the RF precoding matrix $\mathbf{F}_{\mathrm{R}}$ and the RF combining matrix $\mathbf{W}_{\mathrm{R}}$, respectively. In addition, we have power normalization for the precoding at the BS, i.e., $||\mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}}||_{\mathrm{F}}^2 = N_{\mathrm{S}}$.

B. CHANNEL MODEL

Since mmWave channels are expected to have limited scattering [12], [29]–[33], MPCs are mainly generated by reflection. Different MPCs have different physical transmit steering angles and receive steering angles, i.e., physical angles of departure (AoDs) and angles of arrival (AoAs). Consequently, mmWave channels are relevant to the geometry of antenna arrays. While the algorithms and results developed in this paper can be applied to arbitrary antenna arrays, we adopt uniform linear arrays (ULAs) with a half-wavelength antenna space in this paper. Consequently, the channel can be expressed as [16], [24], [29], [33]–[35]

$$\mathbf{H} = \sqrt{N_{\mathrm{A}}M_{\mathrm{A}}}\sum_{\ell=1}^{L}\lambda_{\ell}\mathbf{g}(M_{\mathrm{A}}, \Omega_{\ell})\mathbf{g}(N_{\mathrm{A}}, \psi_{\ell})^{\mathrm{H}}, \tag{2}$$

where λ_{ℓ} is the complex coefficient of the ℓth path, L is the number of MPCs and $L \geq N_{\mathrm{S}}$, $\mathbf{g}(\cdot)$ is the *steering vector function*, Ω_{ℓ} and ψ_{ℓ} are *cos AoD and AoA* of the ℓth path, respectively. Let θ_{ℓ} and φ_{ℓ} denote the *physical AoD and AoA* of the ℓth path, respectively; then we have $\Omega_{\ell} = \cos(\theta_{\ell})$ and $\psi_{\ell} = \cos(\varphi_{\ell})$. Therefore, Ω_{ℓ} and ψ_{ℓ} are within the range $[-1, 1]$. For convenience, in the rest of this paper, Ω_{ℓ} and ψ_{ℓ} are called AoAs and AoDs, respectively, as we design the multi-beam search and hybrid precoding schemes in the cosine angle domain. Similar to [29] and [33], λ_ℓ can be modeled to be complex Gaussian distributed, i.e., $\lambda_\ell \sim \mathcal{CN}(0, 1/L)$, while θ_ℓ and φ_ℓ are modeled to be uniformly distributed within $[0, 2\pi)$. $\mathbf{g}(\cdot)$ is a function of the number of antennas and AoD/AoA, and can be expressed as

$$\mathbf{g}(N, \Omega) = \frac{1}{\sqrt{N}}[e^{j\pi 0\Omega}, e^{j\pi 1\Omega}, \ldots, e^{j\pi(N-1)\Omega}]^{\mathrm{T}}, \tag{3}$$

where N is the number of antennas (N is N_{A} at the BS and M_{A} at the MS), Ω is AoD or AoA. It is easy to find that $\mathbf{g}(N, \Omega)$ is a periodical function which satisfies $\mathbf{g}(N, \Omega) = \mathbf{g}(N, \Omega + 2)$.

Remark 1: Evenly sampling the cosine angle space $[-1, 1]$ with an interval length $2/N$ from $-1 + 1/N$ leads to N steering vectors $\mathbf{g}(N, -1 + (2k-1)/N)$, $k = 1, 2, \ldots, N$. We say each vector of them represents a *basis beam* with a width $2/N$ [36], [37].

C. PROBLEM FORMULATION

Our target is to maximize the total achievable rate for the downlink point-to-point transmission. Thus, the problem can be formulated as [36, Ch. 7 and 8], [38]

$$\begin{aligned} \max_{\mathbf{F}_{\mathrm{B}}, \mathbf{F}_{\mathrm{R}}, \mathbf{W}_{\mathrm{B}}, \mathbf{W}_{\mathrm{R}}, \mathbf{Q}} \quad & R = \log_2 \det\left(\mathbf{I} + \mathbf{K}_{\mathrm{W}}^{-1/2}\mathbf{H}_{\mathrm{E}}\mathbf{Q}\mathbf{H}_{\mathrm{E}}^{\mathrm{H}}\mathbf{K}_{\mathrm{W}}^{-\mathrm{H}/2}\right) \\ \text{subject to } & ||\mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}}||_{\mathrm{F}}^2 = N_{\mathrm{S}}, \\ & \mathrm{Tr}(\mathbf{Q}) = P, \end{aligned} \tag{4}$$

where $\mathbf{H}_{\mathrm{E}} = \mathbf{W}_{\mathrm{B}}^{\mathrm{H}}\mathbf{W}_{\mathrm{R}}^{\mathrm{H}}\mathbf{H}\mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}}$, $\mathbf{K}_{\mathrm{W}} = \mathbf{W}_{\mathrm{B}}^{\mathrm{H}}\mathbf{W}_{\mathrm{R}}^{\mathrm{H}}\mathbf{W}_{\mathrm{R}}\mathbf{W}_{\mathrm{B}}$, and $\mathbf{Q}$ is the power allocation matrix. To solve this problem, we must solve the following two subproblems.

Subproblem 1: Multi-Beam Search: As we do not know the channel matrix, we need to estimate it first, which is the first subproblem. Due to the large antenna size, it would be time consuming to make a full estimation of $\mathbf{H}$. Fortunately, according to the mmWave channel model in (2), it would be time efficient to only search the AoDs and AoAs of several most strong MPCs to capture the majority of the channel energy and dimension, which guarantees the achievable performance.

Subproblem 2: Hybrid Precoding: Assuming the AoAs/AoDs of the most significant MPCs have been estimated, the remaining subproblem is the hybrid precoding, i.e., to obtain $\mathbf{F}_{\mathrm{B}}$, $\mathbf{F}_{\mathrm{R}}$, $\mathbf{W}_{\mathrm{B}}$, $\mathbf{W}_{\mathrm{R}}$ and $\mathbf{Q}$ to maximize R under the CA constraint. Note that due to the CA constraint, it is hard to find a globally optimal solution to this subproblem. On the other hand, it may be not worthy to find the globally optimal solution at a cost of high computational complexity, since the channel estimation is already rather approximate. Hence, we focus on designing a low-complexity hybrid precoding.

III. HIERARCHICAL MULTI-BEAM SEARCH

In mmWave communication, channel estimation is generally realized by estimating the coefficients, AoDs and AoAs of several (N_{S} in the context of this paper) most significant MPCs, as the conventional entry-wise estimation is time consuming due to the large antenna size. Before we introduce the proposed hierarchical multi-beam search method, let us start from the introduction of the bruteforce *sequential search* scheme for better understanding.

A. THE SEQUENTIAL SEARCH SCHEME

The sequential search scheme is straightforward, i.e., sequentially searching the whole Tx/Rx angle plane and finding the N_{S} (AoD AoA) pairs with the highest strengths. Therefore, the codebook for the sequential search consists of steering vectors with evenly sampled angles in the range of $[-1, 1]$, i.e., $\mathbf{g}(N, -1 + \frac{2i-1}{KN})$, $i = 1, 2, \ldots, KN$, where the sampling interval is $\frac{2}{KN}$, and K is the *over-sampling factor*. The larger K is, the smaller the estimation errors of AoDs and AoAs are.

Regarding the considered system in Section II, sequential search is realized by sequentially transmitting training sequences from the BS with codewords $\mathbf{g}(N_{\mathrm{A}}, -1 + \frac{2j-1}{KN_{\mathrm{A}}})$ and receiving the training sequences at the MS with codewords $\mathbf{g}(M_{\mathrm{A}}, -1 + \frac{2i-1}{KM_{\mathrm{A}}})$ for $i = 1, 2, \ldots, KM_{\mathrm{A}}$ and $j = 1, 2, \ldots, KN_{\mathrm{A}}$. Consequently, we can obtain the angle-domain matrix $\mathbf{G}$:

$$\begin{aligned} [\mathbf{G}]_{i,j} &= \mathbf{g}(M_{\mathrm{A}}, -1 + \frac{2i-1}{KM_{\mathrm{A}}})^{\mathrm{H}}\mathbf{H}\mathbf{g}(N_{\mathrm{A}}, -1 + \frac{2j-1}{KN_{\mathrm{A}}}), \\ & i = 1, 2, \ldots, KM_{\mathrm{A}},\ j = 1, 2, \ldots, KN_{\mathrm{A}}. \end{aligned} \tag{5}$$

Afterwards, it is straightforward to find the N_{S} most significant peaks with $\mathbf{G}$ on the Tx/Rx angle plane. If we denote the time period of a training sequence as a time slot (i.e., a measurement), we need $K^2M_{\mathrm{A}}N_{\mathrm{A}}$ time slots to estimate $\mathbf{G}$ with the sequential search approach. Considering that we have N_{R} and M_{R} RF chains at the BS and MS, respectively, in each time slot we can estimate $N_{\mathrm{R}}M_{\mathrm{R}}$ elements of $\mathbf{G}$, by sending different orthogonal training sequences on the N_{R} RF chains with different steering vectors at the BS, and receiving also with different steering vectors on the M_{R} chains at the MS (similar to (14) and (15)). Thus, the total time cost to estimate $\mathbf{G}$ is

$$T_{\mathrm{SS}} = \frac{K^2M_{\mathrm{A}}N_{\mathrm{A}}}{N_{\mathrm{R}}M_{\mathrm{R}}}, \tag{6}$$

which is proportional to $M_{\mathrm{A}}N_{\mathrm{A}}$. While this method is feasible, the time cost would be significantly high for large-array devices.

B. THE HIERARCHICAL MULTI-BEAM SEARCH SCHEME

To reduce the time cost of channel estimation, we propose the hierarchical multi-beam search scheme, where a corresponding hierarchical codebook needs to be designed in advance. Fig. 2 shows a general hierarchial codebook, which has $S+1$ *layers* besides the over-sampling layer. Note that the over-sampling layer differs the hierarchial codebook for multi-beam search from those for single-beam search, because a higher precision of angle estimation is required for multi-beam search. In the codebook, there are M^k codewords with

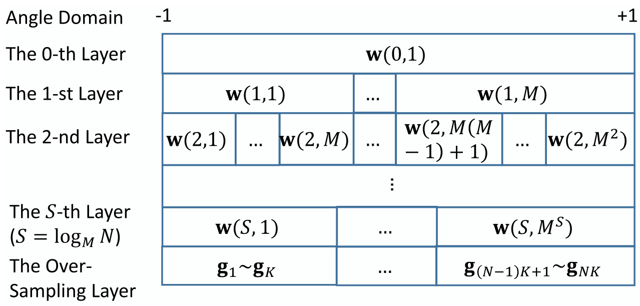

FIGURE 2. Structure of a hierarchical codebook.

the same beam width but different steering angles in the kth layer, where M is the *hierarchical factor* (cf. [1]). $\mathbf{w}(k, n)$ denotes the nth codeword in the kth layer, $n = 0, 1, \ldots, M^k$. For the over-sampling layer, the codewords $\mathbf{g}_i = \mathbf{g}(N, -1 + \frac{2i-1}{KN})$, $i = 1, 2, \ldots, KN$. The KN codewords are in fact the steering vectors to sample the angle domain with an angle interval of $2/(KN)$.

In this paper, we choose to use the joint sub-array and deactivation (JOINT) codebook design [37] for hybrid beamforming. When $M = 2$, the method is introduced as follows.

Codeword Generation with JOINT:

When $k = S = \log_2(N)$, we have $\mathbf{w}(S, n) = \mathbf{g}(N, -1 + \frac{2n-1}{N})$, $n = 1, 2, \ldots, N$.

When $k = S - \ell$, where $\ell = 1, 2, \ldots, S$, we obey the following procedures to compute $\mathbf{w}(k, n)$:

- Separate $\mathbf{w}(k, 1)$ into $m_S = 2^{\lfloor(\ell+1)/2\rfloor}$ sub-arrays with $\mathbf{f}_m = [\mathbf{w}(k, 1)]_{(m-1)n_S+1:mn_S}$, where $\lfloor\cdot\rfloor$ is the flooring integer operation, $n_S = N/m_S$, $m = 1, 2, \ldots, m_S$;
- Set $\mathbf{f}_m$ as (7), where $N_\mathrm{A} = m_S/2$ if ℓ is odd, and $n_\mathrm{A} = M$ if ℓ is even;
- We have $\mathbf{w}(k, n) = \mathbf{w}(k, 1) \circ \sqrt{N}\mathbf{g}(N, \frac{2(n-1)}{N})$, where $n = 2, 3, \ldots, 2^k$, and $\circ$ is the entry-wise product;
- Normalize $\mathbf{w}(k, n)$.

$$\mathbf{f}_m = \begin{cases} e^{jm\pi}\mathbf{g}(n_\mathrm{S}, -1 + \frac{2m-1}{n_\mathrm{S}}), & m = 1, 2, \ldots, n_\mathrm{A}, \\ \mathbf{0}_{n_\mathrm{S}\times 1}, & m = n_\mathrm{A} + 1, n_\mathrm{A} + 2, \ldots, M, \end{cases} \tag{7}$$

where n_A is the number of active sub-arrays. See [37] for more details in the codebook design.

Based on the designed hierarchical codebook, we need to design a multi-beam search scheme. The key of multi-beam search is how to cancel the effect of the already found beams in the on-going beam search. We propose a method to shape an *already found channel response*, which can be subtracted when searching a new beam. The method has been reported in [1], and here we directly list the search algorithm in Algorithm 1.

Algorithm 1 Hierarchical multi-beam search algorithm.

1) Initialization:
$S = \max\{\log_M N_\mathrm{A}, \log_M M_\mathrm{A}\}$.
$i_\mathrm{LY} = 2$. /*The initial layer index. It can be other values depending on practical requirements.*/
$\mathbf{H}_\mathrm{fd} = \mathbf{0}$, /*The already found channel.*/

2) Iteration:
for $\ell = 1 : N_\mathrm{S}$ **do**
  /*Search for the initial Tx/Rx codewords in the i_LYth layer.*/
  for $m = 1 : M^{i_\mathrm{LY}}$ **do**
    for $n = 1 : M^{i_\mathrm{LY}}$ **do**
      $y(m, n) = \mathbf{w}_\mathrm{BS}(i_\mathrm{LY}, n)^\mathrm{H}[\sqrt{P}\mathbf{H}\mathbf{w}_\mathrm{MS}(i_\mathrm{LY}, m) + \mathbf{n}] - \mathbf{w}_\mathrm{BS}(i_\mathrm{LY}, n)^\mathrm{H}\mathbf{H}_\mathrm{fd}\mathbf{w}_\mathrm{MS}(i_\mathrm{LY}, m)$
  $(m_\mathrm{MS}\ m_\mathrm{BS}) = \arg\max_{(m,n)} |y(m, n)|$
  MS feeds back BS m_BS.

  /*Hierarchical refinement.*/
  for $s = i_\mathrm{LY} + 1 : S$ **do**
    for $n = 1 : M$ **do**
      $y_\mathrm{MS}(n) = \mathbf{w}_\mathrm{BS}(s-1, m_\mathrm{BS})^\mathrm{H}[\sqrt{P}\mathbf{H}\mathbf{w}_\mathrm{MS}(s, (m_\mathrm{MS} - 1)M + n) + \mathbf{n}] - \mathbf{w}_\mathrm{BS}(s - 1, m_\mathrm{BS})^\mathrm{H}\mathbf{H}_\mathrm{fd}\mathbf{w}_\mathrm{MS}(s, (m_\mathrm{MS} - 1)M + n)$
    $n_\mathrm{MS} = \arg\max_n |y_\mathrm{MS}(n)|$
    $m_\mathrm{MS} = (m_\mathrm{MS} - 1)M + n_\mathrm{MS}$

    for $n = 1 : M$ **do**
      $y_\mathrm{BS}(n) = \mathbf{w}_\mathrm{BS}(s, (m_\mathrm{BS} - 1)M + n)^\mathrm{H}[\sqrt{P}\mathbf{H}\mathbf{w}_\mathrm{MS}(s, m_\mathrm{MS}) + \mathbf{n}] - \mathbf{w}_\mathrm{BS}(s, (m_\mathrm{BS} - 1)M + n)^\mathrm{H}\mathbf{H}_\mathrm{fd}\mathbf{w}_\mathrm{MS}(s, m_\mathrm{MS})$
    $n_\mathrm{BS} = \arg\max_n |y_\mathrm{BS}(n)|$
    $m_\mathrm{BS} = (m_\mathrm{BS} - 1)M + n_\mathrm{BS}$
    MS feeds back BS m_BS.

  /*High-resolution refinement.*/
  for $m = (m_\mathrm{MS} - 1)K + 1 : m_\mathrm{MS}K$ **do**
    for $n = (m_\mathrm{BS} - 1)K + 1 : m_\mathrm{BS}K$ **do**
      $y(m, n) = \mathbf{w}_\mathrm{BS}(i_\mathrm{LY}, n)^\mathrm{H}[\sqrt{P}\mathbf{H}\mathbf{w}_\mathrm{MS}(i_\mathrm{LY}, m) + \mathbf{n}] - \mathbf{w}_\mathrm{BS}(i_\mathrm{LY}, n)^\mathrm{H}\mathbf{H}_\mathrm{fd}\mathbf{w}_\mathrm{MS}(i_\mathrm{LY}, m)$
  $(I_\ell, J_\ell) = \arg\max_{(m,n)} |y(m, n)|$
  $\beta_\ell = y(I_\ell, J_\ell)$
  MS feeds back BS J_ℓ.
  /*Updating the already found channel response.*/
  $\mathbf{H}_\mathrm{fd} = \mathbf{H}_\mathrm{fd} + \beta_\ell(\mathbf{g}^\mathrm{MS}_{I_\ell})(\mathbf{g}^\mathrm{BS}_{J_\ell})^\mathrm{H}$

3) Result:
The ℓth ($\ell = 1, 2, \ldots, N_\mathrm{S}$) index pair is (I_ℓ, J_ℓ) within the over-sampling layer.

C. TIME COMPLEXITY COMPARISON

As in our system model there are multiple RF chains at the BS and MS, we can observe multiple measurements in each time slot. This can be realized by sending different orthogonal training sequences through different RF chains with different codewords at the BS, and receiving with different codewords at the MS. With this method, the proposed hierarchical search method using Algorithm 1 needs

$$T_{\rm HS} = N_{\rm S}\left(\left(\log_M(N_{\rm A}) + \log_M(M_{\rm A}) - 2i_{\rm LY}\right)\left\lceil \frac{M}{N_{\rm R}M_{\rm R}}\right\rceil + \left\lceil \frac{M^{2i_{\rm LY}}}{N_{\rm R}M_{\rm R}}\right\rceil + \left\lceil \frac{K^2}{N_{\rm R}M_{\rm R}}\right\rceil\right), \quad (8)$$

time slots, where $\lceil\cdot\rceil$ is the ceiling integer operation, $(\log_M(N_{\rm A})+\log_M(M_{\rm A})-2i_{\rm LY})$ and $\lceil\frac{M}{N_{\rm R}M_{\rm R}}\rceil$ are the total number of stages and the number of measurements in each stage at the BS and MS in the hierarchical refinement, respectively, $\lceil\frac{M^{2i_{\rm LY}}}{N_{\rm R}M_{\rm R}}\rceil$ and $\lceil\frac{K^2}{N_{\rm R}M_{\rm R}}\rceil$ are the numbers of measurements in the initial codeword search and the high-resolution refinement, respectively. Since $\lceil\frac{M^{2i_{\rm LY}}}{N_{\rm R}M_{\rm R}}\rceil$ and $\lceil\frac{K^2}{N_{\rm R}M_{\rm R}}\rceil$ are irrelevant to the number of antennas, $T_{\rm HS}$ is roughly proportional to $N_{\rm S}M$.

For a fair comparison, we assume the required number of MPCs L_d is equal to $N_{\rm S}$. Then the required number of measurements of the scheme in [29] is[3]

$$T_{\rm SP} = MN_{\rm S}^2\left\lceil\frac{MN_{\rm S}}{N_{\rm R}}\right\rceil \log_M\left(\frac{KN_{\rm A}}{N_{\rm S}}\right), \quad (9)$$

which is roughly proportional to $N_{\rm S}^3M^2$.

Fig. 3 shows the required time slots for beamforming training between the proposed hierarchical multi-beam search, the sequential search, and the scheme in [29]. It can be found that the proposed hierarchical search scheme has the lowest time complexity among the three. Compared with the scheme in [29], the proposed hierarchical search scheme achieves a further reduction on the required time slots.

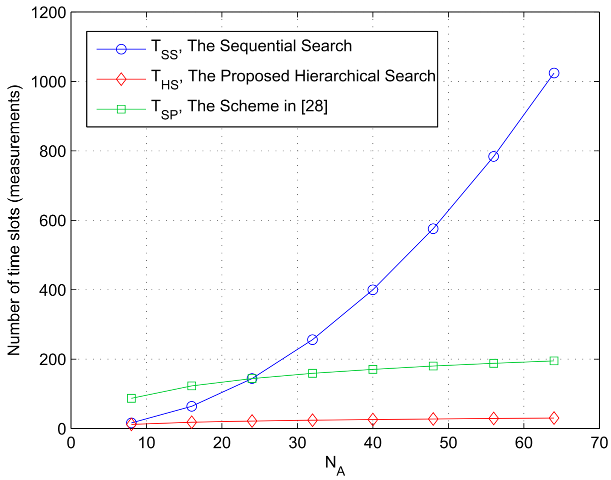

FIGURE 3. Comparison of time complexity between the sequential search, the proposed hierarchical multi-beam search and the scheme in [29], where $M_{\rm A} = N_{\rm A}$, $M = 2$, $K = 2$, $M_{\rm R} = N_{\rm R} = 4$, $N_{\rm S} = 3$, and $i_{\rm LY} = 2$.

IV. LOW-COMPLEXITY HYBRID PRECODING

With the proposed hierarchial multi-beam search approach, we can obtain the index pairs of the $N_{\rm S}$ strongest MPCs, i.e., $(I_1, J_1), (I_2, J_2), \ldots, (I_{N_{\rm S}}, J_{N_{\rm S}})$. In this section, we perform low-complexity hybrid precoding based on the estimated channel information. In specific, we first perform analog precoding without considering the digital precoding, and then compute the digital precoding matrices in the baseband.

[3] In this formula it was assumed that $N_{\rm A} \geq M_{\rm A}$.

A. ANALOG PRECODING

The target of analog precoding is to steer at the $N_{\rm S}$ most significant MPCs/clusters in the angle domain. Hence, the analog precoding and combining matrices are

$$\mathbf{F}_{\rm R} = \Big[\mathbf{g}(N_{\rm A}, -1 + \frac{2J_1 - 1}{KN_{\rm A}}), \mathbf{g}(N_{\rm A}, -1 + \frac{2J_2 - 1}{KN_{\rm A}}), \ldots, \mathbf{g}(N_{\rm A}, -1 + \frac{2J_{N_{\rm S}} - 1}{KN_{\rm A}})\Big], \quad (10)$$

and

$$\mathbf{W}_{\rm R} = \Big[\mathbf{g}(M_{\rm A}, -1 + \frac{2I_1 - 1}{KM_{\rm A}}), \mathbf{g}(M_{\rm A}, -1 + \frac{2I_2 - 1}{KM_{\rm A}}), \ldots, \mathbf{g}(M_{\rm A}, -1 + \frac{2I_{N_{\rm S}} - 1}{KM_{\rm A}})\Big], \quad (11)$$

respectively.

B. DIGITAL PRECODING

While the analog precoding is to steer at the $N_{\rm S}$ most significant MPCs/clusters in the angle domain, the digital precoding is designed to cancel interference between different streams and perform power allocation at the BS.

Provided that $\mathbf{F}_{\rm R}$ and $\mathbf{W}_{\rm R}$ has been designed, we get an $N_{\rm S} \times N_{\rm S}$ baseband effective channel

$$\mathbf{H}_{\rm B} = \mathbf{W}_{\rm R}^{\rm H}\mathbf{H}\mathbf{F}_{\rm R}. \quad (12)$$

Thus,

$$\begin{aligned}[\mathbf{H}_{\rm B}]_{i,j} &= [\mathbf{W}_{\rm R}]_{:,i}{}^{\rm H}\mathbf{H}[\mathbf{F}_{\rm R}]_{:,j} \\ &= \mathbf{g}(M_{\rm A}, -1 + \frac{2I_i}{KM_{\rm A}})^{\rm H}\mathbf{H}\mathbf{g}(N_{\rm A}, -1 + \frac{2J_j}{KN_{\rm A}}). \quad (13)\end{aligned}$$

Since we have obtained I_ℓ and J_ℓ, $\ell = 1, 2, \ldots, L$, the estimation of $\mathbf{H}_{\rm B}$ can be easily realized within a single time slot as follows. Note that this can also be done in the phase of channel estimation after the multi-beam search.

The BS transmits orthogonal training sequences $\mathbf{s}_j$ from the jth RF chain with codeword $\mathbf{g}(N_{\rm A}, -1 + \frac{2J_j-1}{KN_{\rm A}})$, where $j = 1, 2, \ldots, N_{\rm S}$. Then the MS receives with $N_{\rm S}$ RF chains simultaneously, where $\mathbf{g}(M_{\rm A}, -1 + \frac{2I_i-1}{KM_{\rm A}})$ is adopted in the ith RF chain. Thus, at the ith RF chain of the MS, we observe

$$\begin{aligned}\mathbf{r}_i &= \mathbf{g}(M_{\rm A}, -1 + \frac{2I_i - 1}{KM_{\rm A}})^{\rm H}\mathbf{H}\sum_{j=1}^{N_{\rm S}}\mathbf{g}(N_{\rm A}, -1 + \frac{2J_j - 1}{KN_{\rm A}})\mathbf{s}_j \\ &= \sum_{j=1}^{N_{\rm S}}\mathbf{g}(M_{\rm A}, -1 + \frac{2I_i - 1}{KM_{\rm A}})^{\rm H}\mathbf{H}\mathbf{g}(N_{\rm A}, -1 + \frac{2J_j - 1}{KN_{\rm A}})\mathbf{s}_j. \quad (14)\end{aligned}$$

By multiplying with $\mathbf{s}_k$ at the ith RF chain of the MS, where $i = 1, 2, \ldots, N_{\rm S}$ and $k = 1, 2, \ldots, N_{\rm S}$, we get

$$\begin{aligned}
&\mathbf{s}_k^{\rm H}\mathbf{r}_i \\
&= \mathbf{s}_k^{\rm H}\sum_{j=1}^{N_{\rm S}} \mathbf{g}(M_{\rm A}, -1+\frac{2I_i-1}{KM_{\rm A}})^{\rm H}\mathbf{H}\mathbf{g}(N_{\rm A}, -1+\frac{2J_j-1}{KN_{\rm A}})\mathbf{s}_j \\
&= \mathbf{g}(M_{\rm A}, -1+\frac{2I_i-1}{KM_{\rm A}})^{\rm H}\mathbf{H}\mathbf{g}(N_{\rm A}, -1+\frac{2J_k-1}{KN_{\rm A}}) \\
&= [\mathbf{G}]_{I_i,J_k} = [\mathbf{H}_{\rm B}]_{i,k}, \qquad (15)
\end{aligned}$$

where the noise component is neglected.

After the estimation of $\mathbf{H}_{\rm B}$ and the analog precoding, the received signal at the MS can be rewritten as

$$\mathbf{y} = \sqrt{P}\mathbf{W}_{\rm B}^{\rm H}\mathbf{H}_{\rm B}\mathbf{F}_{\rm B}\mathbf{s} + \mathbf{W}_{\rm B}^{\rm H}\mathbf{W}_{\rm R}^{\rm H}\mathbf{n}. \qquad (16)$$

Let $\mathbf{R}_{\rm n} = \mathbf{W}_{\rm R}^{\rm H}\mathbf{W}_{\rm R}$, $\tilde{\mathbf{W}}_{\rm B}^{\rm H} = \mathbf{W}_{\rm B}^{\rm H}\mathbf{R}_{\rm n}^{1/2}$, and $\tilde{\mathbf{H}}_{\rm B} = \mathbf{R}_{\rm n}^{-1/2}\mathbf{H}_{\rm B}$. The received signal is equivalent to

$$\mathbf{y} = \sqrt{P}\tilde{\mathbf{W}}_{\rm B}^{\rm H}\tilde{\mathbf{H}}_{\rm B}\mathbf{F}_{\rm B}\mathbf{s} + \tilde{\mathbf{W}}_{\rm B}^{\rm H}\mathbf{n}. \qquad (17)$$

Regarding the digital operations, since there is no CA constraint on $\mathbf{F}_{\rm B}$ and $\tilde{\mathbf{W}}_{\rm B}$, they can be determined by SVD of $\tilde{\mathbf{H}}_{\rm B}$. Let the SVD of $\tilde{\mathbf{H}}_{\rm B}$ be $\tilde{\mathbf{H}}_{\rm B} = \mathbf{U}_{\rm B}\mathbf{D}_{\rm B}\mathbf{V}_{\rm B}^{\rm H}$, where $\mathbf{U}_{\rm B}$ and $\mathbf{V}_{\rm B}$ are left and right unitary matrices of $\tilde{\mathbf{H}}_{\rm B}$, and $\mathbf{D}_{\rm B}$ is a diagonal matrix with the singular values of $\tilde{\mathbf{H}}_{\rm B}$ listed on the diagonal in a descending order. Then we can determine the digital precoding matrices for the MS and BS immediately, i.e.,

$$\mathbf{W}_{\rm B} = \mathbf{R}_{\rm n}^{-{\rm H}/2}\mathbf{U}_{\rm B}, \qquad (18)$$

and

$$\mathbf{F}_{\rm B} = \mathbf{V}_{\rm B}, \qquad (19)$$

respectively. Note that $\mathbf{W}_{\rm B}$ and $\mathbf{F}_{\rm B}$ require to be normalized according to the power normalization constraint on the precoding and combining matrices.

In addition, the power allocation matrix $\mathbf{Q}$ can be generated by water-filling the total power P on the $N_{\rm S}$ parallel channels with coefficients on the diagonal of $\mathbf{D}_{\rm B}$.

C. COMPUTATIONAL COMPLEXITY COMPARISON

The proposed hybrid precoding consists of analog precoding and digital precoding. The computational complexity of the analog precoding is low, while the digital precoding requires $N_{\rm S} \times N_{\rm S}$ matrix operations, including matrix multiplication, SVD, etc. Thus, the proposed hybrid precoding scheme has an overall computational complexity $\mathcal{O}(N_{\rm S}^3)$ (It is known that the computational complexity of general matrix multiplication, matrix inversion, SVD are on the order of $\mathcal{O}(M^3)$, where M is the matrix size [39]).

In contrast, the hybrid precoding approach proposed in [24] and [29] also requires matrix operations, including matrix multiplication, SVD, etc. Since it jointly designs the analog and digital precodings, the involved matrices are with size $N_{\rm A} \times M_{\rm A}$. Hence, the hybrid precoding in [29] has an overall computational complexity $\mathcal{O}(N_{\rm A}^3)$ (assuming $M_{\rm A} = N_{\rm A}$).

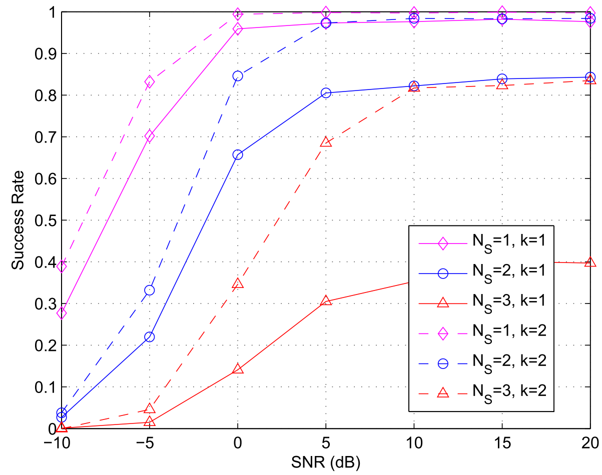

FIGURE 4. **Success rate of the proposed hierarchical multi-beam search scheme with varying K, where $M = 2$, $L = 4$, $N_{\rm R} = M_{\rm R} = 3$, $i_{\rm LY} = 2$.**

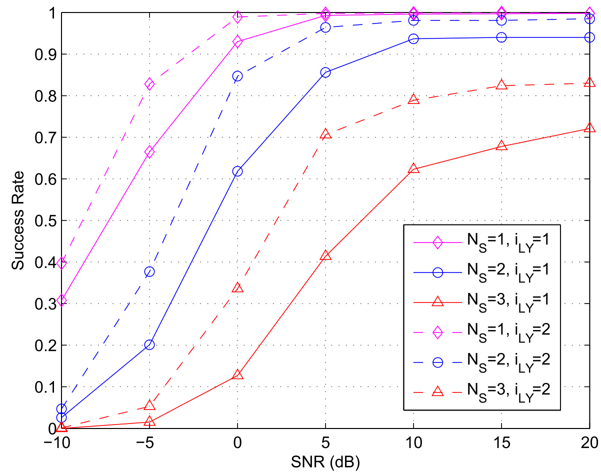

FIGURE 5. **Success rate of the proposed hierarchical multi-beam search scheme with varying $i_{\rm LY}$, where $M = 2$, $L = 4$, $N_{\rm R} = M_{\rm R} = 3$, $K = 2$.**

In brief, the proposed low-complexity hybrid precoding scheme reduces the computational complexity from $\mathcal{O}(N_{\rm A}^3)$ to $\mathcal{O}(N_{\rm S}^3)$, where basically $N_{\rm S} \ll N_{\rm A}$.

V. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed low complexity hybrid precoding (LC-HPC) based on the hierarchical multi-beam search (HIBS). The channel model introduced in Section II is adopted, where the physical angles of the MPCs are randomly generated within $[0, 2\pi)$, and the average strengths of the MPCs are equal, i.e., an NLOS channel model is considered (see Footnote 3). Each performance curve is obtained by averaging 10^3 instantaneous performances with randomly realized channel responses. In all the simulations, we set $N_{\rm A} = M_{\rm A} = 32$, but we note that we have also evaluated the performances with other numbers of antennas, and similar results can be observed. In addition, the other parameters are all set typical values for mmWave communication in the simulations, e.g., the numbers of MPCs and streams (i.e., L and $N_{\rm S}$) are basically small.

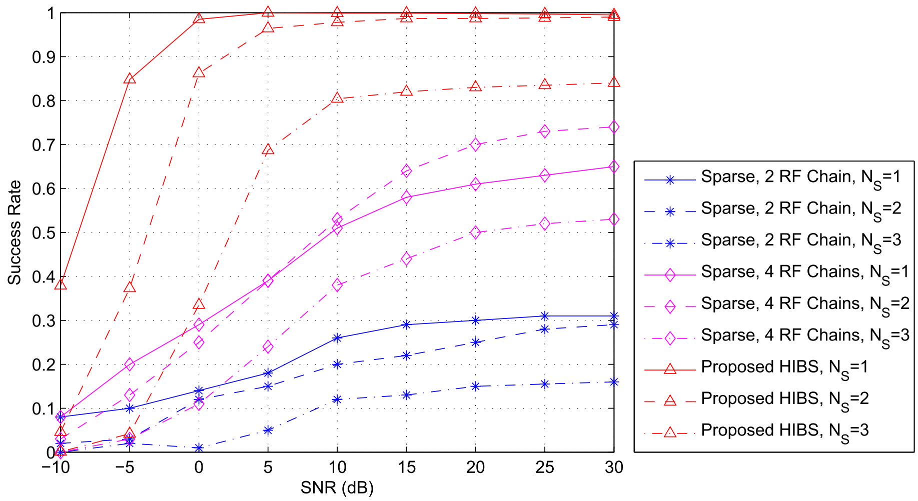

FIGURE 6. Comparison of success rates between the proposed hierarchical multi-beam search scheme with the search scheme in [29] (termed with "Sparse"), where $M = 2$, $L = 4$, $N_R = M_R = 3$, $K = 2$, $i_{LY} = 2$. For the Sparse scheme, the required resolution is set as KN_A, the same as the proposed scheme.

A. HIERARCHIAL MULTI-BEAM SEARCH

First, we demonstrate the performance of the HIBS scheme, for which the most critical figure of merit is the success rate to find the index pairs of multiple beams. According to (2), there are L MPCs with different AoDs and AoAs, and HIBS finds N_S of them. We sequentially decide whether the found N_S MPCs are among the original L MPCs. The decision method is that if there is an ℓ, $1 \leq \ell \leq L$, satisfying $|\psi_\ell - \alpha| < 1/N_A$ and $|\omega_\ell - \beta| < 1/M_A$, where $1/N_A$ and $1/M_A$ are the permitted AoD and AoA errors, α and β are AoD and AoA of an estimated MPC, we say this MPC is successfully searched. Only when all the N_S MPCs are successfully searched, the whole search process succeeds; otherwise it fails.

We note that the search of an arbitrary MPC may fail because of the noise, the effect of the previously found MPCs, and the mutual effect of MPCs, i.e., spatial fading caused by MPCs when measured with wide-beam codewords. In fact, we have simulated the performance with $L = 1$ and $N_S = 1$, i.e., there is only one MPC. The success rate can achieve 100% with high SNR. This is because there is no mutual effect of MPCs when $L = 1$, and there is no effect of the previously found MPCs when $N_S = 1$.

Figs. 4 and 5 show the success rates of the proposed HIBS method with varying K and i_{LY}, respectively, where the SNR refers to the one after correlation operation on the training sequence. From these two figures we can see that: (i) The success rate cannot consistently increase as SNR increases due to the effect of the previously found MPCs and the mutual effect of MPCs. (ii) Basically the success rate is higher when N_S is smaller, because the contributions of the already searched MPCs cannot be completely subtracted off, and they affect the search of the next MPC. The effect gets more severe as N_S increases. (iii) When K becomes larger the success rate becomes higher, because larger K gains more-accurate estimation of the AoDs and AoAs, which means more-accurate contribution subtraction of the already searched MPCs. From Fig. 4 we can observe that when $N_S = 3$, the improvement of success rate by increasing K is more significant than the cases of $N_S = 2$ and $N_S = 1$. (iv) The success rate is higher when i_{LY} is bigger. Indeed, to start from a higher layer not only raises the set-up SNR, but also reduce the possible mutual effect of MPCs when measured with low-layer codewords. However, a bigger i_{LY} means a higher time complexity according to (8).

Fig. 6 shows the comparison of success rates between the proposed HIBS scheme and the scheme in [29] (termed with "Sparse"). We can observe from the comparison that Sparse is sensitive to the number of RF chains. Its performance of success rate with 4 RF chains is even poorer than the proposed scheme with only 1 RF chain for MPC estimation. This performance disadvantage is mainly caused by the hierarchical codebook design in [29], i.e., there may be deep sinks within the beam coverage of the wide-beam codewords when the number of RF chains is not large enough.

Moreover, from these three figures we can observe that although the desired number of MPCs is N_S, the number of actual found MPCs with either HIBS or the Sparse scheme may be less than N_S, especially when N_S is large. In such a case, the actual number of streams will be equal to the number of actual found MPCs, which is less than N_S, resulting a degradation of the achievable rate. However, in practical mmWave communications, the number of independent MPCs/clusters is not large [30], [31]. Thus, these two search schemes are basically suitable.

B. LOW-COMPLEXITY HYBRID PRECODING

Next, we evaluate the performance of achievable rate of the proposed LC-HPC scheme, and learn the effect of K on the performance.

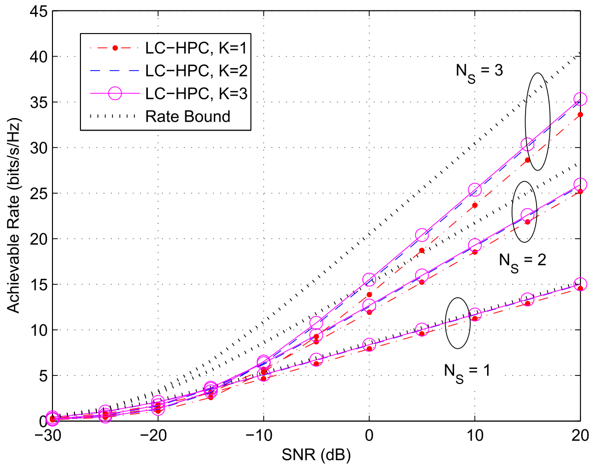

FIGURE 7. Achievable rate of LC-HPC with varying K, where the HIBS scheme is exploited to estimate MPCs. $M = 2$, $L = 4$, $N_R = M_R = 3$, $i_{LY} = 2$.

Fig. 7 shows the achievable rate of LC-HPC with varying K, where the HIBS scheme is exploited to estimate MPCs. The training sequence is assumed long enough to provide sufficiently high SNR for the MPC estimation. From this figure we find that LC-HPC achieves a promising performance. Specifically, compared with the rate bound, which is defined as the achievable rate without the CA constraint, LC-HPC has almost no loss of multiplexing gain, i.e., the slopes of the performance curves of LC-HPC are the same as those of the rate bounds. Although there is increasing SNR loss as N_S increases, which results from the CA constraint, it is basically acceptable in practice where N_S is generally small. Moreover, it is clear that the performance of LC-HPC with $K = 2$ is better than that with $K = 1$, but when $K \geq 2$, further increasing K leads to little performance improvement. Thus, basically $K = 2$ is suitable for practical usage.

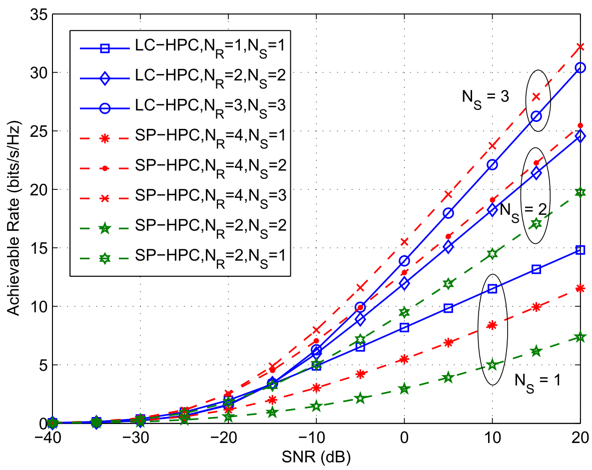

FIGURE 8. Comparison of achievable rates of the proposed low-complexity hybrid precoding (LC-HPC) approach with the sparse precoding approach in [29] (termed as "SP-HPC"). $M = 2$, $L = 3$, $K = 2$, $i_{LY} = 2$. $M_R = N_R$ for all the curves.

Fig. 8 shows the comparison of achievable rates of the proposed LC-HPC approach with the sparse precoding approach in [29] (termed as "SP-HPC"). The HIBS scheme is exploited for the estimation of MPCs for LC-HPC, while the search method proposed in [29] is used for the SP-HPC approach. That is to say, this figure shows the performance comparison of the overall solutions proposed in this paper and [29]. The training sequence is assumed long enough to provide sufficiently high SNR for the MPC estimations in these two approaches. From this figure we find that LC-HPC achieves a close performance to SP-HPC. The performance of SP-HPC gets improved as N_S and the number of RF chains increases, and is basically better than that of LC-HPC when N_R and N_S are not small. That is because SP-HPC selects the steering vectors directly from the optimization of the achievable rate, while LC-HPC just selects several estimated significant MPCs as the analog precoding matrix. However, when N_S or the number of RF chains is small, LC-PHC behaves even better. For instance, when $N_S = 1$, LC-PHC with only 1 RF chain is even better than SP-HPC with 4 RF chains. This is again due to the hierarchical codebook design in [29], where the wide-beam codewords may have deep sinks within the beam coverage when the numbers of RF chains and N_S are not large enough, which may easily result in miss-detection of MPCs.

VI. CONCLUSIONS

In this paper, a low-complexity overall channel estimation and hybrid precoding approach has been proposed. In the channel estimation phase, a new hierarchical multi-beam search scheme, which uses a pre-designed analog hierarchical codebook and the particular channel structure in mmWave communication, was proposed. While in the hybrid precoding phase, the analog precoding is designed to steer along the N_S acquired MPCs/clusters at both Tx/Rx sides, and the digital precoding operates on the $N_S \times N_S$ baseband effective channel. Performance evaluations show that, compared with the approach proposed in [29], the newly proposed approach achieves a close performance to the alternative, or even a better one when the number of RF chains or streams is small. Moreover, the computational complexity of the hybrid precoding is reduced from $\mathcal{O}(N_A^3)$ to $\mathcal{O}(N_S^3)$, where basically $N_S \ll N_A$, while the required time slots for the multi-beam search is reduced from $N_S^3M^2$-proportional to N_SM-proportional.

Acknowledgment

Part of this paper has been reported in [1].

REFERENCES

[1] Z. Xiao, P. Xia, and X.-G. Xia, "Hierarchical multi-beam search for millimeter-wave MIMO systems," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–5.

[2] R. C. Daniels, J. N. Murdock, T. S. Rappaport, and R. W. Heath, Jr., "60 GHz wireless: Up close and personal," *IEEE Microw. Mag.*, vol. 11, no. 7, pp. 44–50, Dec. 2010.

[3] K.-C. Huang and Z. Wang, *Millimeter Wave Communication Systems*. Hoboken, NJ, USA: Wiley, 2011.

[4] E. Perahia, C. Cordeiro, M. Park, and L. L. Yang, "IEEE 802.11 ad: Defining the next generation multi-Gbps Wi-Fi," in *Proc. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Las Vegas, NV, USA, Jan. 2010, pp. 1–5.

[5] S. K. Yong, P. Xia, and A. Valdes-Garcia, *60GHz Technology for Gbps WLAN and WPAN: From Theory to Practice*. West Sussex, U.K.: Wiley, 2011.
[6] Z. Xiao, "Suboptimal spatial diversity scheme for 60 GHz millimeter-wave WLAN," *IEEE Commun. Lett.*, vol. 17, no. 9, pp. 1790–1793, Sep. 2013.
[7] F. Khan and J. Pi, "mmWave mobile broadband (MMB): Unleashing the 3–300GHz spectrum," in *Proc. 34th IEEE Sarnoff Symp.*, Cancún, Mexico, May 2011, pp. 1–6.
[8] A. Alkhateeb, J. Mo, N. González-Prelcic, and R. Heath, Jr., "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–131, Dec. 2014.
[9] J. Choi, "On coding and beamforming for large antenna arrays in mm-Wave systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 2, pp. 193–196, Apr. 2014.
[10] S. Han, C.-L. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, Jan. 2015.
[11] 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.
[12] S. Sun, T. S. Rappaport, R. W. Heath, Jr., 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.
[13] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: Opportunities and challenges," *Wireless Netw.*, vol. 21, no. 8, pp. 2657–2676, Nov. 2015.
[14] J. Wang *et al.*, "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, Oct. 2009.
[15] J. Wang *et al.*, "Beamforming codebook design and performance evaluation for 60 GHz wideband WPANs," in *Proc. IEEE 70th Veh. Technol. Conf. Fall (VTC-Fall)*, Anchorage, AK, USA, Sep. 2009, pp. 1–6.
[16] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, Oct. 2013.
[17] Z. Xiao, L. Bai, and J. Choi, "Iterative joint beamforming training with constant-amplitude phased arrays in millimeter-wave communications," *IEEE Commun. Lett.*, vol. 18, no. 5, pp. 829–832, May 2014.
[18] L. Chen, Y. Yang, X. Chen, and W. Wang, "Multi-stage beamforming codebook for 60 GHz WPAN," in *Proc. 6th Int. ICST Conf. Commun. Netw. China (CHINACOM)*, Harbin, China, Aug. 2011, pp. 361–365.
[19] A. Alkhateeb, G. Leus, and R. W. Heath, Jr., "Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?" in *Proc. IEEE ICASSP*, Apr. 2015, pp. 2909–2913.
[20] Y. Peng, Y. Li, and P. Wang, "An enhanced channel estimation method for millimeter wave systems with massive antenna arrays," *IEEE Commun. Lett.*, vol. 19, no. 9, pp. 1592–1595, Sep. 2015.
[21] P. Wang, Y. Li, L. Song, and B. Vucetic, "Multi-gigabit millimeter wave wireless communications for 5G: From fixed access to cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 168–178, Jan. 2015.
[22] J. Lee, G.-T. Gil, and Y. H. Lee, "Exploiting spatial sparsity for estimating channels of hybrid MIMO systems in millimeter wave communications," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 3326–3331.
[23] Z. Gao, C. Hu, L. Dai, and Z. Wang, "Channel estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1259–1262, Jun. 2016.
[24] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014
[25] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale MIMO systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 2929–2933.
[26] P. Xia and C. Ngo, "System and method for multi-stage antenna training of beamforming vectors," U.S. Patent 8 165 595 B2, Apr. 24, 2012.
[27] S.-H. Wu, K.-Y. Lin, L.-K. Chiu, and M.-C. Chiang, "Hybrid beamforming for two-user SDMA in millimeter wave radio," in *Proc. IEEE 21st Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2010, pp. 1081–1085.
[28] X. Gao, L. Dai, S. Han, C.-L. I, and R. W. Heath, Jr., "Energy-efficient hybrid analog and digital precoding for mmwave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.
[29] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.
[30] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J. N. Murdock, Y. Qiao, and J. I. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Trans. Antennas Propag.*, vol. 61, no. 4, pp. 1850–1859, Apr. 2013.
[31] T. S. Rappaport, Y. Qiao, J. I. Tamir, J. N. Murdock, and E. Ben-Dor, "Cellular broadband millimeter wave propagation and angle of arrival for adaptive beam steering systems," in *Proc. IEEE Radio Wireless Symp. (RWS)*, Santa Clara, CA, USA, Jan. 2012, pp. 151–154.
[32] A. M. Sayeed and V. Raghavan, "Maximizing MIMO capacity in sparse multipath with reconfigurable antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 1, pp. 156–166, Jun. 2007.
[33] Z. Xiao, X. G. Xia, D. Jin, and N. Ge, "Iterative eigenvalue decomposition and multipath-grouping Tx/Rx joint beamformings for millimeter-wave communications," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1595–1607, Mar. 2015.
[34] T. He and Z. Xiao, "Suboptimal beam search algorithm and codebook design for millimeter-wave communications," *Mobile Netw. Appl.*, vol. 20, no. 1, pp. 86–97, Feb. 2015.
[35] J. Nsenga, W. Van Thillo, F. Horlin, V. Ramon, A. Bourdoux, and R. Lauwereins, "Joint transmit and receive analog beamforming in 60 GHz MIMO multipath channels," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Dresden, Germany, Jun. 2009, pp. 1–5.
[36] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge Univ. Press, 2005.
[37] Z. Xiao, T. He, P. Xia, and X.-G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016.
[38] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 684–702, Jun. 2003.
[39] *Computational Complexity of Mathematical Operations*. Accessed on Mar. 1, 2017. [Online]. Available: https://en.wiki-pedia.org/wiki/Computational_complexity_of_mathematical_operations

ZHENYU XIAO (M'11–SM'17) received the B.E. degree with the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2006, and the Ph.D. degree with the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2011. From 2011 to 2013, he held a post-doctoral position with the Electronic Engineering Department, Tsinghua University. From 2013 to 2016, he was a Lecturer with the Department of Electronic and Information Engineering, Beihang University, Beijing, China, where he is currently an Associate Professor.

Dr. Xiao has authored over 50 papers. His research interests are communication signal processing and practical system implementation for wideband communication systems, which cover synchronization, multipath signal processing, diversity, and multiple antenna technology, and millimeter-wave communications and UAV networks. He has been a TPC member of IEEE GLOBECOM'12, IEEE WCSP'12, and IEEE ICC'15. He served as a Reviewer of the IEEE Transactions on Signal Processing, the IEEE Transactions on Wireless Communications, the IEEE Transactions on Vehicular Technology, and the IEEE Communications Letters.

PENGFEI XIA (M'05–SM'10) received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities, MN, USA, in 2005. He is currently a Full Chair Professor with the College of Electronics and Information, Tongji University, Shanghai. He co-edited the first book on 60-GHz millimeter wave communication systems. He has authored over 50 IEEE journal and conference papers, and has more than 70 U.S. patents and/or patent applications. His research interests lie at the intersection of wireless communications, wireless networks, and signal and data processing, including LTE cellular systems, IEEE 802.11 WLANs, millimeter wave communications, transceiver beamforming, and unlicensed band communications. He serves as a Technical Committee Member for the IEEE Signal Processing Society. He received the IEEE Signal Processing Society Best Paper Award 2011. He also serves as the Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the *IEEE Communications Magazines*.

XIANG-GEN XIA (M'97–S'00–F'09) received the B.S. degree in mathematics from Nanjing Normal University, Nanjing, China, in 1983, the M.S. degree in mathematics from Nankai University, Tianjin, China, in 1986, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1992.

He was a Senior/Research Staff Member with Hughes Research Laboratories, Malibu, CA, USA, from 1995 to 1996. In 1996, he joined the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA, where he is currently the Charles Black Evans Professor. His current research interests include space–time coding, MIMO and OFDM systems, digital signal processing, and SAR and ISAR imaging. He is the author of the book *Modulated Coding for Intersymbol Interference Channels* (New York, Marcel Dekker, 2000).

Dr. Xia received the National Science Foundation Faculty Early Career Development Program Award in 1997, the Office of Naval Research Young Investigator Award in 1998, the Outstanding Overseas Young Investigator Award from the National Nature Science Foundation of China in 2001, and the Outstanding Junior Faculty Award of the Engineering School of the University of Delaware in 2001. He is also a Technical Program Chair of the Signal Processing Symposium and Globecom 2007, Washington, DC, USA, and the General Co-Chair of ICASSP 2005, Philadelphia. He is currently serving and has served as an Associate Editor for numerous international journals, including the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON MOBILE COMPUTING, and the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.