Codebook Design for Millimeter-Wave Channel Estimation With Hybrid Precoding Structure

Zhenyu Xiao, *Member, IEEE*, Pengfei Xia, *Senior Member, IEEE*, and Xiang-Gen Xia, *Fellow, IEEE*

***Abstract*— In this paper, we study hierarchical codebook design for channel estimation in millimeter-wave (mmWave) communications with a hybrid precoding structure. Due to the limited saturation power of the mmWave power amplifier, we consider the per-antenna power constraint (PAPC). We first propose a metric, termed generalized detection probability (GDP), to evaluate the quality of *an arbitrary codeword*. This metric not only enables an optimization approach for mmWave codebook design, but also can be used to compare the performance of two different codewords/codebooks. To the best of our knowledge, GDP is the first such metric, particularly for mmWave codebook design. We then propose a heuristic approach to design a hierarchical codebook exploiting beam widening with the multi-RF-chain sub-array (BMW-MS) technique. To obtain crucial parameters of BMW-MS, we provide two solutions, namely, a low-complexity search (LCS) solution to optimize the GDP metric and a closed-form (CF) solution to pursue a flat beam pattern. Performance comparisons show that BMW-MS/LCS and BMW-MS/CF achieve very close performances, and they outperform the existing alternatives under the PAPC.**

***Index Terms*— Millimeter wave, mmWave, mmWave beamforming, mmWave precoding, codebook design, hybrid precoding, hierarchial search.**

I. Introduction

MILLIMETER-WAVE (mmWave) communication is a promising technology for next-generation wireless communications due to the abundant frequency spectrum resource [1]–[11]. To bridge the link budget gap due to the extremely high path loss in the mmWave band, beamforming with large antenna arrays are generally required. Subject to expensive radio-frequency (RF) chains, analog beamforming/combining structure is usually preferred, where all the antennas share a single RF chain and have constant-amplitude constraint on their weights [1], [2], [12]. Meanwhile, a hybrid analog/digital precoding/combining structure has also been proposed to realize multi-stream/multi-user transmission [3], [5], [6], where a small number of RF chains are tied to a large antenna array.

Typically mmWave communications make use of a much smaller number of RF chains than the number of antenna elements. As a result, the conventional channel estimation methods developed for classic multiple-input multiple-output (MIMO) communications is rather inefficient due to high pilot overhead as well as high computational cost [13]. Since mmWave channel is generally sparse in the angle domain, different compressed sensing based channel estimation methods were proposed to estimate the steering angles of multipath components (MPCs) [3], [9], [13]–[15]. On the other hand, a switched beamforming approach was proposed in [2] and [16], where the beam search space (at the transmitter and receiver sides, respectively) is represented by a codebook containing multiple codewords, and the best transmit/receive beams are found within the respective codebooks.

In practice, a coarse sub-codebook may be defined with a small number of coarse sectors (or low-resolution beams) covering the intended angle range, while a fine sub-codebook may be defined with a large number of fine (or high-resolution) beams covering the same intended angle range. A coarse sector may have the same coverage as that of multiple fine beams together [2], [17]–[19]. A divide-and-conquer search may then be carried out across the hierarchical codebook, by finding the best sector first on the low-resolution codebook level, and then finding the best beam on the high-resolution codebook level, while the best high-resolution beam is encapsulated in the best sector [2], [17]–[19]. Such a hierarchical codebook structure and the associated multi-stage beam search have been adopted in many recent works [2], [3], [9], [13]–[16].

Performances of the hierarchical search schemes, including the search time and detection rate of desired MPCs, depend highly on the codebook design. With an analog beamforming structure, [2] proposed to use wider beams to speed up the beam search process, but did not cover how to broaden the beams in detail. In [20], a binary-tree structured hierarchical codebook was designed by using brute-force antenna deactivation (DEACT), where wider beams were generated by turning off part of the antennas. In [21], a hierarchical codebook was also designed, where beam widening is achieved via

Manuscript received March 10, 2016; revised August 4, 2016; accepted October 12, 2016. Date of publication October 20, 2016; date of current version January 6, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61571025, Grant 91538204, Grant 91338106, and Grant 61231013, and in part by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61521091. The associate editor coordinating the review of this paper and approving it for publication was A. Gulliver. *(Corresponding author: Zhenyu Xiao.)*

Z. Xiao is with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China, also with the Beijing Key Laboratory for Network-Based Cooperative Air Traffic Management and the Beijing Laboratory for General Aviation Technology, Beijing 100191, China, and also with the Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China (e-mail: xiaozy@buaa.edu.cn).

P. Xia is with the Key Laboratory of Embedded System and Service Computing, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China.

X.-G. Xia is with the School of Information Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2016.2619705

sub-array technique. Although it was shown to outperform DEACT [21], half of antennas may still need to be turned off for some codewords. In brief, these existing approaches usually rely on turning off part of the antennas. This brings the undesirable effect of reduced maximal transmission power. At the same time, these approaches also require an analog switch for each antenna element path, leading to additional cost and power consumption [22].

In contrast, a hybrid precoding structure with multiple RF chains (typically a few) not only enables multi-stream transmission, but also offers higher flexibility for codebook design. As a result, antenna deactivation and analog switches can be usually avoided. In [13], the hybrid precoding structure was adopted to shape wider beams by exploiting the sparse reconstruction approach (SPARSE). But high-quality wide beams can be shaped only when the number of RF chains is large enough, while deep sinks within the intended angle range appear otherwise. In addition, a phase-shifted discrete Fourier Transform (PS-DFT) method was proposed in [23], where wider beams are shaped by steering multiple RF chains to adjacent equally spaced angles; thus a large number of RF chains are required to shape a very-wide codeword. Although these works [13], [23] are theoretically feasible, they basically need a lot of RF chains for very-wide codewords, which may make them inappropriate for devices with only a few RF chains.

On the other hand, with multiple RF chains the output powers of the antennas may be significantly different from each other, because multiple signals get combined before transmission, and the power fluctuation is expected to be more severe when the number of RF chains is greater. Since in mmWave integrated circuits the saturation power of a power amplifier is usually limited [24], [25], the output power fluctuation may limit the maximal transmission power. In these parallel works [13], [23], the per-antenna power constraint (PAPC), caused by the limited saturation power of the mmWave power amplifiers in each antenna branch, was not taken into account.

In this paper, we design a codebook for mmWave communications with a hybrid precoding structure (typically with a few RF chains), and we take the PAPC into account in the design. When PAPC is considered, we need to pursue two different objects: both flat beam pattern and the maximal transmission power. In such a case, we can hardly find a proper metric to measure the performance of a codeword. Hence, in this paper we first propose a metric, called generalized detection probability (GDP), to evaluate the quality of an arbitrary codeword incorporating both the flatness of beam pattern and the maximal transmission power. This metric not only enables a general optimization approach for mmWave codebook design, but also can be used to compare the performance of two different codewords/codebooks. To the best of our knowledge, GDP is the first metric particularly for mmWave codebook design. We then propose an approach to design a hierarchical codebook for the hybrid structure, where BeaM is Widened via Multi-RF-chain Sub-array technique (BMW-MS). To obtain crucial parameters of BMW-MS, we provide two solutions, namely a low-complexity search (LCS) solution to optimize the GDP metric and a closed-form (CF) solution to pursue a flat beam pattern. Performance comparisons show that BMW-MS/LCS and BMW-MS/CF achieve almost equivalent performances, and they (with only 2 RF chains) outperform the existing alternatives under the PAPC.

The rest of this paper is organized as follows. The system and channel models are introduced in Section II. The channel estimation method is proposed, and the problem of codebook design is formulated in Section III. The GDP metric is proposed in Section IV, and the hierarchical codebook design is presented in Section V. Performance evaluation is conducted in Section VI. Lastly, the conclusions are drawn in Section VII.

Symbol Notations: 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. $\mathbb{E}(\cdot)$ denotes expectation operation. $[\mathbf{a}]_i$ and $[\mathbf{A}]_{ij}$ denote the i-th entry of $\mathbf{a}$ and the i-row and j-column entry of $\mathbf{A}$, respectively. $[\mathbf{a}]_{i:j}$, $[\mathbf{A}]_{:,j}$, and $[\mathbf{A}]_{j,:}$ denote a vector with entries being the i-th to j-th entries of $\mathbf{a}$, the j-th column of $\mathbf{A}$ and the j-th row of $\mathbf{A}$, respectively. $|\cdot|$, $\|\cdot\|$ and $\|\cdot\|_\infty$ denote the absolute value, two-norm and ∞-norm respectively.

II. System and Channel Models

Without loss of generality, we consider a point-to-point mmWave system with a hybrid digital/analog precoding/combining structure,[1] as shown in Fig. 1, where multiple RF chains[2] are tied to a half-wave spaced uniform linear array (ULA) at both the Tx and Rx. Analog precoding/combining refers to signal operations in the analog RF (including only phase shift and signal summation) in comparison with those in the digital baseband. Due to the phase shifters, the analog precoding/combining matrices have constant-amplitude constraints, which challenge the design of a proper beam pattern with desired beam width and steering angle. Relevant system parameters are listed below, where N_{ST} is the number of data streams.

M_{RF}	The number of RF chains at the Tx.
M_{AN}	The number of antennas at the Tx.
N_{RF}	The number of RF chains at the Rx.
N_{AN}	The number of antennas at the Rx.
$\mathbf{F}_{\mathrm{BB}}$	$M_{\mathrm{RF}} \times N_{\mathrm{ST}}$ digital precoding matrix at the Tx.
$\mathbf{F}_{\mathrm{RF}}$	$M_{\mathrm{AN}} \times M_{\mathrm{RF}}$ analog precoding matrix at the Tx.
$\mathbf{W}_{\mathrm{BB}}$	$N_{\mathrm{RF}} \times N_{\mathrm{ST}}$ digital combining matrix at the Rx.
$\mathbf{W}_{\mathrm{RF}}$	$N_{\mathrm{AN}} \times N_{\mathrm{RF}}$ analog combining matrix at the Rx.
$\underline{\mathbf{F}}$	A Tx *composite codeword*, $\underline{\mathbf{F}} \triangleq (\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})$.
$\underline{\mathbf{f}}_i$	A Tx codeword, $\underline{\mathbf{f}}_i \triangleq (\mathbf{F}_{\mathrm{RF}}, [\mathbf{F}_{\mathrm{BB}}]_{:,i})$.
$\underline{\mathbf{W}}$	A Rx *composite codeword*, $\underline{\mathbf{W}} \triangleq (\mathbf{W}_{\mathrm{RF}}, \mathbf{W}_{\mathrm{BB}})$.
$\underline{\mathbf{w}}_i$	A Tx codeword, $\underline{\mathbf{w}}_i \triangleq (\mathbf{W}_{\mathrm{RF}}, [\mathbf{W}_{\mathrm{BB}}]_{:,i})$.

[1] The hybrid structure greatly reduces the hardware complexity in comparison with a fully digital precoding/combining structure, and the hybrid precoding/combining may need only partial instead of full channel state information in mmWave communication. However, the precoding performance of the hybrid structure is not as good as the fully digital structure.

[2] Note that a Tx RF chain consists of signal processing components after digital-to-analog converter but before phase shifters, including filters, amplifiers, up-converter, etc. While a Rx RF chain consists of signal processing components after phase shifters but before analog-to-digital converter, including filters, amplifiers, down-converter, etc.

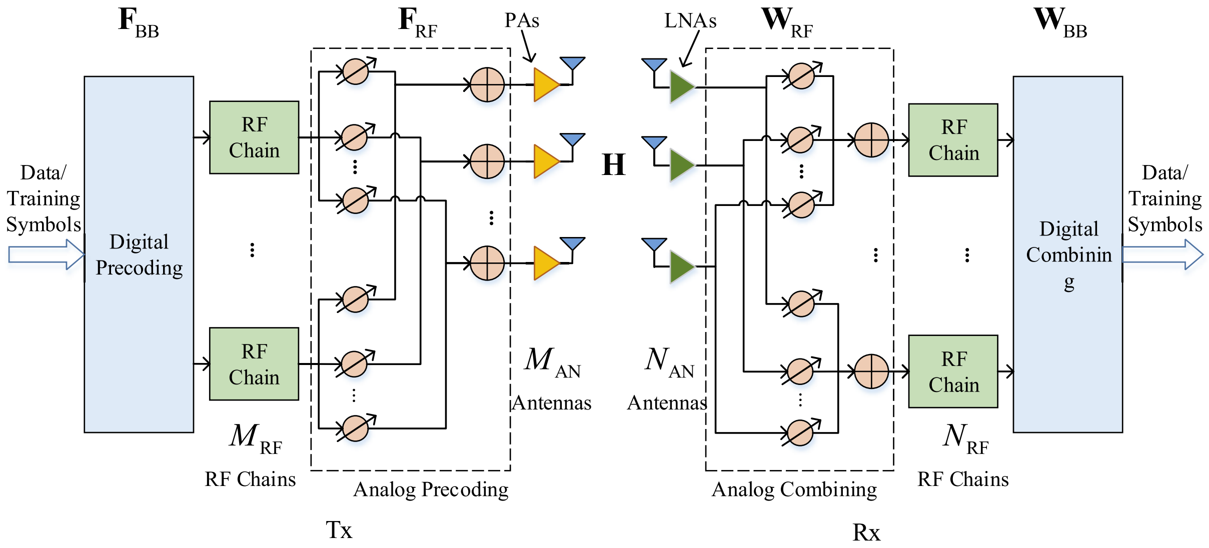


Fig. 1. Illustration of a hybrid analog/digital precoding and combing structure with power amplifiers.

Basically, we have $M_{\text{RF}} \leq M_{\text{AN}}$ and $N_{\text{RF}} \leq N_{\text{AN}}$. In practical mmWave systems, M_{RF} and N_{RF} are much smaller than M_{AN} and N_{AN}, respectively.

In this paper, we propose a channel estimation method[3] based on a hierarchical codebook framework, and design a codebook with the hybrid structure in Fig. 1. A Tx codebook is a collection of *composite codewords*, and a Tx composite codeword is a precoding matrix pair $(\underline{\mathbf{F}_{\text{RF}}}, \underline{\mathbf{F}_{\text{BB}}})$, which can be seen as the composite of M_{RF} Tx codewords $\{(\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}:,i})\}_{i=1,2,\ldots,M_{\text{RF}}}$. The construction of the Rx codebook is similar to the Tx codebook. We emphasize that in this paper *we use underline to indicate a codeword and a composite codeword*, as shown in the above list. Note Tx/Rx codebooks are predesigned, and thus they are irrelevant to an instantaneous channel response. However, they are designed based on the steering feature of an mmWave channel, and used to reduce the training overhead in channel estimation.

We emphasize that there is a single power amplifier in each antenna branch right before the antenna at the Tx. Since the saturation power of an mmWave power amplifier is basically limited [24], [25], we have the PAPC in our model, i.e., the saturation power of a power amplifier is P_{PEP}, which was not considered in [13] and [23].

Without loss of generality, we adopt the same channel model as that in [13], [20], [21], and [26]–[28], which is given by

$$\mathbf{H} = \sqrt{M_{\text{AN}} N_{\text{AN}}} \sum_{\ell=1}^{L} \lambda_\ell \mathbf{a}(N_{\text{AN}}, \Omega_\ell) \mathbf{a}(M_{\text{AN}}, \psi_\ell)^{\text{H}}, \quad (1)$$

where λ_ℓ is the complex coefficient of the ℓ-th path, L is the number of MPCs, $\mathbf{a}(\cdot)$ is the steering vector function, Ω_ℓ and ψ_ℓ are the cosine angle of departure (AoD) and cosine angle of arrival (AoA) of the ℓ-th path, respectively. Let θ_ℓ and φ_ℓ denote the practical AoD and AoA of the ℓ-th path, respectively; then we have $\boldsymbol{\Omega}_\ell = \cos(\theta_\ell)$ and $\psi_\ell = \cos(\varphi_\ell)$. Therefore, $\boldsymbol{\Omega}_\ell$ and ψ_ℓ are within the range $[-1\ 1]$. *For convenience, in the rest of this paper the cosine angles $\boldsymbol{\Omega}_\ell$ and ψ_ℓ are called AoDs and AoAs, respectively. Unless specified otherwise, the angle domain implicitly means the cosine angle domain.* Similar to [13] and [29], λ_ℓ can be modeled to be complex Gaussian distributed, while $\boldsymbol{\Omega}_\ell$ and ψ_ℓ can be modeled to be uniformly distributed within $[-1, 1]$. $\mathbf{a}(\cdot)$ is a function of the number of antennas and AoD/AoA, and with a ULA it can be expressed as

$$\mathbf{a}(N, \Omega) = \frac{1}{\sqrt{N}}[e^{j\pi 0\Omega},\ e^{j\pi 1\Omega}, \ldots, e^{j\pi(N-1)\Omega}]^{\text{T}}, \quad (2)$$

where N is the number of antennas (N is M_{AN} at the transmitter and M_{R} at the receiver), $\boldsymbol{\Omega}$ is AoD or AoA. It is easy to find that $\mathbf{a}(N, \Omega)$ is a periodical function which satisfies $\mathbf{a}(N, \Omega) = \mathbf{a}(N, \Omega + 2)$. The channel matrix $\mathbf{H}$ also has power normalization $\sum_{\ell=1}^{L} \mathbb{E}(|\lambda_\ell|^2) = 1$.

III. Channel Estimation and the Problem of Codebook Design

A. Channel Estimation

Subject to the hardware constraint, i.e., the number of RF chains is far less than that of the antennas, mmWave channel estimation is generally to search the AoDs/AoAs of several strong MPCs one by one via *beam search* in the angle domain [2], [9], [13]–[16]. In this subsection, we propose an improved beam search method to search one MPC with the hybrid precoding structure in Fig. 1. The other MPCs can also be searched similarly.

In order to estimate the AoD/AoA of an MPC, signal measurements must be carried out based on transmission of training sequences. In each measurement, multi-stream orthogonal training sequences are transmitted from Tx to Rx, with precoding matrices selected from a Tx codebook and combining matrices selected from a Rx codebook, respectively.

[3] The proposed channel estimation method has a higher efficiency than the alternatives, as will be shown later, thanks to the parallel transmission enabled by the hybrid structure. However, it is just a by-product to show how a codebook is used for channel estimation in this paper, which focuses on the codebook design.

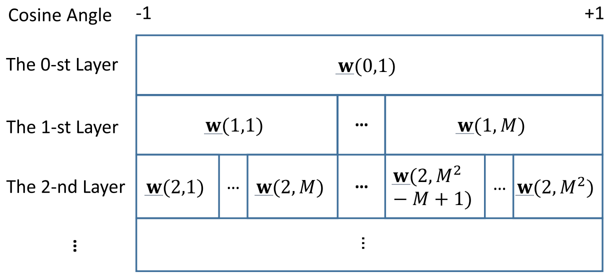


Fig. 2. Beam coverage of a hierarchical codebook.

Algorithm 1 Beam Search Algorithm Based on a Hierarchical Codebook With a Hybrid Structure

```
1) Initialization:
$k = 1$. /*The layer index.*/
$k_{\mathrm{M}} = \max\{\log_{M_{\mathrm{RF}}}(M_{\mathrm{AN}}), \log_{N_{\mathrm{RF}}}(N_{\mathrm{AN}})\}$. /*The maximal layer.*/
Predefine Tx codebook $\mathcal{F}$ and Rx codebook $\mathcal{W}$.
$j_{\mathrm{T}} = i_{\mathrm{R}} = 1$. /*Indices of Tx/Rx composite codewords.*/

2) Iteration:
for $k = 1 : k_{\mathrm{M}}$ do
    Tx/Rx, respectively, selects a composite codeword $\underline{\mathbf{F}}(k, j_{\mathrm{T}})$/$\underline{\mathbf{W}}(k, i_{\mathrm{R}})$ from $\mathcal{F}/\mathcal{W}$, and sets the matrices in the composite codewords to the corresponding precoding/combining matrices.
    Tx sends parallel orthogonal sequences $\mathbf{S}$ as (3), and Rx receives and computes $\rho_{i,j}$ as (5).
    Rx computes the optimal Tx/Rx index pair $(j^\star, i^\star)$ as (6), and sets $i_{\mathrm{R}} = N_{\mathrm{RF}} * (i_{\mathrm{R}} - 1) + i^\star$. Rx feeds back $j^\star$ to Tx, and Tx sets $j_{\mathrm{T}} = M_{\mathrm{RF}} * (j_{\mathrm{T}} - 1) + j^\star$.

3) Result:
The response of the estimated MPC is given by
$\mathbf{H}_1 = \rho_{i^\star,j^\star}\mathbf{a}(N_{\mathrm{AN}}, -1 + \frac{2i_{\mathrm{R}}-1}{N_{\mathrm{AN}}})\mathbf{a}(M_{\mathrm{AN}}, -1 + \frac{2j_{\mathrm{T}}-1}{M_{\mathrm{AN}}})^{\mathrm{H}}$.
```

Hence, we have the following signal model for a measurement:

$$\mathbf{Y} = \sqrt{P}\mathbf{W}_{\mathrm{BB}}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{S} + \mathbf{W}_{\mathrm{BB}}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{Z}, \tag{3}$$

where P is the transmission power per stream, $\mathbf{H}$ is the channel matrix, $\mathbf{Z}$ is a white Gaussian noise matrix with average power N_0, $\mathbf{S}$ has a size of $M_{\mathrm{RF}} \times L_{\mathrm{S}}$ and $[\mathbf{S}]_{j,:}$ is the j-th transmitted training sequence with a length of L_{S} at Tx, $[\mathbf{Y}]_{i,:}$ is the received sequence at the i-th RF chain at Rx, $(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})$ *and* $(\mathbf{W}_{\mathrm{RF}}, \mathbf{W}_{\mathrm{BB}})$ *are a selected Tx composite codeword and a selected Rx composite codeword, respectively.* We have $[\mathbf{S}]_{m,:}[\mathbf{S}]_{n,:}^{\mathrm{H}} = 0$ when $i \neq j$ and $[\mathbf{S}]_{m,:}[\mathbf{S}]_{m,:}^{\mathrm{H}} = L_{\mathrm{S}}$. Let us omit the noise for simplicity. Then we have

$$[\mathbf{Y}]_{i,:} = \sqrt{P}[\mathbf{W}_{\mathrm{BB}}]_{:,i}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}\sum_{m=1}^{M_{\mathrm{RF}}}\mathbf{F}_{\mathrm{RF}}[\mathbf{F}_{\mathrm{BB}}]_{:,m}[\mathbf{S}]_{m,:} \tag{4}$$

Thus,

$$\begin{aligned}\rho_{i,j} &= [\mathbf{Y}]_{i,:}[\mathbf{S}]_{j,:}^{\mathrm{H}}\\ &= L_{\mathrm{S}}\sqrt{P}[\mathbf{W}_{\mathrm{BB}}]_{:,i}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}\mathbf{F}_{\mathrm{RF}}[\mathbf{F}_{\mathrm{BB}}]_{:,j}\\ &\triangleq L_{\mathrm{S}}\sqrt{P}\mathbf{w}_i^{\mathrm{H}}\mathbf{H}\mathbf{f}_j,\end{aligned} \tag{5}$$

where $i = 1, 2, \ldots, N_{\mathrm{RF}}$, and $j = 1, 2, \ldots, M_{\mathrm{RF}}$, $\mathbf{w}_i = \mathbf{W}_{\mathrm{RF}}[\mathbf{W}_{\mathrm{BB}}]_{:,i}$ and $\mathbf{f}_j = \mathbf{F}_{\mathrm{RF}}[\mathbf{F}_{\mathrm{BB}}]_{:,j}$ correspond to the i-th Rx codeword $\underline{\mathbf{w}}_i \triangleq (\mathbf{W}_{\mathrm{RF}}, [\mathbf{W}_{\mathrm{BB}}]_{:,i})$ of the selected Rx composite codeword and the j-th Tx codeword $\underline{\mathbf{f}}_j \triangleq (\mathbf{F}_{\mathrm{RF}}, [\mathbf{F}_{\mathrm{BB}}]_{:,j})$ of the selected Tx composite codeword, respectively.

Afterwards, Rx obtains the optimal Tx/Rx codeword pair as

$$(j^\star, i^\star) = \arg\max_{(j,i)} \ |\rho_{i,j}|^2, \tag{6}$$

and feeds back $j^\star$ to Tx. Hence, in each measurement, Rx in fact finds the best Tx/Rx codeword pair with the highest signal power. If the Tx/Rx codewords are pre-designed to cover different angle ranges, then the AoD/AoA of the MPC will be within the angle coverage of the best Tx/Rx codewords, respectively.

To reduce the training overhead, a hierarchical Tx or Rx codebook, which is a collection of codewords $\underline{\mathbf{f}}$ or $\underline{\mathbf{w}}$, is defined as Fig. 2, The codebook has $\log_M(N) + 1$ layers with indices from $k = 0$ to $k = \log_M(N)$, where M and N are the numbers of RF chains and antennas, respectively. The number of codewords in the k-th layer is M^k, and

$$\mathcal{CV}(\underline{\mathbf{w}}(k, n)) = [-1 + \frac{2n-2}{M^k}, -1 + \frac{2n}{M^k}],$$
$$k = 0, 1, \ldots, \log_M(N), \quad n = 1, 2, \ldots, M^k, \tag{7}$$

where $\mathcal{CV}(\underline{\mathbf{w}})$ denotes the beam coverage in the angle domain of codeword $\underline{\mathbf{w}}$, $\underline{\mathbf{w}}(k, n)$ denotes the n-th codeword in the k-th layer. Note that a set of M adjacent codewords in the same layer, i.e., $\{\underline{\mathbf{w}}(k, (i-1)M + j)\}_{j=1,2,\ldots,M}$ $(i = 1, \ldots, M^{k-1})$, constitute a composite codeword $\underline{\mathbf{W}}(k, i)$.

Based on a hierarchical codebook, an efficient divide-and-conquer search, as shown in Algorithm 1, is launched to fast estimate the response of an MPC. *It is noteworthy that a codebook shown in Fig. 2 and designed in this paper can not only be used in Algorithm 1, but also in other beam search methods.* On the other hand, only one MPC is searched out by launching Algorithm 1 once. For the case that multiple MPCs are needed to be searched out, Algorithm 1 must be launched multiple times to estimate different MPCs, one MPC at a time with new training sequences transmitted. However, extending the one-MPC search to multiple-MPC search is not trivial, because the contribution of the already searched MPC must be subtracted when searching a new MPC. Details may refer to [13] and [30].

Let us next evaluate the training overhead of Algorithm 1. We adopt the duration of a training sequence as a unit to count the training overhead. In each measurement of Algorithm 1, M_{AN} training sequences $\{[\mathbf{S}]_{j,:}\}_{j=1,2,\ldots,M_{\mathrm{AN}}}$ are transmitted in parallel; so the training overhead is 1 unit. Suppose $M_{\mathrm{RF}} = N_{\mathrm{RF}}$ and $M_{\mathrm{AN}} = N_{\mathrm{AN}}$. The total training overhead of Algorithm 1 is $\log_{M_{\mathrm{RF}}}(M_{\mathrm{AN}})$, significantly less than that of the beam search algorithm with an analog beamforming structure in [21], which is $2M_{\mathrm{RF}}\log_{M_{\mathrm{RF}}}(M_{\mathrm{AN}})$ units, as well as that of the beam search algorithm with a hybrid precoding structure in [13] (Algorithm 1 therein, which needs $M_{\mathrm{RF}}\log_{M_{\mathrm{RF}}}(M_{\mathrm{AN}})$

units if $K = M_{\mathrm{RF}}$). In brief, Algorithm 1 reduces the training overhead by a factor of the number of RF chains compared with the alternatives, which benefits from the parallel transmission of multiple training sequences. However, the cost of this benefit is that the codewords within a composite codeword must share the same analog precoding matrix, which will be considered in the codebook design.

B. The Problem of Codebook Design

In this paper we want to design a hierarchical codebook with the beam coverage shown in Fig. 2 based on the hybrid structure shown in Fig. 1. We emphasize that we have both the PAPC on power amplifiers and the constant-amplitude constraint on the analog precoding/combining matrices. Since a Rx codebook design can be the same as a Tx codebook design, we proceed with Tx codebook design.

According to (5), with the hybrid structure an arbitrary codeword $\underline{\mathbf{w}} \triangleq (\mathbf{F}_{\mathrm{RF}}, [\mathbf{F}_{\mathrm{BB}}]_{:,i})$ shapes an antenna weight vector (AWV) $\mathbf{w} = \mathbf{F}_{\mathrm{RF}}[\mathbf{F}_{\mathrm{BB}}]_{:,i}$, and the beam steering and coverage of $\underline{\mathbf{w}}$ are in fact reflected by $\mathbf{w}$. Hence, the codebook design in this paper is to design $\underline{\mathbf{w}}(k,n)$ such that $\mathbf{w}(k,n)$ has the beam coverage $\mathcal{CV}(\mathbf{w}(k,n)) = \mathcal{CV}(\underline{\mathbf{w}}(k,n))$. *For convenience, we also call* $\mathbf{w}$ *a codeword in the remaining of this paper,* but we emphasize that we want to design $\underline{\mathbf{w}} \triangleq (\mathbf{F}_{\mathrm{RF}}, [\mathbf{F}_{\mathrm{BB}}]_{:,i})$ rather than just $\mathbf{w}$ itself, because $\mathbf{w}$ is solely determined by $\underline{\mathbf{w}}$ but not vice versa.

Consequently, a codeword $\mathbf{w}$ has the following structure:

$$\mathbf{w} = \mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB}} = \sum_{j=1}^{M_{\mathrm{RF}}} [\mathbf{f}_{\mathrm{BB}}]_j [\mathbf{F}_{\mathrm{RF}}]_{:,j}, \tag{8}$$

where $\mathbf{f}_{\mathrm{BB}} = [\mathbf{F}_{\mathrm{BB}}]_{:,i}$, $|[\mathbf{F}_{\mathrm{RF}}]_{:,j}| = \frac{1}{\sqrt{M_{\mathrm{AN}}}}\mathbf{1}$ (the constant-amplitude constraint). Note that the codewords belonging to the same composite codeword share the same $\mathbf{F}_{\mathrm{RF}}$.

Given the target beam pattern of $\mathbf{w}(k,n)$ shown in Fig. 2, we need to design ($\mathbf{F}_{\mathrm{RF}}$, $\mathbf{f}_{\mathrm{BB}}$) for each $\mathbf{w}(k,n)$, which is challenging due to the constant-amplitude constraint on $\mathbf{F}_{\mathrm{RF}}$. In [13], this problem is solved by exploiting the sparse reconstruction approach (SPARSE). While in [23], the problem is further constrained by letting $|\mathbf{f}_{\mathrm{BB}}| = \mathbf{1}$, i.e., the transmission power of each RF chain is the same. In such a case, a codeword is a combination of multiple RF vectors with equal power, and it is intuitive that by steering these RF vectors to equally spaced angles, a wide beam can be shaped. This is just the PS-DFT codebook proposed in [23].

In this paper, we let $|\mathbf{f}_{\mathrm{BB}}| = \mathbf{1}$ to simplify the problem, just the same as [23]. However, we propose different methods to design the codewords. In the following, we will first establish a GDP metric to evaluate the quality of an arbitrary Tx codeword $\mathbf{w}$. Then we will design a hierarchical codebook with the target beam coverage shown in Fig. 2 with the codeword structure (8).

IV. The GDP Metric

Given an arbitrary target codeword to cover an angle range $[\psi_0, \psi_0 + B]$, it is clear that the best codeword should have constant absolute beam gain within the covered angle range (i.e., a totally flat beam pattern) [23]. However, due to the constant-amplitude constraint on the analog precoding/combining matrices and the PAPC, an ideal codeword can be hardly designed. Hence, suboptimal designs are of interest, and there have been different approaches [13], [23]. To the best of our knowledge, however, there is no particular metric to directly evaluate the quality of a codeword in the regime of mmWave communications under PAPC.

Let us first define the beam gain of an arbitrary codeword $\mathbf{w}$ along angle Ω ($\Omega \in [-1, 1]$), i.e., $A(\mathbf{w}, \Omega)$:

$$A(\mathbf{w}, \Omega) = \sqrt{N}\mathbf{a}(N, \Omega)^{\mathrm{H}}\mathbf{w} = \sum_{n=1}^{N} [\mathbf{w}]_n e^{-j\pi(n-1)\Omega}. \tag{9}$$

Intuitively, good codewords should have flat beam patterns, and mean square error (MSE) can be adopted to measure how flat a beam pattern is. Moreover, in mmWave communications, the saturation power of a power amplifier is limited. Hence, good codewords should also allow as high as possible maximal transmission power (MTP), which is limited by the saturation power of the power amplifier in each antenna branch. For instance, DEACT is not of high quality, because a lot of antenna elements are turned off, which significantly lowers the MTP. In fact, under the PAPC, the MTP of an arbitrary codeword $\mathbf{w}$ is given by

$$P_{\mathrm{MAX}}(\mathbf{w}) = \frac{P_{\mathrm{PER}}}{\max(\{|[\mathbf{w}]_n|^2\}_{n=1}^{N})} \triangleq \frac{P_{\mathrm{PER}}}{\|\mathbf{w}\|_{\infty}^2}, \tag{10}$$

where P_{PER} is the saturation power for each antenna branch. It is clear that given fixed P_{PER}, $P_{\mathrm{MAX}}(\mathbf{w})$ is maximized when $\mathbf{w}$ (with unit 2-norm) has constant-amplitude elements.

Both MSE and MTP may affect the quality of a codeword and their effects are different. In fact, these two metrics may be contradictory to each other, i.e., a codeword with small MSE tends to have a small MTP as well. We directly bridge the metric to the detection performance in beamforming training. During beamforming training, many Tx/Rx codeword pairs will be selected to detect the AoD/AoA of an MPC. When the AoD/AoA of the MPC locate within the coverage of the codewords, the detection probability is a direct and exact metric incorporating both MSE and MTP. Hence, we can derive the average detection probability, and generalize a metric based on the average detection probability for the Tx codewords.

Suppose that Tx transmits a training sequence with codeword $\mathbf{w}_{\mathrm{T}}$, and Rx receives with codeword $\mathbf{w}_{\mathrm{R}}$, i.e., $\mathbf{w}_{\mathrm{T}}$ and $\mathbf{w}_{\mathrm{R}}$ are fixed. The target beam coverage of $\mathbf{w}_{\mathrm{T}}$ is $[\psi_0, \psi_0 + B]$. We want to develop a metric to evaluate the Tx codeword $\mathbf{w}_{\mathrm{T}}$ based on the average detection probability of a single MPC.

Let $\mathbf{H}_0$ denote the channel response for the MPC to be detected, and it can be defined as

$$\mathbf{H}_0 = \sqrt{M_{\mathrm{AN}}N_{\mathrm{AN}}}\lambda\mathbf{a}(N_{\mathrm{AN}}, \Omega)\mathbf{a}(M_{\mathrm{AN}}, \psi)^{\mathrm{H}}, \tag{11}$$

where λ, Ω and ψ denote the gain, AoA and AoD of the MPC, respectively. Following the channel model in (1), we assume $\lambda \sim \mathcal{CN}(0, 1)$, Ω and ψ are uniformly distributed within $[-1, 1]$.

Given $\mathbf{H}_0$ the detection problem can be formulated as binary hypothesis testing given by [23]

$$y = \begin{cases} \mathbf{w}_{\mathrm{R}}^{\mathrm{H}}\mathbf{n} \sim \mathcal{CN}(0, N_0), & \mathcal{H}_0 \\ \sqrt{P}\mathbf{w}_{\mathrm{R}}^{\mathrm{H}}\mathbf{H}_0\mathbf{w}_{\mathrm{T}}s + \mathbf{w}_{\mathrm{R}}^{\mathrm{H}}\mathbf{n} \sim \mathcal{CN}(S, N_0), & \mathcal{H}_1 \end{cases} \tag{12}$$

where $\mathcal{H}_0$ and $\mathcal{H}_1$ represent the cases when the AoD locates within and outside of $[\psi_0, \psi_0 + B]$, respectively, $S = \sqrt{P}\mathbf{w}_{\mathrm{R}}^{\mathrm{H}}\mathbf{H}_0\mathbf{w}_{\mathrm{T}}s$ denotes the received signal. Given a threshold ΓN_0, the instantaneous detection probability is given by

$$p_{\mathrm{D}}(\Gamma) = \Pr\{|(y|\mathcal{H}_1)|^2 > \Gamma N_0\} = \Pr\{|S + n|^2 > \Gamma N_0\}, \tag{13}$$

where $n = \mathbf{w}_{\mathrm{R}}^{\mathrm{H}}\mathbf{n}$. To derive the average detection probability, we need to average $p_{\mathrm{D}}(\Gamma)$ on all the random variables. Note that S depends on $\mathbf{H}_0$ and $\mathbf{H}_0$ depends on λ, $\mathbf{\Omega}$ and ψ. Hence, we need to average $p_{\mathrm{D}}(\Gamma)$ on n, λ, $\mathbf{\Omega}$ and ψ.

Let us first fix $\mathbf{\Omega}$ and ψ and average $p_{\mathrm{D}}(\Gamma)$ on n and λ. Since $S = \sqrt{P}\mathbf{w}_{\mathrm{R}}^{\mathrm{H}}\mathbf{H}_0\mathbf{w}_{\mathrm{T}}s$, when $\mathbf{\Omega}$ and ψ are fixed, $\mathbf{H}_0$ has only one random parameter $\lambda \sim \mathcal{CN}(0, 1)$. In such a case, S can be seen as a zero-mean complex Gaussian variable, and $(S + n) \sim \mathcal{CN}(0, (1 + \gamma)N_0)$, where γ denotes the average received SNR given by

$$\begin{aligned} \gamma &= \mathbb{E}_{\lambda}\left\{|\sqrt{P}\mathbf{w}_{\mathrm{R}}^{\mathrm{H}}\mathbf{H}_0\mathbf{w}_{\mathrm{T}}s|^2/N_0\right\} \\ &= \frac{PM_{\mathrm{AN}}N_{\mathrm{AN}}}{N_0}|\mathbf{w}_{\mathrm{R}}^{\mathrm{H}}\mathbf{a}(N_{\mathrm{AN}}, \mathbf{\Omega})\mathbf{a}(M_{\mathrm{AN}}, \psi)^{\mathrm{H}}\mathbf{w}_{\mathrm{T}}|^2 \\ &= \frac{P}{N_0}|A(\mathbf{w}_{\mathrm{T}}, \psi)|^2|A(\mathbf{w}_{\mathrm{R}}, \mathbf{\Omega})|^2, \end{aligned} \tag{14}$$

where $|A(\mathbf{w}_{\mathrm{T}}, \psi)|$ and $|A(\mathbf{w}_{\mathrm{R}}, \mathbf{\Omega})|$ are in fact Tx and Rx array gains depending on ψ and $\mathbf{\Omega}$, respectively. According to [31, Ch. 2], $|S + n|^2/N_0$ obeys Chi-square distribution, and its cumulative distribution function (CDF) is $F(y) = 1 - e^{-y/(1+\gamma)}$. Thus, we have

$$\bar{p}_{\mathrm{D0}}(\Gamma) = 1 - F(\Gamma) = e^{-\Gamma/(1+\gamma)}. \tag{15}$$

After averaging over λ and n, we need to further average $\bar{p}_{\mathrm{D0}}(\Gamma)$ in (15) on $\mathbf{\Omega}$ and ψ to obtain the ultimate average detection probability. Note that as we only want to evaluate the quality of the Tx codeword $\mathbf{w}_{\mathrm{T}}$ with angle coverage $[\psi_0, \psi_0 + B]$, we can first get rid of the effects of the Rx codeword $\mathbf{w}_{\mathrm{R}}$ and AoA $\mathbf{\Omega}$. Consequently, we assume the RX gain is fixed for simplicity, and without loss of generality let $|A(\mathbf{w}_{\mathrm{R}}, \mathbf{\Omega})|^2 = 1$. As a result, γ reduces to

$$\gamma = \frac{P}{N_0}|A(\mathbf{w}_{\mathrm{T}}, \psi)|^2. \tag{16}$$

And considering the MTP of $\mathbf{w}_{\mathrm{T}}$, the maximal received SNR is

$$\begin{aligned} \gamma_{\mathrm{MAX}} &= \frac{P_{\mathrm{MAX}}}{\rho^2 N_0}|A(\mathbf{w}_{\mathrm{T}}, \psi)|^2 \\ &= \frac{P_{\mathrm{PER}}}{\|\mathbf{w}_{\mathrm{T}}\|_{\infty}^2\rho^2 N_0}|A(\mathbf{w}_{\mathrm{T}}, \psi)|^2 \\ &\triangleq \frac{\gamma_{\mathrm{PER}}}{\|\mathbf{w}_{\mathrm{T}}\|_{\infty}^2}|A(\mathbf{w}_{\mathrm{T}}, \psi)|^2, \end{aligned} \tag{17}$$

where ρ^2 is the pass loss, γ_{PER} denotes the per-antenna received SNR under the PAPC.

Consequently, the average detection probability is given by

$$\begin{aligned} \bar{p}_{\mathrm{D}}(\Gamma) &= \frac{1}{B}\int_{\psi_0}^{\psi_0+B} e^{-\Gamma/(1+\gamma_{\mathrm{MAX}})}d\psi \\ &= \frac{1}{B}\int_{\psi_0}^{\psi_0+B} \exp\left(-\frac{\Gamma}{1 + \frac{\gamma_{\mathrm{PER}}}{\|\mathbf{w}_{\mathrm{T}}\|_{\infty}^2}|A(\mathbf{w}, \psi)|^2}\right)d\psi. \end{aligned} \tag{18}$$

We can see that $\bar{p}_{\mathrm{D}}(\Gamma)$ depends on both Γ and γ_{PER} in addition to the codeword $\mathbf{w}_{\mathrm{T}}$ itself. Hence, it cannot be directly used as a general metric. Note that the threshold Γ affects only the tradeoff between detection probability in hypothesis $\mathcal{H}_1$ and false-alarm probability in hypothesis $\mathcal{H}_0$ [32]. A small Γ leads to high detection probability but higher false-alarm probability as well. In fact, the threshold itself does not affect the detection capability which involves both detection probability and false-alarm probability [32]. Based on this fact, we can just set $\Gamma = 1$ without loss of generality. When Γ is larger/smaller, detection probability will be lower/higher, but the comparison result of average detection probability between two different codewords basically maintains.

On the other hand, γ_{PER} may affect the comparison result of the average detection probability between two different codewords. Intuitively, when γ_{PER} is sufficiently high, the beam pattern (reflected by $|A(\mathbf{w}, \psi)|$ in (18)) is dominant, but when γ_{PER} is small, the maximal received SNR (reflected by $\|\mathbf{w}_{\mathrm{T}}\|_{\infty}^2$ in (18)) is dominant. Hence, different γ_{PER} should be set for different systems.

A possible way is to set a typical value of γ_{PER} based on the system settings. For instance, the saturation power of a power amplifier (P_{PER}) can be set to 8 dBm [24], [25]. According to the Friis formula, when the wavelength of the carrier frequency is 1 centimeter (30 GHz), and the Tx/Rx distance is 100 meters, the pass loss will be $20\log_{10}(4\pi \times 10000) = 102$ dB. Besides, when the bandwidth $B = 100$ MHz, the noise power can be computed as $N_0 = 10\log_{10}(\kappa TB) = 10\log_{10}(1.38 \times 10^{-23} \times 300 \times 10^8 \times 10^3) = -94$ dBm, where κ and T are the Boltzmann constant and ambient temperature, respectively. Hence, the per-antenna received SNR is $\gamma_{\mathrm{PER}} = 8 - 102 - (-94) = 0$ dB.

It is noteworthy that the above computation is rough. The spreading gain of the training sequence and the receive antenna gain are not included. On the other hand, mmWave circuit losses, like insertion loss, noise figure, etc., are not considered. The overall per-antenna received SNR may have a dynamic range centered on 0 dB. In this paper we prefer to set $\gamma_{\mathrm{PER}} = 0$ dB for conciseness, but it should be clarified that other typical values close to 0 dB are also applicable. It will be shown in Section VI that a small change of γ_{PER} does not affect the comparison result of two codewords.

Based on the above discussions, we propose the metric of GDP for an arbitrary N-entry Tx codeword $\mathbf{w}$ with unit 2-norm and target coverage $[\psi_0, \psi_0 + B]$:

$$\xi(\mathbf{w}, \psi_0, B) = \frac{1}{B}\int_{\psi_0}^{\psi_0+B} \exp\left(-\frac{\|\mathbf{w}\|_{\infty}^2}{\|\mathbf{w}\|_{\infty}^2 + |A(\mathbf{w}, \psi)|^2}\right)d\psi, \tag{19}$$

where $\|\mathbf{w}\|_{\infty}^2 = \max(\{|[\mathbf{w}]_n|^2\}_{n=1}^{N})$.

Although (19) is defined for Tx codewords, it also can be used for Rx codewords, because small input fluctuation of low-noise amplifier (LNA) is also favored in mmWave communications, where the linearity of LNA may be not perfect due to the high frequency and large signal bandwidth. In the case that the linearity of LNA is good enough, (19) can be modified by replacing $\|\mathbf{w}\|_\infty^2$ with constant 1 for Rx codewords. In this paper, we use (19) for both Tx/Rx codeword designs. In addition, although in the derivation of (19) we have assumed Gaussian-distributed path gain, this metric can also be applied to the cases when the path gain obeys other distributions or is even unknown. In such a case, the GDP metric in (19) is suboptimal.

It is clear that GDP depends on both $\|\mathbf{w}\|_\infty^2$ and $|A(\mathbf{w}, \psi)|^2$; thus in fact both MTP of $\mathbf{w}$ and MSE of the beam pattern ($|A(\mathbf{w}, \psi)|$) are incorporated in the GDP. According to (19), a good codeword should have elements with close amplitudes, such that the MTP will be higher. In addition, a good codeword should also have a flat beam pattern; thus deep sinks within the beam coverage should be avoided.

Moreover, one significance of GDP lies in that it enables a general optimization approach to design the codewords. In particular, if we want to design an arbitrary codeword $\mathbf{w}$ with target beam coverage $[\psi_0, \psi_0 + B]$, we can formulate the following problem

$$\begin{aligned}&\underset{\mathbf{b}}{\text{maximize}}\ \xi(\mathbf{w}(\mathbf{b}), \psi_0, B),\\ &\text{subject to Constraints on } \mathbf{b},\end{aligned} \tag{20}$$

where $\mathbf{b}$ is a parameter vector to be determined, and the constraints can include other desired structure constraints to simplify the search complexity in addition to the constant-amplitude constraint. As it is difficult to further simplify the expression and obtain a closed-form GDP, it is thus difficult to obtain a solution of (20) through analytical approach. Fortunately, by appropriately designing the structure constraints, the problem can be solved through numerical search methods. In fact, the proposed BMW-MS/LCS codebook is just obtained with the optimization approach (cf. (25)) by adopting the sub-array structure.

Another significance of GDP lies in that it provides an additional way to compare two different codewords/codebooks besides simulation. With the same target beam coverage, a codeword with higher GDP has better performance. For two different codebooks with the same coverage structure, its performance is basically determined by the codewords with the widest beams, because the widest beams have the lowest beam gain in general. Hence, by comparing the GDPs of the widest codewords of two different codebooks, we can evaluate which one is better. In Section VI we show that the results of GDP comparison agree with those of the success rate and achievable rate comparisons.

V. Hierarchical Codebook Design

In this section we propose the BMW-MS approach to design a Tx hierarchical codebook based on multi-RF-chain sub-array technique.[4] In order to obtain the coefficients for each sub-array, we propose two candidate solutions. The first one is a low-complexity search (LCS) solution to optimize the GDP metric, and the second one is a closed-form (CF) solution to pursue flat beam patterns. Hence, the BMW-MS approach with the two solutions are termed as BMW-MS/LCS and BMW-MS/CF, respectively.

It is noteworthy that when letting $|\mathbf{f}_{\text{BB}}| = \mathbf{1}$ in (8) the structure of the Tx codeword can be further written as

$$\mathbf{w} = \sum_{i=1}^{M_{\text{RF}}} \mathbf{v}_i, \tag{21}$$

where $\mathbf{v}_i = [\mathbf{F}_{\text{RF}}]_{:,i}$ is the RF weight vector (RWV) of the i-th RF chain, and the phases of $\mathbf{f}_{\text{BB}}$ have been absorbed into those of $\mathbf{v}_i$; thus we have in fact let $\mathbf{f}_{\text{BB}} = \mathbf{1}$ here.

A. The BMW-MS Approach

A critical challenge to design the hierarchical codebook shown in Fig. 2 is beam widening, i.e., to design the low-layer codewords which have wide beam widths. Intuitively, if M_{RF} is sufficiently large, wide beams can be shaped by steering these RF RWVs towards equally spaced angles within the beam coverage. This is just the PS-DFT approach [23]. However, in practice M_{RF} may be rather small. In such a case, we consider to use the sub-array technique to shape a wide beam. In particular, a large RWV of each RF chain can be divided into multiple sub-vectors (called sub-arrays), and these sub-arrays can point at different directions, such that a wider beam can be shaped.

To illustrate this, let us separate the N-element RWV of each RF chain into M_{S} sub-arrays with N_{S} elements in each sub-array, which means $N = M_{\text{S}} N_{\text{S}}$. In addition, letting $\mathbf{f}_{i,m} = [\mathbf{v}_i]_{(m-1)N_{\text{S}}+1:mN_{\text{S}}}$, we have $[\mathbf{f}_{i,m}]_n = [\mathbf{v}_i]_{(m-1)N_{\text{S}}+n}$, $m = 1, 2, \ldots, M_{\text{S}}$, $n = 1, 2, \ldots, N_{\text{S}}$, and $i = 1, 2, \ldots, M_{\text{RF}}$. $\mathbf{f}_{i,m}$ can be seen as the sub-RWV of the m-th sub-array of the i-th RF chain. Therefore, the beam gain of $\mathbf{w}$ writes

$$\begin{aligned}A(\mathbf{w}, \omega) &= \sum_{n=1}^{N} [\sum_{i=1}^{M_{\text{RF}}} \mathbf{v}_i]_n e^{-j\pi(n-1)\omega}\\ &= \sum_{n=1}^{N} \sum_{i=1}^{M_{\text{RF}}} [\mathbf{v}_i]_n e^{-j\pi(n-1)\omega}\\ &= \sum_{m=1}^{M_{\text{S}}} \sum_{n=1}^{N_{\text{S}}} \sum_{i=1}^{M_{\text{RF}}} [\mathbf{v}_i]_{(m-1)N_{\text{S}}+n} e^{-j\pi((m-1)N_{\text{S}}+n-1)\omega}\\ &= \sum_{i=1}^{M_{\text{RF}}} \sum_{m=1}^{M_{\text{S}}} \sum_{n=1}^{N_{\text{S}}} e^{-j\pi(m-1)N_{\text{S}}\omega} [\mathbf{f}_{i,m}]_n e^{-j\pi(n-1)\omega}\\ &= \sum_{i=1}^{M_{\text{RF}}} \sum_{m=1}^{M_{\text{S}}} e^{-j\pi(m-1)N_{\text{S}}\omega} A(\mathbf{f}_{i,m}, \omega),\end{aligned} \tag{22}$$

where we can find that the beam coverage of $\mathbf{w}$ can be controlled by controlling the $M_{\text{RF}} M_{\text{S}}$ sub-arrays $\mathbf{f}_{i,m}$. It is noteworthy that the coefficient between different sub-arrays is $e^{-j\pi(m-1)N_{\text{S}}\omega}$. As the coefficient depends on m and ω, it induces coupling effect between different sub-arrays of the same RF chain. When the angle gap of two adjacent sub-arrays

[4] Rx codebook design is similar.

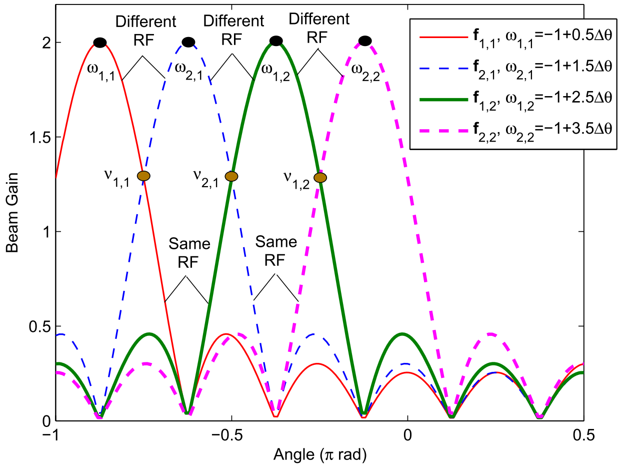


Fig. 3. Beam patterns of the sub-arrays, where $N_{\rm S} = 8$, $M_{\rm RF} = M_{\rm S} = 2$, $\Delta\theta = 2/N_{\rm S} = 0.25$, $B = 1$, and $\Omega_0 = -1$.

of the same RF chain is not wide enough, the coupling effect will be significant. In contrast, the coefficient does not depend on i. Hence, there is no coupling effect between different sub-arrays of different RF chains, which means that the steering angles of two sub-arrays of different RF chains can be close without affecting each other.

Based on the above observation, we propose the BMW-MS approach for beam widening, i.e., to cover an arbitrary angle range $[\Omega_0, \Omega_0 + B]$ with $M_{\rm RF}$ RF chains, where each RF chain is decomposed into $M_{\rm S}$ sub-arrays, and the sub-RWVs $\mathbf{f}_{i,m}$ are set to steer along the angles

$$\omega_{i,m} = \Omega_0 + (i - 1/2)\Delta\theta + (m-1)M_{\rm RF}\Delta\theta, \tag{23}$$

where $\Delta\theta = B/(M_{\rm RF}M_{\rm S})$, i.e., $\mathbf{f}_{i,m}$ satisfies

$$\mathbf{f}_{i,m} = \sqrt{\frac{N_{\rm S}}{N}} e^{j\theta_{i,m}} \mathbf{a}(N_{\rm S}, \omega_{i,m}), \tag{24}$$

where $\theta_{i,m}$ are phase parameters (in the angle domain instead of cosine angle domain) to be determined. Since the beam width of a sub-array is $2/N_{\rm S}$, $\Delta\theta$ should be no larger than $2/N_{\rm S}$; otherwise there will be sink between two adjacent sub-arrays.

An example of the beam patterns of the sub-arrays is shown in Fig. 3, where $N_{\rm S} = 8$, $M_{\rm RF} = M_{\rm S} = 2$, $\Delta\theta = 2/N_{\rm S} = 0.25$, $B = 1$, and $\Omega_0 = -1$. The intuition of this approach is explained as follows. As we want to cover an angle interval of B, and there are $M_{\rm RF}M_{\rm S}$ controllable sub-arrays in total, we can evenly steer these sub-arrays with an angle gap $\Delta\theta = B/(M_{\rm RF}M_{\rm S})$ over the desired angle range. Moreover, in order to reduce the coupling effect between adjacent sub-arrays of the same RF chain, we set their angle gaps as wide as possible; hence we try to set sub-arrays of different RF chains to steering along adjacent angles. For instance, the steering angles from small to large are $\omega_{1,1}$, $\omega_{2,1}$, $\omega_{1,2}$, $\omega_{2,2}$ in turn, such that $\omega_{1,1}$ and $\omega_{1,2}$, as well as $\omega_{2,1}$ and $\omega_{2,2}$, are largely spaced.

B. Low-Complexity Search and Closed-Form Solutions

A remaining critical issue is to determine the coefficients $\theta_{i,m}$ for the BMW-MS approach in (24). We propose two solutions as follows, i.e., a low-complexity search (LCS) and a closed-form (CF) solutions.

1) A Low-Complexity Search Solution: According to (20), the following optimization problem can be formulated:

$$\begin{aligned} &\underset{\theta_{i,m}}{\text{maximize}} \;\; \xi(\mathbf{w}, \Omega_0, B), \\ &\text{subject to } [\mathbf{v}_i]_{(m-1)N_{\rm S}+1:mN_{\rm S}} \\ &\qquad = \mathbf{f}_{i,m} = \sqrt{\frac{N_{\rm S}}{N}} e^{j\theta_{i,m}} \mathbf{a}(N_{\rm S}, \omega_{i,m}), \end{aligned} \tag{25}$$

which is a non-convex problem. Although the exhaustive grid search can be directly adopted to search over the feasible domains of $\theta_{i,m}$, it has a high computational complexity which grows exponentially with the total number of the sub-arrays. To lower the search complexity, we assume equal-difference phase sequences for the sub-arrays of the same RF chain and the sub-arrays of different RF chains, respectively, i.e., we let

$$\theta_{i,m} = m\phi_1 + i\phi_2, \tag{26}$$

where $\phi_1 \in [0, 2\pi]$ is the phase difference of the phase sequence for the sub-arrays of the same RF chain, and $\phi_2 \in [0, 2\pi]$ is the phase difference of the phase sequence for the sub-arrays of different RF chains. With this assumption, the problem (25) reduces to a 2-parameter search problem, which does not grow with the total number of sub-arrays. Thus, the computational complexity becomes affordable.

2) A Closed-Form Solution: According to (23), the sub-arrays are set to steer along $\omega_{i,m}$ with an angle gap $\Delta\theta$. This can only guarantee that the beam gains along these directions are high. To pursue a flat beam pattern, we also hope that the beam gains along the other angles between adjacent $\omega_{i,m}$ are high, such that the beam pattern is flatter. Thus, we can design $\theta_{i,m}$ to maximize the beam gains along the middle angles of adjacent $\omega_{i,m}$, i.e.,

$$\nu_{i,m} = \Omega_0 + i\Delta\theta + (m-1)M_{\rm RF}\Delta\theta, \tag{27}$$

where $im \neq M_{\rm RF}M_{\rm S}$. Fig. 3 also shows the locations of $\nu_{i,m}$.

Since $\mathbf{f}_{i,m} = \sqrt{\frac{N_{\rm S}}{N}} e^{j\theta_{i,m}} \mathbf{a}\left(N_{\rm S}, \omega_{i,m}\right)$, according to (22) the beam gain of $\mathbf{w}$ along angles $\nu_{i,m}$ can be derived as

$$\begin{aligned} A(\mathbf{w}, \nu_{k,n}) = &\frac{N_{\rm S}}{\sqrt{N}} \sum_{i=1}^{M_{\rm RF}} \sum_{m=1}^{M_{\rm S}} e^{-j\pi(m-1)N_{\rm S}\nu_{k,n}} e^{j\theta_{i,m}} \\ &\times \mathbf{a}(N_{\rm S}, \nu_{k,n})^{\rm H} \mathbf{a}\left(N_{\rm S}, \omega_{i,m}\right). \end{aligned} \tag{28}$$

It is clear that it is still complicated to determine $e^{\theta_{i,m}}$ by optimizing the absolute beam gain in (28). Notice that $|\mathbf{a}(N_{\rm S}, \omega_1)^{\rm H}\mathbf{a}(N_{\rm S}, \omega_2)|$ becomes smaller when $|\omega_1 - \omega_2|$ becomes greater from 0 to $2/N_{\rm S}$, and can be neglected when $|\omega_1 - \omega_2| > 2/N_{\rm S}$. This means that the two sub-arrays with steering angles closest to $\nu_{k,n}$ have the most significant effects on the beam gain along $\nu_{k,n}$, while the sub-arrays with steering angles far from $\nu_{k,n}$ have a little effect on the beam gain along $\nu_{k,n}$ (see Fig. 3). This motivates us to consider only the two close sub-arrays when optimizing the beam gain for simplicity. With this idea, we can finally obtain

$$\begin{aligned} \theta_{i,m} = &\,\pi m(m-1)N_{\rm S}M_{\rm RF}\Delta\theta/2 \\ &- \pi(mM_{\rm RF} + i)(N_{\rm S} - 1)\Delta\theta/2, \end{aligned} \tag{29}$$

where $\Delta\theta = B/(M_{\mathrm{RF}}M_{\mathrm{S}})$, $i = 1, 2, \ldots, M_{\mathrm{RF}}$, and $m = 1, 2, \ldots, M_{\mathrm{S}}$. The detailed derivation can be found in Appendix.

C. Codebook Generation

Up to now we have assumed that M_{RF}, M_{S} and N_{S} are known in priori. However, in practice M_{RF} is given by the system setting, while M_{S} and N_{S} are in fact determined by the beam width B of the codeword to be designed. In other words, M_{S} and N_{S} may be different for different codewords with different beam widths. Since when M_{S} is smaller N_{S} will be bigger and a higher beam gain can be provided, M_{S} should be as small as possible. As $\Delta\theta = B/(M_{\mathrm{RF}}M_{\mathrm{S}}) \leq 2/N_{\mathrm{S}}$ and $N = M_{\mathrm{S}}N_{\mathrm{S}}$, we can obtain

$$M_{\mathrm{S}} = \left\lceil \sqrt{BN/(2M_{\mathrm{RF}})} \right\rceil, \tag{30}$$

where $\lceil\cdot\rceil$ is the ceiling operation.

Recall that we need to design $\underline{\mathbf{w}}(k, n)$ instead of just $\mathbf{w}(k, n)$ itself. By exploiting the BMW-MS approach we can design $\underline{\mathbf{w}}(k, n) \triangleq (\mathbf{F}_{\mathrm{RF}(k,n)}, \mathbf{f}_{\mathrm{BB}(k,n)} = \mathbf{1})$. Recall again that different codewords within the same composite codeword share the same $\mathbf{F}_{\mathrm{RF}}$, i.e., the same $\{\mathbf{v}_i\}_{i=1,2,\ldots,M_{\mathrm{RF}}}$. This can be satisfied by using [21, Corollary 1] for beam rotation, i.e.

Corollary 1: Given the first codeword in the k-th layer $\mathbf{w}(k, 1)$*, all the other codewords in the k-th layer can be found through rotating* $\mathbf{w}(k, 1)$ *by* $\frac{2n-2}{M^k}$, $n = 2, 3, \ldots, M^k$*, respectively, i.e.,* $\mathbf{w}(k, n) = \mathbf{w}(k, 1) \circ \sqrt{N}\mathbf{a}(N, \frac{2n-2}{M^k})$.

According to Corollary 1, we have

$$\begin{aligned}
\mathbf{w}(k, n) &= \mathbf{w}(k, 1) \circ \sqrt{N}\mathbf{a}(N, \frac{2(n-1)}{M_{\mathrm{RF}}^k}) \\
&= \mathbf{F}_{\mathrm{RF}(k,1)}\mathbf{1} \circ \sqrt{N}\mathbf{a}(N, \frac{2(n-1)}{M_{\mathrm{RF}}^k}) \\
&= \mathbf{F}_{\mathrm{RF}(k,1)}\sqrt{N}\mathbf{a}(N, \frac{2(n-1)}{M_{\mathrm{RF}}^k}),
\end{aligned} \tag{31}$$

which means that all the codewords within the same layer can share the same $\mathbf{F}_{\mathrm{RF}}$.

In summary, the codebook is generated as follows, where $k = 1, \ldots, \log_{M_{\mathrm{RF}}}(M_{\mathrm{AN}})$, $N = M_{\mathrm{AN}}$.

- Split each RF chain into $M_{\mathrm{S}} = \lceil \sqrt{B_k N/(2M_{\mathrm{RF}})} \rceil$ sub-arrays, where $B_k = 2/M_{\mathrm{RF}}^k$. Let the number of antennas of each sub-array be $N_{\mathrm{S}} = N/M_{\mathrm{S}}$.
- Compute $\underline{\mathbf{w}}(k, 1) = (\mathbf{F}_{\mathrm{RF}(k,1)} = \{\mathbf{v}_i\}_{i=1,2,\ldots,M_{\mathrm{RF}}}, \mathbf{1})$, where $[\mathbf{v}_i]_{(m-1)N_{\mathrm{S}}+1:mN_{\mathrm{S}}} = \sqrt{\frac{N_{\mathrm{S}}}{N}} e^{j\theta_{i,m}} \mathbf{a}(N_{\mathrm{S}}, \omega_{i,m})$. $\omega_{i,m}$ is computed as (23) ($\Omega_0 = -1$, $B = 2/2^k$). $\theta_{i,m}$ can either be computed by solving (25) with the low-complexity search method (BMW-MS/LCS) or according to the closed-form expression (29) (BMW-MS/CF).
- Compute $\underline{\mathbf{w}}(k, n) = (\mathbf{F}_{\mathrm{RF}(k,1)}, \sqrt{N}\mathbf{a}(N, \frac{2(n-1)}{M_{\mathrm{RF}}^k}))$, where $n = 2, 3, \ldots, M_{\mathrm{RF}}^k$.

We emphasize that the main purpose of BMW-MS is to design a full hierarchical codebook shown in Fig. 2 with as less as possible RF chains, because in reality the number of available RF chains in an mmWave device would be small, e.g., typically only 2, 4 or 8. In fact, we recommend to select $M_{\mathrm{RF}} = 2$ to realize BMW-MS, because fewer RF chains help to reduce the input/output fluctuation of the power amplifiers, and with 2 RF chains BMW-MS can already achieve promising performance as we shall see from simulations later, but a larger number of RF chains can improve the efficiency of channel estimation as shown in Section III.

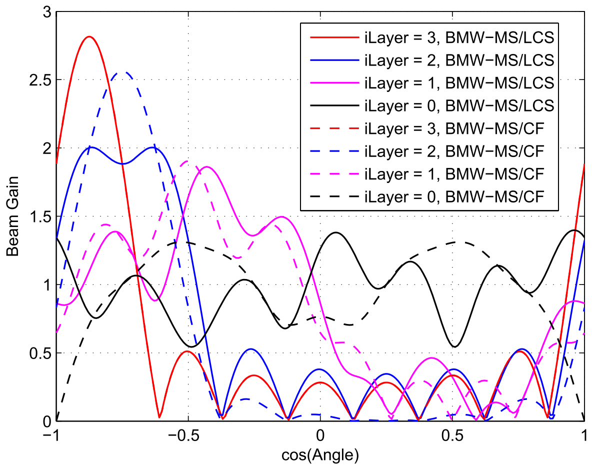


Fig. 4. Beam pattern comparison between BMW-MS/LCS and BMW-MS/CF, where $N = 8$, $M_{\mathrm{RF}} = 2$, $iLayer$ is the layer index. The dashed red line superimposes the solid red line.

In addition, we adopt a ULA model in this paper. There are also other types of antenna arrays in practice, e.g., uniform planar array (UPA) and uniform circular array (UCA). Different from ULA, both UPA and UCA are 2-dimensional arrays with both azimuth and elevation steering angles. Although the idea, i.e., multi-RF-chain sub-array, behind the proposed BMW-MS approach can be also used for the UPA and UCA models, the corresponding derivations are non-trivial and may be difficult. In particular, both the definitions of the steering vector in (2) and the beam gain in (9) must be redefined in a 2-dimensional angle domain if UPA or UCA is adopted, and all the derivations in Section V must be updated based on the new definitions of steering vector and beam gain. Subject to the structure of the 2-dimensional arrays, especially UCA, the derivations may be much more difficult than those for ULA in this paper. For this reason, we will particularly consider the extension to UPA and ULA models in our future work.

VI. Performance Evaluation

In this section we evaluate the performance of BMW-MS. We will first show the beam patterns of BMW-MS and compare them with those of the alternatives. Afterwards, we will perform extensive performance comparisons between these candidates in terms of the GDP metric, success (detection) rate and achievable rate.

Fig. 4 shows the beam pattern comparison between BMW-MS/LCS and BMW-MS/CF, where $N = 8$, $M_{\mathrm{RF}} = 2$ and $iLayer$ is the layer index. From this figure we can observe that both approaches have realized the beam coverage shown in Fig. 2. In addition, BMW-MS/CF basically provides similar beam patterns to BMW-MS/LCS, which means that the closed-form solution is also promising.

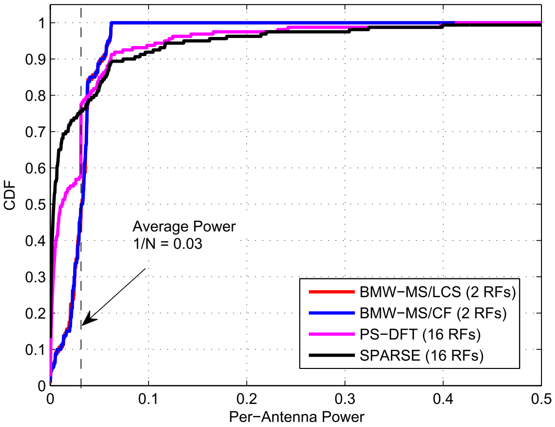


Fig. 5. CDF comparison between different codebooks. $N = 32$.

Next, we compare the performances of different codebooks. We know that when the saturation power of a power amplifier is limited, the input fluctuation significantly affects the average output power. Basically the less the fluctuation is, the higher the average output power is. Hence, we first evaluate the fluctuation of the input power of the antennas with different codebooks. To do so, as each codebook has $\log_2(N) + 1$ layers, we select one codeword from each layer of a codebook except the 0th layer, because SPARSE did not provide the 0th layer in [13]. Since each codeword has N elements, we now have $N \log_2(N)$ elements in total. Hence, we can calculate the statistics on these elements and obtain the CDF curve. Fig. 5 shows the CDF comparison between different codebooks, where $N = 32$. Note that in this figure all the codewords have unit 2-norm, i.e., the PAPC is not applied yet. We can find that for BMW-MS/LCS and BMW-MS/CF, most of the elements locate around the average power $1/N$, and the strongest element has a power about 0.06 (corresponding to $\|\mathbf{w}\|_\infty^2$ in (10)). However, for PS-DFT and SPARSE, the power of the elements disperses within a large range from 0 to more than 0.5. Hence, it is clear that BMW-MS/LCS and BMW-MS/CF have lower power fluctuation than PS-DFT and SPARSE, and according to (10) we can deduce that under the PAPC BMW-MS/LCS and BMW-MS/CF have higher MTP than PS-DFT and SPARSE.

Fig. 6 shows beam pattern comparison between different schemes with/without PAPC, where $N = 32$ and the codeword is $\mathbf{w}(1, 1)$ with coverage $[-1, 0]$ for all the schemes. When without PAPC the 2-norm of a codeword is normalized to 1, while when with PAPC the entry with the largest absolute value of the codeword is normalized to 1 according to (10). From this figure we can find that without PAPC, PS-DFT and SPARSE have flatter beam patterns than BMW-MS/LCS and BMW-MS/CF (the upper figure), but with PAPC, BMW-MS/LCS and BMW-MS/CF have higher beam gains than PS-DFT and SPARSE (the bottom figure).

Fig. 7 shows the beam pattern comparison between BMW-MS/CF and PS-DFT with/without PAPC, where $N = 64$. (a) and (b) are without PAPC, and the beam patterns match the results in [23, Fig. 4 (b: Level 1) and (b: Level 2)]. (c) and (d) are with PAPC. From this figure we can find that PS-DFT outperforms BMW-MS when there is no PAPC, because PS-DFT has flatter beam patterns. Note that PS-DFT achieves the superiority at the cost of a larger number of RF chains. The numbers of RF chains are $N/2^{iLayer}$ and 2 for PS-DFT and BMW-MS, respectively. In contrast, when PAPC is considered, which is practically reasonable in mmWave communications due to the limited performance of power amplifier, BMW-MS/CF can offer an higher beam gain.

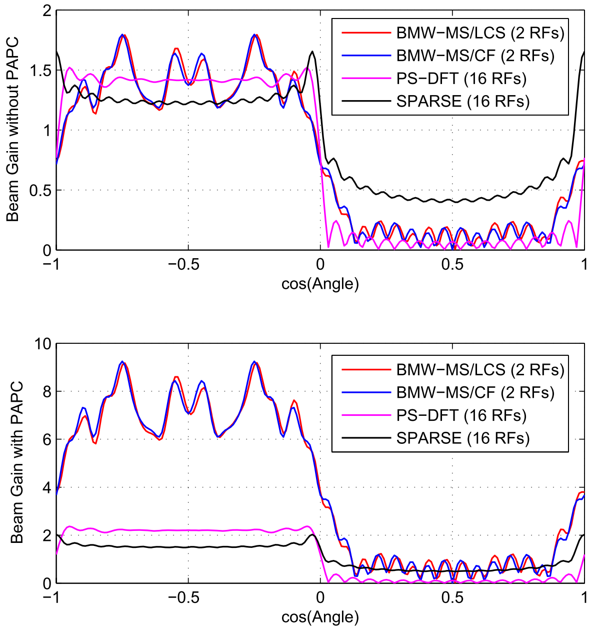
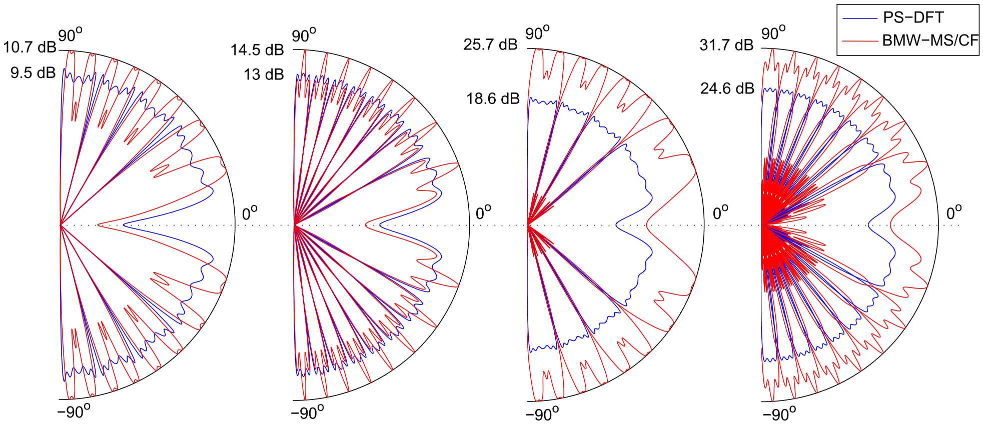


Fig. 6. Beam pattern comparison between different schemes with/without PAPC. $N = 32$, and the codeword is $\mathbf{w}(1, 1)$ with coverage $[-1, 0]$ for all the schemes.

Fig. 8 shows the GDP comparison between different schemes under PAPC. The codewords of the 1st layer is considered, because the performance of a codebook is basically determined by the widest codeword.[5] Both cases of $\gamma_{\text{PEP}} = 0$ dB (the left hand side figure) and $\gamma_{\text{PEP}} = 2$ dB (the right hand side figure) are considered. From them we can observe that BMW-MS/LCS and BMW-MS/CF have equivalent GDP performance, which is significantly better than PS-DFT and SPARSE. For SPARSE, there is a peak GDP as N increases. This is because when N is small the received SNR plays a cardinal role to determine the GDP, and thus the GDP increases with N, which increases the beam gain. However, when N is large, the received SNR is already high enough. In such a case, the beam pattern plays a cardinal role instead. Since the number of RF chains is fixed, there will appear sinks as N increases according to [13]. Thus, the GDP decreases on the contrary as N increases. Moreover, although the GDP grows when $\gamma_{\text{PEP}} = 2$ dB compared with the case of $\gamma_{\text{PEP}} = 0$ dB, the comparison results maintain, which demonstrates that the GDP metric in (19) is reasonable and robust.

Figs. 9 and 10 show the comparisons of success rate and achievable rate between different schemes under PAPC, where $M_{\text{AN}} = N_{\text{AN}} = 32$, and the channel model shown in (1) is

[5] The 0th layer codeword was not provided in [13]. Hence we prefer to compare the 1st-layer codewords.



(a) iLayer=3, Without PAPC (b) iLayer=4, Without PAPC (c) iLayer=3, With PAPC (d) iLayer=4, With PAPC

Fig. 7. Beam pattern comparison between BMW-MS/CF and PS-DFT with/without PAPC. $N = 64$. (a) and (b) are without PAPC, while (c) and (d) are with PAPC.

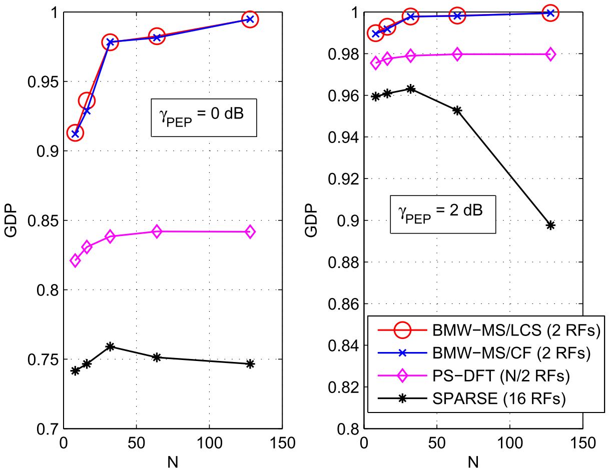


Fig. 8. GDP comparison of the 1st layer codewords between different schemes under PAPC.

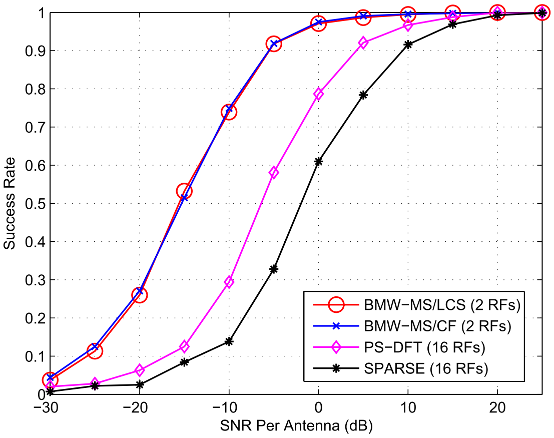


Fig. 9. Comparison of success rate between different schemes under PAPC, where $M_{\text{AN}} = N_{\text{AN}} = 32$.

adopted. $L = 1$ in the simulations, and similar results can be observed when L is set to other values. Success rate refers to the rate that an MPC is successfully acquired by Algorithm 1, while achievable rate is computed by using the best Tx/Rx precoding/combining codewords with Algorithm 1. These two figures show again that BMW-MS/LCS and BMW-MS/CF have similar overall performances, which are significantly better than PS-DFT and SPARSE. The results of these two figures have a good agreement with the GDP results shown in Fig. 8, which again demonstrates the rationality of the established GDP metric in (19).

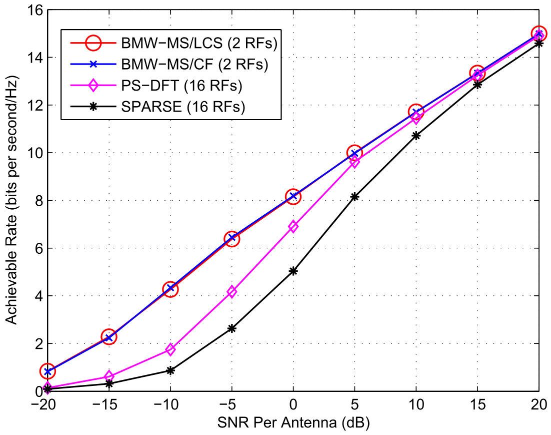


Fig. 10. Comparison of achievable rate between different schemes under PAPC, where $M_{\text{AN}} = N_{\text{AN}} = 32$.

VII. Conclusions

In this paper we have designed a hierarchical codebook taking the per-antenna power constraint into account for mmWave channel estimation with a hybrid precoding/combining structure, where multiple RF chains are used. We first propose a GDP metric to measure the performance of an arbitrary codeword incorporating both the flatness of beam pattern and the maximal transmission power. The metric not only enables a general optimization approach for codebook design, but also provides an additional way to compare the performance of different codewords/codebooks. As expected, the GDP metric positively reflects both the success detection rate performance as well as the achievable rate performance. Then BWM-MS/LCS and BMW-MS/CF have been proposed to design a hierarchical codebook by exploiting the multi-RF-chain sub-array technique, where BWM-MS/LCS optimizes the GDP metric using a simplified search method under

$$\begin{aligned}
&A(\mathbf{w}, \nu_{k,n})|_{k<M_{\mathrm{RF}},\ \mathbf{f}_{i,m}} \\
&\quad = e^{j\theta_{i,m}}\mathbf{a}\left(N_{\mathrm{S}}, \omega_{i,m}\right) = \frac{N_{\mathrm{S}}}{\sqrt{N}}\sum_{i=1}^{M_{\mathrm{RF}}}\sum_{m=1}^{M_{\mathrm{S}}} e^{-j\pi(m-1)N_{\mathrm{S}}\nu_{k,n}}e^{j\theta_{i,m}}\mathbf{a}(N_{\mathrm{S}}, \nu_{k,n})^{\mathrm{H}}\mathbf{a}\left(N_{\mathrm{S}}, \omega_{i,m}\right) \\
&\quad \approx \frac{N_{\mathrm{S}}}{\sqrt{N}}e^{-j\pi(n-1)N_{\mathrm{S}}\nu_{k,n}}e^{j\theta_{k,n}}\mathbf{a}(N_{\mathrm{S}}, \nu_{k,n})^{\mathrm{H}}\mathbf{a}\left(N_{\mathrm{S}}, \omega_{k,n}\right) + \frac{N_{\mathrm{S}}}{\sqrt{N}}e^{-j\pi(n-1)N_{\mathrm{S}}\nu_{k,n}}e^{j\theta_{k+1,n}}\mathbf{a}(N_{\mathrm{S}}, \nu_{k,n})^{\mathrm{H}}\mathbf{a}\left(N_{\mathrm{S}}, \omega_{k+1,n}\right) \\
&\quad = \frac{1}{\sqrt{N}}e^{-j\pi(n-1)N_{\mathrm{S}}\nu_{k,n}}e^{j\theta_{k,n}}e^{-j\pi(N_{\mathrm{S}}-1)\Delta\theta/4}\frac{\sin(-\pi\,\Delta\theta N_{\mathrm{S}}/4)}{\sin(-\pi\,\Delta\theta/4)} \\
&\qquad + \frac{1}{\sqrt{N}}e^{-j\pi(n-1)N_{\mathrm{S}}\nu_{k,n}}e^{j\theta_{k+1,n}}e^{j\pi(N_{\mathrm{S}}-1)\Delta\theta/4}\frac{\sin(\pi\,\Delta\theta N_{\mathrm{S}}/4)}{\sin(\pi\,\Delta\theta/4)} \\
&\quad = \frac{1}{\sqrt{N}}\frac{\sin(\pi\,\Delta\theta N_{\mathrm{S}}/4)}{\sin(\pi\,\Delta\theta/4)}e^{-j\pi(n-1)N_{\mathrm{S}}\nu_{k,n}}e^{j\theta_{k,n}}e^{-j\pi(N_{\mathrm{S}}-1)\Delta\theta/4}\left(1 + e^{j(\theta_{k+1,n}-\theta_{k,n}+\pi(N_{\mathrm{S}}-1)\Delta\theta/2)}\right). \qquad (32)
\end{aligned}$$

$$\begin{aligned}
&A(\mathbf{w}, \nu_{M_{\mathrm{RF}},n})|_{\mathbf{f}_{i,m}} \\
&\quad = e^{j\theta_{i,m}}\mathbf{a}\left(N_{\mathrm{S}}, \omega_{i,m}\right) = \frac{N_{\mathrm{S}}}{\sqrt{N}}\sum_{i=1}^{M_{\mathrm{RF}}}\sum_{m=1}^{M_{\mathrm{S}}} e^{-j\pi(m-1)N_{\mathrm{S}}\nu_{k,n}}e^{j\theta_{i,m}}\mathbf{a}(N_{\mathrm{S}}, \nu_{k,n})^{\mathrm{H}}\mathbf{a}\left(N_{\mathrm{S}}, \omega_{i,m}\right) \\
&\quad \approx \frac{N_{\mathrm{S}}}{\sqrt{N}}e^{-j\pi(n-1)N_{\mathrm{S}}\nu_{k,n}}e^{j\theta_{k,n}}\mathbf{a}(N_{\mathrm{S}}, \nu_{k,n})^{\mathrm{H}}\mathbf{a}\left(N_{\mathrm{S}}, \omega_{k,n}\right) + \frac{N_{\mathrm{S}}}{\sqrt{N}}e^{-j\pi nN_{\mathrm{S}}\nu_{k,n}}e^{j\theta_{1,n+1}}\mathbf{a}(N_{\mathrm{S}}, \nu_{k,n})^{\mathrm{H}}\mathbf{a}\left(N_{\mathrm{S}}, \omega_{1,n+1}\right) \\
&\quad = \frac{1}{\sqrt{N}}e^{-j\pi(n-1)N_{\mathrm{S}}\nu_{k,n}}e^{j\theta_{k,n}}e^{-j\pi(N_{\mathrm{S}}-1)\Delta\theta/4}\frac{\sin(-\pi\,\Delta\theta N_{\mathrm{S}}/4)}{\sin(-\pi\,\Delta\theta/4)} \\
&\qquad + \frac{1}{\sqrt{N}}e^{-j\pi nN_{\mathrm{S}}\nu_{k,n}}e^{j\theta_{1,n+1}}e^{j\pi(N_{\mathrm{S}}-1)\Delta\theta/4}\frac{\sin(\pi\,\Delta\theta N_{\mathrm{S}}/4)}{\sin(\pi\,\Delta\theta/4)} \\
&\quad = \frac{1}{\sqrt{N}}\frac{\sin(\pi\,\Delta\theta N_{\mathrm{S}}/4)}{\sin(\pi\,\Delta\theta/4)}e^{-j\pi(n-1)N_{\mathrm{S}}\nu_{k,n}}e^{j\theta_{k,n}}e^{-j\pi(N_{\mathrm{S}}-1)\Delta\theta/4}\left(1 + e^{j[(\theta_{1,n+1}-\theta_{k,n}+\pi(N_{\mathrm{S}}-1)\Delta\theta/2)-\pi N_{\mathrm{S}}\nu_{k,n}]}\right). \qquad (33)
\end{aligned}$$

the sub-array structure, while BMW-MS/CF provides closed-form codewords to pursue flat beam patterns. Performance comparisons show that BMW-MS/LCS and BMW-MS/CF achieve very close performances, and they (with only 2 RF chains) outperform PS-DFT and SPARSE under the per-antenna power constraint.

Appendix
Derivation of (29)

As shown in Fig. 3, there are two different types of positions of $\nu_{k,n}$. The first one is $\nu_{k,n}$ with $k = 1, 2, \ldots, M_{\mathrm{RF}} - 1$. The closest steering angles to it are $\omega_{k,n}$ and $\omega_{k+1,n}$, i.e., the two corresponding sub-arrays have adjacent RF indices and the same sub-array index. The other one is $\nu_{M_{\mathrm{RF}},n}$. The closest steering angles to it are $\omega_{M_{\mathrm{RF}},n}$ and $\omega_{1,n+1}$, i.e., the two corresponding sub-arrays have adjacent sub-array indices but the RF index switches from M_{RF} to 1. The beam gain of the first type of $\nu_{k,n}$ is derived as in (32), shown at the top of this page, where we have used

$$\begin{aligned}
\sum_{i=1}^{N} e^{j(i-1)\theta} &= \frac{1-e^{jN\theta}}{1-e^{j\theta}} = \frac{e^{jN\theta/2}(e^{-jN\theta/2}-e^{jN\theta/2})}{e^{j\theta/2}(e^{-j\theta/2}-e^{j\theta/2})} \\
&= e^{j(N-1)\theta/2}\frac{\sin(N\theta/2)}{\sin(\theta/2)}. \qquad (34)
\end{aligned}$$

From (32) we can find that to optimize the absolute gain, we have

$$\theta_{k+1,n} - \theta_{k,n} = -\pi(N_{\mathrm{S}} - 1)\Delta\theta/2. \qquad (35)$$

In addition, the beam gain of the other type of $\nu_{k,n}$ is derived as in (33) on the top of this page, where we can find that to optimize the absolute gain, we have

$$\begin{aligned}
&\theta_{1,n+1} - \theta_{M_{\mathrm{RF}},n} \\
&\quad = -\pi(N_{\mathrm{S}} - 1)\Delta\theta/2 + \pi N_{\mathrm{S}}(M_{\mathrm{RF}}\Delta\theta + (n-1)M_{\mathrm{RF}}\Delta\theta) \\
&\quad = -\pi(N_{\mathrm{S}} - 1)\Delta\theta/2 + \pi N_{\mathrm{S}} n M_{\mathrm{RF}}\Delta\theta \qquad (36)
\end{aligned}$$

Based on (35) and (36), we finally obtain (29).

References

[1] 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.

[2] 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.

[3] A. Alkhateeb, J. Mo, N. Gonzalez-Prelcic, and R. Heath, "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–131, Dec. 2014.

[4] S. Han, I. Chih-Lin, 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.

[5] 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.

[6] 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.

[7] 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.

[8] P. Wang, Y. Li, X. Yuan, L. Song, and B. Vucetic, "Tens of gigabits wireless communications over e-band los MIMO channels with uniform linear antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3791–3805, Jul. 2014.

[9] 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.

[10] X. Gao, L. Dai, S. Han, C.-L. I, and R. W. Heath, "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.

[11] Z. Gao, L. Dai, Z. Wang, and S. Chen, "Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6169–6183, Dec. 2015.

[12] P. Xia, H. Niu, J. Oh, and C. Ngo, "Practical antenna training for millimeter wave MIMO communication," in *Proc. IEEE 68th Veh. Technol. Conf. (VTC)*, Sep. 2008, pp. 1–5.

[13] 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.

[14] A. Alkhateeb, G. Leus, and R. W. Heath, Jr., (May 2015). "Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?" [Online]. Available: https://arxiv.org/abs/1505.00299

[15] 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.

[16] M. Kokshoorn, P. Wang, Y. Li, and B. Vucetic, "Fast channel estimation for millimetre wave wireless systems using overlapped beam patterns," in *Proc. IEEE Int. Conf. Commun.(ICC)*. London, U.K., Jun. 2015, pp. 1304–1309.

[17] 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.

[18] S. K. Yong, H.-R. Shao, X. Qin, P. Xia, and C. Ngo, "System and method for antenna training of beamforming vectors by selective use of beam level training," U.S. Patent 8 280 445 B2, Oct. 2, 2012.

[19] *mmWave Multi-Resolution Beamforming*, document IEEE 802.15-08-0182-00-003c, Jan. 2008.

[20] 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, Jan. 2015.

[21] 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.

[22] E. A. Firouzjaei, "mm-wave phase shifters and switches," Ph.D. dissertation, Elect. Eng. Comput. Sci., Univ. California, Berkeley, Berkeley, CA, USA, 2010.

[23] S. Noh, M. D. Zoltowski, and D. J. Love, "Multi-Resolution Codebook Based Beamforming Sequence Design in Millimeter-Wave Systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.

[24] Y. Jin, M. A. T. Sanduleanu, E. A. Rivero, and J. R. Long, "A millimeter-wave power amplifier with 25dB power gain and +8dBm saturated output power," in *Proc. Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2007, pp. 276–279.

[25] Y. Zhao and J. R. Long, "A Wideband, dual-path, millimeter-wave power amplifier with 20 dBm output power and PAE above 15% in 130 nm SiGe-BiCMOS," *IEEE J. Solid-State Circuits*, vol. 47, no. 9, pp. 1981–1997, Sep. 2012.

[26] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. Heath, Jr, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[27] 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.

[28] 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.

[29] 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.

[30] Z. Xiao, P. Xia, and X.-G. Xia. (Mar. 2016) "Low complexity hybrid precoding and channel estimation based on hierarchical multi-beam search for millimeter-wave MIMO systems." [Online]. Available: https://arxiv.org/abs/1603.01634

[31] J. G. Proakis, *Digital communications*. 5th ed. New York, NY, USA: McGraw Hill, 2007.

[32] Z. Xiao, C. Zhang, D. Jin, and N. Ge, "GLRT Approach for Robust Burst Packet Acquisition in Wireless Communications," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1127–1137, Mar. 2013.

Zhenyu Xiao (M'14) 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 Electrical Engineering Department, Tsinghua University. From 2013 to 2016 he was a Lecturer with the Department of Electronic and Information Engineering, Beihang University, Beijing, where he is currently an Associate Professor.

He 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, multiple antenna technology, millimeter-wave communications, and UAV networks. He has been a TPC member of the IEEE GLOBECOM'12, the IEEE WCSP'12, and the 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 (S'03–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 has authored over 50 IEEE journal and conference papers. He holds over 70 U.S. patents and/or patent applications. His research interests lie at the intersection of wireless communications, wireless networks, 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 co-edited the first book on 60 GHz millimeter wave communication systems. He also serves as an Editor of the IEEE Transactions on Signal Processing and the *IEEE Communications Magazine*. He received the IEEE Signal Processing Society Best Paper Award in 2011.

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

He was a Senior/Research Staff Member with Hughes Research Laboratories, Malibu, CA, USA, from1995 to 1996. In 1996, he joined the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA, where he is currently a 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 authored the book *Modulated Coding for Intersymbol Interference Channels* (Marcel Dekker, 2000).

Dr. Xia is a Technical Program Chair of the Signal Processing Symposium, Globecom 2007, Washington, DC, USA, and the General Co-Chair of ICASSP 2005, Philadelphia, PA, USA. He is currently serving and has served as an Associate Editor for numerous international journals, including the IEEE Wireless Communications Letters, 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. He received the National Science Foundation Faculty Early CareerDevelopment Program Award in 1997, the Office of Naval Research Young 1071 Investigator Award in 1998, and the Outstanding Overseas Young Investigator Award from the National Nature Science Foundation of China in 2001. He also received the Outstanding Junior Faculty Award of the Engineering School of the University of Delaware in 2001.