Joint Activity Detection and Channel Estimation for Massive IoT Access Based on Millimeter-Wave/Terahertz Multi-Panel Massive MIMO

Hanlin Xiu, Zhen Gao, Anwen Liao, Yikun Mei, Dezhi Zheng, Shufeng Tan, Marco Di Renzo, *Fellow, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—The multi-panel array, as a state-of-the-art antenna-in-package technology, is very suitable for millimeter-wave (mmWave)/terahertz (THz) systems, due to its low-cost deployment and scalable configuration. But in the context of non-uniform array structures it leads to intractable signal processing. Based on such an array structure at the base station, this paper investigates a joint active user detection (AUD) and channel estimation (CE) scheme based on compressive sensing (CS) for application to the massive Internet of Things (IoT). Specifically, by exploiting the structured sparsity of mmWave/THz massive IoT access channels, we firstly formulate the multi-panel massive multiple-input multiple-output (mMIMO)-based joint AUD and CE problem as a multiple measurement vector (MMV)-CS problem. Then, we harness the expectation maximization (EM) algorithm to learn the prior parameters (i.e., the noise variance and the sparsity ratio) and an orthogonal approximate message passing (OAMP)-EM-MMV algorithm is developed to solve this problem. Our simulation results verify the improved AUD and CE performance of the proposed scheme compared to conventional CS-based algorithms.**

***Index Terms*—Massive IoT access, multi-panel mMIMO, active user detection, channel estimation, millimeter-wave, terahertz.**

I. Introduction

Multi-panel massive multiple-input multiple-output (mMIMO) is a viable array configuration to realize the future millimeter-wave (mmWave)/terahertz (THz) communications [1], [2]. Specifically, the antenna elements are integrated into a uniform planar array (UPA) to create a panel, and multiple panels are juxtaposed to form the multi-panel mMIMO array shown in Fig. 1. As a partially-connected hybrid MIMO architecture relying on a modest number of RF chains, multi-panel mMIMO schemes exhibit high energy efficiency [2]. Moreover, compared to conventional mMIMO arrays having half-wavelength antenna spacing, multi-panel arrays have advantages of low-cost deployment and flexible configurations [2]. However, the resultant non-uniformly spaced arrays pose challenging on signal processing [3].

In addition, the next-generation communications are expected to support the high-throughput uplink transmission, including the applications of Internet of Things (IoT), Internet of Vehicles (IoV), and meta-universe, where efficient massive IoT access protocols are a prerequisite [4]–[6]. Sophisticated techniques have been proposed in the literature [7]–[10], [12]–[14] for the joint active user detection (AUD) and channel estimation (CE) in support of massive IoT access. In [7], by exploiting both the active user sparsity and the joint sparsity structures observed at multiple receive antennas, a modified Bayesian compressive sensing (CS) algorithm was proposed for joint AUD and CE. The authors of [8] proposed an orthogonal matching pursuit (OMP)-based joint AUD and time-domain CE technique for grant-free massive IoT access. Similar to other greedy algorithms, this detector fails to effectively harness any *a priori* information, and the associated high-dimensional matrix inversion imposes excessive complexity. To reduce the complexity, an approximate message passing (AMP) algorithm based joint AUD and CE scheme was developed in [9], but this AMP design requires the prior distributions of wireless channels and the noise variance to be known, which are hard to acquire in practice. In [10], by exploiting both the active user sparsity and the joint sparsity observed at the multiple receive antennas, an efficient low-complexity expectation propagation-based algorithm was proposed under the Bayesian framework for joint AUD and CE. In [11], the authors proposed a deep learning based AUD and CE in the grant-free non-orthogonal multiple access (NOMA) systems, where deep learning figured out the direct mapping between the received NOMA signal and the indices of active devices and associated channels using the long short-term memory. However, the schemes in [5]–[11] have not considered mMIMO systems. As a further advance, the authors of [12] designed an mMIMO-based three-phase transmission protocol, which consist of joint AUD and CE conceived for uplink and downlink data transmission in massive cellular IoT access. To solve the joint AUD and CE problem in grant-free random access over a given coherence interval, the authors of [13] proposed a logarithmic smoothening method for handling a non-smooth objective function. Based on the structured sparsity of the channel matrix, a generalized multiple measurement vector (GMMV)-AMP algorithm was proposed for the uplink of broadband massive IoT access systems [14]. However, the fully-digital mMIMO considered in [12]–[14] suffer from prohibitively high hardware cost and power consumption. We provide a brief summary of the related literature in Table I. Furthermore, when the sensing matrices are ill-conditioned, the mean square error (MSE) performance and the convergence speed of the orthogonal AMP (OAMP) algorithm proposed in [15] outperforms the existing AMP algorithms. However, the conventional OAMP algorithm is restricted to the single measurement vector (SMV) CS problem. Moreover, the OAMP algorithm requires the *a priori* distribution to be known, whose parameters are difficult to obtain in the realistic

The work of Z. Gao is supported by Natural Science Foundation of China (NSFC) under Grant 62071044. L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/W016605/1 and EP/P003990/1 (COALESCE) as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028) (*corresponding author: Zhen Gao*).

Hanlin Xiu, Zhen Gao, Anwen Liao, Yikun Mei, Shufeng Tan, and Dezhi Zheng are with the School of Information and Electronics and the Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: gaozhen16@bit.edu.cn).

Marco Di Renzo is with the Laboratoire des Signaux et Systmes, CNRS, CentraleSuplec, University Paris Sud, Universit Paris-Saclay, 91192 Paris, France (e-mail: marco.direnzo@centralesupelec.fr)

Lajos Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

Table I: A brief comparison of the related literature

Contents \ Literature		[2], [3]	[7]–[11]	[12]–[14]	Proposed
BS	1/2/4 Antennas		✓		
	Fully-digital mMIMO			✓	
	Multi-panel mMIMO	✓(Linear Array)			✓(Planar Array)
Processing at BS	CE	✓	✓	✓	✓
	AUD		✓	✓	✓

communication systems.

In this paper, we study the multi-panel mMIMO operating at mmWave/THz frequency for high-throughput massive IoT access. Specifically, a CS-based joint AUD and CE scheme is proposed in support of the high-efficient uplink access, where the multi-panel MIMO array at the BS adopts a partially-connected hybrid architecture. We introduce the mmWave/THz multi-panel mMIMO channel model for the first time. By exploiting the structured sparsity of massive IoT access channels, the joint AUD and CE problem can be formulated as a multiple measurements vector (MMV) problem under the CS framework. To solve this MMV-CS problem in the massive IoT access based on the multi-panel mMIMO system, we develop an OAMP-expectation maximization (EM)-MMV algorithm, where the EM algorithm can adaptively learn some unknown parameters, i.e., the noise variance and the sparsity ratio. Moreover, the sensing matrix of the multi-panel system can be easily designed to be a partially unitary matrix, so that the computational complexity of the proposed OAMP-EM-MMV algorithm can be reduced and the signal processing challenges of the associated non-uniform array can be mitigated. Finally, our simulation results verify that the OAMP-EM-MMV algorithm proposed for joint AUD and CE has a better performance than conventional CS-based algorithms.

Notations: Boldface lower and upper-case symbols denote column vectors and matrices, respectively. The superscripts $(\cdot)^{\mathrm{T}}$, $(\cdot)^{\mathrm{H}}$, and $(\cdot)^{-1}$ denote the transpose, conjugate transpose, and matrix inversion operators, respectively; $\|\mathbf{a}\|_2$ and $\|\mathbf{A}\|_F$ are the ℓ_2-norm of $\mathbf{a}$ and the Frobenius norm of $\mathbf{A}$, respectively; $\otimes$ denotes the Kronecker product operation; $\mathbf{0}_N$ and $\mathbf{I}_N$ represent the vector of size N with all the elements being 0 and the $N\times N$ identity matrix, respectively; $\mathrm{vec}[\mathbf{A}]$ stacks the columns of $\mathbf{A}$ on top of each other; $\mathrm{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$ that calculates the sum of the diagonal elements of $\mathbf{A}$; $\mathcal{CN}(x;m,\sigma^2)$ denotes the complex Gaussian distribution with expectation m and covariance σ^2. $\mathbf{D}_N$ denotes the $N\times N$ discrete Fourier transform matrix with (m,n)th element equal to $e^{-j2\pi(m-1)(n-1)/N}$. Finally, $\mathbb{E}(\cdot)$, $\mathrm{var}[\cdot]$, and $\Re\{\cdot\}$ denote the expectation, the variance, and the real part of the argument, respectively.

II. System Model

We consider a multi-panel mmWave/THz mMIMO system, where the BS equipped with a rectangular array serves K potential single antenna UEs in uplink massive IoT access scenarios, as shown in Fig. 1. The BS adopts the multi-panel structure in conjunction with a partially-connected hybrid MIMO. The specific configuration of the rectangular antenna array is as follows. The number of subarray panels is $N_{\mathrm{P}}=I_{\mathrm{h}}I_{\mathrm{v}}$ with each of the subarray panels being a UPA, where I_{h} and I_{v} are the numbers of panels in the horizontal and vertical directions, respectively. We define N_{h} (M_{h}) and

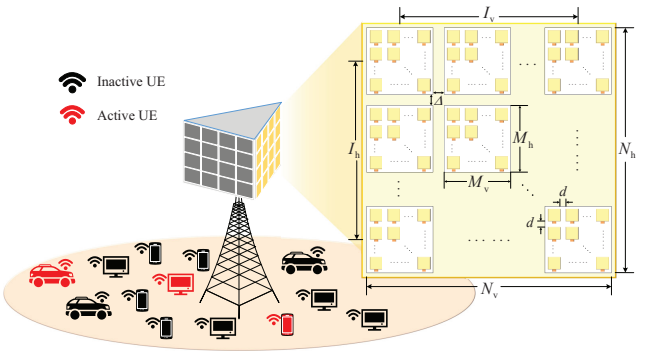

Fig. 1. Multi-panel mMIMO based massive IoT access system.

N_{v} (M_{v}) as the numbers of antennas in the horizontal and vertical directions of the rectangular array (subarray panel), respectively, i.e., $N_{\mathrm{h}}=I_{\mathrm{h}}M_{\mathrm{h}}$ and $N_{\mathrm{v}}=I_{\mathrm{v}}M_{\mathrm{v}}$. Therefore, the total number of antennas of the rectangular array is $N_{\mathrm{BS}}=N_{\mathrm{h}}N_{\mathrm{v}}$ ($M_{\mathrm{BS}}=M_{\mathrm{h}}M_{\mathrm{v}}$). The BS is equipped with N_{P} radio frequency (RF) chains, and each of them connects the corresponding subarray panel via the partially-connected phase shift network. Furthermore, the adjacent antenna spacing d within each panel of Fig. 1 is equal to $\lambda/2$, where λ is the wavelength, and the adjacent panel spacing Δ is equal to an integer multiple of d, yielding $\Delta=Dd$ for $D\geq 2$.

To combat the multipath effect at the BS caused by different scatterers in the communication environment, an orthogonal frequency-division multiplexing (OFDM) scheme having N_{c} subcarriers is applied for massive IoT access. Explicitly, P subcarriers uniformly selected from the N_{c} available subcarriers can be utilized to transmit pilot signals for joint AUD and CE. Taking the special multi-panel mMIMO structure into consideration, the mmWave/THz channel $\mathbf{h}_{p,k}\in\mathbb{C}^{N_{\mathrm{BS}}}$ between the BS and the kth UE at the pth pilot subcarrier can be formulated as

$$\mathbf{h}_{p,k}=\sum_{l=1}^{L}\beta_{k,l}\mathbf{a}_{\mathrm{MP}}(\mu_{k,l},\nu_{k,l})e^{-\mathrm{j}2\pi\varpi_{k,l}\left(-\frac{B_s}{2}+\left(\frac{pN_{\mathrm{c}}}{P}-1\right)\frac{B_s}{N_{\mathrm{c}}}\right)}, \tag{1}$$

where $1\leq k\leq K$, $1\leq p\leq P$, L is the total number of paths, $\mathbf{a}_{\mathrm{MP}}(\mu_{k,l},\nu_{k,l})\in\mathbb{C}^{N_{\mathrm{BS}}}$ is the array response vector evaluated at the horizontal and vertical virtual angles $\mu_{k,l}$ and $\nu_{k,l}$. Furthermore, $\beta_{k,l}\sim\mathcal{CN}(0,1)$ and $\varpi_{k,l}$ denote the complex gain and path delay associated with the lth path, respectively, B_{s} is system bandwidth, and N_{c}/P is an integer. Specifically, by defining the horizontal and vertical virtual angles $\mu_{k,l}=\pi\sin\theta_{k,l}\cos\phi_{k,l}$ and $\nu_{k,l}=\pi\sin\phi_{k,l}$ with $\theta_{k,l}$ and $\phi_{k,l}$ being the azimuth and elevation angles, respectively, $\mathbf{a}_{\mathrm{MP}}(\mu_{k,l},\nu_{k,l})$ in (1) can be acquired by the vectorization of $\mathbf{A}(\mu_{k,l},\nu_{k,l})=\mathbf{a}_{\mathrm{h}}(\mu_{k,l})\mathbf{a}_{\mathrm{v}}^{\mathrm{T}}(\nu_{k,l})$. Explicitly, we have $\mathbf{a}_{\mathrm{MP}}(\mu_{k,l},\nu_{k,l})=\mathrm{vec}\left[\mathbf{A}(\mu_{k,l},\nu_{k,l})\right]=\mathbf{a}_{\mathrm{v}}(\nu_{k,l})\otimes\mathbf{a}_{\mathrm{h}}(\mu_{k,l})$, while $\mathbf{a}_{\mathrm{h}}(\mu_{k,l})=\mathbf{a}_{\mathrm{h}}^{\mathrm{I}}(\mu_{k,l})\otimes\mathbf{a}_{\mathrm{h}}^{\mathrm{M}}(\mu_{k,l})\in\mathbb{C}^{N_{\mathrm{h}}}$ and $\mathbf{a}_{\mathrm{v}}(\nu_{k,l})=\mathbf{a}_{\mathrm{v}}^{\mathrm{I}}(\nu_{k,l})\otimes\mathbf{a}_{\mathrm{v}}^{\mathrm{M}}(\nu_{k,l})\in\mathbb{C}^{N_{\mathrm{v}}}$ are the horizontal and vertical steering vectors, respectively, in which the vectors $\mathbf{a}_{\mathrm{h}}^{\mathrm{I}}(\mu_{k,l})\in\mathbb{C}^{I_{\mathrm{h}}}$, $\mathbf{a}_{\mathrm{h}}^{\mathrm{M}}(\mu_{k,l})\in\mathbb{C}^{M_{\mathrm{h}}}$, $\mathbf{a}_{\mathrm{v}}^{\mathrm{I}}(\nu_{k,l})\in\mathbb{C}^{I_{\mathrm{v}}}$, and $\mathbf{a}_{\mathrm{v}}^{\mathrm{M}}(\nu_{k,l})\in\mathbb{C}^{M_{\mathrm{v}}}$ can be further written as

$$\mathbf{a}_{\mathrm{h}}^{\mathrm{I}}(\mu_{k,l})=\left[1,e^{\mathrm{j}(M_{\mathrm{h}}+D-1)\mu_{k,l}},\cdots,e^{\mathrm{j}(I_{\mathrm{h}}-1)(M_{\mathrm{h}}+D-1)\mu_{k,l}}\right]^{\mathrm{T}},$$
$$\mathbf{a}_{\mathrm{h}}^{\mathrm{M}}(\mu_{k,l})=\left[1,e^{\mathrm{j}\mu_{k,l}},\cdots,e^{\mathrm{j}(M_{\mathrm{h}}-1)\mu_{k,l}}\right]^{\mathrm{T}},$$

$$\mathbf{a}_{\rm v}^{\rm I}(\nu_{k,l}) = \left[1, e^{{\rm j}(M_{\rm v}+D-1)\nu_{k,l}}, \cdots, e^{{\rm j}(I_{\rm v}-1)(M_{\rm v}+D-1)\nu_{k,l}}\right]^{\rm T},$$
$$\mathbf{a}_{\rm v}^{\rm M}(\nu_{k,l}) = \left[1, e^{{\rm j}\nu_{k,l}}, \cdots, e^{{\rm j}(M_{\rm v}-1)\nu_{k,l}}\right]^{\rm T}.$$

Due to the inherently sporadic traffic pattern of typical massive IoT access, only a small fraction of the total UE population K is activated, where the number of active UEs is $K_{\rm a}$ (usually $K_{\rm a} \ll K$). We define a binary activity indicator flag α_k as the activity of the kth UE, i.e., $\alpha_k = 1$ when the kth UE is active, and $\alpha_k = 0$ otherwise. The signal vector $\mathbf{y}_p^{(g)} \in \mathbb{C}^{N_{\rm P}}$ received at the BS from the K UEs at the pth pilot subcarrier of the gth OFDM symbol can be expressed as

$$\begin{aligned}\mathbf{y}_p^{(g)} &= \left(\mathbf{W}_{\rm RF}^{(g)}\mathbf{W}_{\rm BB}\right)^{\rm H}\sum\nolimits_{k=1}^{K}\alpha_k\mathbf{h}_{p,k}s_{p,k}^{(g)} + \mathbf{n}_p^{(g)} \\ &= \left(\mathbf{W}_{\rm RF}^{(g)}\mathbf{W}_{\rm BB}\right)^{\rm H}\mathbf{H}_p\mathbf{s}_p^{(g)} + \mathbf{n}_p^{(g)}, \end{aligned} \tag{2}$$

where $\mathbf{W}_{\rm RF}^{(g)} \in \mathbb{C}^{N_{\rm BS}\times N_{\rm P}}$ and $\mathbf{W}_{\rm BB} \in \mathbb{C}^{N_{\rm P}\times N_{\rm P}}$ denote the analog and digital combining matrices, respectively, $\mathbf{H}_p = [\alpha_1\mathbf{h}_{p,1}, \alpha_2\mathbf{h}_{p,2}, \cdots, \alpha_K\mathbf{h}_{p,K}] \in \mathbb{C}^{N_{\rm BS}\times K}$ is the channel matrix, $\mathbf{s}_p^{(g)} = \left[s_{p,1}^{(g)}, s_{p,2}^{(g)}, \cdots, s_{p,K}^{(g)}\right]^{\rm T} \in \mathbb{C}^K$ denotes the pilot signal vector, which is randomly selected from the columns of $\mathbf{D}_K$. and $\mathbf{n}_p^{(g)} = \left(\mathbf{W}_{\rm RF}^{(g)}\mathbf{W}_{\rm BB}\right)^{\rm H}\bar{\mathbf{n}}_p^{(g)}$ is the noise vector with $\bar{\mathbf{n}}_p^{(g)} \in \mathbb{C}^{N_{\rm BS}}$ being the additive white Gaussian noise (AWGN), i.e., $\bar{\mathbf{n}}_p^{(g)} \sim \mathcal{CN}(\mathbf{0}_{N_{\rm BS}}, \sigma^2\mathbf{I}_{N_{\rm BS}})$. Observe that when $\alpha_k = 1$, the elements of the kth column of $\mathbf{H}_p$ are nonzero. With the definition of the binary activity indicator flag α_k and the combination between α_k and $\mathbf{h}_{p,k}$ in $\mathbf{H}_p$, the activity of UEs can be fully embedded in the channel matrix $\mathbf{H}_p$, which inspires us to jointly estimate the channel and detect the UEs' activity simultaneously.

We assume the digital combining matrix to be an identity matrix, i.e., $\mathbf{W}_{\rm BB} = \mathbf{I}_{N_{\rm P}}$. To design $\mathbf{W}_{\rm RF}^{(g)}$, we first construct a partial unitary matrix $\mathbf{Z}^{(g)} = \mathbf{D}_{N_{\rm v}} \otimes \mathbf{D}_{N_{\rm h}}\mathbf{P} \in \mathbb{C}^{N_{\rm BS}\times N_{\rm P}}$, where the modulus of the elements in $\mathbf{Z}^{(g)}$ is 1 and $\mathbf{P}$ is a permutation matrix which consists of $N_{\rm P}$ columns randomly extracted from $\mathbf{I}_{N_{\rm BS}}$. For our partially-connected multi-panel array architecture at the BS, we initialize the n_pth column of $\mathbf{W}_{\rm RF}^{(g)}$ that corresponds to the n_pth RF chain as $\mathbf{w}_{n_p}^{(g)} = \mathbf{0}_{N_{\rm BS}}$, then let $[\mathbf{w}_{n_p}^{(g)}]_{\mathcal{I}_{n_p}} = \frac{1}{\sqrt{M_{\rm BS}}}[\mathbf{z}_{n_p}^{(g)}]_{\mathcal{I}_{n_p}}$, where the ordered set $\mathcal{I}_{n_p}$ having a cardinality of $M_{\rm BS}$ denotes the antenna index of the n_pth subarray panel. Note that the design of fully-digital MIMO architecture does not have the constraints of $\mathbf{W}_{\rm RF}^{(g)}$. By contrast, this paper considers the multi-panel mMIMO with partially-connected hybrid MIMO architecture, which leads to the extra hardware constrains and poses the challenging on algorithm design. In Section III, we will formulate the joint AUD and CE scheme in the massive IoT access with multi-panel mMIMO system.

III. Proposed Joint AUD and CE Scheme

In this section, we will formulate the joint AUD and CE scheme as a CS-based MMV problem with the utilization of the structured sparsity of massive IoT access channels. Furthermore, to solve this MMV-CS problem, the OAMP-EM-MMV algorithm is conceived where the EM algorithm learns the unknown parameters, i.e., the noise variance and the sparsity ratio.

A. Formulation of Massive IoT Access in Multi-Panel mMIMO

We firstly focus on the received signal vector $\mathbf{y}_p^{(g)}$ in (2). By applying the vectorization rule $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^{\rm T}\otimes\mathbf{A})\cdot\mathrm{vec}(\mathbf{B})$, the signal vector $\mathbf{y}_p^{(g)}$ can be rewritten as

$$\mathbf{y}_p^{(g)} = \mathbf{F}_p^{(g)}\mathbf{h}_p + \mathbf{n}_p^{(g)}, \tag{3}$$

where $\mathbf{F}_p^{(g)} = (\mathbf{s}_p^{(g)})^{\rm T}\otimes\left(\mathbf{W}_{\rm RF}^{(g)}\right)^{\rm H} \in \mathbb{C}^{N_{\rm P}\times J}$, $\mathbf{h}_p = \mathrm{vec}(\mathbf{H}_p) \in \mathbb{C}^J$, and $J = KN_{\rm BS}$. The structured sparsity of the pth subchannel $\mathbf{H}_p$ is preserved in the vector $\mathbf{h}_p$. Note that when the kth UE is active, the elements in $\mathbf{h}_p$ having indices from the $((k-1)N_{\rm BS}+1)$th to the $kN_{\rm BS}$th are nonzero, which inspires us that UEs' activity can be detected according to the position of non-zero elements and the structured sparsity of channel. Furthermore, we consider the same signal vector used at all pilot subcarriers, i.e., $\mathbf{s}_p^{(g)} = \mathbf{s}^{(g)}$ and thus $\mathbf{F}_p^{(g)} = \mathbf{F}^{(g)}$ for $1 \le p \le P$. By aggregating the received signals at the P pilot subcarriers of the gth OFDM symbol as $\mathbf{Y}^{(g)} \in \mathbb{C}^{N_{\rm P}\times P}$, we have

$$\mathbf{Y}^{(g)} = \left[\mathbf{y}_1^{(g)}, \mathbf{y}_2^{(g)}, \cdots, \mathbf{y}_P^{(g)}\right] = \mathbf{F}^{(g)}\mathbf{H} + \mathbf{N}^{(g)}, \tag{4}$$

where $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_P] \in \mathbb{C}^{J\times P}$ and $\mathbf{N}^{(g)}$ denote the aggregated channel and noise matrices, respectively.

It can be observed from (4) that, according to the identical UE activity α_k, for $1 \le k \le K$, observed at all subchannels, the aggregated channel matrix $\mathbf{H}$ exhibits the intrinsically structured sparsity. More explicitly, its columns, i.e., $\{\mathbf{h}_p\}_{p=1}^P$, have a common sparsity pattern (a. k. a. sparse support set) in the frequency domain, given by

$$\mathrm{supp}\{\mathbf{h}_1\} = \mathrm{supp}\{\mathbf{h}_2\} = \cdots = \mathrm{supp}\{\mathbf{h}_P\}, \tag{5}$$

where $\mathrm{supp}\{\cdot\}$ denotes an ordered set consisting of the non-zero elements of the argument. Note that the support of $\mathbf{h}_p$ does not vary with the index of different subcarriers p, which can facilitate better CE performance.

Due to the limited observations in multi-panel mMIMO system relying on a partially-connected structure, we stack the received signal matrices in G OFDM symbols, i.e., $\mathbf{Y}^{(g)}$ for $1 \le g \le G$, to improve the joint AUD and CE performance. The stacked signal matrix $\mathbf{Y} \in \mathbb{C}^{Q\times P}$ can be expressed as

$$\mathbf{Y} = \left[(\mathbf{Y}^{(1)})^{\rm T}, (\mathbf{Y}^{(2)})^{\rm T}, \cdots, (\mathbf{Y}^{(G)})^{\rm T}\right]^{\rm T} = \mathbf{F}\mathbf{H} + \mathbf{N}, \tag{6}$$

where $Q = GN_{\rm P}$, while $\mathbf{F} = \left[(\mathbf{F}^{(1)})^{\rm T}, \cdots, (\mathbf{F}^{(G)})^{\rm T}\right]^{\rm T} \in \mathbb{C}^{Q\times J}$ and $\mathbf{N}$ represent the sensing matrix and the stacked noise matrix, respectively. The sensing matrix $\mathbf{F}$ is a partial unitary matrix, which prompts us to design our solution developed from OAMP algorithm [15]. Since $\mathbf{H}$ exhibits the structured sparsity, the joint AUD and CE based on (6) is an MMV-CS problem associated with $Q \ll J$, which can be solved by the proposed OAMP-EM-MMV algorithm introduced in the next subsection. With the estimated channel $\widehat{\mathbf{H}}$, the support of $\widehat{\mathbf{H}}$ can be utilized to detect the activity of UEs, so the proposed solution is termed as a joint AUD and CE scheme.

B. Proposed OAMP-EM-MMV Algorithm

The OAMP algorithm is developed from the AMP algorithm for solving the considered sparse signal recovery problem, while imposing a relaxed requirement on the sensing matrices [15]. When the sensing matrices are ill-conditioned transform matrices or partial unitary matrices, the performance of the AMP algorithm is not guaranteed, while the OAMP algorithm

has improved robustness and performs still well as demonstrated in [15]. Specifically, the OAMP algorithm includes both a linear estimation (LE) module and a non-linear estimation (NLE) module, which are activated iteratively. The output of the NLE module is the MMSE estimate. Next, we elaborate on the proposed OAMP-EM-MMV algorithm.

For the sparse channel matrix $\mathbf{H}$ in (4), the entries $h_{j,p}$ can be reasonably assumed to follow the Bernoulli-Gaussian distribution [15], and $\lambda_{j,p}$ denotes the sparsity ratio representing the non-zero probability of $h_{j,p}$. The proposed OAMP-EM-MMV algorithm involves T iterations between the LE and NLE modules, and we focus our attention on the tth iteration. The linear MMSE (LMMSE) estimator and the mean error variance estimator of the LE module are listed in the 5th and 6th lines of **Algorithm 1**, respectively.

The NLE module assumes that $\mathbf{h}_p$ is corrupted by an AWGN vector $\mathbf{z}_p$, i.e., we have $\mathbf{r}_p = \mathbf{h}_p + \tau_p \mathbf{z}_p$, where $\mathbf{z}_p \sim \mathcal{CN}(\mathbf{0}_J, \mathbf{I}_J)$ is independent of $\mathbf{h}_p$. The mean error variance of the NLE module at the tth iteration $(v^2)^t$ can be further calculated as

$$(v_p^2)^t = \left(\frac{1}{\overline{\omega}_p^t} - \frac{1}{(\tau_p^2)^t}\right)^{-1}, \tag{7}$$

where $\overline{\omega}_p^t = \frac{1}{J}\sum_{j=1}^{J} \mathrm{var}\left[h_{j,p}|r_{j,p}^t\right]$, and $r_{j,p}^t$ is the jth entry of $\mathbf{r}_p^t$. According to the *a priori* distribution of $h_{j,p}$ and the NLE model, the *a posteriori* distribution of $h_{j,p}$ can be represented as $p(h_{j,p}|r_{j,p}) = (1-\eta_{j,p}^t)\delta(h_{j,p}) + \eta_{j,p}^t \mathcal{CN}(h_{j,p}; 0, (\psi^2)^t)$, where $(\psi^2)^t = \frac{\rho^2 (v_p^2)^t}{\rho^2 + (v_p^2)^t}$, and

$$\eta_{j,p}^t = b_{j,p}^t / (a_{j,p}^t + b_{j,p}^t), \tag{8}$$

with $a_{j,p}^t = \frac{1-\lambda_{j,p}}{\pi((v_p^2)^t)} e^{-\frac{|r_{j,p}|^2}{(v_p^2)^t}}$ and $b_{j,p}^t = \frac{\lambda_{j,p}}{\pi(\rho^2 + (v_p^2)^t)} e^{-\frac{|r_{j,p}|^2}{\rho^2 + (v_p^2)^t}}$. When $\eta_{j,p}^t$ tends to zero, $p(h_{j,p}|r_{j,p})$ can be approximately regarded as a Dirac function, and $h_{j,p}$ tends to zero. When $\eta_{j,p}^t$ tends to one, by contrast, $h_{j,p}$ tends to be nonzero. Therefore, $\eta_{j,p}^t$ is termed as the belief indicator (BI). The *a posteriori* mean and variance can be expressed as

$$\xi_{j,p}^t = \mathbb{E}\left[h_{j,p}|r_{j,p}^t\right] = \frac{b_{j,p}^t}{a_{j,p}^t + b_{j,p}^t} \kappa_{j,p}^t, \tag{9}$$

$$\omega_{j,p}^t = \mathrm{var}\left[h_{j,p}|r_{j,p}^t\right] = \frac{b_{j,p}^t (\psi^2)^t}{a_{j,p}^t + b_{j,p}^t} + \frac{a_{j,p}^t b_{j,p}^t \left|\kappa_{j,p}^t\right|^2}{(a_{j,p}^t + b_{j,p}^t)^2}, \tag{10}$$

where $\kappa_{j,p}^t = \frac{\rho^2}{\rho^2 + (v_p^2)^t} r_{j,p}$.

As mentioned above, we revealed the theoretical basis process of the OAMP algorithm. The value of the noise variance σ^2 and the sparsity ratio $\lambda_{j,p}$ are required by the conventional OAMP algorithm. However, the exact values of these two parameters are difficult to obtain in practice, which motivates us to design adaptive parameter learning for enhancing the performance of the OAMP algorithm. Based on the above considerations, we integrate the EM algorithm into the OAMP algorithm. The EM algorithm is applied to estimate the unknown noise variance and sparsity ratio using the E step and M step, respectively,

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^t\right) = \mathbb{E}\left[\ln p(\mathbf{H}, \mathbf{Y})|\mathbf{Y}; \boldsymbol{\theta}^t\right], \tag{11}$$

$$\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^t\right), \tag{12}$$

where $\mathbb{E}\left[(\cdot)|\mathbf{Y}; \boldsymbol{\theta}^t\right]$ denotes the expectation conditioned on $\mathbf{Y}$ in conjunction with the parameters $\boldsymbol{\theta}^t = \{(\sigma^2)^t, \lambda_{j,p}^t, \forall j, p\}$.

Algorithm 1: OAMP-EM-MMV Algorithm

Require: Received signal matrix $\mathbf{Y}$, sensing matrix $\mathbf{F}$, and maximum iterations T

Ensure: Estimated channel $\widehat{\mathbf{H}}$, BIs $\eta_{j,p}, \forall j, p$

1: $\forall j, p$: Calculate $\lambda_{j,p}^0$ in (15) and $(\sigma^2)^0$ in (16);
2: $\forall p$: Initialize $\mathbf{r}_p^0 = \mathbf{0}_J$ and $(v_p^2)^0 = 1$;
3: **for** $t = 1, \cdots, T$ **do**
4: % LE module
5: LMMSE: $\forall p$: $\mathbf{r}_p^t = \mathbf{u}_p^{t-1} + \frac{J}{Q}\mathbf{F}^H\left(\mathbf{y}_p - \mathbf{F}\mathbf{u}_p^{t-1}\right)$;
6: The mean error variance estimator: $\forall p$: $(\tau_p^2)^t = \frac{J-Q}{Q}(v_p^2)^{t-1} + \frac{J}{Q}(\sigma^2)^{t-1}$;
7: % NLE module
8: $\forall j, p$: Calculate the *a posteriori* mean $\xi_{j,p}^t$ in (9) and variance $\omega_{j,p}^t$ in (10);
9: $\forall p$: Calculate the mean error variance of the NLE $(v_p^2)^t$ in (7);
10: $\forall j, p$: Update BI $\eta_{j,p}^t$ in (8);
11: % EM module
12: $\forall j, p$: Update the parameters in (14) and (17);
13: **end for**
14: $\forall j, p$: $\hat{h}_{j,p} = \xi_{j,p}^T$, and $\hat{h}_{j,p}$ is the (j,p)th element of $\widehat{\mathbf{H}}$.

The exact *a posteriori* distribution required in (11) is intractable, but we can approximate it from the OAMP algorithm. However, due to the multiple elements contained in $\boldsymbol{\theta}^t$ of (12), its joint optimization with $\boldsymbol{\theta}$ is difficult. Therefore, we adopt the so-called incremental EM algorithm, which estimates only a single parameter at each iteration, while keeping the others fixed. By taking the partial derivative of (11) with respect to each element of $\boldsymbol{\theta}$ and setting the derivatives to zero, we obtain the update rules of $\boldsymbol{\theta}$ as

$$\lambda_{j,p}^t = \eta_{j,p}^{t-1}, \forall j, p, \tag{13}$$

$$(\sigma^2)^t = \frac{1}{P}\{\sum_{p=1}^{P} \frac{1}{J}\{\sum_{j=1}^{J} |r_{j,p} - \sum_{p=1}^{P} f_{q,j}\xi_{j,p}^{t-1}|^2\} + \overline{\omega}_p^{t-1}\}, \tag{14}$$

where $f_{q,j}$ is the (q,j)th element of $\mathbf{F}$. For the initialization of (13) and (14) [16], the following expressions can be shown to be suitable

$$\lambda_{j,p}^0 = \frac{Q}{J} \max_{c>0} \frac{1 - 2J[(1+c)^2\Phi(-c) - c\phi(c)]/Q}{1 + c^2 - 2[(1+c)^2\Phi(-c) - c\phi(c)]}, \forall j, p, \tag{15}$$

$$(\sigma^2)^0 = \frac{1}{P}\sum_{p=1}^{P} \frac{||\mathbf{y}_p||_2^2}{\left(\mathrm{SNR}^0 + 1\right)Q}, \tag{16}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function and probability distribution function of the standard normal distribution, respectively. Given that the initial signal-to-noise-ratio (i.e., SNR^0) is usually unknown in practice, we set $\mathrm{SNR}^0 = 100$, which is an appropriate empirical value.

The OAMP algorithm assisted by the aforementioned EM algorithm is capable of solving the SMV problem. Furthermore, to solve the MMV problem in (6), the sparsity of $\mathbf{H}$ can be exploited and we adopt an innovative update rule to learn the structured sparsity. Since $\lambda_{j,p}$ represents the non-zero probability of $h_{j,p}$ and it is independently updated in (17), it is plausible that the sparsity of (5) cannot be exploited. In view of this fact, we can refine $\lambda_{j,p}$ as follows

$$\lambda_{j,1}^{t} = \cdots = \lambda_{j,P}^{t} = \frac{1}{P}\sum\nolimits_{p=1}^{P} \eta_{j,p}^{t-1}, \tag{17}$$

for exploiting the joint sparsity. Based on the aforementioned derivation and analysis, we summarize our OAMP-EM-MMV solution at a glance in **Algorithm 1**.

After obtaining the CE result $\widehat{\mathbf{H}}$, we propose a pair of AUD detectors based on $\widehat{\mathbf{H}}$ and $\eta_{j,p}$, respectively. Since the P subchannels share the same support over all the P subcarriers, we opt for the channel of arbitrary subcarrier, e.g., $p = 1$, to detect the UEs' activity. Given the CE result $\widehat{\mathbf{H}}$, we may readily obtain the channel $\widehat{\mathbf{H}}_1 \in \mathbb{C}^{N_{\rm BS}\times K}$, whose element is $\hat{h}_{n_{\rm BS},k}$. For the AUD, firstly a threshold function $r(x;\epsilon)$ is defined beforehand, where $r(x;\epsilon)$ equals 1 if $|x| > \epsilon$ and 0 otherwise.

In accordance with the structured sparsity of the estimated channel matrix $\widehat{\mathbf{H}}$, we define the channel gain based activity detector (CG-AD) for AUD as follows

$$\widehat{\alpha}_k = \begin{cases} 1, \frac{1}{J}\sum\limits_{n_{\rm BS}}\sum\limits_{k} r(\hat{h}_{n_{\rm BS},k};\epsilon_{\rm cg}) \geq p_{\rm cg}, \\ 0, \frac{1}{J}\sum\limits_{n_{\rm BS}}\sum\limits_{k} r(\hat{h}_{n_{\rm BS},k};\epsilon_{\rm cg}) < p_{\rm cg}, \end{cases} \tag{18}$$

where $\epsilon_{\rm cg} = 0.01\max\{|\hat{h}_{j,k}|, \forall j,k\}$ and $p_{\rm cg} = 0.9$ [14].

Furthermore, we define a BI based activity detector (BI-AD) as follows

$$\widehat{\alpha}_k = \begin{cases} 1, \frac{1}{J}\sum\limits_{n_{\rm BS}}\sum\limits_{k} r(\eta_{n_{\rm BS},k};\epsilon_{\rm bi}) \geq p_{\rm bi}, \\ 0, \frac{1}{J}\sum\limits_{n_{\rm BS}}\sum\limits_{k} r(\eta_{n_{\rm BS},k};\epsilon_{\rm bi}) < p_{\rm bi}, \end{cases} \tag{19}$$

where $\{\eta_{n_{\rm BS},k}, \forall n_{\rm BS}, k\}$ can be obtained from $\eta_{1,1}$ to $\eta_{J,1}$.

For our channel model, we set $\epsilon_{\rm bi}$ to 0.5 for convenience[1].

IV. Simulation Results

In this section, we evaluate the performance of the proposed joint AUD and CE scheme based on multi-panel mMIMO aided massive IoT access. In our simulations, the carrier frequency, bandwidth, and the number of subcarriers are 30 GHz, $B_s = 1$ GHz, and $N_{\rm c} = 256$, respectively. For the multi-panel mMIMO array at the BS, we use $I_{\rm v} = I_{\rm h} = 4$, that is $N_{\rm P} = 16$ panels, and $M_{\rm h} = M_{\rm v} = 2$ for each panel, so that the total number of antennas in this multi-panel mMIMO is $N_{\rm BS} = 64$. The adjacent panel spacing is $\Delta = 6d$, i.e., $D = 6$. Furthermore, in the channel, $L = 4$, and the path delay $\varpi_{k,l}$ follows the uniform distribution $\mathcal{U}[0, 32/B_s]$. The maximum number of iterations in **Algorithm 1** is $T = 100$ and SNR $= 30$ dB. The AUD error probability and the CE MSE defined in [14] are used as our performance metrics. Based on our simulation parameters, the transmission delay of an OFDM symbol is equal to 0.288 microsecond (μs).

Fig. 2 compares the AUD performance of different schemes versus the number of OFDM symbols G. In Fig. 2 and Fig. 3, we set $K = 500$ and $K_a = 50$, and we consider the cases of $P = 8$ and $P = 16$. We observe from Fig. 2 that the proposed OAMP-EM-MMV algorithm outperforms the other three greedy algorithms (namely the SAMP, SP, and SWOMP algorithms utilized in [14] as baseline schemes), despite using less pilot subcarriers, and has a significant advantage over the GMMV-AMP algorithm [14]. Furthermore, for the proposed OAMP-EM-MMV algorithms relying on the CG-AD and BI-AD, the AUD performance of BI-AD is distinctly better than

[1]The choice of $\epsilon_{\rm bi}$ can be further optimized according to the cost of missed detection and false alarm required by the practical communication systems.

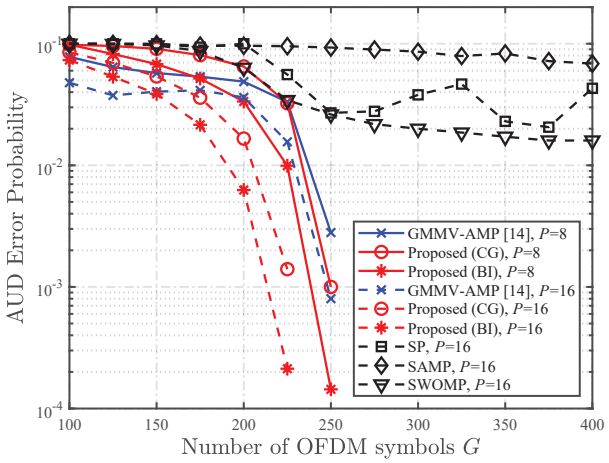

Fig. 2. AUD performance comparison of different schemes versus G, where $K = 500$ and $K_a = 50$, and we consider the cases of $P = 8$ and $P = 16$.

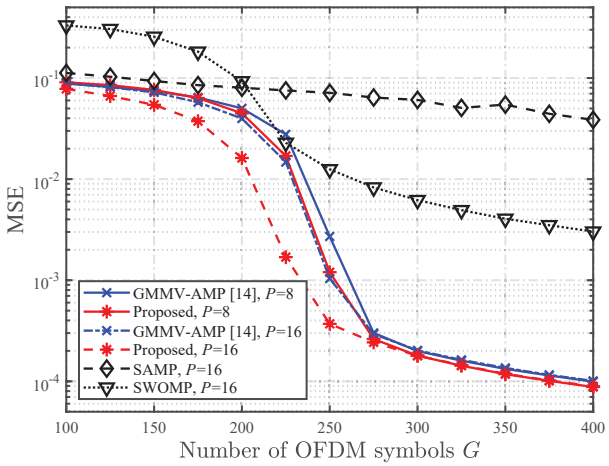

Fig. 3. CE performance comparison of different schemes versus G, where $K = 500$ and $K_a = 50$, and we consider the cases of $P = 8$ and $P = 16$.

that of CG-AD. When $G \geq 225$, the AUD performance of the CG-AD and BI-AD for $P = 16$ tends to zero quite rapidly. In the case of $P = 16$, the AUD performance of the proposed BI-AD tends to zero rapidly when $G \geq 250$. Hence, all the UEs can be detected correctly within the access latency of 72 μs.

Fig. 3 compares the MSE performance of the CE versus the number of OFDM symbols G. In Fig. 3, the MSE performance of the proposed OAMP-EM-MMV algorithm is seen to be superior to the other baseline algorithms, especially when $P = 16$. The CE accuracy of the proposed algorithm relying on less pilot subcarriers, i.e., $P = 8$, will be better than that of the baseline algorithms using $P = 16$. When $200 \leq G \leq 275$, observe from Fig. 3 that the MSE curves of the algorithms based on the message passing method decays rapidly, while these MSE curves will almost overlap when G is large enough (e.g., $G > 275$). It becomes clear from Fig. 2 and Fig. 3 that the access latency to achieve reliable joint AUD and CE performance is less than 79.2μs, which can meet the latency requirements of the IoV.

Fig. 4 compares the AUD performance of different schemes versus the number of OFDM symbols G with different ratios of active UEs. In Fig. 4 and Fig. 5, we consider $P = 16$ and $K = 400$, and K_a is set to 40, 60 and 80 so the sparsity ratio is 10%, 15%, and 20%, respectively. In the cases of $K_a = 40$ and $K_a = 60$, the AUD error probability becomes very small when the number of OFDM symbols G exceeds

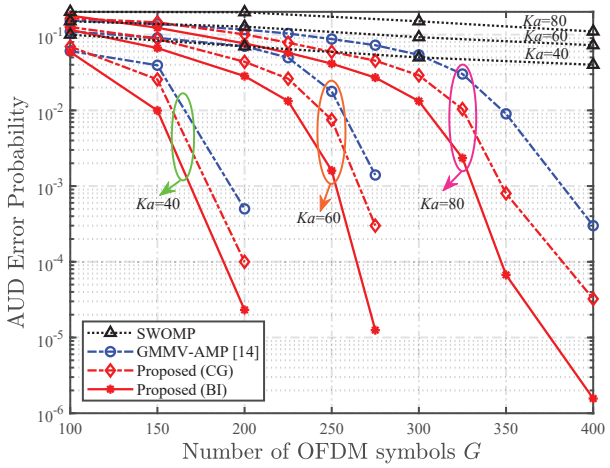

Fig. 4. AUD performance comparison of different schemes versus G, where $P=16$ and $K=400$, and we consider the cases of $K_a=40$, 60, and 80.

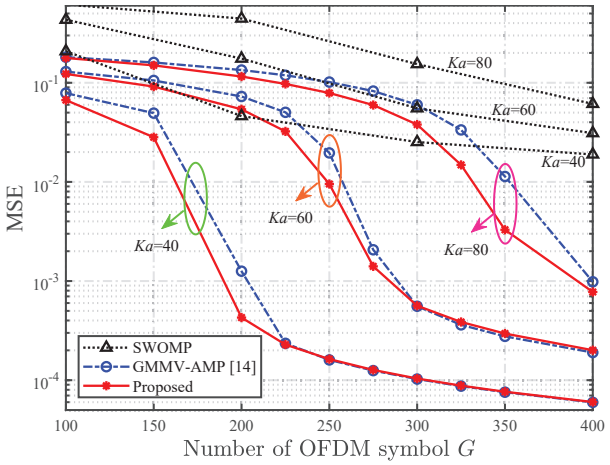

Fig. 5. CE performance comparison of different schemes versus G, where $P=16$ and $K=400$, and we consider the cases of $K_a=40$, 60, and 80.

150 and 275, respectively. It can be observed from Fig. 4 that the AUD performance of each algorithm deteriorates as the number of the active UEs and the sparsity ratio increase. While given one specific value of K_a, the proposed OAMP-EM-MMV algorithm is obviously superior to other baseline algorithms, which demonstrates the robustness of the proposed OAMP-EM-MMV algorithm. Furthermore, for the proposed OAMP-EM-MMV algorithm relying on the CG-AD and BI-AD, the AUD performance of BI-AD is better than that of CG-AD in the case of different numbers of the active UEs, which indicates that EM algorithm can update the sparsity ratio robustly when the sparsity level changes.

Fig. 5 compares the MSE performance of the CE versus the number of OFDM symbols G with different numbers of the active UEs. In the cases of $K_a = 40$ and $K_a = 60$, the MSE declines rapidly when $150 \leqslant G \leqslant 200$ and $225 \leqslant G \leqslant 300$, respectively. Given one specific value of K_a, the MSE performance of the proposed OAMP-EM-MMV algorithm is superior to other baseline algorithms in the cases of different numbers of the active UEs. Furthermore, the simulation results of Fig. 4 and Fig. 5 demonstrate the superiority of the combination between the OAMP algorithm and the EM algorithm.

V. Conclusions

In this paper, we have proposed a CS-based joint AUD and CE scheme for massive IoT access relying on mmWave/THz multi-panel mMIMO. Since the multi-panel mMIMO is a kind of partially-connected hybrid MIMO, the existing AUD and CE schemes designed for fully-digital MIMO can not perform well. Specifically, by designing the uplink combining matrix and exploiting the structured sparsity of the uplink massive IoT access channels, the joint AUD and CE problem can be formulated as an MMV-CS problem. We further develop an OAMP-EM-MMV algorithm to solve this problem by utilizing the EM algorithm to learn the *a priori* parameters, i.e., the noise variance and the sparsity ratio. Our simulation results have demonstrated that the proposed OAMP-EM-MMV algorithm based joint AUD and CE scheme achieves better AUD and CE performance than the state-of-the-art schemes.

References

[1] Y. Huang, Y. Li, H. Ren, J. Lu, and W. Zhang, "Multi-panel MIMO in 5G," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 56-61, Mar. 2018.

[2] W. Wang and W. Zhang, "Orthogonal projection-based channel estimation for multi-panel millimeter wave MIMO," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2173-2187, Apr. 2020.

[3] Y. Zhang, Y. Huo, D. Wang, X. Dong and X. You, "Channel estimation and hybrid precoding for distributed phased arrays based MIMO wireless communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 12921-12937, Nov. 2020.

[4] J. Wang, Z. Zhang, and L. Hanzo, "Joint active user detection and channel estimation in massive access systems exploiting Reed-Muller sequences," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 3, pp. 739-752, Jun. 2019.

[5] Y. Liu, L. Yang, and L. Hanzo, "Sparse space-time-frequency-domain spreading for large-scale non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12327-12332, Oct. 2020.

[6] Y. Liu, L. Yang, and L. Hanzo, "Joint user-activity and data detection for grant-free spatial-modulated multi-carrier non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11673-11684, Oct. 2020.

[7] X. Xu, X. Rao, and V. K. N. Lau, "Active user detection and channel estimation in uplink C-RAN systems," *IEEE Int. Conf. Commun. (ICC)*, 2015, pp. 2727-2732.

[8] S. Park, H. Seo, H. Ji, and B. Shim, "Joint active user detection and channel estimation for massive machine-type communications," *IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2017, pp. 1-5.

[9] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890-1904, Apr. 2018.

[10] J. Ahn, B. Shim, and K. B. Lee, "EP-based joint active user detection and channel estimation for massive machine-type communications," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5178-5189, Jul. 2019.

[11] Y. Ahn, W. Kim and B. Shim, "Active user detection and channel estimation for massive machine-type communication: Deep learning approach," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 11904-11917, Jul. 2022.

[12] X. Shao, X. Chen, C. Zhong, J. Zhao, and Z. Zhang, "A unified design of massive access for cellular Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3934-3947, Apr. 2019.

[13] X. Shao, X. Chen and R. Jia, "A dimension reduction-based joint activity detection and channel estimation algorithm for massive access," *IEEE Trans. Signal Process.*, vol. 68, pp. 420-435, 2020.

[14] M. Ke, Z. Gao, Y. Wu, X. Gao and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764-779, 2020.

[15] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020-2033, 2017.

[16] J. P. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE Trans. Signal Proccess.*, vol. 61, pp. 4658-4672, 2013.