Active Terminal Identification, Channel Estimation, and Signal Detection for Grant-Free NOMA-OTFS in LEO Constellation Internet-of-Things

Xingyu Zhou, Keke Ying, Zhen Gao, Yongpeng Wu, Zhenyu Xiao, Symeon Chatzinotas, Jinhong Yuan, and Björn Ottersten

Abstract

This paper investigates the massive connectivity of low Earth orbit (LEO) satellite constellations based Internet-of-Things (IoT) for seamless global coverage. Specifically, we propose to integrate the grant-free non-orthogonal multiple access (NOMA) paradigm for random access with the emerging orthogonal time frequency space (OTFS) modulation for accommodating the massive IoT access. So that the long round-trip latency and severe Doppler effect of terrestrial–satellite links (TSLs) can be mitigated. Furthermore, we put forward a two-stage joint active terminal identification (ATI) and channel estimation (CE) scheme as well as a low-complexity multi-user signal detection (SD) method. At the first stage, the proposed training sequence aided OTFS (TS-OTFS) data frame structure facilitates the ATI and CE, whereby both the traffic sparsity of terrestrial IoT terminals and the sparse channel impulse response are leveraged for enhanced performance. Moreover, based on the single Doppler shift property for each TSL and sparsity of delay-Doppler domain channel, we develop a parametric approach to further refine the CE performance. Finally, a least squares based parallel time domain SD is developed for demodulating the OTFS signals with relatively low complexity. Simulation results demonstrate the superiority of the proposed methods over the state-of-the-art solutions in the case of the long round-trip latency and severe Doppler effect.

Index Terms

Internet of Things (IoT), low Earth orbit (LEO) satellite, orthogonal time frequency space (OTFS), grant-free non-orthogonal multiple access (NOMA).

I. Introduction

With the advent of the 5G era, Internet of Things (IoT) based on terrestrial cellular networks has developed rapidly and has been widely used in various aspects of human life [1]. In the coming beyond 5G and even 6G, IoT is expected to revolutionize the way we live and work, by means of a wealth of new services based on the seamless interactions of massive heterogeneous terminals [2]. However, in many application scenarios, IoT terminals are widely distributed. Particularly, a considerable percentage of IoT terminals are located in remote areas, which indicates that these IoT applications can not be well supported by conventional terrestrial cellular infrastructures. In recent years, low Earth orbit (LEO) satellite communication systems have attracted considerable research interest, and dense LEO constellations are expected to complement and extend existing terrestrial communication networks for the goal of seamless global coverage [3]. To date, several projects on LEO constellations, e.g., OneWeb and Starlink, have begun the commercial exploration [3]. As an indispensable component of the 6G space-air-ground-sea integrated networks, the LEO constellations are envisioned to provide a promising solution for enabling wide area IoT services [4]. Nevertheless, distinct from the terrestrial communication environment, satellite communications usually suffer from harsh channel conditions such as long round-trip delay, severe Doppler effects, and etc. Besides, in sharp contrast to the conventional downlink-dominated human-type communication (HTC) systems, IoT is mainly driven by the uplink massive machine-type (mMTC) communications with the characteristics of sporadic traffic behavior, since IoT terminals usually intermittently transmit short packets with low rates [5]. Consequently, the design of efficient random access (RA) paradigm for massive IoT terminals based on LEO constellations is a challenging problem.

A. Related Work

The traditional grant-based RA protocols adopted by terrestrial cellular networks usually suffer from the complicated control signaling exchanges and scheduling for requesting uplink access resources [6], [7]. In the case of the extremely long terrestrial–satellite link (TSL) and the resulting large round-trip signal propagation delay, this type of solutions will further aggravate the unfordable access latency. To this end, the ALOHA protocols arise as a better option and are widely used in existing satellite communications for RA [8]. The original ALOHA protocol allows the terminals to transmit their data packets without any coordination. To improve the RA throughput, more advanced ALOHA techniques are developed, such as contention resolution

TABLE I

A BRIEF COMPARISON OF THE RELATED LITERATURE WITH OUR WORK

Reference	**Channel model**	**Bandwidth**	**Transmit signal waveform**	**Signal processing at the receiver**			**Algorithm**
				ATI	**CE**	**SD**	
[10]	Frequency selective Rayleigh fading	Broadband	OFDM	✓		✓	Structured iterative support detection
[11]	Rayleigh fading	Narrowband	Single-carrier	✓		✓	Block sparse modified subspace pursuit (SP)
[12]	Frequency selective Rayleigh fading	Broadband	OFDM	✓		✓	Modified orthogonal matching pursuit (OMP)
[13]	Rayleigh fading	Narrowband	Single-carrier	✓		✓	Prior-information aided adaptive SP
[14]	Rayleigh fading	Narrowband	Single-carrier	✓		✓	Maximum a posteriori probability (MAP)
[15]	Frequency selective Rayleigh fading	Broadband	OFDM	✓		✓	Approximate message passing (AMP) and expectation maximization (EM)
[16]	Frequency selective Rayleigh fading (Pre-equalized)	Broadband	OFDM	✓		✓	Orthogonal AMP with multiple measurement vectors (MMV)
[17]	Frequency fading	Broadband	OFDM	✓	✓		Iterative identified user cancellation
[18]	Rayleigh fading	Narrowband	Single-carrier	✓	✓		Modified Bayesian CS
[19]	Rayleigh fading	Narrowband	Single-carrier	✓	✓		AMP
[20]	Frequency selective fading	Broadband	OFDM	✓	✓		Generalized MMV (GMMV)-AMP-EM
[21]	Land mobile satellite	Narrowband	Single-carrier	✓	✓		Bernoulli–Rician MP-EM
[33]	Double-dispersive	Broadband	OTFS		✓	✓	Three-dimensional simultaneous-OMP
[34]	Double-dispersive	Broadband	OTFS		✓		EM-variational Bayesian (VB)
Our work	TSL channel (Double-dispersive)	Broadband	OTFS	✓	✓	✓	Two-stage ATI & CE and LS-based parallel SD

diversity ALOHA (CRDSA) and enhanced spread spectrum ALOHA (E-SSA), and etc. Despite the aforementioned efforts, the current ALOHA-based RA protocols mainly depend on orthogonal multiple access (OMA) technique and may suffer from the network congestion when the number of terrestrial IoT terminals becomes massive [8].

Recently, grant-free non-orthogonal multiple access (GF-NOMA) schemes have been emerging. These schemes allow IoT terminals to directly transmit their non-orthogonal preambles followed by data packets over the uplink and avoid complicated access requests for resource scheduling [9]. By exploiting the intrinsic sporadic traffic, the receiver of the base station (BS) can separate the non-orthogonal preambles transmitted by different terminals and thus identify the active terminal set (ATS) with compressive sensing (CS) techniques. Benefitting from the non-orthogonal resource allocation, the GF-NOMA schemes can improve the system throughput with limited radio resources. To date, the state-of-the-art CS-based GF-NOMA study mainly focuses on two typical problems: 1) joint active terminal identification (ATI) and signal detection (SD); 2) joint ATI and channel estimation (CE).

The former category is developed by assuming the perfect channel state information (CSI) known at the BS [10]–[15] or the perfect pre-equalization at the terminals (e.g., based on the beacons periodically broadcast by the BS [16]), where CSI is usually regarded to be quasi-static. In particular, [10] and [11] proposed a structured iterative support detection algorithm and a block sparsity based subspace pursuit (SP) algorithm, respectively, to jointly perform ATI and SD in one signal frame (consists of multiple continuous time slots), where the terminals' activity is assumed to remain unchanged. [12] and [13] further relaxed the assumption, i.e., the ATS may vary in several continuous time slots, and developed a modified OMP algorithm and a priori information aided adaptive SP algorithm, respectively, to perform dynamic ATI and SD, where the estimated ATS is exploited as a priori knowledge for the detection in the following time slots. Moreover, to fully exploit the a priori information of the transmit constellation symbols for enhanced accuracy, some Bayesian inference-based detection algorithms were proposed in [14]–[16]. In [14], based on the maximum a posteriori probability (MAP) criterion, the proposed algorithm calculated a posteriori activity probability and soft symbol information to identify the active terminals and detect the payload data, respectively. To overcome the challenge that the perfect a priori information could be unavailable in practical systems, an approximate message passing (AMP)-based detection scheme was proposed in [15], where the hyper-parameters of terminals' activity and noise variance can be adaptively learned through the expectation-maximization (EM)

algorithm. The above literature is mainly based on the assumption that the CSI is perfectly known at the BS and requires the elements of the adopted spreading sequences to be independent and identically distributed (i.i.d), which can be impractical in practice. Therefore, [16] developed an orthogonal AMP (OAMP)-based ATI and SD algorithm for orthogonal frequency division multiplexing (OFDM) systems, where the CSI can be pre-equalized at the terminals according to the beacon signals broadcast by the BS, and the spreading sequences are selected from the partial discrete Fourier transformation (DFT) matrix.

Another category can be applied to time-varing channels, where perfect CSI at the BS or perfect pre-compensation at terminals is unrealistic [17]–[21]. An iterative joint ATI and CE scheme was proposed in [17], where the sparsity of delay-domain channel impulse response (CIR) was exploited and an identified user cancellation approach was proposed for enhanced performance. By exploiting not only the sparse traffic behavior of IoT terminals, but also the innate heterogeneous path loss effects and the joint sparsity structures in multi-antenna systems, the authors in [18] developed a modified Bayesian CS algorithm. With the full knowledge of the a priori distribution of the channels and the noise variance, the authors in [19] developed an AMP-based scheme for massive access in massive multiple-input multiple-output (MIMO) systems. For more challenging massive MIMO-OFDM systems, the authors in [20] proposed a generalized multiple measurement vector (GMMV)-AMP algorithm, where the structured sparsity of spatial-frequency domain and angular-frequency domain channels was leveraged with EM algorithm incorporated. Moreover, a Bernoulli–Rician message passing with expectation–maximization (BR-MP-EM) algorithm was proposed for the LEO constellations-based narrowband massive access using single-carrier in [21]. However, these aforementioned works [17]–[20] usually assume the channels to be slowly time-varying, which can not be applied to the highly dynamic TSLs due to the high-mobility of LEO satellites.

B. Motivation

An emerging two-dimensional modulation scheme, orthogonal time frequency space (OTFS), has been widely considered as a promising alternative to the dominant OFDM. Particularly, OTFS is expected to support reliable communications under high-mobility scenarios in the next-generation mobile communications [22]–[28], [31], [32], [36]. OTFS multiplexes information symbols on a lattice in the delay-Doppler (DD) domain and utilizes a compact DD channel model, where the channel in the DD domain is considered to exhibit more stable, separable, and

sparse features than that in the TF domain. Consequently, OTFS can achieve more robust signal processing with additional diversity gain in the presence of Doppler effect. In fact, [24]–[26] have integrated the OTFS waveform with OMA based on the grant-based access protocols and investigated some new resource allocation schemes. Besides, [27], [28] further amalgamated the OTFS modulation scheme with the NOMA technique. However, [24]–[28] adopt the grant-based RA schemes, which may not cater to the requirements of stringent access latency and massive connectivity for LEO constellations based IoT.

C. Contributions

In this paper, we propose a GF-NOMA paradigm that incorporates OTFS modulation for LEO constellations-based IoT for RA, and investigate the challenging joint ATI, CE, and SD problems. The main contributions of this paper are summarized as follows.

- **GF NOMA-OTFS paradigm**: We amalgamate the GF-NOMA scheme with the OTFS waveform, i.e., GF NOMA-OTFS paradigm, tailored for the LEO constellations-based massive IoT RA. By allowing the uncoordinated IoT terminals to transmit the data packets directly, reusing the limited delay-Doppler resources, and exploiting the stability, sparsity, and separability of TSLs represented in the DD domain, the proposed GF NOMA-OTFS paradigm can reap the benefit of high RA throughput and Doppler-robustness.
- **Training sequences (TSs) aided OTFS modulation/demodulation architecture**: Existing CE solutions for OTFS systems embed the pilots and guard symbols in the DD domain [32]–[34]. However, in the case of highly dynamic TSLs with extremely severe Doppler shift, the the compactness of the DD domain channel would no longer maintain, which would give rise to the dramatical increase of guard symbols. Moreover, the low-resolution of Doppler lattices could lead to the severe Doppler spreading even each TSL's Doppler shift is a single value, which would further deteriorate the performance and effectiveness of signal processing in the DD domain. To circumvent these challenges, we utilize the time domain TSs to replace the conventional DD domain pilot and guard symbols for performing joint ATI and CE, and further propose a TSs aided OTFS (TS-OTFS) modulation/demodulation architecture.
- **Time domain ATI, CE and SD method**: Furthermore, we put forward a two-stage joint ATI and CE scheme as well as a following low-complexity multi-user SD for the GF NOMA-OTFS paradigm. Specifically, for the ATI and CE, at the first stage, the proposed time

domain TSs facilitate the ATI and CE, whereby both the traffic sparsity of terrestrial IoT terminals and the structural sparse CIR are leveraged. On this basis, a parametric approach is developed to further refine the CE performance, whereby the single Doppler shift property for each TSL and sparsity of DD domain channel are exploited. Finally, a least squares (LS)-based parallel time domain SD is developed for demodulating the OTFS signals with relatively low complexity.

D. Organization

The remainder of this paper is organized as follows. In Section II, we introduce the TSL model of the LEO constellations-based IoT. The GF NOMA-OTFS paradigm and TS-OTFS modulation/demodulation architecture are proposed in Section III. In Section IV, the proposed joint ATI and CE scheme for the GF NOMA-OTFS paradigm is presented. Then, in Section V, we further propose a multi-user signal detector based on the previous results of ATI and CE. The effectiveness of our proposed scheme is demonstrated by simulation results in Section VI. Finally, our conclusions are drawn in Section VII.

E. Notations

Throughout this paper, scalar variables are denoted by normal-face letters, while boldface lower and upper-case letters denote column vectors and matrices, respectively. The transpose, conjugate, Hermitian transpose, inversion, and pseudo-inversion for matrix are denoted by $(\cdot)^{\mathrm{T}}$, $(\cdot)^{*}$, $(\cdot)^{\mathrm{H}}$, $(\cdot)^{-1}$, and $(\cdot)^{\dagger}$ respectively. Besides, $|\cdot|$, $\|\cdot\|_1$, and $\|\cdot\|_2$ represent modulus, ℓ_1-norm, and ℓ_2-norm, respectively. $\mathbf{X}_{[m,n]}$ is the (m,n)-th element of matrix $\mathbf{X}$; $\mathbf{X}_{[m,:]}$ and $\mathbf{X}_{[:,n]}$ are the m-th row vector and the n-th column vector of matrix $\mathbf{X}$, respectively. $\mathbf{X}_{[:,\mathcal{I}]}$ and $\mathbf{X}_{[:,\mathcal{I}]}$ denote the submatrix consisting of the columns and rows of $\mathbf{X}$ indexed by the ordered set $\mathcal{I}$, respectively. $\mathbf{x}_{[n]}$ denotes the n-th element of $\mathbf{x}$. Furthermore, $|\mathcal{A}|_c$ is the cardinality of the set $\mathcal{A}$, and $\mathrm{supp}(\cdot)$ is the support set of a vector or a matrix. The operators $\odot$ and $\otimes$ represent the Hadamard product and Kronecker product, respectively. The operator $\mathrm{vec}(\mathbf{X})$ stacks the columns of $\mathbf{X}$ on top of each another, and $\mathrm{mat}(\mathbf{x};m,n)$ converts the vector $\mathbf{x}$ of size mn into the matrix of size $m \times n$ by successively selecting every m elements of $\mathbf{x}$ as its columns. $< \mathbf{x}_1, \mathbf{x}_2 >$ represents the inner product of $\mathbf{x}_1$ and $\mathbf{x}_2$. Finally, $\mathbf{I}_n$ is the identity matrix of size $n \times n$, $\mathbf{0}_{n\times m}$ is the $n \times m$ zero matrix, $\emptyset$ denotes the empty set, and $\delta(x)$ is the Dirac function.

II. TERRESTRIAL-SATELLITE LINK MODEL

As illustrated in Fig. 1, we consider that the LEO constellations consisting of a large number of LEO satellites can provide ubiquitous connections for massive IoT terminals. Each LEO satellite is equipped with a uniform planar array (UPA) composed of $P = P_x \times P_y$ antennas, where P_x and P_y are the number of antennas on the x-axis and y-axis, respectively. Meanwhile, single-antenna is assumed to be employed at the IoT terminals without loss of generality [1]. Due to the sporadic traffic behavior in typical IoT [5], within a given time interval, the number of active IoT terminals K_a can be much smaller than the number of all potential IoT terminals K, i.e., $K_a \ll K$. The active IoT terminals transmit RA signals consisting of a preamble as well as the following payload data, and the inactive remain silent. To reflect the activity status of all potential IoT terminals, we define an activity indicator α_k, which is equal to 1 when the k-th IoT terminal is active and 0 otherwise. Meanwhile, the ATS is defined as $\mathcal{A} = \{k|\alpha_k = 1, 1 \leq k \leq K\}$ and the cardinality of it is $K_a = |\mathcal{A}|_c$.

Since the TSLs connecting the LEO satellites and terrestrial IoT terminals experience few propagation scatterers and the links are rarely blocked by obstacles, it is reasonable to assume there coexist the line-of-sight (LoS) and few non-LoS (NLoS) paths. Therefore, the DD domain uplink channel between the LEO satellite and the served k-th IoT terminal can be expressed as [23], [29], [30]

$$\mathbf{h}_k^{\mathrm{DD}}(\tau,\nu) = \sqrt{\frac{\gamma_k}{\gamma_k+1}}\delta(\tau-\tau_k^{\mathrm{LoS}})\delta(\nu-\nu_k^{\mathrm{LoS}})\mathbf{v}_k + \sqrt{\frac{1}{\gamma_k+1}}\sum_{q=1}^{Q_k} g_k^q\delta(\tau-\tau_k^q)\delta(\nu-\nu_k^q)\mathbf{v}_k, \quad (1)$$

where the first term corresponds to the LoS path and the NLoS paths contribute to the other Q_k terms. ν_k^{LoS} and ν_k^q respectively denote the Doppler shift of the LoS and the q-th NLoS path, τ_k^{LoS} and τ_k^q respectively denote the remanent relative time of arrive (RToA) and delay of the q-th NLoS path, γ_k and g_k^q are respectively the Rician factor and the small-scale fading factor of the q-th NLoS path, $\mathbf{v}_k \in \mathbb{C}^{P\times 1}$ denotes the receive array steering vector at the LEO satellite. The further explanations of these parameters are presented as follows.

- **Array steering vector**: Since the TSL's distance is far larger than the distances between the terminal and its surrounding scatterers, the angles of arrival (AoAs), i.e., the zenith angle

[1]Without loss of generality, we assume that the single-antenna is employed at the IoT terminals. Note that if the phased array with analog beamforming is deployed, the subsequent mathematical formulation is equivalent from the signal processing perspective, since analog beamforming at the IoT terminals can be easily implemented with the predictable trajectory of LEO satellites in theory.

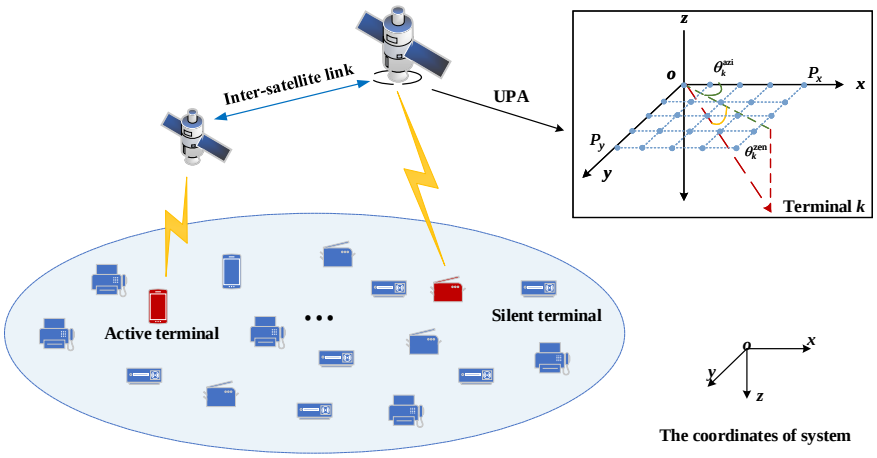


Fig. 1. Illustration of the LEO constellations-based IoT based on the proposed GF NOMA-OTFS scheme.

$\theta_{k,\mathrm{LoS}}^{\mathrm{zen}}$ ($\theta_{k,q}^{\mathrm{zen}}$) and the azimuth angle $\theta_{k,\mathrm{LoS}}^{\mathrm{azi}}$ ($\theta_{k,q}^{\mathrm{azi}}$) as illustrated in Fig. 1, for the k-th terminal can be assumed to be almost identical [29], i.e., $\theta_{k,\mathrm{LoS}}^{\mathrm{zen}} \approx \theta_{k,q}^{\mathrm{zen}} = \theta_k^{\mathrm{zen}}$ and $\theta_{k,\mathrm{LoS}}^{\mathrm{azi}} \approx \theta_{k,q}^{\mathrm{azi}} = \theta_k^{\mathrm{azi}}$. Therefore, the array steering vector of the widely considered UPA can be simplified as

$$\mathbf{v}_k = \frac{1}{\sqrt{P}} \left[e^{-j2\pi \frac{d}{\lambda} \sin(\theta_k^{\mathrm{zen}}) \cos(\theta_k^{\mathrm{azi}}) \mathbf{p}_1} \right] \otimes \left[e^{-j2\pi \frac{d}{\lambda} \sin(\theta_k^{\mathrm{zen}}) \sin(\theta_k^{\mathrm{azi}}) \mathbf{p}_2} \right], \tag{2}$$

where $\mathbf{p}_1 = [0, 1, \ldots, P_x - 1]^{\mathrm{T}}$, $\mathbf{p}_2 = [0, 1, \ldots, P_y - 1]^{\mathrm{T}}$, λ is the wavelength of carrier frequency and d is the antenna spacing. Without loss of generality, the elements of the UPA are assumed to be separated by one-half wavelength in both the x-axis and y-axis.

- **Doppler shift**: The Doppler shift ν_k^{LoS} (ν_k^q) includes two independent components: $\nu_k^{\mathrm{LoS,sat}}$ ($\nu_k^{q,\mathrm{sat}}$) and $\nu_k^{\mathrm{LoS,UT}}$ ($\nu_k^{q,\mathrm{UT}}$) caused by the mobility of LEO satellite and terrestrial IoT terminals, respectively. Since LEO satellite moves much faster than IoT terminals, the motion of LEO satellite mainly determines ν_k^{LoS} (ν_k^q), i.e., $\nu_k^{\mathrm{LoS,sat}} \gg \nu_k^{\mathrm{LoS,UT}}$ ($\nu_k^{q,\mathrm{sat}} \gg \nu_k^{q,\mathrm{UT}}$). Besides, combined with the fact that the AoAs of LoS and NLoS links related to the k-th terminal are alomst identical, it is reasonable to assume that the Doppler shift of the TSL is single-valued, i.e., $\nu_k^{\mathrm{LoS}} \approx \nu_k^q \approx \nu_k^{\mathrm{LoS,sat}} \approx \nu_k^{q,\mathrm{sat}}$.
- **Remanent RToA and multipath components' (MPCs') delay**: Since IoT terminals' locations are geographically distributed, the ToA of signals received from different terminals may undergo severe time offsets, and we consider the major part of time offsets can be compensated, while the remanent RToA is denoted as τ_k^{LoS}. Meanwhile, in the case of MPCs, the relative delay of the q-th NLoS path can be denoted as τ_k^q.

Note that (1) can be transformed into time-varying CIR as

$$\mathbf{h}_k(t,\tau) = \int \mathbf{h}_k(\tau,\nu) e^{j2\pi\nu(t-\tau)} d\nu. \tag{3}$$

III. Proposed TS-OTFS Transmission Scheme

In this section, we introduce the GF NOMA-OTFS paradigm and the transceiver structure of the proposed TS-OTFS scheme, which is illustrated in Fig. 2.

A. Modulation of the Proposed TS-OTFS at Transmitter

For the active IoT terminals, the input information bits are first mapped to quadrature amplitude modulation (QAM) symbols and then rearranged in the DD domain plane as $\mathbf{X}_k^{\mathrm{DD}} \in \mathbb{C}^{M\times N}, \forall k$. Here, N and M are the dimensions of the latticed resource units in the Doppler domain and delay domain, respectively. On this basis, the DD domain $\mathbf{X}_k^{\mathrm{DD}}$ is parallel-to-serial converted to the transmit signal vector $\mathbf{s}_k$ in the time domain via a cascade of TS-OTFS transformations, which are constituted by a pre-processing module and time-frequency (TF) modulator.

Specifically, the pre-processing module is consistent with that of the traditional OFDM-based OTFS architecture [31], [33], i.e., the DD domain data $\mathbf{X}_k^{\mathrm{DD}}$ is transformed into the TF domain data matrix $\mathbf{X}_k^{\mathrm{TF}} \in \mathbb{C}^{M\times N}$ by applying the *inverse symplectic finite Fourier transform (ISFFT)* and a transmit windowing function [22]. This ISFFT processing for $\mathbf{X}_k^{\mathrm{DD}}$ can be written as

$$\mathbf{X}_k^{\mathrm{ISFFT}} = \mathbf{F}_M \mathbf{X}_k^{\mathrm{DD}} \mathbf{F}_N^{\mathrm{H}}, \forall k, \tag{4}$$

where both $\mathbf{F}_M \in \mathbb{C}^{M\times M}$ and $\mathbf{F}_N \in \mathbb{C}^{N\times N}$ are the DFT matrices. Furthermore, the transmit windowing matrix $\mathbf{W}^{\mathrm{tx}} \in \mathbb{C}^{M\times N}$ multiplies $\mathbf{X}_k^{\mathrm{ISFFT}}$ element-wise to generate the TF domain data matrix $\mathbf{X}_k^{\mathrm{TF}}$ as

$$\mathbf{X}_k^{\mathrm{TF}} = \mathbf{X}_k^{\mathrm{ISFFT}} \odot \mathbf{W}^{\mathrm{tx}}, \forall k. \tag{5}$$

For simplicity and without loss of generality, a rectangular window, namely $\mathbf{W}^{\mathrm{tx}}$ with all elements equal to one, is adopted in this paper.

Based on the acquired TF domain data matrix $\mathbf{X}_k^{\mathrm{TF}}$, the subsequent TF modulator transforms $\mathbf{X}_k^{\mathrm{TF}}$ into the transmit signal vector $\mathbf{s}_k$. In particular, firstly, *Heisenberg transform* [22] is applied to each column of $\mathbf{X}_k^{\mathrm{TF}}$ to produce the time domain data matrix $\tilde{\mathbf{S}}_k \in \mathbb{C}^{M\times N}$ as

$$\tilde{\mathbf{S}}_k = \mathbf{F}_M^{\mathrm{H}} \mathbf{X}_k^{\mathrm{TF}}, \forall k, \tag{6}$$

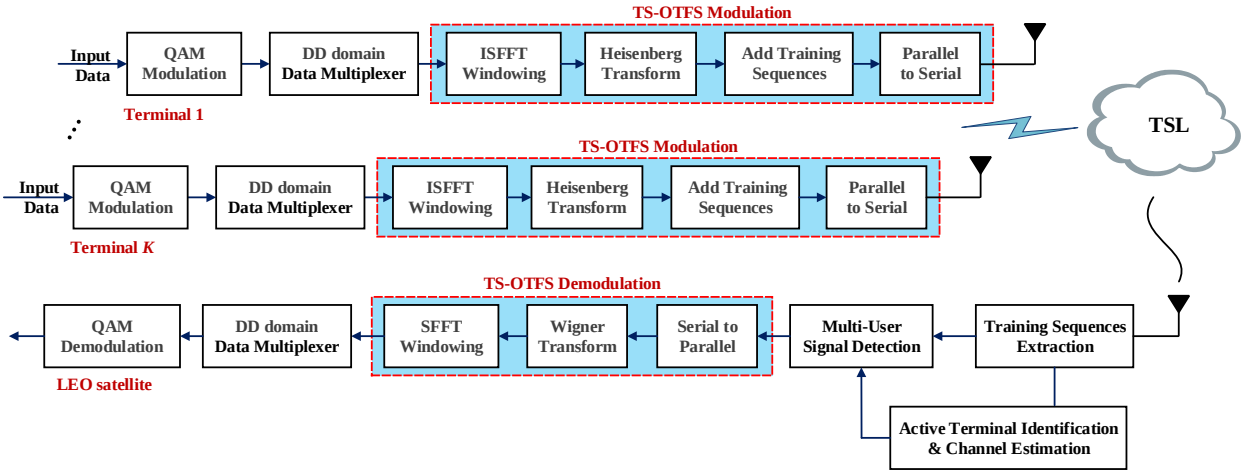


Fig. 2. The transceiver structure of the proposed TS-OTFS scheme for the GF NOMA-OTFS paradgim.

where its vector form is denoted as $\tilde{\mathbf{s}}_k = \mathrm{vec}\{\tilde{\mathbf{S}}_k\} \in \mathbb{C}^{MN\times 1}$. Furthermore, for the traditional OFDM-based OTFS architecture, a cyclic prefix (CP) is added to the front of each time domain OTFS symbol $\tilde{\mathbf{s}}_k^i \in \mathbb{C}^{M\times 1}$ ($\tilde{\mathbf{s}}_k^i$ is the i-th column vector of $\tilde{\mathbf{S}}_k$). By contrast, for the proposed TS-OTFS scheme, $N+1$ duplicate TSs with the length of M_t, denoted by $\mathbf{c}_k = [c_{k,0}\ c_{k,1}\ldots\ c_{k,M_t-1}]^{\mathrm{T}} \in \mathbb{C}^{M_t\times 1}$, are appended to the front and rear of the OTFS payload data as illustrated in Fig. 3. These time domain TSs are known by the transceiver, and they can not only be utilized to avoid ISI over time dispersive channels, but also perform ATI and CE (will be detailed in Section IV). Finally, the transmit TS-OTFS signal consisting of TSs and time domain OTFS payload data $\mathbf{s}_k = \left[\mathbf{c}_k^{\mathrm{T}}, \tilde{\mathbf{s}}_k^{1^{\mathrm{T}}}, \mathbf{c}_k^{\mathrm{T}}, \tilde{\mathbf{s}}_k^{2^{\mathrm{T}}}, \ldots, \mathbf{c}_k^{\mathrm{T}}, \tilde{\mathbf{s}}_k^{N^{\mathrm{T}}}, \mathbf{c}_k^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{(M_tN+MN+M_t)\times 1}$ is obtained through the parallel-to-serial conversion.

B. Proposed TS-OTFS Demodulation at Receiver

In fact, the discrete form of (1) can be rewritten as

$$h_{k,p}^{\mathrm{DD}}[\ell,\upsilon] = h_{k,p}^{\mathrm{DD}}(\tau,\nu)\,|_{\tau=\frac{\ell}{M\Delta f},\,\nu=\frac{\upsilon}{NT}}, \tag{7}$$

where $h_{k,p}^{\mathrm{DD}}(\tau,\nu)$ is the p-th element of $\mathbf{h}_k^{\mathrm{DD}}(\tau,\nu)$, Δf is the frequency spacing between adjacent sub-carriers, and T is the duration of one TS-OTFS symbol. Moreover, the discrete form of (3) can be denoted as

$$h_{k,p}[\kappa,\ell] = h_{k,p}(t,\tau)\,|_{t=\kappa T_s,\,\tau=\ell T_s}, \tag{8}$$

where T_s is the sampling interval of the system.

Therefore, the κ-th element of the signal $\mathbf{r}_p \in \mathbb{C}^{(M_tN+MN+M_t)\times 1}$ received at the p-th antenna is the superposition of the signals received from all active terminals, which can be expressed as

$$r_p(\kappa) = \sum_{k=1}^{K} \sum_{l=0}^{L-1} \alpha_k h_{k,p}\left[\kappa, \ell\right] s_k\left[\kappa - \ell\right] + w_p(\kappa), \forall p, \tag{9}$$

where L represents the maximum of remanent RToA and MPCs' delay, and $w_p(\kappa) \sim \mathcal{CN}(0, \sigma_w^2)$ denotes the additive white Gaussian noise (AWGN) at the receiver.

The receiver of LEO satellite consists of two cascaded modules: the first one performs ATI, CE, and multi-user SD, and the others demodulate the OTFS payload data. For the first one, the receiver of LEO satellites firstly extracts TSs from the received signals to perform ATI and CE. With the identified active terminal set (ATS) $\hat{\mathcal{A}}$ and their corresponding CSI, the proposed multi-user signal detector detects the payload data for the ATS to obtain $\hat{\tilde{\mathbf{s}}}_k \in \mathbb{C}^{MN\times 1}, k \in \hat{\mathcal{A}}$. The above ATI, CE, and SD modules will be discussed in detail in the following Sections IV and V.

For the TS-OTFS demodulation, it is equivalent to the inverse operation of the modulation, which consists of a TF demodulator and a post-processing module, and transforms the detected time domain OTFS payload data $\hat{\tilde{\mathbf{s}}}_k$ to the original DD domain $\hat{\mathbf{X}}_k^{\mathrm{DD}}$. In particular, $\hat{\tilde{\mathbf{s}}}_k$ can be rewritten as time domain 2D data matrix $\hat{\mathbf{S}}_k \in \mathbb{C}^{M\times N}$ through serial-to-parallel conversion, i.e.,

$$\hat{\mathbf{S}}_k = \mathrm{mat}\left(\hat{\tilde{\mathbf{s}}}_k, M, N\right), \forall k, \tag{10}$$

Then, the *Wigner transform* [22] is applied to recover the TF data $\hat{\mathbf{X}}_k^{\mathrm{TF}}$ as

$$\hat{\mathbf{X}}_k^{\mathrm{TF}} = \mathbf{F}_M \hat{\mathbf{S}}_k, \forall k. \tag{11}$$

In the post-processing module, the receive windowing matrix $\mathbf{W}^{\mathrm{rx}}$ multiplies $\hat{\mathbf{X}}_k^{\mathrm{TF}}$ element-wise as

$$\hat{\mathbf{X}}_k^{\mathrm{TF,W}} = \hat{\mathbf{X}}_k^{\mathrm{TF}} \odot \mathbf{W}^{\mathrm{rx}}, \forall k, \tag{12}$$

where a rectangular window $\mathbf{W}^{\mathrm{rx}}$ is adopted similar to the transmitter in (5). Finally, *symplectic finite Fourier transform (SFFT)* is applied to $\hat{\mathbf{X}}_k^{\mathrm{TF,W}}$ for restoring the TF domain OTFS data to DD domain as

$$\hat{\mathbf{X}}_k^{\mathrm{DD}} = \mathbf{F}_M^{\mathrm{H}} \hat{\mathbf{X}}_k^{\mathrm{TF,W}} \mathbf{F}_N, \forall k. \tag{13}$$

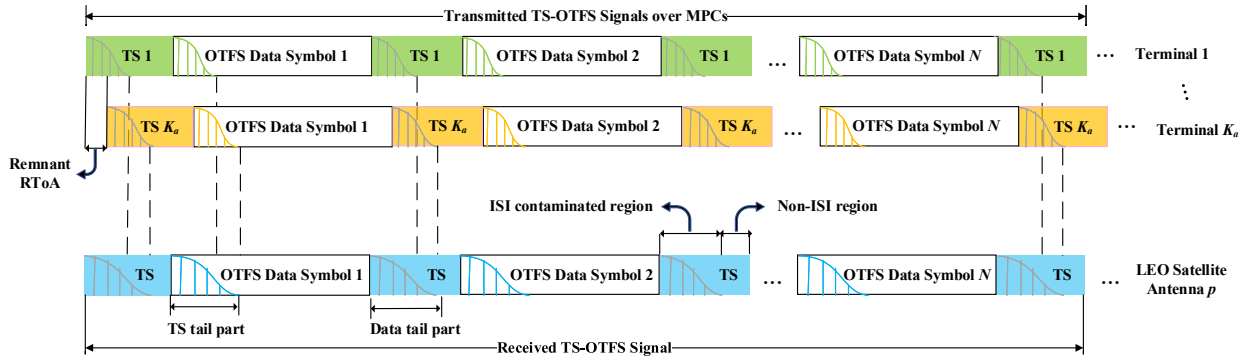


Fig. 3. Transmit and receive signal structures of the proposed TS-OTFS scheme at the transceiver.

IV. Proposed Active Terminal Identification and Channel Estimation

To handle the challenging ATI and CE over TSLs with severe Doppler effect, we propose a two-stage joint ATI and CE scheme for the proposed GF NOMA-OTFS paradigm.

A. Problem Formulation of Joint ATI and CE

The structure of time domain TS-OTFS signal vector $\mathbf{s}_k\,(1 \le k \le K)$ and its received version at the p-th receive antenna $\mathbf{r}_p\,(1 \le p \le P)$ are illustrated in Fig. 3, both of which consist of the OTFS payload data and the embedded TS's part. One significant challenge to perform ATI and CE based on the received TSs, lies in the fact that each received TS is contaminated by the previous OTFS data symbol due to the time dispersive CIR of each TSL and the remnant RToA among different terminals' TSLs. An effective approach is to utilize the non inter symbol interference (non-ISI) region as illustrated in Fig. 3, which is the rear part of the TSs and immune from the influence of the previous OTFS data symbol [37]. Therefore, the TS's length M_t is designed to be longer than the maximum of remnant RToA and MPCs' delay L in order to ensure the non-ISI region with sufficient length, and thus the length of non-ISI region can be denoted as $G \triangleq M_t - L + 1$.

In this way, according to (9), the non-ISI region of the i-th $(1 \le i \le N+1)$ TS $\mathbf{r}^i_{\mathrm{TS},p} \in \mathbb{C}^{G\times 1}$ can be expressed as

$$\mathbf{r}^i_{\mathrm{TS},p} = \sum_{k=1}^{K} \alpha_k \left(\mathbf{\Delta}^{\mathrm{LoS}}_k \mathbf{\Psi}_k \mathbf{h}^{i,\mathrm{LoS}}_{\mathrm{TS},k,p} + \sum_{q=1}^{Q_k} \mathbf{\Delta}^{q}_k \mathbf{\Psi}_k \mathbf{h}^{i,q}_{\mathrm{TS},k,p} \right) + \mathbf{w}^i_{\mathrm{TS},p}, \forall i, p, \tag{14}$$

where $\mathbf{h}_{\mathrm{TS},k,p}^{i,\mathrm{LoS}} \in \mathbb{C}^{L\times 1}$ and $\mathbf{h}_{\mathrm{TS},k,p}^{i,q} \in \mathbb{C}^{L\times 1}$ denote the LoS and NLoS components of the vector form of CIR $\mathbf{h}_{\mathrm{TS},k,p}^{i}$ (aligned with the instant of the beginning of the i-th non-ISI region) as

$$\mathbf{h}_{\mathrm{TS},k,p}^{i} = \mathbf{h}_{\mathrm{TS},k,p}^{i,\mathrm{LoS}} + \sum_{q=1}^{Q_k} \mathbf{h}_{\mathrm{TS},k,p}^{i,q}, \forall i, p, k, \tag{15}$$

$\mathbf{w}_{\mathrm{TS},p}^{i} \in \mathbb{C}^{G\times 1}$ is the vector form of AWGN, $\mathbf{\Psi}_k \in \mathbb{C}^{G\times L}$ is a Toeplitz matrix given by [37]

$$\mathbf{\Psi}_k = \begin{bmatrix} c_{k,L-1} & c_{k,L-2} & \cdots & c_{k,0} \\ c_{k,L} & c_{k,L-1} & \cdots & c_{k,1} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k,M_t-1} & c_{k,M_t-2} & \cdots & c_{k,M_t-L} \end{bmatrix}, \tag{16}$$

and $\mathbf{\Delta}_k^{\mathrm{LoS}}$ ($\mathbf{\Delta}_k^{q}$) $\in \mathbb{C}^{G\times G}$ is the diagonal Doppler shift matrix associated with the LoS (the q-th NLoS) path as

$$\mathbf{\Delta}_k^{\mathrm{LoS}} = \mathrm{diag}\left\{ e^{\frac{j2\pi v_k^{\mathrm{LoS}}}{N(M+M_t)}\cdot\left[(-\ell_k^{\mathrm{LoS}}),\ldots,0,\ldots,(G-\ell_k^{\mathrm{LoS}}-1)\right]^{\mathrm{T}}} \right\}, \tag{17}$$

$$\mathbf{\Delta}_k^{q} = \mathrm{diag}\left\{ e^{\frac{j2\pi v_k^{q}}{N(M+M_t)}\cdot\left[(-\ell_k^{q}),\ldots,0,\ldots,(G-\ell_k^{q}-1)\right]^{\mathrm{T}}} \right\}. \tag{18}$$

Since both $\mathbf{\Delta}_k^{\mathrm{LoS}}$ and $\mathbf{\Delta}_k^{q}$ are unknown matrices for the receiver of LEO satellites, it would be infeasible to recover the sparse CIR vectors in (14) with the unknown sensing matrices. Fortunately, on the one hand, the duration of each non-ISI region is much shorter than the span of the entire data frame, and thus the TSLs can be assumed to be approximately unchanged without causing obvious errors for the sparse CIR vector recovery. On the other hand, it will be proved in Remark 2 that the ambiguity caused by this approximation can be counteracted through the following CE refinement. In this case, both $\mathbf{\Delta}_k^{\mathrm{LoS}}$ and $\mathbf{\Delta}_k^{q}$ are approximate to the identity matrices, and (14) can be rewritten as

$$\begin{aligned} \mathbf{r}_{\mathrm{TS},p}^{i} &= \sum_{k=1}^{K} \alpha_k \mathbf{\Psi}_k \underbrace{\left(\mathbf{h}_{\mathrm{TS},k,p}^{i,\mathrm{LoS}} + \sum_{q=1}^{Q_k} \mathbf{h}_{\mathrm{TS},k,p}^{i,q} \right)}_{\mathbf{h}_{\mathrm{TS},k,p}^{i}} + \tilde{\mathbf{w}}_{\mathrm{TS},p}^{i} \\ &= \mathbf{\Psi}\tilde{\mathbf{h}}_{\mathrm{TS},p}^{i} + \tilde{\mathbf{w}}_{\mathrm{TS},p}^{i}, \end{aligned} \tag{19}$$

where $\tilde{\mathbf{h}}_{\mathrm{TS,p}}^{i} = \left[\tilde{\mathbf{h}}_{\mathrm{TS},1,p}^{i^{\mathrm{T}}}, \tilde{\mathbf{h}}_{\mathrm{TS},2,p}^{i^{\mathrm{T}}}, \ldots, \tilde{\mathbf{h}}_{\mathrm{TS},K,p}^{i^{\mathrm{T}}} \right]^{\mathrm{T}} \in \mathbb{C}^{KL\times 1}$, $\tilde{\mathbf{h}}_{\mathrm{TS},k,p}^{i} = \alpha_k \mathbf{h}_{\mathrm{TS},k,p}^{i}$, $\mathbf{\Psi} = [\mathbf{\Psi}_1, \mathbf{\Psi}_2, \ldots, \mathbf{\Psi}_K] \in \mathbb{C}^{G\times KL}$, the approximation error and the AWGN are collectedly considered as the effective noise term $\tilde{\mathbf{w}}_{\mathrm{TS},p}^{i}$.

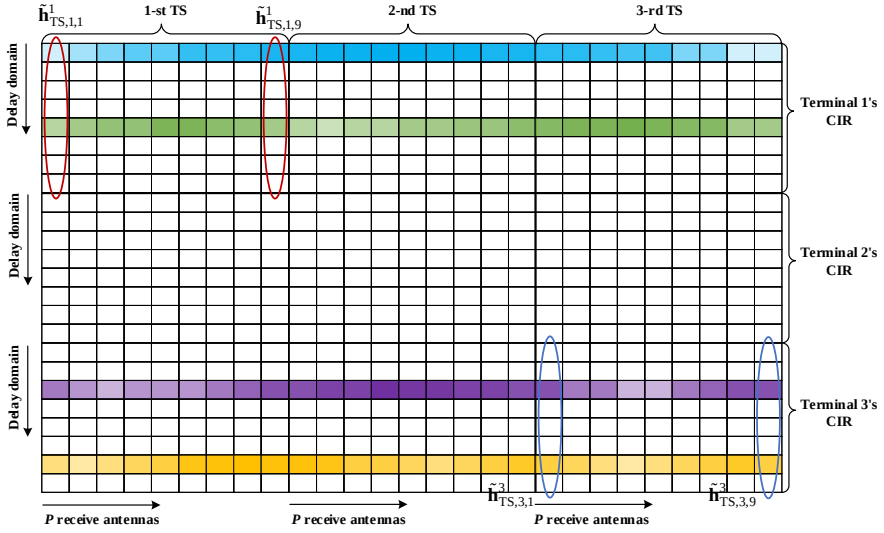


Fig. 4. The illustration of the common sparsity of different CIR vectors resulting from the spatial and temporal correlations in the TSLs, where $P = 9$, $N = 2$, $K = 3$, and $L = 8$.

Due to the severe path loss of TSLs, the energy of NLoS paths reflected by scatterers around the terminals could be weak and the number of non-negligible NLoS paths is limited. Therefore, the delay domain sparsity of the CIR vector $\mathbf{h}^i_{\mathrm{TS},k,p}$ can be represented as

$$\left|\mathrm{supp}\left\{\mathbf{h}^i_{\mathrm{TS},k,p}\right\}\right|_c = Q_k + 1 \ll L, \forall i, p, k. \tag{20}$$

Moreover, combined with the sporadic traffic behaviors, $\tilde{\mathbf{h}}^i_{\mathrm{TS},p}$ exhibits the sparsity as

$$\left|\mathrm{supp}\left\{\tilde{\mathbf{h}}^i_{\mathrm{TS},p}\right\}\right|_c = \sum_{k\in\mathcal{A}}(Q_k + 1) = Q \ll KL, \forall i, p. \tag{21}$$

It indicates that (19) is a typical sparse signal recovery problem with the single measurement vector (SMV) form.

B. Exploiting TSL's Spatial and Temporal Correlations for Enhanced Performance

To further enhance the system performance, we will leverage the structured common sparsity inherent in the TSLs. *1) Spatial correlation:* Specially, for different receive antennas, the RToA, MPCs' delay, and Doppler shift of the signals received from the same terminal are approximately identical. This implies that the support sets can be treated as common for sparse CIR vectors $\{\mathbf{h}^i_{\mathrm{TS},k,p}\}_{p=1}^P$, whereas the non-zero coefficients could be distinct. *2) Temporal correlation:* Additionally, although the TSLs vary continuously with time due to the high mobility of LEO satellites, within the duration of one frame, the relative positions of the IoT terminals and LEO

satellite will not change dramatically. This fact implies that it can also be reasonable to assume the RToA, the propagation delay, and the Doppler shift of signals received from the same terminal are approximately identical for multiple adjacent OTFS symbols within one TS-OTFS frame. Hence, the support sets can also be regarded as identical for sparse CIR vectors $\{\mathbf{h}^{i}_{\mathrm{TS},k,p}\}_{i=1}^{N+1}$.

We will further illustrate the TSL's spatial and temporal correlations with the following example. Specifically, we consider the carrier frequency is $f_c = 10\,\mathrm{GHz}$, the bandwidth is $B_w = 100\,\mathrm{MHz}$, the orbital altitude of LEO satellites is $d_o = 500\,\mathrm{km}$, the size of TS-OTFS frame is $(M, N, M_t) = (512, 8, 50)$, a 5×5 UPA is equipped at the LEO satellite receiver, and the antenna elements are separated by one-half wave length. Furthermore, we assume the mobile direction of the LEO satellite is perpendicular to the line connecting the LEO satellite with the terminal, where the Doppler shift rate of change gets its maximum value. The distance that the LEO satellite travels in the duration of one TS-OTFS frame $T = 25\,\mathrm{ms}$ is $d_l = 190.2\,\mathrm{m}$. Then, the resulting maximum delay jitter (between the central antenna element and the corner one), can be calculated as

$$\Delta\tau = \frac{\sqrt{{d_o}^2 + (\lambda + d_l)^2 + \lambda^2} - d_o}{c} = 1.2 \times 10^{-10}\mathrm{s}, \tag{22}$$

where c is the velocity of light. Such tiny delay jitter is negligible compared with the sampling interval $T_s = 1/B_w$, and so the same with the RToA and MPCs' delay. Besides, the Doppler jitter can be calculated similarly as

$$\Delta\nu = \frac{v_s}{c} \cdot f_c \sin\Delta\theta^{\mathrm{zen}} = 96.4\,\mathrm{Hz}, \tag{23}$$

where v_s is the velocity of LEO satellites and $\Delta\theta^{\mathrm{zen}}$ is the zenith angle's jitter and the sine value of $\Delta\theta^{\mathrm{zen}}$ is $\sin\Delta\theta^{\mathrm{zen}} \approx \tan\Delta\theta^{\mathrm{zen}} = 1.8 \times 10^{-4}$. Such tiny Doppler jitter is also negligible compared with the Doppler sampling interval $1/NT = 10\,\mathrm{kHz}$.

Remark 1: The above analysis implies that the stability of TSLs in the DD domain can maintain in the one TS-OTFS frame since both the delay and Doppler jitters are negligible, and thus it determines that the DD domain signal processing of the OTFS technique is effective in this case.

With the above discussion in mind, we come to the conclusion that the CIR vectors display a common sparsity pattern as illustrated in Fig. 4. On this basis, we propose to extend (19) to MMV to jointly process the received signal from multiple antennas and multiple TSs. Specifically, by

collecting the received signal $\{\mathbf{r}_{\mathrm{TS},p}^{i}\}_{p=1}^{P}$ from multiple antennas (i.e., different subscript p), we can obtain

$$\mathbf{R}_{\mathrm{TS}}^{i}=\mathbf{\Psi}\tilde{\mathbf{H}}_{\mathrm{TS}}^{i}+\tilde{\mathbf{W}}_{\mathrm{TS}}^{i},\forall i, \tag{24}$$

where $\mathbf{R}_{\mathrm{TS}}^{i}=\left[\mathbf{r}_{\mathrm{TS},1}^{i},\mathbf{r}_{\mathrm{TS},2}^{i},\ldots,\mathbf{r}_{\mathrm{TS},P}^{i}\right]\in\mathbb{C}^{G\times P}$, $\tilde{\mathbf{H}}_{\mathrm{TS}}^{i}=\left[\tilde{\mathbf{h}}_{\mathrm{TS},1}^{i},\tilde{\mathbf{h}}_{\mathrm{TS},2}^{i},\ldots,\tilde{\mathbf{h}}_{\mathrm{TS},P}^{i}\right]\in\mathbb{C}^{KL\times P}$, and $\tilde{\mathbf{W}}_{\mathrm{TS}}^{i}=\left[\tilde{\mathbf{w}}_{\mathrm{TS},1}^{i},\tilde{\mathbf{w}}_{\mathrm{TS},2}^{i},\ldots,\tilde{\mathbf{w}}_{\mathrm{TS},P}^{i}\right]\in\mathbb{C}^{G\times P}$. Moreover, by stacking the received signals from multiple adjacent TSs (i.e., different superscript i), we can further obtain

$$\mathbf{R}_{\mathrm{TS}}=\mathbf{\Psi}\tilde{\mathbf{H}}_{\mathrm{TS}}+\tilde{\mathbf{W}}_{\mathrm{TS}}, \tag{25}$$

where $\mathbf{R}_{\mathrm{TS}}=\left[\mathbf{R}_{\mathrm{TS}}^{(1)},\mathbf{R}_{\mathrm{TS}}^{(2)},\ldots,\mathbf{R}_{\mathrm{TS}}^{(N+1)}\right]\in\mathbb{C}^{G\times P(N+1)}$, $\tilde{\mathbf{H}}_{\mathrm{TS}}=\left[\tilde{\mathbf{H}}_{\mathrm{TS}}^{(1)},\tilde{\mathbf{H}}_{\mathrm{TS}}^{(2)},\ldots,\tilde{\mathbf{H}}_{\mathrm{TS}}^{(N+1)}\right]\in\mathbb{C}^{KL\times P(N+1)}$, and $\tilde{\mathbf{W}}_{\mathrm{TS}}=\left[\tilde{\mathbf{W}}_{\mathrm{TS}}^{(1)},\tilde{\mathbf{W}}_{\mathrm{TS}}^{(2)},\ldots,\tilde{\mathbf{W}}_{\mathrm{TS}}^{(N+1)}\right]\in\mathbb{C}^{G\times P(N+1)}$.

C. Joint ATI and CE Based on MMV-CS Theory

Generally speaking, the dimension of non-ISI region G is expected to be as small as possible to reduce the TSs overhead, so that G could be usually far smaller than the dimension of $\tilde{\mathbf{H}}_{\mathrm{TS}}$. Nevertheless, based on (25), estimating the high-dimensional $\tilde{\mathbf{H}}_{\mathrm{TS}}$ from the low-dimensional non-ISI region is difficult, and conventional LS and linear minimum mean square error (LMMSE) estimators would fail. Fortunately, the CS theory has proved that high-dimension signals can be accurately reconstructed by low-dimensional uncorrelated observations if the target signal is sparse or approximately sparse [38]. For the joint sparse signal recovery of (25), various signal recovery algorithms have been developed, which aim to exploit the inherent common sparstiy to jointly recover a set of sparse vectors for enhanced performance [38].

We propose to utilize the simultaneous matching pursuit (SOMP) algorithm [39] for fully exploiting the spatial-temporal joint sparsity of the CIR and the sparse traffic behavior of terrestrial IoT terminals, which is listed in the stage 1 part of **Algorithm 1**. Specifically, step 3-step 7 heuristically find the most correlated atoms in each iteration by calculating the correlation coefficients in step 3 and augment the support set of non-zero elements in step 4. According to the current support set, the locally optimal solution is calculated in step 5. Then the residual is updated in step 6 for the next iteration until the stop condition meets.

The estimated support set of $\hat{\tilde{\mathbf{H}}}_{\mathrm{TS}}$ is denoted as $\mathcal{I}$ and the individual index of support set divided for each IoT terminal can be denoted as $\Omega_k=\{\omega_k^q|\omega_k^q\in\mathcal{I},(k-1)L\leq\omega_k^q<kL\}$, where

ω_k^q is the q-th $(1 \leq q \leq |\Omega_k|_c)$ element of the set Ω_k. On the basis of the estimated support set and CIR vectors, a channel gain-based activity identificator is proposed for ATI as follows

$$\hat{\alpha}_k = \begin{cases} 1, & \frac{1}{P(N+1)} \sum_p \sum_{l=(k-1)L+1}^{kL} |\hat{\tilde{\mathbf{H}}}_{\mathrm{TS}_{[l,p]}}|^2 \geq \xi \\ 0, & \frac{1}{P(N+1)} \sum_p \sum_{l=(k-1)L+1}^{kL} |\hat{\tilde{\mathbf{H}}}_{\mathrm{TS}_{[l,p]}}|^2 < \xi \end{cases}, \tag{26}$$

where $\xi = \beta \max\{\frac{1}{P(N+1)} \sum_p \sum_{l=(k-1)L+1}^{kL} |\hat{\tilde{\mathbf{H}}}_{\mathrm{TS}_{[l,p]}}|^2, \forall k\}$ and $\beta = 0.1$ is the threshold factor[2]. As a result, the ATS can be represented by $\hat{\mathcal{A}} = \{k|\hat{\alpha}_k = 1, 1 \leq k \leq K\}$ and the cardinality of $\hat{\mathcal{A}}$ is denoted as $\hat{K}_a = |\hat{\mathcal{A}}|_c$.

D. CE Refinement with Parametric Approach

From the above discussion and (1), it can be observed that the separability, stability, and sparsity of the DD domain channels maintain in the TSLs, which motivates us to leverage the parametric approach to acquire the accurate estimation of the DD domain channel parameters and further refine the CE results. Specifically, the remanent RToA among different terminals and MPCs' delay for each terminal's TSL can be acquired from the index of support set of $\hat{\tilde{\mathbf{H}}}_{\mathrm{TS}}$ as [3]

$$\hat{\ell}_k^q = \omega_k^q - (k-1)L,\ k \in \hat{\mathcal{A}},\ 1 \leq q \leq |\Omega_k|_c. \tag{27}$$

Besides, the acquired $N+1$ sampled values of the time-varying CIR from $N+1$ adjacent TSs can facilitate the super-resolution estimation of the Doppler shift. Specifically, we can use the one-dimensional estimating signal parameters via rotational invariance techniques (ESPRIT) algorithm [40], which is a class of harmonic analysis algorithms by exploiting an underlying rotational invariance among signal subspaces.

For the convenience of the following estimation, we define a path-gain matrix $\hat{\Upsilon}_k^{q^*} \in \mathbb{C}^{(N+1)\times P}$ for the MPC with maximum energy as follows

$$\hat{\Upsilon}_k^{q^*} = \mathrm{mat}\left(\hat{\tilde{\mathbf{H}}}_{\mathrm{TS}_{[\omega_k^{q^*},:]}}, P, N+1\right)^{\mathrm{T}}, \forall k, \tag{28}$$

where $q^* = \arg\max_q \|\hat{\tilde{\mathbf{H}}}_{\mathrm{TS}_{[\omega_k^q,:]}}\|_2^2$. In fact, each column vector of $\hat{\Upsilon}_k^{q^*}$ is composed of CIR associated with multiple TSs, and different column vectors of it originate from different receive

[2]If the channel gain of the k-th IoT terminal is decided to be below the threshold ξ, the k-th IoT terminal is declared to be inactive.

[3]Here, we no longer distinguish LoS and NLoS paths, and uniformly treat them as MPCs with different subscript q.

Algorithm 1 Proposed Two-Stage Joint ATI and CE.

Input: Measurement signals $\mathbf{R}_{\mathrm{TS}}$ and sensing matrix $\mathbf{\Psi}$.

Output: Estimated activity indicator $\hat{\alpha}_k, \forall k$, the ATS $\hat{\mathcal{A}}$, and the correspoding CIR $\hat{h}_{k,p}[\kappa,\ell], \hat{h}^{\mathrm{DD}}_{k,p}[\ell,\upsilon], k \in \hat{\mathcal{A}}, \forall p$;

Stage 1 (Joint ATI and CE)

Initialize $t=1$, the residual $\mathbf{R}_0 = \mathbf{R}_{\mathrm{TS}}$, the index of support set $\mathcal{I} = \emptyset$, and define $I_{\max}$ as the maximum number of iterations;

while $t \leq I_{\max}$ **do**

  $i^* = \arg\max_{i=0,\ldots,KL-1} \sum_{k=1}^{(N+1)P} |\langle [\mathbf{R}_{t-1}]_{:,k}, \boldsymbol{\psi}_i \rangle|$;

  $\mathcal{I} = \mathcal{I} \cup \{i^*\}$;

  $\hat{\mathbf{H}}^{\mathrm{temp}}_{\mathrm{TS}} = \mathbf{\Psi}^{\dagger}\mathbf{R}_{\mathrm{TS}}$;

  $\mathbf{R}_t = \mathbf{R}_{\mathrm{TS}} - \mathbf{\Psi}_{\mathcal{I}}\hat{\mathbf{H}}^{\mathrm{temp}}_{\mathrm{TS}}$;

  $t = t+1$;

end while

$\hat{\tilde{\mathbf{H}}}_{\mathrm{TS}_{[\mathcal{I},:]}} = \hat{\mathbf{H}}^{\mathrm{temp}}_{\mathrm{TS}}$;

Compute the estimate of the activity indicator according to (26) and obtain the ATS $\hat{\mathcal{A}}$;

Stage 2 (CE refinement)

for $k \in \hat{\mathcal{A}}$ **do**

  Compute the estimate of the RToA and MPCs' delay according to (27);

  Estimate the Doppler shift according to (29)-(33);

  Compute the effective channel coefficients according to (37);

  Refine the results of CE by reconstructing CIR according to (38) and (39);

end for

antennas, where the Doppler shift is approximately identical and thus they can be regarded as multiple snapshots to mitigate the effects of noise. The main steps of Doppler estimation based on the ESPRIT algorightm are detailed as follows.

First of all, we divide two subarrays for each snapshot, which consist of CIR from the first N TSs and the last N TSs, respectively, as

$$\mathbf{x}^1_{k,p} = \hat{\Upsilon}^{q^*}_{k_{[1:N,p]}}, \mathbf{x}^2_{k,p} = \hat{\Upsilon}^{q^*}_{k_{[2:N+1,p]}}, \forall k,p, \tag{29}$$

and their combination $\mathbf{x}_{k,p} = \left[\mathbf{x}^{1^{\mathrm{T}}}_{k,p}\ \mathbf{x}^{2^{\mathrm{T}}}_{k,p}\right]^{\mathrm{T}} \in \mathbb{C}^{2N\times 1}$. In the presence of noise, the low rank

property of autocorrelation matrix

$$\mathbf{R}_{xx}^{k} = E\left[\mathbf{x}_{k,p}\mathbf{x}_{k,p}^{\mathrm{H}}\right] \approx \frac{1}{P}\sum_{p=1}^{P}\mathbf{x}_{k,p}\mathbf{x}_{k,p}^{\mathrm{H}}, \tag{30}$$

is destroyed. To mitigate the impact of noise, the eigenvalue decomposition (EVD) is utilized to distinguish the signal subspace and noise subspace, and we take the minimum eigenvalue $\hat{\sigma}_k^2$ as the estimate of the noise's variance. As a result, the noise cancelled autocorrelation matrix $\hat{\mathbf{R}}_{xx}^{k}$ can be calculated as

$$\hat{\mathbf{R}}_{xx}^{k} = \mathbf{R}_{xx}^{k} - \hat{\sigma}_k^2\mathbf{I}, \tag{31}$$

Then, the subspace of subarray $\mathbf{x}_{k,p}^{1}$ and $\mathbf{x}_{k,p}^{2}$ can be obtained by performing EVD on $\hat{\mathbf{R}}_{xx}^{k}$ as

$$\hat{\mathbf{R}}_{xx}^{k} = \hat{\mathbf{U}}_k\hat{\mathbf{\Sigma}}_k\hat{\mathbf{U}}_k^{\mathrm{H}}, \tag{32}$$

and the first column of the eigenvector matrix $\hat{\mathbf{U}}_k^{s}$ can approximate the dominant signal subspace of them, i.e., $\mathbf{e}_k^1 = \hat{\mathbf{U}}_{k_{[1:N,1]}}^{s}$, $\mathbf{e}_k^2 = \hat{\mathbf{U}}_{k_{[N+1:2N,1]}}^{s}$. In fact, $\mathbf{e}_k^1$ and $\mathbf{e}_k^2$ are characterized by rotational invariance [40]. Therefore, based on the LS criterion, the estimated Doppler shift can be calculated by

$$\hat{\upsilon}_k = \frac{N}{2\pi}\arg(\mathbf{e}^{1\dagger}_{k}\mathbf{e}_k^2). \tag{33}$$

Before proceeding with the estimation of channel coefficients of the MPCs, we introduce a lemma as follows.

Lemma 1: We assume that the support set of $\hat{\tilde{\mathbf{H}}}_{\mathrm{TS}}$ is estimated perfectly. The non-zero elements of the recoverd sparse CIR vector associated with the i-th TS and p-th receive antenna are defined as $\hat{\mathbf{h}}_{\mathrm{TS},p}^{\mathrm{eff},i} = \hat{\tilde{\mathbf{H}}}_{\mathrm{TS}_{[\mathcal{I},p+(i-1)P]}} \in \mathbb{C}^{Q\times 1}$, and the effective channel coefficient of the LoS and NLoS paths are denoted as

$$g_{k,p}^{\mathrm{eff,LoS}} = \sqrt{\frac{\gamma_k}{\gamma_k+1}}\left[\mathbf{v}_k\right]_p, \tag{34}$$

$$g_{k,p}^{\mathrm{eff},q} = \sqrt{\frac{1}{\gamma_k+1}}g_k^q\left[\mathbf{v}_k\right]_p. \tag{35}$$

Their relationship can be written as

$$\hat{\mathbf{h}}_{\mathrm{TS},p}^{\mathrm{eff},i} = \mathbf{\Psi}_{[:,\mathcal{I}]}^{\dagger}\mathbf{\Gamma}\boldsymbol{\eta}^{i-1}\odot\mathbf{g}_p^{\mathrm{eff}} + \mathbf{\Psi}_{[:,\mathcal{I}]}^{\dagger}\mathbf{w}_{\mathrm{TS},p}^{i}, \tag{36}$$

where the specific expression of $\mathbf{\Gamma}$, $\boldsymbol{\eta}^{i-1}$, and $\mathbf{g}_p^{\mathrm{eff}}$ can be referred to the Appendix.

Proof 1: Please refer to Appendix.

In line with the Lemma 1, the effective channel coefficients related to the p-th receive antenna can be mathematically calculated as

$$\hat{\mathbf{g}}_p^{\text{eff}} = \frac{1}{N+1}\sum_{i=1}^{N+1}\left[(\boldsymbol{\Psi}_{[:,\mathcal{I}]}^{\dagger}\hat{\boldsymbol{\Gamma}})^{-1}\hat{\mathbf{h}}_{\text{TS},p}^{\text{eff},i} \odot \frac{1}{\hat{\boldsymbol{\eta}}^{i-1}}\right]. \tag{37}$$

Up to now, the dominated DD domain channel's parameters have been acquired, and the results of CE refinement can be expressed as

$$\hat{h}_{k,p}^{\text{DD}}[\ell,\upsilon] = \sum_{q=1}^{|\Omega_k|_c}\hat{g}_{k,p}^{\text{eff},q}\delta[\ell-\hat{\ell}_k^q]\delta[\upsilon-\hat{\upsilon}_k], \forall k,p, \tag{38}$$

where $\hat{g}_{k,p}^{\text{eff},q}$ is the element of $\hat{\mathbf{g}}_p^{\text{eff}}$ related to the q-th path and the k-th terminal. Besides, the estimate of time-varying CIR can be represented by

$$\hat{h}_{k,p}[\kappa,\ell] = \sum_{q=1}^{|\Omega_k|_c}\hat{g}_{k,p}^{\text{eff},q}e^{j2\pi\frac{\hat{\upsilon}_k(\kappa-\hat{\ell}_k^q)}{N(M+M_t)}}\delta[\ell-\hat{\ell}_k^q], \forall k,p. \tag{39}$$

Remark 2: In fact, ignoring the noise term, (36) can be rewritten as

$$\hat{\mathbf{h}}_{\text{TS},p}^{\text{eff},i} \approx \boldsymbol{\zeta} \odot \boldsymbol{\eta}(i,\{\upsilon_k\}_{k\in\mathcal{A}}), \tag{40}$$

where $\boldsymbol{\zeta} = \boldsymbol{\Psi}_{[:,\mathcal{I}]}^{\dagger}\boldsymbol{\Gamma}\mathbf{g}_p^{\text{eff}} \in \mathbb{C}^{Q\times 1}$ is a Doppler-invariant and i-invariant vector, and η is a function of i and Doppler shift $\{\upsilon_k\}_{k\in\mathcal{A}}$. The linear relationship between $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ ensures that the proposed Doppler shift estimation method is immune from the uncertainty of $\boldsymbol{\zeta}$, i.e., the approximation error in (19), which also guarantees the effectiveness of the proposed CE refinement.

Consequently, we have finished the discussion of the proposed two-stage joint ATI and CE scheme, and the complete procedures are listed in **Algorithm 1**.

V. Signal Detection

Based on the above ATI and CE results, we develop a LS-based parallel time domain multi-user SD for demodulating the OTFS signals with relatively low computational complexity in this section.

A. Received Signal Preprocessing

As illustrated in Fig. 3, on the one hand, the received OTFS payload data symbols over the time dispersive channels would be contaminated by trailing of the preceding TS; on the other hand, the data symbols can also contaminate the following TS part. These can lead to severe ISI. Fortunately, TSs are known by the transceiver. With the estimated CSI, the aforementioned ISI

can be eliminated to facilitate the following SD. The details of the preprocessing of the received signals before SD are presented as follows.

First, (9) can be rewritten as a vector form as

$$\mathbf{r}_p = \sum_{k=1}^{K} \alpha_k \underbrace{\left(\mathbf{\Pi}_{k,p}^{\mathrm{LoS}} + \sum_{q=1}^{Q_k} \mathbf{\Pi}_{k,p}^{q} \right)}_{\mathbf{\Pi}_{k,p}} \mathbf{s}_k + \mathbf{w}_p, \ \forall p, \tag{41}$$

where $\mathbf{\Pi}_{k,p}^{\mathrm{LoS}}$ (also $\mathbf{\Pi}_{k,p}^{q}$) $\in \mathbb{C}^{(MN+M_tN+M_t)\times(MN+M_tN+M_t)}$ consists of elements of time-varying CIR in (7), and its (m,n)-th element is given by

$$\mathbf{\Pi}_{k,p_{[m,n]}}^{\mathrm{LoS}} = \begin{cases} h_{k,p}[m-1, \ell_k^{\mathrm{LoS}}], & m = n + \ell_k^{\mathrm{LoS}} \\ 0, & \text{otherwise} \end{cases}, \tag{42}$$

and

$$\mathbf{\Pi}_{k,p_{[m,n]}}^{q} = \begin{cases} h_{k,p}[m-1, \ell_k^{q}], & m = n + \ell_k^{q} \\ 0, & \text{otherwise} \end{cases}. \tag{43}$$

Consequently, the ISI in the OTFS payload data symbol caused by the tail of the preceeding TS can be estimated as

$$\mathbf{r}_p^{\mathrm{ISI}} = \sum_{k=1}^{K} \hat{\alpha}_k \hat{\mathbf{\Pi}}_{k,p} \mathring{\mathbf{s}}_k, \ \forall p, \tag{44}$$

where $\mathring{\mathbf{s}}_k$ consists of TSs and zero sequences, i.e., $\mathring{\mathbf{s}}_k = \left[\mathbf{c}_k^{\mathrm{T}}, \mathbf{0}_{M\times 1}^{\mathrm{T}}, \mathbf{c}_k^{\mathrm{T}}, \mathbf{0}_{M\times 1}^{\mathrm{T}}, \ldots, \mathbf{c}_k^{\mathrm{T}}, \mathbf{0}_{M\times 1}^{\mathrm{T}}, \mathbf{c}_k^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{(M_tN+MN+M_t)\times 1}$, $\hat{\mathbf{\Pi}}_{k,p}$ is the estimate of $\mathbf{\Pi}_{k,p}$ with its elements padded by the estimated CIR in (39).

Besides, to form the "cyclic convolution" relationship between the OTFS data signal and the CIR (here we first assume the CIR is time-invariant, and then this assumption will be relaxed), the data tail part (cause the ISI to the following TS) will be superposed onto the header of each OTFS data symbol region. In fact, such a data tail part can be acquired and shifted by

$$\mathbf{r}_p^{\mathrm{tail}} = \mathbf{R}_t^{\mathrm{T}} (\mathbf{I}_N \otimes \mathbf{R}_s) \mathbf{R}_t (\mathbf{r}_p - \mathbf{r}_p^{\mathrm{ISI}}), \ \forall p \tag{45}$$

where $\mathbf{R}_t = \left[\mathbf{0}_{(M+M_t)N\times M_t} \ \mathbf{I}_{(M+M_t)N}\right] \in \mathbb{C}^{(M+M_t)N\times(MN+M_tN+M_t)}$ and the (m,n)-th element of $\mathbf{R}_s \in \mathbb{C}^{(M+M_t)\times(M+M_t)}$ is defined as

$$\mathbf{R}_{s_{[m,n]}} = \begin{cases} 1 & m-n = M, \ n \in [1, M_t] \\ 0 & \text{otherwise} \end{cases}. \tag{46}$$

Therefore, the preprocessed OTFS payload data symbols $\hat{\mathbf{r}}_p \in \mathbb{C}^{MN\times 1}$ can be finally acquired after removing the TSs as

$$\hat{\mathbf{r}}_p = (\mathbf{I}_N \otimes \mathbf{R}_r)\mathbf{R}_t(\mathbf{r}_p - \mathbf{r}_p^{\mathrm{ISI}} + \mathbf{r}_p^{\mathrm{tail}}),\ \forall p, \tag{47}$$

where $\mathbf{R}_r = [\mathbf{I}_M\ \mathbf{0}_{M\times M_t}] \in \mathbb{C}^{M\times(M+M_t)}$.

B. LS-Based Parallel Time Domain Multi-User SD

It has been shown that with the fractional Doppler shift, the TF domain and DD domain effective channel matrices could be not very sparse due to the Doppler spreading with the limited Doppler resolution, while the sparsity of time domain channel remains to hold [35]. Therefore, we are motivated to perform multi-user SD in the time domain to exploit its sparse pattern for lower computational complexity. In fact, based on (41) and (44), we have

$$\begin{aligned}\mathbf{r}_p - \mathbf{r}_p^{\mathrm{ISI}} &= \sum_{k\in\hat{\mathcal{A}}} \hat{\mathbf{\Pi}}_{k,p}\left(\mathbf{s}_k - \mathring{\mathbf{s}}_k\right) + \sum_{k\in\hat{\mathcal{A}}} \left(\mathbf{\Pi}_{k,p} - \hat{\mathbf{\Pi}}_{k,p}\right)\mathbf{s}_k \\ &\quad + \sum_{k\in(\mathcal{A}-\hat{\mathcal{A}})} \mathbf{\Pi}_{k,p}\mathbf{s}_k + \mathbf{w}_p \\ &= \sum_{k\in\hat{\mathcal{A}}} \hat{\mathbf{\Pi}}_{k,p}\left(\mathbf{s}_k - \mathring{\mathbf{s}}_k\right) + \hat{\mathbf{w}}_p,\end{aligned} \tag{48}$$

where $\hat{\mathbf{w}}_p = \sum_{k\in\hat{\mathcal{A}}} \left(\mathbf{\Pi}_{k,p} - \hat{\mathbf{\Pi}}_{k,p}\right)\mathbf{s}_k + \sum_{k\in(\mathcal{A}-\hat{\mathcal{A}})} \mathbf{\Pi}_{k,p}\mathbf{s}_k + \mathbf{w}_p$ is the effective noise vector including errors in the signal preprocessing and AWGN. Furthermore, by stacking (45), (47), and (48), we have

$$\begin{aligned}\hat{\mathbf{r}}_p &= \sum_{k\in\hat{\mathcal{A}}} \left[\mathbf{I}_N \otimes \mathbf{R}_r(\mathbf{I}_{M+M_t} + \mathbf{R}_s)\right]\mathbf{R}_t\left(\mathbf{r}_p - \mathbf{r}_p^{\mathrm{ISI}}\right) \\ &= \sum_{k\in\hat{\mathcal{A}}} \left[\mathbf{I}_N \otimes \mathbf{R}_r(\mathbf{I}_{M+M_t} + \mathbf{R}_s)\right]\mathbf{R}_t\hat{\mathbf{\Pi}}_{k,p}\mathbf{R}_t^{\mathrm{T}}(\mathbf{I}_N \otimes \mathbf{A}_t) \\ &\quad \tilde{\mathbf{s}}_k + \left[\mathbf{I}_N \otimes \mathbf{R}_r(\mathbf{I}_{M+M_t} + \mathbf{R}_s)\right]\mathbf{R}_t\hat{\mathbf{w}}_p,\end{aligned} \tag{49}$$

where $\mathbf{A}_t = \left[\mathbf{0}^{\mathrm{T}}_{(M+M_t)N\times M_t}\ \mathbf{I}^{\mathrm{T}}_{(M+M_t)N}\right]^{\mathrm{T}}$.

With the aid of the TSs and the preprocessing aforementioned, the ISI between adjacent OTFS data symbols can be avoided. Hence, the SD in the time domain can be performed in parallel for

N OTFS payload data symbols, which can significantly reduce the computational complexity. As a result, (49) can be further decomposed into

$$\hat{\mathbf{r}}_p^i = \sum_{k\in\hat{\mathcal{A}}} \underbrace{\mathbf{R}_r(\mathbf{I}_{M+M_t}+\mathbf{R}_s)\hat{\mathbf{\Pi}}_{k,p}^i\mathbf{A}_t}_{\hat{\mathbf{U}}_{p,k}^i}\tilde{\mathbf{s}}_k^i + \hat{\mathbf{w}}_p^i,\ \forall i,p, \tag{50}$$

where $\hat{\mathbf{\Pi}}_{k,p}^i = \left(\mathbf{R}_t\hat{\mathbf{\Pi}}_{k,p}\mathbf{R}_t^T\right)_{[(i-1)(M+M_t)+1:i(M+M_t)]} \in \mathbb{C}^{(M+M_t)\times(M+M_t)}$, $\hat{\mathbf{r}}_p^i = \hat{\mathbf{r}}_{p_{[(i-1)M+1:iM]}} \in \mathbb{C}^{M\times 1}$, and $\hat{\mathbf{w}}_p^i \in \mathbb{C}^{M\times 1}$ is the corresponding noise vector.

Moreover, we intend to extend (50) to jointly process the received signal from P receive antennas as

$$\hat{\mathbf{r}}^i = \hat{\mathbf{U}}^i\tilde{\mathbf{s}}^i + \hat{\mathbf{w}}^i,\ \forall i, \tag{51}$$

where $\hat{\mathbf{U}}^i \in \mathbb{C}^{PM\times\hat{K}_aM}$ is a block matrix

$$\hat{\mathbf{U}}^i = \begin{bmatrix} \hat{\mathbf{U}}_{1,k_1}^i & \hat{\mathbf{U}}_{1,k_2}^i & \cdots & \hat{\mathbf{U}}_{1,k_{\hat{K}_a}}^i \\ \hat{\mathbf{U}}_{2,k_1}^i & \hat{\mathbf{U}}_{2,k_2}^i & \cdots & \hat{\mathbf{U}}_{2,k_{\hat{K}_a}}^i \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{U}}_{P,k_1}^i & \hat{\mathbf{U}}_{P,k_2}^i & \cdots & \hat{\mathbf{U}}_{P,k_{\hat{K}_a}}^i \end{bmatrix}, \tag{52}$$

$k_1,k_2,\ldots,k_{\hat{K}_a}$ are the elements of the set $\hat{\mathcal{A}}$, $\hat{\mathbf{r}}^i = \left[\hat{\mathbf{r}}_1^{i\mathrm{T}},\hat{\mathbf{r}}_2^{i\mathrm{T}},\ldots,\hat{\mathbf{r}}_P^{i\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{MP\times 1}$, $\tilde{\mathbf{s}}^i = \left[\tilde{\mathbf{s}}_{k_1}^{i\mathrm{T}},\tilde{\mathbf{s}}_{k_2}^{i\mathrm{T}},\ldots,\tilde{\mathbf{s}}_{k_{\hat{K}_a}}^{i\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{M\hat{K}_a\times 1}$, and $\hat{\mathbf{w}}^i$ denotes the noise vectors from different received antennas.

Therefore, given $P \geq K_a$, the time domain OTFS signals of different active terminals can be detected by calculating the LS solution of (51). Benefitting from the sparsity of TSLs, $\hat{\mathbf{U}}^i$ displays favorable sparse pattern, where we can further utilize the fast method, such as LS QR-factorization (LSQR) [41], to facilitate the approximate solution of sparse linear equations and make a tradeoff between the computational complexity and detection accuracy.

VI. Performance Evaluation

A. Simulation Setup

In this section, we carry out extensive simulation investigations to prove the effectiveness of our proposed scheme under different parameter configurations, and compare it with the state-of-the-art solutions. First of all, we define the error probability P_e for ATI as

$$P_e = \frac{1}{K}\sum_{k=1}^{K}|\hat{\alpha}_k - \alpha_k|, \tag{53}$$

the normalized mean square error (NMSE) for CE as

$$\mathrm{NMSE} = \frac{\sum_{k=1}^{K}\sum_{p=1}^{P}||\hat{\alpha}_k\hat{\mathbf{\Pi}}_{k,p} - \alpha_k\mathbf{\Pi}_{k,p}||_{\mathrm{F}}^2}{\sum_{k=1}^{K}\sum_{p=1}^{P}||\alpha_k\mathbf{\Pi}_{k,p}||_{\mathrm{F}}^2}, \tag{54}$$

and the uncoded bit error rate (BER) for SD as

$$\mathrm{BER} = \frac{E_u NMM_b + B_a}{K_a NMM_b}, \tag{55}$$

where E_u is the number of active terminals missed to be detected, B_a is the overall error bits for the correctly identified active terminals, $K_a NMM_b$ is the total bits transmitted by K_a active terminals, and M_b is the modulation order.

For the massive MIMO based massive connectivity in IoT, the performance of payload data demodulation highly depends on the channel statistics of the simultaneously served terminals. As for the LEO constellations-based IoT, the channel characteristics are mainly determined by the simultaneously served terminals' AoA observed at the satellite receiver, i.e., θ_k^{zen} and $\theta_k^{\mathrm{azi}}, \forall k \in \mathcal{A}$ [29]. Therefore, channels of IoT terminals owning minor AoA differences are highly correlated, which inevitably leads the multi-user MIMO channel matrix to be ill-conditioned with considerably performance deterioration if they are allocated with the same TF or DD resources. To this end, we adopt a space angle user grouping (SAUG) strategy as [29] did. In brief, for K terminals scheduled into the same DD resources for RA, their zenith and azimuth spacing satisfy

$$|\theta_{k_1}^{\mathrm{zen}} - \theta_{k_2}^{\mathrm{zen}}| \geq \Delta_z, \tag{56}$$

$$|\theta_{k_1}^{\mathrm{azi}} - \theta_{k_2}^{\mathrm{azi}}| \geq \Delta_a, \tag{57}$$

where $k_1, k_2 \in \mathcal{A}$ and $k_1 \neq k_2$, Δ_z and Δ_e are the preset minimum zenith and azimuth spacing to avoid the MIMO channel matrix to be ill-conditioned. The principle of SAUG is well investigated in [29] and the detailed procedure can refer to it.

Additionally, to meet the mutual coherence property (MCP) [38] of the sensing matrix for reliable recovery of sparse vectors in (25), we assume the time domain TS associated with the k-th terminal is generated from a standard complex Gaussian distribution, i.e., $\mathbf{c}_k \sim \mathcal{CN}(0, 1)$. Other detailed system parameters for the following simulations are summarized in Table II. Besides, the number of potential and active IoT terminals is fixed as $K = 100$ and $K_a = 10$, respectively, and the signal-to-noise ratio (SNR) of the received signals is set as $\mathrm{SNR} = 20\,\mathrm{dB}$ unless otherwise mentioned.

Since the typical value of the sampling time T_s in the delay domain is usually sufficiently small, the impact of fractional delays in typical broadband systems can be neglected, while the fractional part of normalized Doppler can not be neglected due to large value of the sampling time T in the Doppler domain [35].

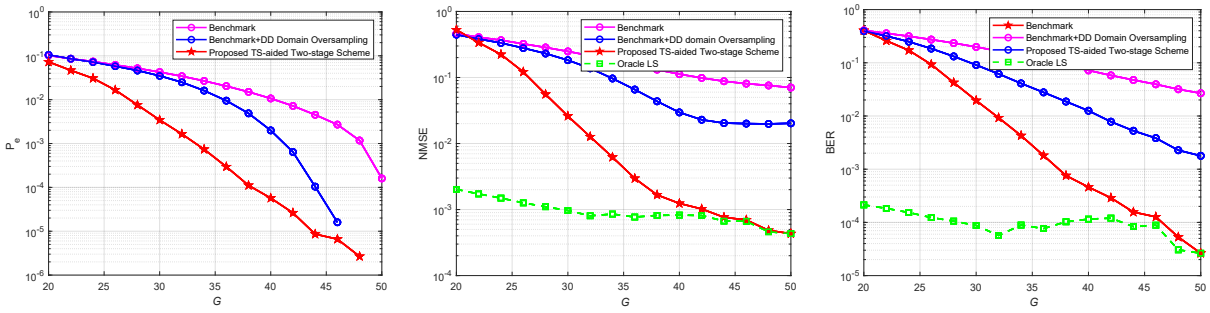
TABLE II

SIMULATION PARAMETERS

Contents	**Parameters**	**Values**
System	Carrier frequency (GHz)	10
	Subcarrier spacing (kHz)	480
	Bandwidth (MHz)	122.88
	Size of OTFS frame (M,N)	(256,8)
	Modulation scheme	QPSK
	Number of satellite antennas (P_x,P_y)	(5,5)
	Number of terminal antennas	1
TSL	LEO satellite velocity (km/s)	7.58
	IoT terminals velocity (m/s)	$0 \sim 10$
	RToA τ_k^{LoS} and MPCs' delay τ_k^q (ms)	$0 \sim 0.067$
	Doppler shift of TSL (kHz) ν_k	$0 \sim 178.2$
	Range of zenith angle θ_k^{zen}	$[-44.7^\circ, 44.7^\circ]$
	Range of azimuth angle θ_k^{azi}	$[0, 360^\circ)$
	Preset minimum angular spacing (Δ_z,Δ_e)	$(14.4^\circ, 14.3^\circ)$

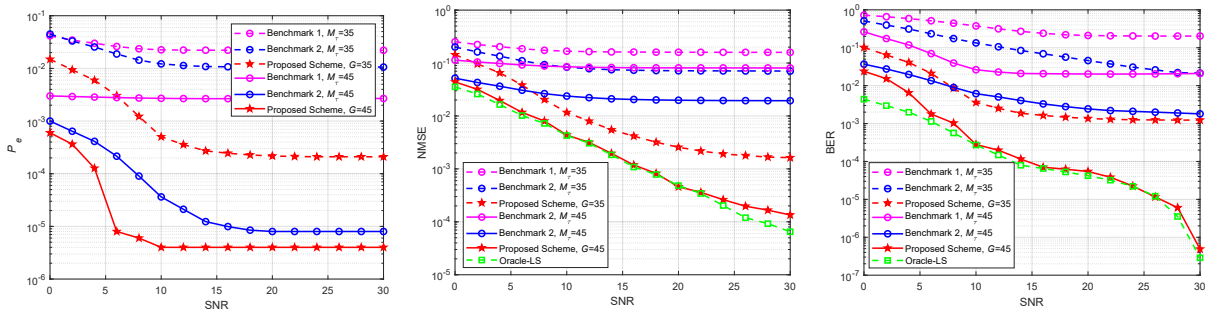
B. Performance under LoS TSL

As a fledgling concept, the GF RA has first been integrated with OTFS waveform in this paper, and there has been little work dedicated to the field of ATI, CE, and multi-user SD in the framework of OTFS. We take one of the most representative schemes proposed in [33] as the Benchmark 1 for comparison, which embeds the guard and non-orthogonal pilot symbols in the DD domain to facilitate downlink massive MIMO-OTFS channel estimation, and it is a dual problem of uplink ATI and CE with massive connectivity. The size of embedded DD domain pilots along the Doppler dimension and the delay dimension are denoted as N_ν and M_τ, respectively, and N is fixed as $N = N_\nu$. Besides, the size of embedded DD domain guard symbols along the Doppler dimension and the delay dimension, which are utilized to eliminate ISI, is set as $N_g = 0$ and $M_g = L$, respectively. Besides, we set Benchmark 2, where the virtual



(a) (b) (c)

Fig. 5. Performance comparision versus the proportion of effective pilot symbols overhead: (a) P_e performance comparison; (b) NMSE performance comparison; (c) BER performance comparison.



(a) (b) (c)

Fig. 6. Performance comparision versus SNR: (a) P_e performance comparison; (b) NMSE performance comparison; (c) BER performance comparison.

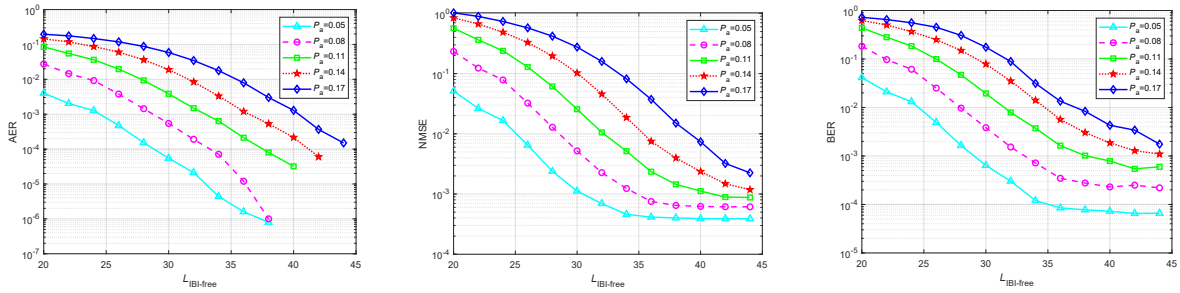
sampling grid in the DD space [36] is attached for the Benchmark 1 for further comparison, and its virtual sampling grid size in the Doppler domain is fixed as $N' = 2N$.

Fig. 5 provides P_e, NMSE, and BER performance comparison versus the size of effective pilot symbols overhead. To evaluate the performance fairly, we define the proportion of effective pilot overheads for the benckmark and the proposed scheme as $\varepsilon_b = \frac{M_\tau}{M}$ and $\varepsilon_p = \frac{G(N+1)}{(M+M_t)N}$, respectively. On the one hand, it can be observed from Fig. 5, that Benchmark 1 suffers serious performance degradation and the performance gain of the proposed method over the Benchmark 1 is particularly noticeable. To figure out the rationality behind this phenomenon, the result of Benchmark 2 is presented. As Fig. 5 exhibits, the superiority of Benchmark 2 over Benchmark 1 is self-evident and it indicates that it's the low-resolution of Doppler domain that severely holds back the performance of Benchmark 1 especially when the small size of Doppler dimension

N is adopted. However, oversize N is prohibitive in the LEO constellations for the intolerable computational complexity and signal processing latency. And more importantly, the quasi-static property of TSLs in the DD domain could no longer hold as N increases. Therefore, the proposed scheme with the Doppler domain super-solution enabled by the time domain TSs and parametric approach is rewarding in this kind of harsh channel conditions. On the other hand, the NMSE and BER performance of the proposed method is very close to Oracle-LS when $\varepsilon_p \geq 13.7\,\%$, which manifests that the approximation error of (19) only leads to a slight increase of TSs overhead to ensure the performance of sparse signal recovery. Meanwhile, the indisputable superiority of the proposed method even with lower TSs overhead demonstrates that the impact of approximation error on the following CE refinement is negligible in contrast to the low-resolution of Doppler domain of the Benchmark 1 and Benchmark 2.

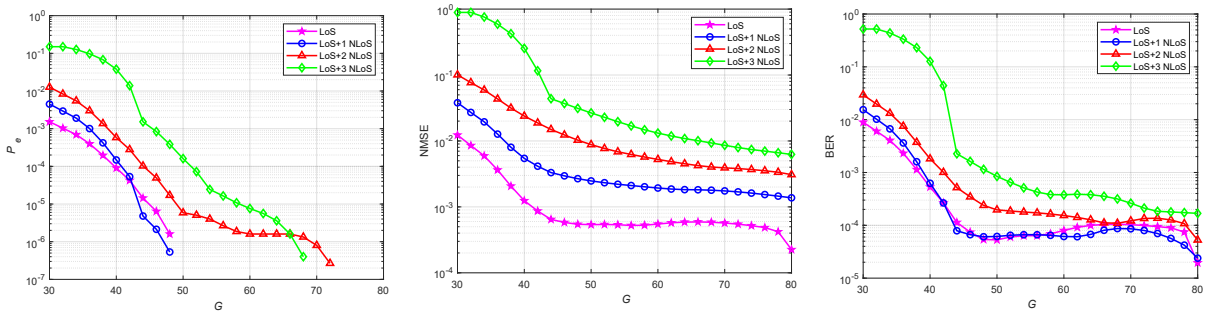
Fig. 6 exhibits P_e, NMSE, and BER performance comparisons versus the variation of SNR. It can be observed that when the effective pilot overheads are close, the noticeable gain of our proposed scheme hold in the almost whole regime of SNR (0-30 dB) in contrast to Benchmark 1 and Benchmark 2. This can be interpreted that in the range of low SNR, the effective utilization of both the spatial and temporal correlations in the TSLs considerably promotes the accuracy of sparse signal recovery, and as a result, our proposed scheme outperforms the benchmarks. Moreover, in the range of high SNR, the system performance is mainly dominated by the CE performance. In spite of the approximation error, our proposed scheme overcomes the problem of low-resolution in the Doppler domain and reaps more satisfactory performance.

Moreover, in order to show the applicability of our proposed scheme in various IoT applications, we investigate P_e, NMSE, and BER performance comparisons versus the activation probability $P_a = K_a/K$ and the numerical results are illustrated in Fig. 7. As the results show, the performance of ATI, CE, and SD exhibit a similar deteriorating trend with an increasing number of active IoT terminals trying to access the LEO satellite in the same DD resources. Although the performance degrades at a fast pace with the fixed effective pilot overheads, the superiority of our proposed scheme over Benchmark 1 and Benchmark 2 is still retained. As a matter of fact, this loss can be compensated to some extent when a larger size of TSs is employed.



(a) (b) (c)

Fig. 7. Performance comparision versus activation probability P_a: (a) P_e performance comparison; (b) NMSE performance comparison; (c) BER performance comparison.



(a) (b) (c)

Fig. 8. Performance comparision versus the channel condition of the proposed scheme: (a) P_e performance; (b) NMSE performance; (c) BER performance.

C. Performance under Different Channel Conditions

To further demonstrate the robustness of the proposed method, we investigate its performance under different TSL conditions. Fig. 8 displays P_e, NMSE, and BER performance comparisons with the variation of MPCs, while the Rician factor is fixed at $\gamma_k = 15\,\mathrm{dB}, \forall k$. It can be observed that with the increase of MPCs, there is a slight rise of the effective pilot overheads to guarantee constant performance. This can be interpreted that the increase of MPCs leads to more observations to recover the increasing non-zero elements of sparse CIR vectors. In fact, despite the fact that the performance of NMSE deteriorates at a relatively rapid rate, the performance of P_e and BER degrades sluggishly. It verifies the system performance is mainly determined by the accuracy of estimation of the LoS path and those low-energy NLoS paths have negligible impact on the system performance. Besides, it is noteworthy that the increase of MPCs could

contribute to the enhancement of ATI, which could be treated as a diversity gain.

VII. Conclusion

This paper investigates an effective RA paradigm for accommodating massive IoT access based on the LEO constellations. Specifically, we first propose the GF NOMA-OTFS scheme to mitigate the access scheduling overheads and latency, and combat the severe Doppler effect of TSLs. On this basis, to handle the challenging problem of ATI, CE, and SD, we further develop a TS-OTFS transmission scheme and a two-stage joint ATI and CE method. At the first stage, the time domain TSs facilitate us to leverage the traffic sparsity of IoT terminals and the sparse CIR to jointly perform ATI and CE. Furthermore, a parametric approach is introduced to refine the CE performance based on the sparsity of TSLs in the DD domain. With the results of ATI and CE, we are further motivated to propose a time-domain parallel multi-user SD with relatively low computational complexity to circumvent the channel spreading in the DD or TF domain. Simulation results of the effectiveness and superiority of our proposed paradigm particularly for LEO constellations-based scenario.

Appendix

Proof of Lemma 1

In fact, under the assumption that the support set of $\hat{\tilde{\mathbf{H}}}_{\mathrm{TS}}$ is perfectly recovered, the non-zero elements $\hat{\mathbf{h}}_{\mathrm{TS},p}^{\mathrm{eff},i}$ of $\hat{\tilde{\mathbf{H}}}_{\mathrm{TS}_{[:,p+(i-1)P]}}$ can be derived from

$$\hat{\mathbf{h}}_{\mathrm{TS},p}^{\mathrm{eff},i} = \mathbf{\Psi}_{[:,\mathcal{I}]}^{\dagger}\mathbf{r}_{\mathrm{TS},p}^{i}. \tag{58}$$

From (14), it can be further written as

$$\hat{\mathbf{h}}_{\mathrm{TS},p}^{\mathrm{eff},i} = \mathbf{\Psi}_{[:,\mathcal{I}]}^{\dagger}\sum_{k\in\mathcal{A}}\left(\mathbf{\Delta}_k^{\mathrm{LoS}}\boldsymbol{\psi}_k^{\mathrm{LoS}}h_{\mathrm{TS},k,p}^{i}(\ell_k^{\mathrm{LoS}}+1)+\sum_{q=1}^{Q_k}\mathbf{\Delta}_k^{q}\boldsymbol{\psi}_k^{q}h_{\mathrm{TS},k,p}^{i}(\ell_k^{q}+1)\right)+\underbrace{\mathbf{\Psi}_{[:,\mathcal{I}]}^{\dagger}\mathbf{w}_{\mathrm{TS},p}^{i}}_{\text{noise}}, \tag{59}$$

where $\boldsymbol{\psi}_k^{\mathrm{LoS}} = \mathbf{\Psi}_{k_{[:,\ell_k^{\mathrm{LoS}}+1]}}$ and $\boldsymbol{\psi}_k^{q} = \mathbf{\Psi}_{k_{[:,\ell_k^{q}+1]}}$. Ignoring the noise term, (59) can be further approximate to the vector form as

$$\begin{aligned}\hat{\mathbf{h}}_{\mathrm{TS},p}^{\mathrm{eff},i} \approx \mathbf{\Psi}_{[:,\mathcal{I}]}^{\dagger}\sum_{k\in\mathcal{A}}&\left[\mathbf{\Delta}_k^{\mathrm{LoS}}\boldsymbol{\psi}_k^{\mathrm{LoS}},\mathbf{\Delta}_k^{1}\boldsymbol{\psi}_k^{1},\ldots,\mathbf{\Delta}_k^{Q_k}\boldsymbol{\psi}_k^{Q_k}\right]\times\\&\left[h_{\mathrm{TS},k,p}^{i}(\ell_k^{\mathrm{LoS}}+1),h_{\mathrm{TS},k,p}^{i}(\ell_k^{1}+1),\ldots,h_{\mathrm{TS},k,p}^{i}(\ell_k^{Q_k}+1)\right]^{\mathrm{T}}.\end{aligned} \tag{60}$$

According to the CIR model in (8), (60) can be further expressed as

$$\hat{\mathbf{h}}_{\mathrm{TS},p}^{\mathrm{eff},i} \approx \boldsymbol{\Psi}_{[:,\mathcal{I}]}^{\dagger} \sum_{k\in\mathcal{A}} \left[\boldsymbol{\Delta}_k^{\mathrm{LoS}}\boldsymbol{\psi}_k^{\mathrm{LoS}}, \boldsymbol{\Delta}_k^{1}\boldsymbol{\psi}_k^{1}, \ldots, \boldsymbol{\Delta}_k^{Q_k}\boldsymbol{\psi}_k^{Q_k}\right] \times \left[\underbrace{\sqrt{\frac{\gamma_k}{\gamma_k+1}}\left[\mathbf{v}_k\right]_p}_{g_{k,p}^{\mathrm{eff,LoS}}} e^{j2\pi\upsilon_k^{\mathrm{LoS}}\left[\frac{i-1}{N}+\frac{(L-\ell_k^{\mathrm{LoS}})}{N(M+M_t)}\right]}, \right.$$
$$\left. \underbrace{\sqrt{\frac{1}{\gamma_k+1}}g_k^{1}\left[\mathbf{v}_k\right]_p}_{g_{k,p}^{\mathrm{eff},1}} e^{j2\pi\upsilon_k^{1}\left[\frac{i-1}{N}+\frac{(L-\ell_k^{1})}{N(M+M_t)}\right]}, \ldots, \underbrace{\sqrt{\frac{1}{\gamma_k+1}}g_k^{Q_k}\left[\mathbf{v}_k\right]_p}_{g_{k,p}^{\mathrm{eff},Q_k}} e^{j2\pi\upsilon_k^{Q_k}\left[\frac{i-1}{N}+\frac{(L-\ell_k^{Q_k})}{N(M+M_t)}\right]}\right]^{\mathrm{T}}, \tag{61}$$

Finally, by extracting the effective channel coefficients, (61) can be decomposed into

$$\hat{\mathbf{h}}_{\mathrm{TS},p}^{\mathrm{eff},i} \approx \boldsymbol{\Psi}_{[:,\mathcal{I}]}^{\dagger} \sum_{k\in\mathcal{A}} \underbrace{\left[\boldsymbol{\Delta}_k^{\mathrm{LoS}}\boldsymbol{\psi}_k^{\mathrm{LoS}}, \boldsymbol{\Delta}_k^{1}\boldsymbol{\psi}_k^{1}, \ldots, \boldsymbol{\Delta}_k^{Q_k}\boldsymbol{\psi}_k^{Q_k}\right]}_{\boldsymbol{\Gamma}_k} \underbrace{\left[g_{k,p}^{\mathrm{eff,LoS}}, g_{k,p}^{\mathrm{eff},1}, \ldots, g_{k,p}^{\mathrm{eff},Q_k}\right]^{\mathrm{T}}}_{\mathbf{g}_{k,p}^{\mathrm{eff}}}$$
$$\odot \underbrace{\left[e^{j2\pi\upsilon_k^{\mathrm{LoS}}\left[\frac{i-1}{N}+\frac{(L-\ell_k^{\mathrm{LoS}})}{N(M+M_t)}\right]}, e^{j2\pi\upsilon_k^{1}\left[\frac{i-1}{N}+\frac{(L-\ell_k^{1})}{N(M+M_t)}\right]}, \ldots, e^{j2\pi\upsilon_k^{Q_k}\left[\frac{i-1}{N}+\frac{(L-\ell_k^{Q_k})}{N(M+M_t)}\right]}\right]^{\mathrm{T}}}_{\boldsymbol{\eta}_k^{i-1}}. \tag{62}$$

Therefore, the mathematical relationship between $\hat{\mathbf{h}}_{\mathrm{TS},p}^{\mathrm{eff},i}$ and the effective channel coefficients can be represented by

$$\hat{\mathbf{h}}_{\mathrm{TS},p}^{\mathrm{eff},i} = \boldsymbol{\Psi}_{[:,\mathcal{I}]}^{\dagger} \sum_{k\in\mathcal{A}} \boldsymbol{\Gamma}_k\boldsymbol{\eta}_k^{i-1} \odot \mathbf{g}_{k,p}^{\mathrm{eff}} + \boldsymbol{\Psi}_{[:,\mathcal{I}]}^{\dagger}\mathbf{w}_{\mathrm{TS},p}^{i}. \tag{63}$$

Furthermore, by collecting the vectors and matrices with different subscripts k, (63) can be vectorized to

$$\hat{\mathbf{h}}_{\mathrm{TS},p}^{\mathrm{eff},i} = \boldsymbol{\Psi}_{[:,\mathcal{I}]}^{\dagger}\boldsymbol{\Gamma}\boldsymbol{\eta}^{i-1} \odot \mathbf{g}_{p}^{\mathrm{eff}} + \boldsymbol{\Psi}_{[:,\mathcal{I}]}^{\dagger}\mathbf{w}_{\mathrm{TS},p}^{i}, \tag{64}$$

where $\boldsymbol{\Gamma} = [\boldsymbol{\Gamma}_{k_1}, \boldsymbol{\Gamma}_{k_2}, \ldots, \boldsymbol{\Gamma}_{K_a}] \in \mathbb{C}^{G\times Q}$, $\boldsymbol{\eta}^{i-1} = \left[\boldsymbol{\eta}_{k_1}^{i-1,\mathrm{T}}, \boldsymbol{\eta}_{k_2}^{i-1,\mathrm{T}}, \ldots, \boldsymbol{\eta}_{k_{Ka}}^{i-1,\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{Q\times 1}$, and $\mathbf{g}_{p}^{\mathrm{eff}} = \left[\mathbf{g}_{k_1,p}^{\mathrm{eff,T}}, \mathbf{g}_{k_2,p}^{\mathrm{eff,T}}, \ldots, \mathbf{g}_{k_{Ka},p}^{\mathrm{eff,T}}\right]^{\mathrm{T}} \in \mathbb{C}^{Q\times 1}$ with $k_1, k_1, \ldots, k_{K_a} \in \mathcal{A}$.

It's clear that $\hat{\mathbf{h}}_{\mathrm{TS},p}^{\mathrm{eff},i}$ and $\mathbf{g}_{p}^{\mathrm{eff}}$ have linear relationship. Since $\boldsymbol{\Gamma}$ and $\boldsymbol{\eta}^{i-1}$ can be reconstructed with the estimated Doppler shift, RToA and MPCs' delay, $\mathbf{g}_{p}^{\mathrm{eff}}$ can be calculated mathematically based on the LS criterion according to (64) as well. This completes the proof of Lemma 1.

REFERENCES

[1] L. Chettri and R. Bera, "A comprehensive survey on Internet of Things (IoT) toward 5G wireless systems," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 16-32, Jan. 2020.

[2] F. Guo, F. R. Yu, H. Zhang, X. Li, H. Ji and V. C. M. Leung, "Enabling Massive IoT Toward 6G: A Comprehensive Survey," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 11891-11915, 1 Aug.1, 2021.

[3] S. Liu, Z. Gao, Y. Wu, D. W. K. Ng, X. Gao, K. Wong, S. Chatzinotas, and B. Ottersten, "LEO satellite constellations for 5G and beyond: How will they reshape vertical domains?", *to appear in IEEE Commun. Mag.*

[4] O. Kodheli et al., "Satellite communications in the new space era: A survey and future challenges," *IEEE Commun. Surv. Tutor.,*, vol. 23, no. 1, pp. 70-109, 1st Quart. 2021.

[5] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59-65, Sep. 2016.

[6] M. Hasan, E. Hossain, and D. Niyato, "Random access for machineto- machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.

[7] 3GPP TS 38.300 V16.0.0, "Technical specification group radio access network; NR; NR and NG-RAN overall description," Dec. 2019.

[8] R. De Gaudenzi, O. Del Rio Herrero, G. Gallinaro, S. Cioni, and P.-D. Arapoglou, "Random access schemes for satellite networks, from VSAT to M2M: A survey," *Int. J. Satellite Commun. Netw.*, vol. 36, no. 1, pp. 66–107, 2018. [Online]. Available: https:// onlinelibrary.wiley.com/doi/abs/10.1002/sat.1204

[9] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1805-1838, 3rd Quart. 2020.

[10] B. Wang, L. Dai, T. Mir, and Z. Wang, "Joint user activity and data detection based on structured compressive sensing for NOMA," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1473–1476, Jul. 2016.

[11] Y. Du *et al.*, "Block-sparsity-based multiuser detection for uplink grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7894-7909, Dec. 2018.

[12] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2320–2323, Nov. 2016

[13] Y. Du *et al.*, "Efficient multi-user detection for uplink grant-free NOMA: Prior-information aided adaptive compressive sensing perspective," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2812-2828, Dec. 2017.

[14] B. K. Jeong, B. Shim, and K. B. Lee, "MAP-based active user and data detection for massive machine-type communications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8481-8494, Sept. 2018.

[15] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640-643, Mar. 2017.

[16] Y. Mei *et al.*, "Compressive sensing based joint activity and data detection for grant-free massive IoT access," *IEEE Trans. Wireless Commun.*, early access, Sept. 1, 2021, doi: 10.1109/TWC.2021.3107576.

[17] S. Park, H. Seo, H. Ji, and B. Shim, "Joint active user detection and channel estimation for massive machine-type communications," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun.*, Sapporo, Japan, Jul. 2017, pp. 1–5.

[18] X. Xu, X. Rao, and V. K. N. Lau, "Active user detection and channel estimation in uplink C-RAN systems," in *Proc. Int. Conf. Commun.*, Jun. 2015, pp. 2727–2732.

[19] L. Liu and W. Yu, "Massive connectivity with massive MIMO – Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, 2018.

[20] 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, Jan. 2020.

[21] Z. Zhang *et al.*, "User activity detection and channel estimation for grant-free random access in LEO satellite-enabled Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8811-8825, Sept. 2020.

[22] R. Hadani *et al.*, "Orthogonal time frequency space modulation," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2017, pp. 1–6.

[23] Z. Wei *et al.*, "Orthogonal time-frequency space modulation: A promising next-generation waveform," *IEEE Wireless Commun.*, vol. 28, no. 4, pp. 136-144, Aug. 2021..

[24] V. Khammammetti and S. K. Mohammed, "OTFS-based multiple-access in high doppler and delay spread wireless channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 528-531, Apr. 2019.

[25] A. K. Sinha, S. K. Mohammed, P. Raviteja, Y. Hong and E. Viterbo, "OTFS based random access preamble transmission for high mobility scenarios," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15078-15094, Dec. 2020.

[26] M. Li, S. Zhang, F. Gao, P. Fan and O. A. Dobre, "A new path division multiple access for the massive MIMO-OTFS networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 903-918, Aug. 2020.

[27] Z. Ding, R. Schober, P. Fan and H. Vincent Poor, "OTFS-NOMA: An efficient approach for exploiting heterogenous user mobility profiles," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7950-7965, Nov. 2019.

[28] A. Chatterjee, V. Rangamgari, S. Tiwari and S. S. Das, "Nonorthogonal multiple access with orthogonal time-frequency space signal transmission," *IEEE Syst. J.*, vol. 15, no. 1, pp. 383-394, Mar. 2021.

[29] L. You, K. -X. Li, J. Wang, X. Gao, X. -G. Xia and B. Ottersten, "Massive MIMO transmission for LEO satellite communications," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1851–1865, Aug. 2020

[30] A. G. Kanatas and A. D. Panagopoulos, *Radio Wave Propagation and Channel Modeling for Earth-Space Systems*. New York, NY, USA: CRC Press, 2016.

[31] A. Farhang, A. RezazadehReyhani, L. E. Doyle, and B. Farhang Boroujeny, "Low complexity modem structure for OFDM-based orthogonal time frequency space modulation," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 344–347, Jun. 2018.

[32] S. Wang, J. Guo, X. Wang, W. Yuan and Z. Fei, "Pilot design and optimization for OTFS modulation," *IEEE Wireless Commun. Lett.*, vol. 10, no. 8, pp. 1742-1746, Aug. 2021.

[33] W. Shen, L. Dai, J. An, P. Fan and R. W. Heath, "Channel estimation for orthogonal time frequency space (OTFS) massive MIMO," *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4204–4217, Aug. 2019.

[34] Y. Liu, S. Zhang, F. Gao, J. Ma and X. Wang, "Uplink-aided high mobility downlink channel estimation over massive MIMO-OTFS system," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 1994-2009, Sept. 2020.

[35] S. Li, W. Yuan, Z. Wei and J. Yuan, "Cross domain iterative detection for orthogonal time frequency space modulation," *IEEE Trans. Wireless Commun.*, early access, Sept. 13, 2021, doi: 10.1109/TWC.2021.3110125.

[36] Z. Wei *et al.*, "Off-grid channel estimation with sparse bayesian learning for OTFS systems," [Online], arXiv:2101.05629, 2021.

[37] Z. Gao, C. Zhang, Z. Wang and S. Chen, "Priori-information aided iterative hard threshold: A low-complexity high-accuracy compressive sensing based channel estimation for TDS-OFDM," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 242–251, Jan. 2015.

[38] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4053–4085, Sep. 2011.

[39] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, “Simultaneous sparse approximation via greedy pursuit,” in *Proc. Acoust., Speech, Signal Process. (ICASSP)*, 2005, pp. V-721–V-724.

[40] R. Roy and T. Kailath,“Esprit-estimation of signal parameters via rotational invariance techniques,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-37, no. 7, pp. 984-995, Jul. 1989.

[41] Paige, C. C. and M. A. Saunders, “LSQR: An algorithm for sparse linear equations and sparse least squares,” *ACM Trans. Math. Soft.*, vol. 8, 1982, pp. 43-71.