

CONTENTS

ZTE COMMUNICATIONS December 2021 Vol. 19 No. 4 (Issue 76)

Special Topic

OTFS Modulation for 6G and Future High Mobility Communications

Editorial 01

YUAN Jinhong, FAN Pingzhi, BAI Baoming, AI Bo

A Survey on Low Complexity Detectors for OTFS Systems 03

The authors review low complexity OTFS detectors and provide some insights on future research. They firstly present the OTFS system model and basic principles, and then survey the principles of OTFS detection algorithms. Furthermore, they discuss the design of hybrid OTFS and OFDM detectors in single user and multi-user multi-waveform communication systems. Finally, they address the main challenges in designing low complexity OTFS detectors and identify some future research directions.

ZHANG Zhengquan, LIU Heng, WANG Qianli, FAN Pingzhi

Signal Detection and Channel Estimation in OTFS 16

This paper presents an overview of the state-of-the-art approaches in OTFS signal detection and DD channel estimation. The authors classify the signal detection approaches into three categories. Similarly, they classify the DD channel estimation approaches into three categories. They compile and present an overview of some of the key algorithms under these categories and illustrate their performance and complexity attributes.

Ashwitha NAIKOTI, Ananthanarayanan CHOCKALINGAM

Message Passing Based Detection for Orthogonal Time Frequency Space Modulation 34

The authors provide an overview of some recent message passing based OTFS detectors, compare their performance, and shed some light on potential research on the design of message passing based OTFS receivers.

YUAN Zhengdao, LIU Fei, GUO Qinghua, WANG Zhongyong

45 Performance of LDPC Coded OTFS Systems over High Mobility Channels

This paper evaluates the performance of coded OTFS systems. The authors consider 5G LDPC codes for OTFS systems based on 5G OFDM frame structures over high mobility channels. They show the performance of the OTFS systems with 5G LDPC codes when sum-product detection algorithm and iterative detection and decoding are employed. They also illustrate the effect of channel estimation error on the performance of the LDPC coded OTFS systems.

ZHANG Chong, XING Wang, YUAN Jinhong, ZHOU Yiqing

54 Coded Orthogonal Time Frequency Space Modulation

The coded OTFS modulation system is considered and introduced in detail. Furthermore, the performance of the uncoded/coded OTFS system and OFDM system is analyzed with different relative speeds, modulation schemes, and iterations.

LIU Mengmeng, LI Shuangyang, ZHANG Chunqiong, WANG Boyu, BAI Baoming

63 OTFS Enabled NOMA for mMTC Systems over LEO Satellite

The application of OTFS enabled NOMA for mMTC over the LEO satellite is investigated in this paper. The LEO satellite based mMTC system and the OTFS-NOMA schemes are described. Subsequently, the challenges of applying OTFS and NOMA into LEO satellite mMTC systems are discussed. Finally, the potential technologies for the systems are investigated.

MA Yiyang, MA Guoyu, WANG Ning, ZHONG Zhangdui, AI Bo

Submission of a manuscript implies that the submitted work has not been published before (except as part of a thesis or lecture note or report or in the form of an abstract); that it is not under consideration for publication elsewhere; that its publication has been approved by all co-authors as well as by the authorities at the institute where the work has been carried out; that, if and when the manuscript is accepted for publication, the authors hand over the transferable copyrights of the accepted manuscript to *ZTE Communications*; and that the manuscript or parts thereof will not be published elsewhere in any language without the consent of the copyright holder. Copyrights include, without spatial or timely limitation, the mechanical, electronic and visual reproduction and distribution; electronic storage and retrieval; and all other forms of electronic publication or any other types of publication including all subsidiary rights.

Responsibility for content rests on authors of signed articles and not on the editorial board of *ZTE Communications* or its sponsors.

All rights reserved.

CONTENTS

ZTE COMMUNICATIONS December 2021 Vol. 19 No. 4 (Issue 76)

Orthogonal Time Frequency Space Modulation **71** in Multiple-Antenna Systems

The application of the OTFS modulation in multiple-antenna systems is investigated. The authors provide two classes of OTFS-based multiple-antenna approaches for both the open-loop and the closed-loop systems. In the closed-loop system, they adopt the Tomlinson-Harashima precoding in our derived delay-Doppler equivalent transmission model. Numerical evaluations demonstrate the advantages of applying the multiple-antenna techniques to the OTFS. At last, several challenges and opportunities are presented.

WANG Dong, WANG Fanggang, LI Xiran, YUAN Pu, JIANG Dajie

Review

Study on Security of 5G and Satellite **79** Converged Communication Network

This paper proposes the security architecture of the 5G SCCN and systematically sorts out the key protection technologies and improvement directions. In particular, unique thinking on the security of lightweight data communication and design reference for the 5G SCCN network architecture is presented. It is expected to provide a piece of reference for the follow-up 5G SCCN security technology research, standard evolution, and industrialization.

YAN Xincheng, TENG Huiyun, PING Li, JIANG Zhihong, ZHOU Na

Research Paper

Payload Encoding Representation from **90** Transformer for Encrypted Traffic Classification

The authors propose a method named PERT to perform automatic traf-

fic feature extraction using a state-of-the-art dynamic word embedding technique. By implementing traffic classification experiments on a public encrypted traffic data set and the captured Android HTTPS traffic, the authors prove the proposed method can achieve an obvious better effectiveness than other compared baselines.

HE Hongye, YANG Zhiguo, CHEN Xiangning

98 AI-Based Optimization of Handover Strategy in Non-Terrestrial Networks

A new handover strategy is proposed based on the convolutional neural network. Firstly, the handover process is modeled as a directed graph. Secondly, a convolutional neural network is used to extract the underlying regularity of the best handover strategies of different users. Numerical simulation shows that the proposed handover strategy can efficiently reduce the handover number while ensuring the signal strength.

ZHANG Chenchen, ZHANG Nan, CAO Wei, TIAN Kaibo, YANG Zhen

105 Truly Grant-Free Technologies and Protocols for 6G

This paper analyzes the problems of existing access protocols and provides novel access technologies to solve them. These technologies include contention-based NOMA, data features, enhanced pilot design and SIC of diversity. With these key enablers, truly grant-free access can be realized, and some potential modifications of protocols are then analyzed. Finally, this paper uses massive and critical scenarios in digital transformations to show the great necessity of introducing novel access technologies into future communication protocols.

MA Yihua, YUAN Zhifeng, LI Weimin, LI Zhigang

Roundup

I Table of Contents for Volume 19, 2021

Serial parameters: CN 34-1294/TN*2003*q*16*110*en*P*¥ 20.00*2000*12*2021-12

Statement

This magazine is a free publication for you. If you do not want to receive it in the future, you can send the "TD unsubscribe" mail to magazine@zte.com.cn. We will not send you this magazine again after receiving your email. Thank you for your support.



Editorial: Special Topic on OTFS Modulation for 6G and Future High Mobility Communications



Guest Editor

YUAN Jinhong received the B.E. and Ph.D. degrees in electronics engineering from the Beijing Institute of Technology, China in 1991 and 1997, respectively. From 1997 to 1999, he was a research fellow with the School of Electrical Engineering, University of Sydney, Australia. In 2000, he joined the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia, where he is currently a professor and the Head of Telecommunication Group with the school. He has published two books, five book chapters, over 300

papers in telecommunications journals and conference proceedings, and 50 industrial reports. He is a co-inventor of one patent on MIMO systems and four patents on low-density-parity-check codes. He has co-authored four Best Paper Awards and one Best Poster Award, including the Best Paper Award from the IEEE International Conference on Communications, USA in 2018, the Best Paper Award from IEEE Wireless Communications and Networking Conference in 2011, and the Best Paper Award from the IEEE International Symposium on Wireless Communications Systems in 2007. He is an IEEE Fellow and currently serving as an Associate Editor for the IEEE Transactions on Wireless Communications and IEEE Transactions on Communications. His current research interests include error control coding and information theory, communication theory, and wireless communications.



Guest Editor

BAI Baoming received the B.S. degree from the Northwest Telecommunications Engineering Institute, China in 1987, and the M.S. and Ph.D. degrees in communication engineering from Xidian University, China in 1990 and 2000, respectively. From 2000 to 2003, he was a senior research assistant at the Department of Electronic Engineering, City University of Hong Kong, China. Since April 2003, he has been with the State Key Laboratory of Integrated Services Networks (ISN), School of Telecommunication Engineering, Xidian University, China, where he is currently a professor. In 2005, he was with the University of California, USA, as a visiting scholar. In 2018, he spent one month as a Senior Visiting Fellow at McMaster University, Canada. Dr. BAI co-authored the book *Channel Coding for 5G* (in Chinese, 2020). His research interests include information theory and channel coding, wireless communication, and quantum communication. He received the Best Paper Award from the CIC/IEEE *China Communications*, in 2018. He is a senior member of IEEE.



Guest Editor

FAN Pingzhi received the M.Sc. degree in computer science from Southwest Jiaotong University, China in 1987, and the Ph.D. degree in electronic engineering from Hull University, U.K. in 1994. He is currently the director and distinguished professor of the Institute of Mobile Communications, Southwest Jiaotong University, China, and a visiting professor of Leeds University, UK (1997-), a guest professor of Shanghai Jiaotong University (1999-). He served as an EXCOM member for the IEEE Region 10, IET (IEE) Council and the IET Asia Pacific Region.

He was a recipient of the UK ORS Award (1992), the National Science Fund for Distinguished Young Scholars (1998, NSFC), IEEE VT Society Jack Neubauer Memorial Award (2018), IEEE SP Society SPL Best Paper Award (2018), IEEE WCSP 10-Year Anniversary Excellent Paper Award (2009-2019), and IEEE/CIC ICC Best Paper Award (2020). He served as a chief scientist of the National 973 Plan Project between 2012.1 – 2016.12. His research interests include high mobility wireless communications, massive random-access techniques, signal design & coding, etc. He is an IEEE VTS Distinguished Speaker (2019-2022), a fellow of IEEE, IET, CIE and CIC.



Guest Editor

AI Bo is a professor and doctoral supervisor of Beijing Jiaotong University. He is also the deputy director of the State Key Laboratory of Rail Traffic Control and Safety. Prof. AI has been awarded the National Science Fund for Distinguished Young Scholars, the Outstanding Youth Science Fund, the Newton Advanced Fellowship, the National Central Organization Department's 10 000-Person Plan Leading Talents, the Chinese Academy of Engineering Distinguished Young Investigator of China Frontiers of Engineering. Prof. AI has published 6 Chinese academic books, 3 English books, 150 IEEE journal articles. He has obtained 13 international paper awards include IEEE VTS Neil Shepherd Memorial Best Propagation Award and IEEE GLOBE-COM 2018 Best Paper Award, 32 invention patents; 21 proposals adopted by the ITU, 3GPP, etc., and 9 provincial and ministerial-level science and technology awards. His research results have been involved in 4 national standards. He is mainly engaged in the research and application of the theory and core technology of broadband mobile communication and rail transit dedicated mobile communication systems (GSM-R, LTE-R, 5G-R, LTE-M). He is the president of IEEE BTS Xi'an Branch, vice president of IEEE VTS Beijing Branch, and IEEE VTS distinguished lecturer. He is a fellow of IEEE.

Future wireless networks are expected to provide high speed and ultra-reliable communications for a wide range of emerging mobile applications, including real-time online gaming, vehicle-to-everything (V2X), un-

manned aerial vehicle (UAV) communications, and high-speed railway systems. Communications in high mobility scenarios suffer from severe channel Doppler spreads as well as delay spread, which deteriorates the performance of the widely adopted orthogonal frequency division multiplexing (OFDM) modulation in the current 4G and 5G networks.

Recently, a new two-dimensional (2D) modulation scheme referred to as orthogonal time frequency space (OTFS) modulation was proposed, where the information symbols are multiplexed in the delay-Doppler (DD) domain rather than the

DOI: 10.12142/ZTECOM.202104001

Citation (IEEE format): J. H. Yuan, P. Z. Fan, B. M. Bai, et al., "Editorial: special topic on OTFS modulation for 6G and future high mobility communications," *ZTE Communications*, vol. 19, no. 4, pp. 1-2, Dec. 2021. doi: 10.12142/ZTECOM. 202104001

time-frequency (TF) domain as in the traditional modulation techniques. The DD domain multiplexing provides the possibility to embrace the channel impairments and to provide the benefits of delay- and Doppler-resilience. OTFS enjoys the full time-frequency diversity of the channel, a key to provide reliable communications. Since it was introduced in 2017, OTFS has been recognized globally for its great potential to achieve high-speed and high-reliable communications in a high-mobility environment. While some initial works on the concepts and the implementations of OTFS have been investigated, there are still several challenges and open problems to be addressed.

In this special issue, we have invited seven active and leading research groups who have been working on this topic in the last few years to provide tutorials, surveys or technical papers on various important issues facing OTFS system designs, including efficient DD domain channel estimation, low complexity OTFS signal detection, coded OTFS systems performance evaluations, iterative receiver designs, multiple-antenna OTFS systems designs, OTFS based multiple access for satellite communications, etc. These papers provide an overview of the latest OTFS research and innovations as well as their applications. Thereby, it is expected that these papers will motivate and inspire further work amongst researchers, engineers, and Ph.D students working on OTFS and related areas.

One of the challenges facing OTFS system design is how to detect the transmitted symbols with a low complexity receiver. In this special issue, three papers investigate OTFS receiver designs, particularly on the signal detection and channel estimations for OTFS systems. ZHANG Zhengquan et al. in the paper entitled “A Survey on Low Complexity Detectors for OTFS Systems” provide a survey on low complexity OTFS detectors and their related insights on future researches. OTFS detector structures and classifications are compared and discussed. Motivated by the principles of OTFS detection algorithms, the authors propose the design of hybrid OTFS and OFDM detector in single user and multi-user systems.

The paper entitled “Signal Detection and Channel Estimation in OTFS” by CHOCKALINGAM et al. presents an overview of the state-of-the-art approaches in the OTFS signal detection and DD domain channel estimation. Three signal detection methods (linear detection, approximate maximum a posteriori detection, and deep neural network based detection) and three DD channel estimations (separate pilot, embedded pilot, and superimposed pilot) are discussed. The main challenges and future research directions are identified.

Considering that an efficient detector is paramount to har-

vesting the time and frequency diversities promised by OTFS, GUO Qinghua et al. in their paper “Message Passing Based Detection for Orthogonal Time Frequency Space Modulation” offer an overview of some recent message passing based OTFS detectors, which can exploit the features of the OTFS channel matrices, compare their performance, and shed some light on potential research on the design of OTFS receivers.

While most of the existing OTFS work deals with uncoded systems, this special issue has two papers contributing to the design of coded OTFS systems. In the paper entitled “Performance of LDPC Coded OTFS Systems over High Mobility Channels” by ZHANG Chong et al., the performance of coded OTFS systems with 5G LDPC codes and 5G OFDM frame structure over high mobility channels is evaluated. Various iterative detection and decoding algorithms are proposed. The effect of channel estimation error on the LDPC coded OTFS system performance is discussed.

The work “Coded Orthogonal Time Frequency Space Modulation” by LIU Mengmeng et al. analyses the performance of the uncoded/coded OTFS system and compare them with OFDM systems with different relative speeds, modulation schemes and iterations. They show that the OTFS system has the potential of full diversity gain and better robustness under high mobility scenarios.

The paper “OTFS Enabled NOMA for mMTC Systems over LEO Satellite” by MA Yiyang et al. suggests one potential application of OTFS for the massive machine type communications (mMTC) in low earth orbit (LEO) satellite networks with notable Doppler shifts. OTFS-NOMA schemes are described for the systems. The challenges of applying OTFS and NOMA into the LEO satellite mMTC systems and the potential technologies for the system are investigated.

Finally, in the work “Orthogonal Time Frequency Space Modulation in Multiple-Antenna Systems” by WANG Dong et al., the application of OTFS modulation in multiple-antenna systems is investigated. Two classes of OTFS-based multiple-antenna approaches for both the open-loop and the closed-loop (with Tomlinson-Harashima precoding) systems are proposed. Key challenges and opportunities for applying OTFS to multiple-antenna systems are presented.

In summary, this special issue covers the state-of-the-art OTFS technologies for channel estimation, detection, code design, iterative receiver development, and its applications for LEO satellite and multiple-antenna systems. We thank all authors, reviewers, editorial staff who have contributed to this issue. It is our expectation that these papers will inspire further research and development for future 6G and emerging wireless communications.

A Survey on Low Complexity Detectors for OTFS Systems



ZHANG Zhengquan^{1,2}, LIU Heng¹, WANG Qianli¹, FAN Pingzhi¹

(1. Key Lab of Information Coding and Transmission, Southwest Jiaotong University, Chengdu 611756, China;
2. State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China)

Abstract: The newly emerging orthogonal time frequency space (OTFS) modulation can obtain delay-Doppler diversity gain to significantly improve the system performance in high mobility wireless communication scenarios such as vehicle-to-everything (V2X), high-speed railway and unmanned aerial vehicles (UAV), by employing inverse symplectic finite Fourier transform (ISFFT) and symplectic finite Fourier transform (SFFT). However, OTFS modulation will dramatically increase system complexity, especially at the receiver side. Thus, designing low complexity OTFS receiver is a key issue for OTFS modulation to be adopted by new-generation wireless communication systems. In this paper, we review low complexity OTFS detectors and provide some insights on future researches. We firstly present the OTFS system model and basic principles, followed by an overview of OTFS detector structures, classifications and comparative discussion. We also survey the principles of OTFS detection algorithms. Furthermore, we discuss the design of hybrid OTFS and orthogonal frequency division multiplexing (OFDM) detectors in single user and multi-user multi-waveform communication systems. Finally, we address the main challenges in designing low complexity OTFS detectors and identify some future research directions.

Keywords: high mobility wireless communications; OTFS; ISFFT; SFFT; delay-Doppler diversity; iterative maximum ratio combining (MRC) detection; message passing detection

DOI: 10.12142/ZTECOM.202104002

<https://kns.cnki.net/kcms/detail/34.1294.TN.20211129.1724.002.html>, published online November 30, 2021

Manuscript received: 2021-10-18

Citation (IEEE Format): Z. Q. Zhang, H. Liu, Q. L. Wang, et al., "A survey on low complexity detectors for OTFS systems," *ZTE Communications*, vol. 19, no. 4, pp. 03 - 15, Dec. 2021. doi: 10.12142/ZTECOM.202104002.

1 Introduction

The new-generation mobile communication systems^[1] are the key enabler for the digital society in the next ten years and are expected to satisfy the requirements for high mobility applications such as vehicle-to-everything (V2X) services^[2-3], high-speed railway services^[4-5], as

well as unmanned aerial vehicles (UAV), which require the support of high mobility up to 500 - 1 000 km/h with acceptable quality of service (QoS)^[1,6].

However, high mobility wireless communications suffer from high Doppler spread, and the transmitted signals experience time-frequency doubly selective channel^[7]. High Doppler spread will result in very serious inter-carrier interference (ICI), especially in orthogonal frequency division multiplexing (OFDM) systems. Another challenge is to perform channel estimation to obtain exact channel state information (CSI) of fast time-variant channels, even to the extent that the reported CSI is outdated. These challenges will seriously reduce the performance of conventional OFDM systems. To tackle the challenge

This work is supported in part by the NSFC Project under Grant No. 61871334, in part by the open research fund of the State Key Laboratory of Integrated Services Networks, Xidian University under Grant No. ISN21-15, and in part by the Fundamental Research Funds for the Central Universities, SWJTU under Grant No. 2682020CX79. FAN Pingzhi's work is also supported by the NSFC project under Grant No. 61731017 and the "111" project under Grant No. 111-2-14.

es of high mobility, learning-based channel estimation, flexible subcarrier spacing and length of cyclic prefix (CP), double demodulation reference signals (DMRS), i. e., front-loaded DMRS and additional DMRS with configurable time-domain density, have been studied. However, these methods still treat high mobility as a negative factor, which results in very limited performance improvements of OFDM systems.

Recently, the orthogonal time frequency space (OTFS) modulation technology^[8-9] has been proposed for high mobility wireless communications, and attracted increasing attention due to its excellent performance. This new two-dimensional (2D) modulation transforms high mobility into a positive factor by introducing inverse symplectic finite Fourier transform (ISFFT)-based pre-processing before OFDM modulation and symplectic finite Fourier transform (SFFT)-based post-processing after OFDM demodulation. With ISFFT/SFFT transforms, delay-Doppler (DD) domain is introduced in OTFS systems and the modulated symbols are transmitted in DD domain rather than time-frequency (TF) domain. The equivalent DD channel exhibits excellent features of separability, stability, compactness, and possible sparsity^[9], which enables OTFS systems to obtain delay-Doppler diversity gain. Additionally, these excellent features are also beneficial for performing channel estimation under high mobility environments. OTFS modulation has also been submitted to 3GPP as a candidate waveform for 5G systems^[10-12], and is regarded as a promising waveform for next-generation wireless communications^[13].

However, since each modulated symbol is spread to the whole TF resource grid by ISFFT operation in OTFS systems, the number of equivalent DD channel dimensions is larger than that of OFDM systems, which dramatically increases the complexity of signal detection. To address this challenge, some efforts have been devoted to the research of low complexity OTFS detector structures such as decision feedback equalizer (DFE)^[14], iterative maximum ratio combining (MRC) detector^[15-16], non-iterative joint TF- and DD-domain detector^[17], iterative joint time- and DD-domain detector^[18], non-iterative MRC detector with compensation^[19], learning-based detector^[20-23], and separate low complexity OTFS detector^[24]. Several OTFS detection algorithms, including linear minimum mean square error (MMSE) and zero-forcing (ZF)^[25-29], message passing (MP)^[30-35] and its variants like approximate message passing (AMP)^[34-36], MRC^[15-16], joint MP and MRC^[37], hybrid maximum a posteriori (MAP) and parallel interference cancellation (PIC)^[38], expectation propagation (EP)^[39], variational Bayes (VB)^[40], and iterative least squares minimum residual (LSMR)^[41], have been studied.

In this paper, a comprehensive survey on OTFS detector structures and detection algorithms is provided. We compare the advantages and disadvantages of each OTFS detector structure and detection algorithm, which can provide some insights for future research. We also provide classifications for OTFS detectors from different dimensions. Furthermore, we

study a hybrid OFDM-OTFS multi-waveform detection framework. Finally, we discuss some challenges for low complexity OTFS detectors, and identify some future research directions. The rest of the paper is organized as follows. A brief discussion on the OTFS system model and the principles of OTFS modulation are given in Section 2. In Section 3, a survey on the state-of-the-art OTFS detector structures is provided, while the research progress on OTFS detection algorithms is given in Section 4. In Section 5, a hybrid OTFS-OFDM multi-waveform detection framework is discussed briefly, while Section 6 discusses the research challenges and identifies some future research directions, followed with conclusions.

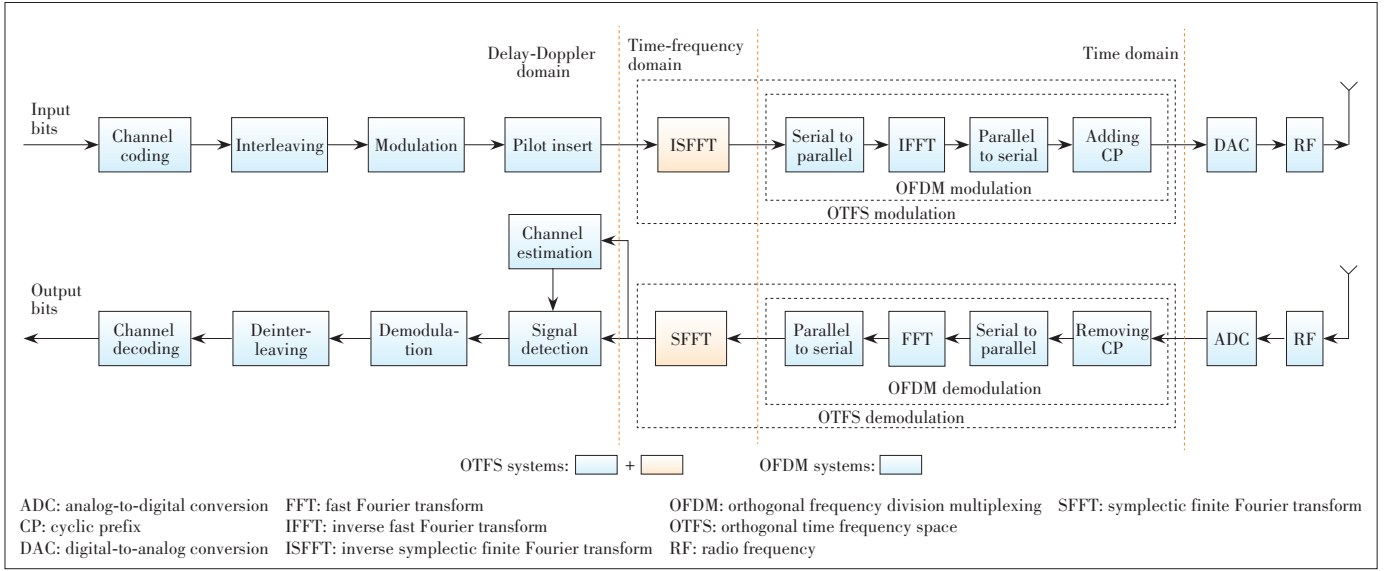
2 Basic Principles of OTFS Modulation

The OTFS system model is shown in Fig. 1, which includes OTFS transmitter and receiver structures. Compared with OFDM systems, OTFS systems add ISFFT-based transform precoding before OFDM modulation at the transmitter side, while SFFT-based post-processing is employed after OFDM demodulation at the receiver side. From the perspective of system structures, OTFS systems can be regarded as a type of precoded OFDM systems and can be easily compatible with OFDM systems. With the introduction of ISFFT/SFFT transform, a new domain, i. e., DD domain, is introduced. As a result, there are three domains in OTFS systems: DD domain, TF domain and time domain, while OFDM systems only have TF and time domains.

Considering an OTFS system with an $N \times M$ DD resource grid, at the OTFS transmitter side, the modulated symbols and pilots are mapped to the DD resource elements. The signal carried by the (k, l) -th DD resource element is denoted by $x^{DD}[k, l]$ for $k = 0, 1, \dots, N-1, l = 0, 1, \dots, M-1$. Then, the symbols $x^{DD}[k, l]$ in the DD domain are converted to the symbols $x^{TF}[n, m]$ in the TF domain using the ISFFT as

$$x^{TF}[n, m] = \text{ISFFT}(x^{DD}[k, l]) = \frac{1}{\sqrt{MN}} \sum_{k=0}^{N-1} \sum_{l=0}^{M-1} x^{DD}[k, l] e^{j2\pi(\frac{nk}{N} - \frac{ml}{M})}$$

for $n = 0, 1, \dots, N-1, m = 0, 1, \dots, M-1$. Next, the signals $x^{TF}[n, m]$ in the TF domain is converted to the symbols in the time domain signal as $x(t) = \text{IFFT}(x^{TF}[n, m]) = \frac{1}{\sqrt{MN}} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} x^{TF}[n, m] g_{tx}(t - nT) e^{j2\pi m \Delta f (t - nT)}$ and is transmitted through the channel. At the OTFS receiver side, the received signal in the time domain is $y(t) = \int_v \int_\tau h(\tau, \nu) x(t - \tau) e^{j2\pi \nu (t - \tau)} d\tau d\nu$. After OFDM demodulation (i. e., FFT transform), the symbols in the TF domain are denoted by $y^{TF}[n, m]$. Then, applying SFFT on $y^{TF}[n, m]$, the symbols in the DD domain can be obtained as $y^{DD}[k, l] = \text{SFFT}(y^{TF}[n, m]) = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} y^{TF}[n, m] e^{-j2\pi(\frac{nk}{N} - \frac{ml}{M})}$. Finally, the transmitted symbols $x^{DD}[k, l]$ can be recovered from $y^{DD}[k, l]$ through the OTFS detector.

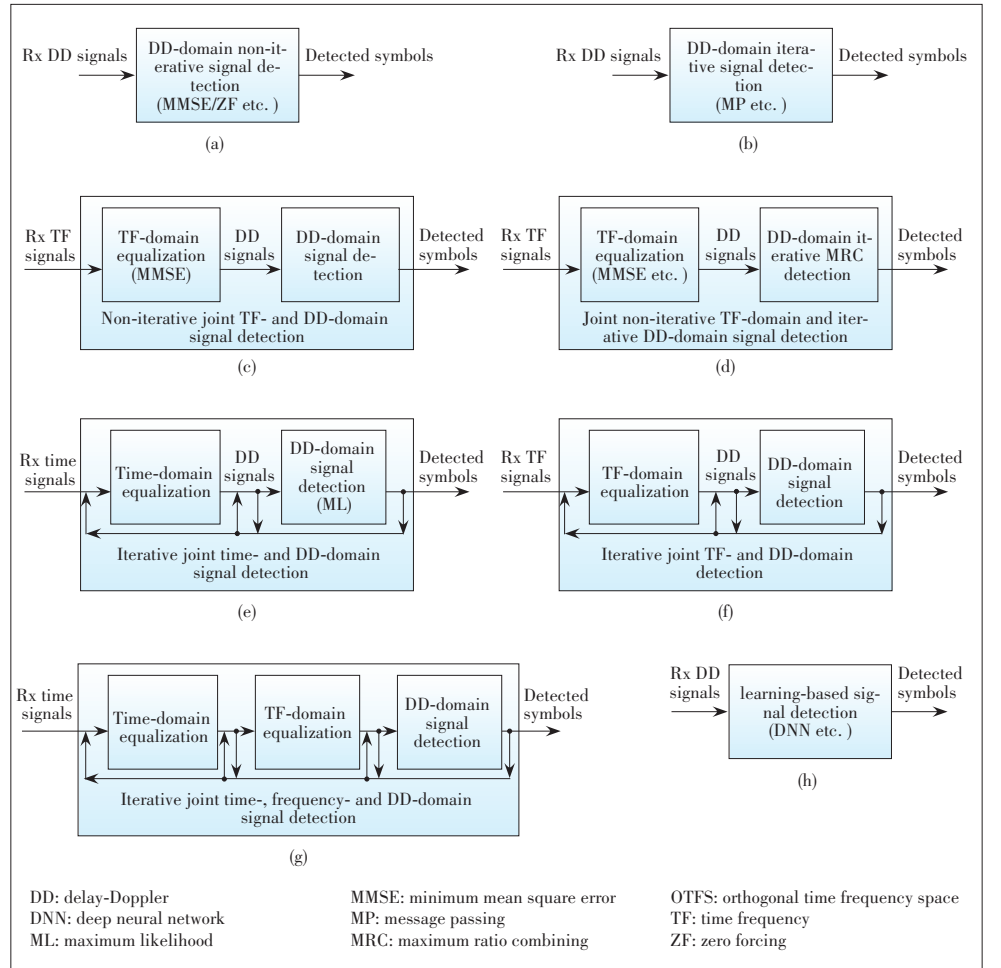


▲ Figure 1. OTFS system model

As shown in the expression for ISFFT, a DD symbol carried by a DD resource element is spread to all the TF resource elements, which enables OTFS systems to obtain full diversity. Further, since the pilots inserted in DD domain are also spread to all the TF resource elements, the equivalent DD channels obtained by channel estimation have the average channel gain. Due to delay and Doppler spread, the received symbols in DD domain are interfered with the neighboring symbols. Therefore, the main challenges for OTFS systems focus on the OTFS receiver, which needs to design very low complexity detectors, while the OTFS transmitter is relatively simple, as it only needs to add ISFFT operation before OFDM modulation.

3 OTFS Signal Detector Structures

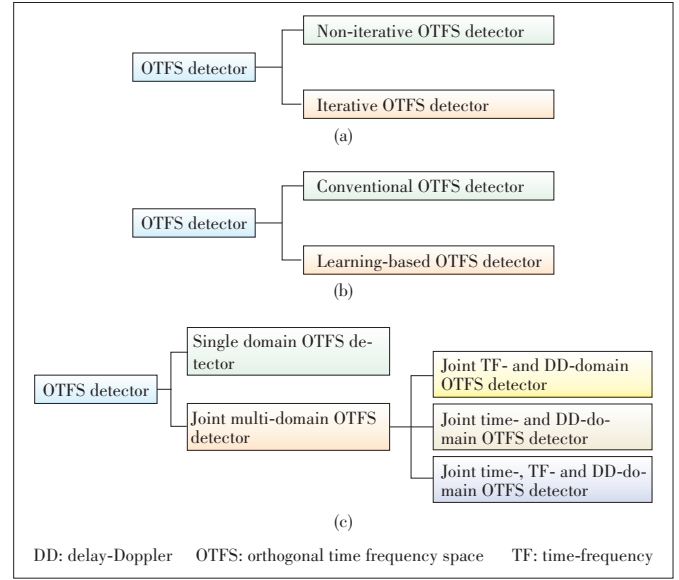
Several works have been devoted to studying low complexity OTFS detectors. Fig. 2 illustrates several popular OTFS detector structures, including DD-domain non-iterative detector^[25–29], DD-



▲ Figure 2. OTFS detector structures: (a) DD-domain non-iterative OTFS detector; (b) DD-domain iterative OTFS detector; (c) non-iterative joint TF- and DD-domain OTFS detector; (d) joint non-iterative TF-domain and iterative DD-domain OTFS detector; (e) iterative joint time- and DD-domain OTFS detector; (f) iterative joint TF- and DD-main OTFS detector; (g) iterative joint time-, TF- and DD-main OTFS detector; (h) learning-enabled OTFS detector

domain iterative detector^[30–35], non-iterative joint TF- and DD-domain detector^[17], joint non-iterative TF-domain and iterative DD-domain detector^[15–16], iterative joint Time- and DD-domain detector^[18], iterative joint TF- and DD-main detector, and learning-enabled detector^[20–23]. According to the number of domains involved in detection processing, these OTFS detectors can be divided into two categories: the single-domain OTFS detector and joint multi-domain OTFS detector. With the need of iteration, these OTFS detectors can be divided into the non-iterative OTFS detector and iterative OTFS detector. These OTFS detectors can also be divided into the conventional OTFS detector and learning-based OTFS detector. The detailed classifications of OTFS detectors are shown in Fig. 3. A summary of OTFS detectors is illustrated in Table 1.

The DD-domain non-iterative OTFS detector shown in Fig. 2(a) achieves signal detection in the DD domain by using non-iterative detection algorithms like MMSE/ZF, spherical detection, maximum likelihood (ML) detection, etc., where MMSE/ZF is popular and has also been adopted by 4G/5G systems due to its low complexity, while spherical detection and ML



▲ Figure 3. OTFS detector classifications: (a) non-iterative and iterative OTFS detectors; (b) single domain and multi-domain OTFS detectors; (c) conventional and learning-based OTFS detectors

▼ Table 1. Summary of OTFS detector structures

Ref.	Detector Structure	Detector Structure Type	Domain	Basic Idea	Advantage	Disadvantage
Refs. [25]–[29]	Single-domain OTFS detector	DD-domain non-iterative OTFS detection	DD domain	Adopting non-iterative detection algorithms (e.g., MMSE/ZF) in DD domain	Signal detection is only performed in DD domain; Non-iterative signal detection algorithms are relatively low complexity.	Non-iterative signal detection algorithms suffer from some performance loss.
Refs. [30]–[35]	OTFS detector	DD-domain iterative OTFS detection	DD domain	Adopting iterative detection algorithms, like MP/AMP and MRC etc., in DD domain	Iterative detection algorithms can achieve better performance.	Iterative detection will increase the complexity of algorithm design; the convergence of algorithms should be analyzed and ensured.
Ref. [17]	Joint multi-domain OTFS detector	Non-iterative joint TF- and DD-domain OTFS detection	TF domain and DD domain	Joint TF- and DD- domain processing with non-iterative detection algorithms	Joint multi-dimension processing can achieve better detection performance; Joint multi-dimension processing can relax the processing requirements in DD domain.	Joint multi-dimension processing increases the complexity of designing OTFS detector.
Refs. [15] and [16]		Joint non-iterative TF-domain and iterative DD-domain OTFS detection	TF domain and DD domain	Employing TF MMSE equalizer to provide good initials for DD-domain iterative MRC detector	Introducing non-iterative TF MMSE equalizer can accelerate the convergence of DD-domain iterative MRC detector; iterative MRC detector can fully merge separable taps to obtain better performance.	TF detection is needed to provide initial estimates; iteration processing increases the complexity; need to add null symbols to construct full channel matrix.
Ref. [18]		Iterative joint time- and DD-domain OTFS detection	Time domain and DD domain	Joint processing of time and DD domains to form a large iterative detection loop.	Iterative joint time- and DD-domain detection can achieve better performance and faster convergence by fully utilizing time- and DD-domain information.	Iterative joint time- and DD-domain detection increases the complexity of designing OTFS detector; a large amount external information exchange is inevitable.
Ref. [20]–[23]		Learning-based OTFS detection	TF domain and DD domain	Iterative joint TF- and DD-main OTFS detection	Iterative joint TF- and DD-domain detection can achieve better performance and faster convergence by fully utilizing TF- and DD-domain information.	Iterative joint TF- and DD-domain detection increases the complexity of designing OTFS detector; a large amount external information exchange is inevitable.
Refs. [20]–[23]		Learning-based OTFS detection	DD domain	Using machine learning techniques to perform signal detection in DD domain or estimate some parameters in conventional OTFS detector.	It is relatively simple to design learning-based signal detection as a black box without understanding expert knowledge of OTFS detection; better detection performance is achieved.	Learning-based detection is un-explainable; more computing capability is required; massive training and testing datasets are necessary.

AMP: approximate message passing DD: delay-Doppler MMSE: minimum mean square error MP: message passing MRC: maximum ratio combining
OTFS: orthogonal time frequency space TF: time-frequency ZF: zero forcing

detection are very complex. In general, the DD-domain non-iterative signal detector adopts MMSE/ZF algorithms. Without iteration operation, the computational complexity and processing delay of MMSE/ZF are small, but at the cost of detection performance loss.

The DD-domain iterative OTFS detector shown in Fig. 2(b) also achieves signal detection in the DD domain, but uses iterative detection algorithms like MP and its improved algorithms, and the EP algorithm^[39]. These algorithms iteratively update information to achieve better detection performance. However, the iteration operation brings some extra computational complexity. Additionally, the convergence of iterative detection algorithms needs to be considered. In Ref. [39], the iterative EP algorithm and its improvement named Approximate EP (AEP) were studied. They exhibit better bit error rate (BER) performance than MMSE, MP, MRC rank and VB algorithms.

The non-iterative joint TF- and DD-domain OTFS detector shown in Fig. 2(c) can be considered as an improvement of the DD-domain non-iterative OTFS detector, which utilizes both TF- and DD-domain information to improve the detection performance. In Ref. [17], a sliding window-assisted MMSE (SW-MMSE) equalization in the TF domain was studied, and a DD equalizer like decision feedback equalizer (DFE) was introduced. The computation complexity of this non-iterative two-stage equalizer is lower than conventional MMSE, and the BER performance is also better than conventional MMSE.

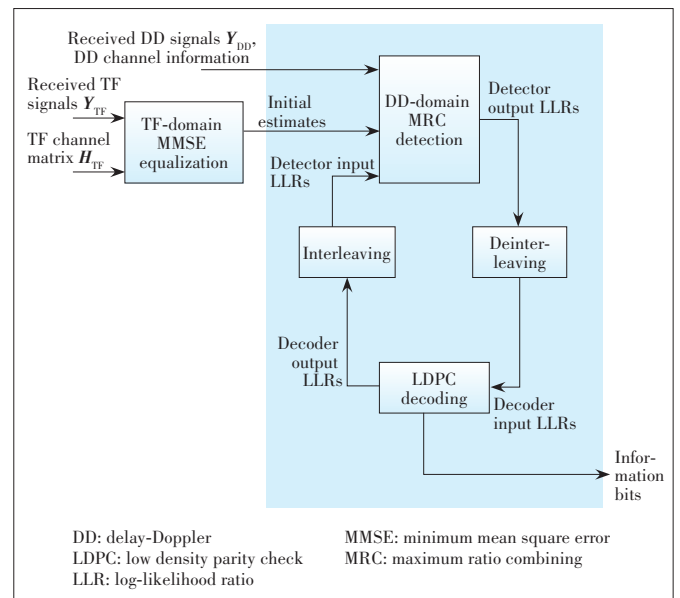
The joint non-iterative TF-domain and iterative DD-domain OTFS detector shown in Fig. 2(d) can be regarded as an improvement of the DD-domain iterative OTFS detector, in which the non-iterative TF-domain equalizer provides good initials for the iterative DD-domain OTFS detector to improve its convergence performance. In Refs. [15] and [16], an iterative MRC detector with initial estimates from the output of TF-domain MMSE equalizer was studied, as shown in Fig. 4. The results show that the iterative MRC detector with initial estimates can achieve better BER performance than that without TF-domain MMSE equalization, iterative MPA or MMSE. Considering spatial correlation at the receiver antennas, a sample-based method to estimate such correlation and the optimized combining weights for MRC from the estimated correlation matrix were studied in Ref. [42].

The iterative joint time- and DD-domain OTFS detector shown in Fig. 2(e) forms a large iteration loop among the time domain and DD domain, which is expected to obtain better performance and lower computational complexity by exploiting time domain channel sparsity and DD domain symbol constellation constraints. In Ref. [18], the iterative joint time- and DD-domain signal detector was studied, which adopted an L-MMSE estimator in the time domain and a symbol-by-symbol detection in the DD domain. The results show that this iterative joint time- and DD-domain signal detector could achieve almost the same error performance as the maximum-likelihood

sequence detection even in the presence of fractional Doppler shifts, and the computational complexity associated with the domain transformation was low.

The iterative joint TF- and DD-main OTFS detector shown in Fig. 2(f) can be regarded as an improvement of non-iterative joint TF- and DD-main signal detector. Similar as the iterative joint time- and DD-domain OTFS detector shown in Fig. 2(e), the iterative joint TF- and DD-main signal detector forms a large iteration loop among the TF domain and DD domain, which is expected to obtain better performance and faster convergence by utilizing TF- and DD-domain information. Furthermore, based on the OTFS detector shown in Fig. 2(f), an iterative joint time-, TF- and DD-main OTFS detector with time-domain equalization is shown in Fig. 2(g).

The learning-enabled OTFS detector shown in Fig. 2(h) uses advanced machine learning method to improve detection performance. In Ref. [20], to reduce the complexity of conventional MP detector in OTFS systems, a damped generalized approximate message passing (GAMP) algorithm was studied and deep learning (DL) was introduced to optimize damping factors. Its BER performance can outperform the classical GAMP algorithm and MP algorithm. In Ref. [21], a two-dimensional convolutional neural network (2D-CNN) based detector was studied to replace the conventional OTFS detector, and an MP-based data augmentation (DA) tool was employed to enlarge the training features of the input dataset and mitigate the effect of the channel variations to some degree, leading to improvement of the robustness and learning ability of the deep neural network (DNN). This 2D-CNN based detector can achieve superior performance compared with the MP detector and similar performance as the MAP detector with a very low complexity. In Ref. [22], a DD-domain symbol-level DDN detector was studied, which could achieve similar BER perfor-



▲ Figure 4. Iterative MRC detector^[15]

mance as the full DDN detector and ML detector in static multipath channel with Gaussian noise, while it achieved better BER performance than the full DDN detector and ML detector in static multipath channel with non-Gaussian noise. In Ref. [23], a reservoir computing (RC)-based OTFS detector was studied, in which one-shot online learning was sufficiently flexible to cope with channel variations among different OTFS frames and explicit CSI was not required.

4 OTFS Detection Algorithms

OTFS detection algorithms include linear MMSE/ZF, MP and its improvements, MRC, MAP, EP, and VB algorithms. A summary of these detection algorithms including their computational complexity and BER performance is presented in Table 2.

4.1 Linear MMSE/ZF Detection Algorithm

Linear signal detection mainly includes MMSE and ZF, while MMSE has been adopted by 4G/5G OFDM systems, due to its low complexity. The detection matrices of classical MMSE and ZF in OTFS systems are $\mathbf{G}_{MMSE} = (\mathbf{H}^H \mathbf{H} + \sigma^2 \mathbf{I})^{-1} \mathbf{H}^H$ and $\mathbf{G}_{ZF} = (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H$, respectively. However, when these classical MMSE and ZF detection algorithms are used for OTFS systems directly, they suffer from very high complexity $O(M^3 N^3)$. This is because the number of dimensions of equivalent DD channel matrix is $MN \times MN$ in OTFS

systems, which results in $MN \times MN$ matrix inversion. To reduce the complexity of linear signal detection in OTFS systems, considering the sparsity and the block circulant nature of equivalent DD channel, some low complexity linear signal detection schemes have been studied.

In Ref. [27], the eigenvalues of \mathbf{G}_{MMSE} was computed from the eigenvalues of DD channel matrix \mathbf{H} , which can significantly reduce the complexity. This MMSE with low complexity is summarized as follows:

- 1) Compute the eigenvalues of each block of \mathbf{H} , by computing DFTs of the first row of each circulant block;
- 2) Compute the eigenvalues of \mathbf{H} ;
- 3) Compute the eigenvalues of \mathbf{G}_{MMSE} , by using the eigenvalues of \mathbf{H} ;
- 4) Compute $\mathbf{G}_{MMSE} \mathbf{y}$.

This idea was also adopted by Ref. [28] to study the detection in MIMO-OTFS systems. Unlike the SISO-OTFS channel, the eigenvalue matrix \mathbf{D} in MIMO-OTFS channel is not diagonal, however, the inverse of the \mathbf{D}_A constructed by the matrix \mathbf{D} can be performed block-wise by two steps: matrix partitioning and backtracking^[28].

The computational complexity of MMSE is mainly caused by large matrix inversion. Considering the sparsity of equivalent DD channel matrix and quasi-banded structure of matrices in MMSE detection, a lower-upper (LU) factorization-based low complexity MMSE detection algorithm was studied for OTFS systems with reduced CP^[25] and full CP^[29], in which high complexity channel inversion is replaced by low complexity LU factorization operation. Further, the final estimate symbols step can be performed by Fast Fourier Transform (FFT). Its detailed procedure is illustrated in Fig. 5.

There are some other low complexity MMSE/ZF detection algorithms. For example, the one-tap MMSE detection algorithm studied in Ref. [26] achieved low complexity detection in pulse-shaped OTFS systems over doubly-dispersive channels, which only estimated the channel main diagonal and the self-interference power instead of interference cancellation and considered the power of the channel estimation error and self-interference as additional tuning variance parameters.

4.2 MRC Detection Algorithm

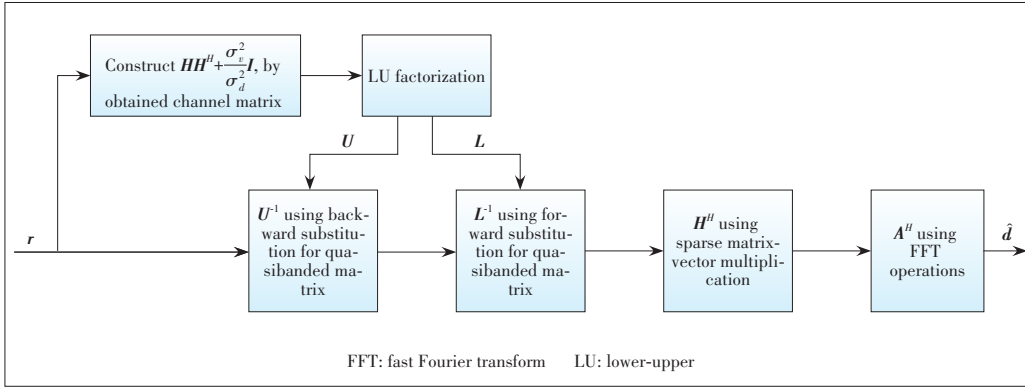
The MRC detection algorithm extracts received multipath components of the transmitted symbols in the delay-Doppler grid and combines them by using MRC to improve the signal-to-noise ratio (SNR) of the combined signal. The detailed steps of MRC algorithm are shown as follows^[15–16]:

- 1) Construct circulant matrix having element $\mathbf{K}_{m,l}$, according to the channel Doppler spread vector at each delay tap;
- 2) Construct matrix \mathbf{R} by using a circulant matrix, where $\mathbf{R}_m = \sum_{l \in \{L\}} \mathbf{K}_{m+l,l}^H \mathbf{K}_{m+l,l}$, $m = 0, 1, \dots, M - l_{\max}$;
- 3) Construct the equations for the symbol vector estimates \mathbf{b}_m^l , by using the estimates of symbol vectors from previous iteration;

▼ Table 2. Summary of computational complexity and performance

Reference	Detection Algorithm	Algorithm Characteristic	Computational Complexity	Performance
Ref. [39]	Classical MMSE	Non-iterative	$O(M^3 N^3)$	
Ref. [27]	Low complexity MMSE	Non-iterative	$O(MN \log(MN))$	
Ref. [25]	lower-upper factorization-based MMSE	Non-iterative	$O(MN \log(N))$	
Refs. [31] and [32]	MP	Iterative	$O(2^Q IMNS)$	UAMP>EP>AEP>MRC-rake>VB>MP
Ref. [33]	MF-MP-PC	Iterative	$O(IMN(2^{Q/2} + S))$	>Classical MMSE
Ref. [34]	GAMP	Iterative	$O(2^Q IMNS)$	\geq low complexity MMSE
Ref. [35]	UAMP	Iterative	$O(IMN \log(MN)) + O(2^Q IMN)$	
Ref. [36]	ICMP	Iterative	$O(2^Q IMNS)$	
Refs. [15] and [16]	MRC-rake	Iterative	$O(IMN(L + \log(N))) + O(MN(L + \log(M)))$	
Ref. [39]	EP	Iterative	$O(IMN(2^Q + S))$	
	AEP	Iterative	$O(IMN(2^Q + S))$	
Ref. [40]	VB	Iterative	$O(2^Q IMNS)$	

AEP: approximate expectation propagation
 EP: expectation propagation
 GAMP: generalized approximate message passing
 ICMP: iterative combining message passing
 MF: matched filtering
 MMSE: minimum mean square error
 MP: message passing
 MRC: maximum ratio combining
 PC: probability clipping
 UAMP: unitary approximate message passing
 VB: variational Bayes



▲ Figure 5. LU factorization-based low complexity minimum-mean-square-error (MMSE) detection^[25]

- 4) Construct \mathbf{g}_m according to \mathbf{K}_{m+L_d} and \mathbf{b}_m^l ;
- 5) Perform MRC of the estimates and obtain the output of the maximal ratio combiner, $\mathbf{c}_m = \mathbf{R}_m^{-1} \cdot \mathbf{g}_m$;
- 6) Estimate all information symbol vectors by ML criterion;
- 7) Stop criteria by stopping iteration, when some conditions are satisfying, e.g., the number of iterations is up to the maximum number of iterations.

4.3 MP Detection Algorithm and Its Improvements

The MP algorithm^[30–32] uses graphical models to decompose a hard problem into several easy sub-problems and iteratively solve them by passing messages between different types of nodes. The detailed processing steps of an MP algorithm are shown as follows^[32]:

- 1) Message passings from observation nodes to variable nodes: Observation nodes compute the mean and variances of Gaussian random variables and pass them to variables nodes;
- 2) Message passings from variable nodes to observation nodes: Variable nodes update the probability mass function (PMF) of the alphabet and pass them to observation;
- 3) Convergence indicator: The convergence indicator is computed;
- 4) Update decision: The decision on the transmitted symbols is updated, if needed;
- 5) Stopping criteria: The iteration is stopped when some conditions are satisfying. Note that different stopping criteria will affect convergence and the number of iterations.

Popular MPA detectors still suffer from high complexity and high storage requirements, as well as error floor of BER performance at high SNRs. To further improve the MPA detector, a matched-filtering based message passing detector with probability clipping (MF-PC-MPD) for OTFS systems was studied in Ref. [33]. MF-PC-MPD first performs matched filtering on received OTFS signals, and then uses probability clipping to redistribute the probability if the probability distribution satisfies a certain condition, which makes the symbol variance fluctuate within a certain range and close to each other; in this way, the Gaussianity is retained.

In Ref. [34], a Gaussian approximate message passing (GA-

MP) detection was studied, which aimed at overcoming the performance degradation of MP detectors caused by non-ideal Gaussian interference due to the limited number of interfering symbols for a certain symbol. The GA-MP detector modeled the individual transmit signals by Gaussian distributions, rather than approximating the ISI. This detector outperforms the classical

MP detector by at least 1.5 dB at a BER of 10^{-4} , with the same complexity order.

To overcome the performance loss of MP detector in the case of rich scattering environments or fractional Doppler shifts, a unitary approximate message passing (UAMP) detector was studied in Ref. [35]. Considering the equivalent DD channel is a block circulant matrix with circulant blocks which can be diagonalized using a 2D DFT matrix, the UAMP detector performs unitary transform by using a unitary matrix after receiving DD signals. As a result, the UAMP detector allows more efficient implementation with the FFT algorithm, and can achieve better BER performance than VB, MRC, MP, and AMP algorithms.

In Ref. [36], fractionally spaced sampling (FSS) was introduced to the OTFS receiver, which can be equivalent to a SI-MO system, and then iterative combining message passing (IC-MP) and turbo message passing (TMP) detectors were studied, by exploiting the sparsity of DD channel and the channel diversity gain via FSS. The ICMP detector combines two receiving channels and then performs message passing iteratively with the Gaussian approximation of the interference components. Considering there are two receiving channels in the FSS receiver, the TMP detector uses two individual MP equalizers with extrinsic log-likelihood ratios (LLRs) exchanging to form a turbo receiver.

The MP detection algorithm is based on the factor graph between variable nodes and observation nodes, and is very efficient for the sparse channel. However, its complexity will be increased when there are a large number of paths such as multi-antenna transmission. To overcome this challenge for OTFS systems with multi-antennas, a joint MP and MRC detection algorithm was studied in Ref. [37], which separated the Doppler frequency offsets (DFOs) in the spatial domain with a beamforming network to ensure the equivalent sparsity and obtained the best diversity by employing MRC to combine all beamforming branches. The main steps in each iteration of the joint MP-MRC algorithm are: 1) Each observation node passes the mean and variance of the interference terms to the connected variable nodes; 2) Each variable node updates the

PMF of alphabet symbols and then passes it back to the connected observation nodes; 3) The joint convergence indicator of all beamforming branches is calculated in the MRC fashion after each iteration. Finally, when the convergence is satisfying, the soft output of each transmitted symbol is computed, followed with hard decision.

4.4 MAP Detection Algorithm

The MAP detection algorithm uses all received signals to estimate all transmitted symbols, which can be formulated as $\hat{\mathbf{x}} = \arg \max_{\mathbf{x} \in \mathcal{A}^{NM \times 1}} \Pr(\mathbf{x}|\mathbf{y}, \mathbf{H})$. Obviously, its complexity increases with exponent in NM ^[32]. To reduce the complexity, a near-optimal symbol-wise MAP detection algorithm was studied in Ref. [38], and its detection rule is expressed as $x(k, l) = \arg \max_{x(k, l) \in \mathcal{A}} \Pr(x(k, l)|\mathbf{y}, \mathbf{H})$.

4.5 EP Detection Algorithm

The OTFS system can be represented by a sparsely-connected factor graph where each variable node (VN) is connected to factor nodes D . The main idea of EP algorithm^[39] is to use a Gaussian distribution through distribution projection to approximate the sophisticated posterior distribution in the message updating steps, which leads to the complicated belief computation being replaced by means and variances computation. The detailed steps of EP algorithm are represented as follows^[39]:

- 1) Compute the joint distribution $p(\mathbf{x}_{DD}, \mathbf{y}_{DD})$;
- 2) Compute the likelihood function $p(\mathbf{y}_b|\mathbf{x}_{DD})$;
- 3) Compute the means and variances passed from FNs and VNs as $u_{f_b \rightarrow x_a}^i$ and $v_{f_b \rightarrow x_a}^i$;
- 4) Compute the means and variances passed from VNs and FNs as $u_{x_a \rightarrow f_b}^i$ and $v_{x_a \rightarrow f_b}^i$;
- 5) Compute the a posteriori LLR of each coded bit as c_a^q ;
- 6) Stop criteria by stopping iteration, when some conditions are satisfied.

Note that the main computational complexity of the EP algorithm depends on the number of non-zero elements D of channel matrix. In case of rich scattering scenarios and fractional Doppler shift, D is relatively large. To further reduce the computational complexity, small channel coefficients can be approximated to a fixed value (e.g., the median value of these small elements) during the message passing from FNs to VNs, which is named channel coefficients-aware approximate EP (AEP) algorithm^[39].

4.6 VB Detection Algorithm

The optimal MAP detection algorithm suffers from very high complexity, which increases exponentially with the size of data symbol vector. To reduce the complexity of MAP algorithm, a variational Bayes algorithm was studied in Ref. [40]. The main idea of VB algorithm is to find a distribution $q(\mathbf{d})$ from a tractable distribution family as an optimized approxima-

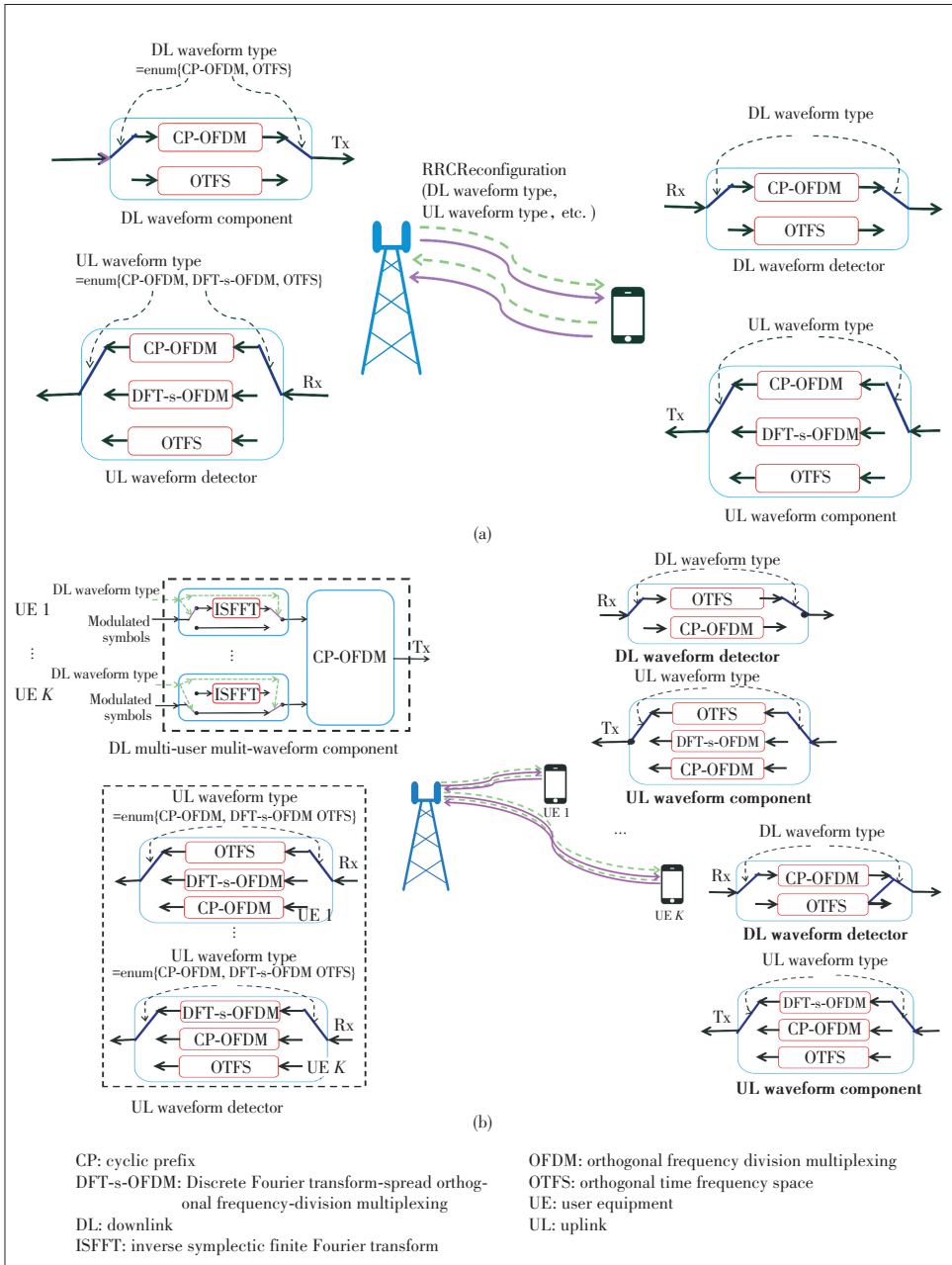
tion of the a posteriori distribution $p(\mathbf{d}|\mathbf{y})$. The detailed procedures are shown as follows:

- 1) Formulate the approximation $q^*(\mathbf{d})$ as an optimal problem by minimizing the Kullback-Leibler divergence;
- 2) Construct the approximation $q(\mathbf{d})$ by mean field approximation as $q(\mathbf{d}) = \prod_{k,l} q_{k,l}(d_{k,l})$. Note that in this form, all variables are mutually independent;
- 3) Transform $p(\mathbf{d}|\mathbf{y})$ into a pairwise form;
- 4) Obtain the variational function in the optimization problem in Step 1, by substituting $q(\mathbf{d})$ and $p(\mathbf{d}|\mathbf{y})$ into the optimization problem;
- 5) Find a stationary point of the variational function, by iteratively updating each local function $q_{k,l}(d_{k,l})$;
- 6) Approximate a posteriori distributions for all the data symbols iteratively, resulting in the approximate marginals $q_{k,l}^*(d_{k,l})$;
- 7) Estimate the transmitted symbols by finding the maximum of marginal distribution $q_{k,l}^*(d_{k,l})$.

5 Hybrid OFDM-OTFS Multi-Waveform Detector Structure

To satisfy the requirements for various scenarios and applications, mobile communication systems have evolved from single waveform to multi-waveform systems. For example, in 4G systems, high-spectrum efficiency CP-OFDM is adopted by the downlink, while the uplink adopts single-carrier frequency division multiple access (SC-FDMA) with a low peak to average power ratio (PAPR). In 5G systems, the downlink adopts CP-OFDM, while CP-OFDM and discrete Fourier transform-spread orthogonal frequency-division multiplexing (DFT-s-OFDM) with low-PAPR are adopted by the uplink. In general, when UE is in the cell center, UE can still obtain the expected QoS with low transmit power, thus UE can adopt CP-OFDM waveform with higher spectrum efficiency. When UE is at the cell edge, UE should increase transmit power to obtain the expected QoS, which requires UE to adopt DFT-s-OFDM waveform with low PAPR. Since OTFS exhibits excellent performance in high mobility environments, if OTFS is accepted by future mobile communication systems (FMCS), its downlink waveform will be CP-OFDM or OTFS, while its uplink waveform will be CP-OFDM, DFT-s-OFDM or OTFS. To determine each user's uplink (UL)/downlink (DL) waveform, the base station shall call UL and DL waveform decision algorithms with the input of user mobility speed and user type. Further, the base station also needs to dynamically switch user's downlink and uplink waveform type if some conditions are triggered.

Fig. 6 shows the hybrid OFDM-OTFS multi-waveform detector structure, in which Fig. 6(a) is for single user OTFS systems and Fig. 6(b) is for multi-user OTFS systems. The base station first determines the DL and UL waveform types accord-



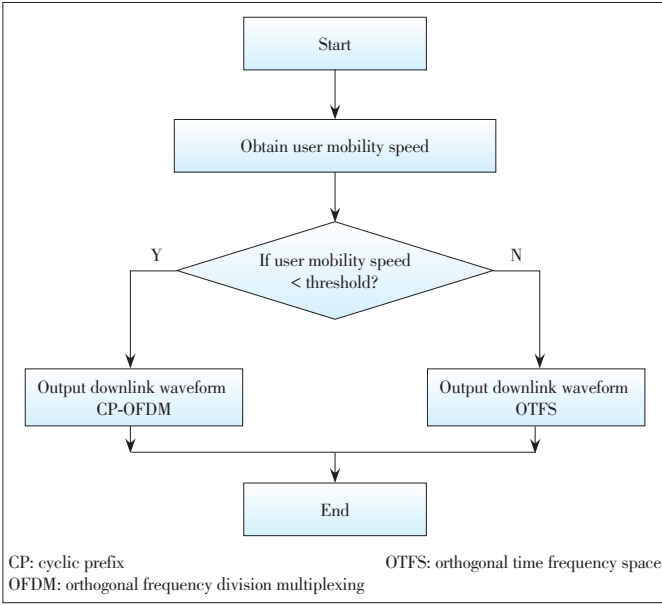
▲ Figure 6. Hybrid OTFS-OFDM multi-waveform detector structure: (a) single user OTFS system; (b) multi-user OTFS system

ing to certain algorithms with the input of user mobility speed and user type. And then such waveform information is carried by the RRCReconfiguration message and is configured to UE through the air interface. As a result, the base station and UE perform the same waveform processing. Comparing single user hybrid OTFS-OFDM systems and multi-user hybrid OTFS-OFDM systems, the main difference and difficulty are in the base station. In multi-user hybrid OTFS-OFDM systems, since the base station supports multi-user transmission and users may adopt different waveforms, the base station should have the capability of processing multiple waveform in parallel.

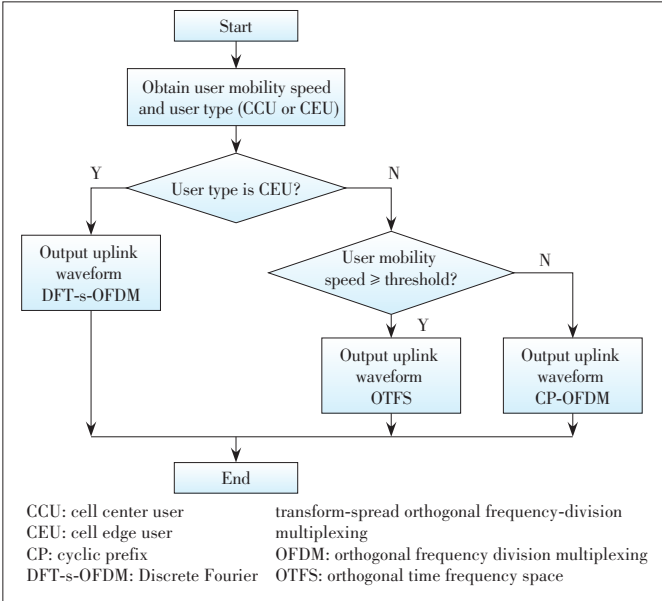
Fig. 7 shows the downlink waveform selection procedure, which will select CP-OFDM or OTFS according to user's mobility speed. Since OTFS does not show excellent performance in low mobility, the downlink adopts CP-OFDM waveform when user's mobility speed is lower than a certain speed threshold, otherwise, the downlink adopts OTFS waveform.

Fig. 8 shows the uplink waveform selection procedure, which will select CP-OFDM, DFT-s-OFDM or OTFS according to user's mobility speed and user type. Since UE has strict requirements for waveform's PAPR, the uplink adopts DFT-s-OFDM with low PAPR when UE is a cell edge user (CEU). When UE is a cell center user (CCU), the uplink adopts CP-OFDM waveform if user mobility speed is lower than a certain speed threshold, otherwise, the uplink adopts OTFS waveform.

Considering multi-user and multi-waveform communication systems, the base station needs to simultaneously process multiple users with different waveform types, which requires to design multiple access for multi-waveform multi-user systems. Taking downlink transmission for an example, Fig. 9 shows two multiple access schemes for downlink hybrid OFDM-OTFS multi-user systems. In Fig. 9(a), the resource of each user allocated in the TF domain is orthogonal and OTFS users' resource in the DD domain is also orthogonal, which can effectively avoid inter-user interference. In Fig. 9(b), the T-F resource are firstly divided into two parts, in which one part is for CP-OFDM users and the other part is for OTFS users. Then, CP-OFDM users occupy different T-F resources; while OTFS users are spread in the total T-F resources allocated to all the OTFS users, each OTFS user is orthogonal in the DD domain. Comparing the schemes shown in Figs. 9(a) and 9(b), OTFS users in Fig. 9(a) suffer from less inter-user interference, as they are orthogonal in both DD and T-F domains. However, their diversity gain is also lower than that in Fig. 9



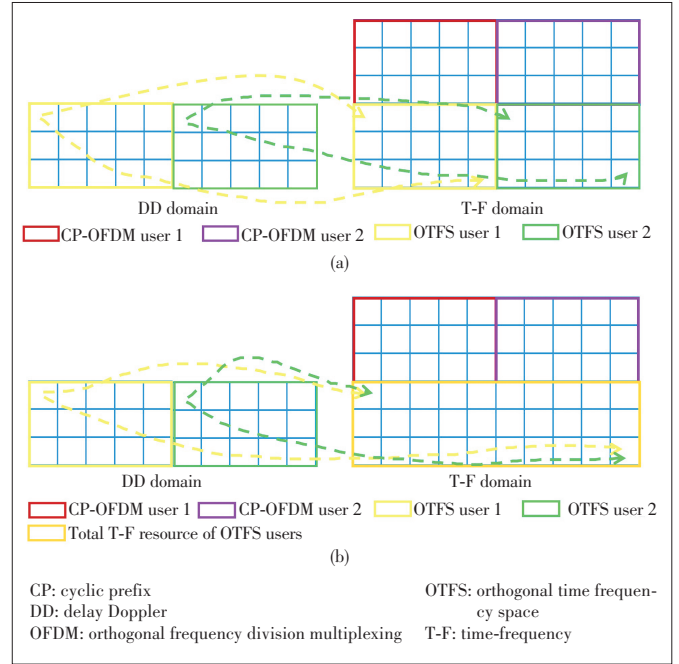
▲ Figure 7. Downlink waveform selection procedure



▲ Figure 8. Uplink waveform selection procedure

(b), as diversity gain of OTFS modulation is positively related with the number of resources. Obviously, the number of resources allocated to each OTFS user in Fig. 9(a) is less than that in Fig. 9(b). For UL multi-waveform multi-user systems, they are similar as the downlink situation.

Fig. 10 shows the block error rate (BLER) performance of two-user OTFS systems without inter-user interference (IUI) and two-user hybrid OTFS-OFDM systems in fractional Doppler channels, where the speed of User 1 is 500 km/h, while the speed of User 2 is 10 km/h. In Fig. 10(a), both User 1 and User 2 adopt OTFS modulation, while in Fig. 10(b), User 1 adopts OTFS modulation but User 2 adopts OFDM modulation. The results show that the OTFS modulation cannot



▲ Figure 9. Multiple access schemes for downlink hybrid OFDM-OTFS systems in multi-user scenario: (a) hybrid orthogonal frequency division multiple access (OFDMA) and orthogonal time frequency space multiple access (OTFSMA) in both DD and TF domains; (b) hybrid OFDMA and OTFSMA with overlap in the TF domain

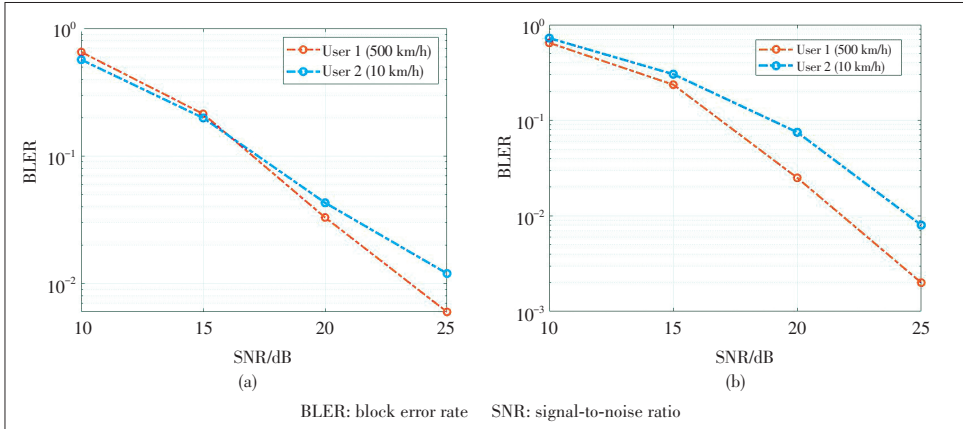
achieve dramatic BLER performance gain in a low-speed scenario while some BLER performance gain can be obtained in a high-speed scenario. Therefore, it is suggested that low-complexity OFDM modulation is used to low-speed users, while high-speed users adopt OTFS modulation. Since different users with different speeds coexist in the base station, hybrid OTFS and OFDM systems should be considered.

6 Main Challenges and Future Research Directions

6.1 Main Challenges

6.1.1 Low Complexity OTFS Detection Algorithms

In OTFS systems, NM symbols in the DD domain are spread to the TF domain by employing ISFFT transform, which results in the large number of dimensions of the equivalent DD channel. Furthermore, with the introduction of multi-antenna transmission, the complexity of OTFS detection will also increase dramatically. According to current research results, the minimum computational complexity of OTFS detector is $O(MN \log(N))$. Obviously, the computational complexity of OTFS detector is still far higher than that of OFDM detector. As a result, current OTFS detection algorithms cannot satisfy the requirements for OTFS systems. Additionally, many works assume that the DD channel matrix is block circulant and sparse. However, as shown in some works, if there are



▲ Figure 10. BLER performance: (a) two-user orthogonal time frequency space (OTFS) systems without inter-user interference (IUI); (b) two-user hybrid OTFS-orthogonal frequency division multiplexing (OFDM) systems

rich scattering or a large number of paths such as MIMO-OTFS systems, the block circulant and sparsity will not be satisfying. Furthermore, integer Doppler shifts are assumed in many works, while the assumption of fractional Doppler shifts is more reasonable in practical OTFS systems. However, fractional Doppler shifts will increase computational complexity and result in more serious inter-Doppler interference. Therefore, the research on low complexity OTFS detection algorithm is a great challenge.

6.1.2 Decoupling Between MIMO-OTFS Detector and Precoder

Current research on OTFS receivers mainly focuses on SISO-OTFS systems, while just a few works study MIMO-OTFS systems. However, when extending SISO-OTFS detection algorithms to MIMO-OTFS systems, it will face some new problems. For example, when the MRC detector is used to MIMO-OTFS systems, it needs to obtain the precoding matrix. However, when a non-codebook-based precoding scheme is adopted, it is difficult for the OTFS receiver to obtain the precoding matrix. That is, in order to match the detection algorithms, some detectors require special design at the MIMO-OTFS transmitter side. This strong coupling design between the MIMO-OTFS receiver and transmitter reduces the flexibility of MIMO-OTFS system design and processing. Therefore, the research on MIMO-OTFS detectors and detection algorithms, which is decoupled with the MIMO-OTFS transmitter, is another challenge.

6.1.3 Multi-Waveform Hybrid OTFS Detector

OTFS modulation can obtain delay-Doppler diversity gain, and thus it can achieve better performance than conventional OFDM systems in high mobility scenarios. However, OTFS modulation cannot obtain obvious performance gain in low mobility scenarios. When users experience different scenarios, it would be better to switch waveform to obtain better performance. The coexistence of multiple waveforms, such as OFDM and OTFS, requires that the receiver supports multi-

waveform processing and waveform switching, which increases the complexity of the receiver. Therefore, designing a low complexity unified multi-waveform receiver is also a challenge.

6.2 Further Research Directions

6.2.1 Advanced Low Complexity OTFS Detectors and Detection Algorithms

The computational complexity of current OTFS detection algorithms has been reduced to $O(MN \log(N))$, but it is still

much higher than the acceptable complexity in practical systems. Therefore, in the future, the first important work is to study advanced low complexity OTFS detector structures and detection algorithms. As for OTFS detector structures, even though some single domain and joint multi-domain OTFS detectors have been studied, there are still some novel OTFS detector structures to be studied such as joint channel estimation and detection. As for OTFS detection algorithms, some non-iterative and iterative detection algorithms have been studied, but their computation complexities are still very high, up to $O(MN \log(N))$. The properties of DD channel matrix together with some simplified and approximated matrix operations should be further exploited to develop novel OTFS detection algorithms. Furthermore, iterative detection algorithms can achieve better performance, but they are needed to further analyze the convergence by employing some tools such as extrinsic information transfer (EXIT) chart and design efficient iteration stopping schemes to reduce the number of iterations.

6.2.2 Learning-Based OTFS Detectors

Several OTFS detection algorithms have been studied, but they reduce the computational complexity by exploiting the block circulant and sparsity, as well as simplified and approximated mathematical methods. It will become more difficult to find more low complexity OTFS detection algorithms. A learning-based method provides a new way for the OTFS detector, which considers OTFS detection processing as a black box and performs OTFS detection by deploying online learning model trained offline. Currently, there are few research works on learning-based OTFS detector, and thus many efforts are needed to study learning models and performance verification. Therefore, the learning-based OTFS detector is a future research direction.

6.2.3 Unified Multi-Waveform Detector Design

If OTFS modulation is adopted, the downlink and uplink of future mobile communication systems will be multi-waveform.

However, there are few works to study the receiver design to support coexistence of multiple waveform. To support multi-waveform systems, the receiver should support signal detection of each waveform. A simple way is to deploy multiple signal detection modules and switch among these detection modules according to configured waveform type. However, this way is inefficient and will increase the processing complexity. A better way is to design a unified multi-waveform receiver, which can flexibly and efficiently support signal detection of different waveform. Therefore, unified multi-waveform receiver design is another future research direction.

7 Conclusions

In this paper, the research works on low complexity OTFS detectors have been surveyed comprehensively. Firstly, we present the OTFS system model and basic principles. And then, we focus on low complexity OTFS detector structures, and give the categories and discussions of all surveyed OTFS detector structures. According to different classification regulations, two classifications of OTFS detectors are the single-domain OTFS detector and joint multi-domain OTFS detector; the non-iterative OTFS detector and iterative OTFS detector. As for their performance, the joint multi-domain OTFS detector is superior to the single-domain one, while the iterative OTFS detector is better than the non-iterative one. We also provide an overview on the principles of popular OTFS detection algorithms, and discuss them in terms of complexity and performance. Furthermore, considering the coexistence of multiple waveforms such as OTFS and OFDM, we discuss the design for hybrid multi-waveform detectors in single user and multi-user OTFS systems, and waveform switching procedures and algorithms. Finally, we present main challenges for low complexity OTFS detectors and identify some future research directions.

References

- [1] ITU-R. Framework and overall objectives of the future development of IMT for 2020 and beyond: ITU-R M.2083-0 [R]. 2015
- [2] 3GPP. Study on NR vehicle-to-everything (V2X) (Release 16): 3GPP TR 38.885 [R]. 2019
- [3] CHEN S Z, HU J L, SHI Y, et al. Vehicle-to-everything (v2x) services supported by LTE-based systems and 5G [J]. IEEE communications standards magazine, 2017, 1(2): 70 – 76. DOI: 10.1109/MCOMSTD.2017.1700015
- [4] 3GPP. Mobile communication system for railways (Release 17): 3GPP TS 22.289 [S]. 2019
- [5] AI B, GUAN K, RUPP M, et al. Future railway services-oriented mobile communications network [J]. IEEE communications magazine, 2015, 53(10): 78 – 85. DOI: 10.1109/MCOM.2015.7295467
- [6] ZHANG Z Q, XIAO Y, MA Z, et al. 6G wireless networks: vision, requirements, architecture, and key technologies [J]. IEEE vehicular technology magazine, 2019, 14(3): 28 – 41. DOI: 10.1109/MVT.2019.2921208
- [7] FAN P Z, ZHAO J, I C L. 5G high mobility wireless communications: challenges and solutions [J]. China communications, 2016, 13 (Supplement 2): 1 – 13. DOI: 10.1109/CC.2016.7833456
- [8] HADANI R, RAKIB S, TSATSANIS M, et al. Orthogonal time frequency space modulation [C]/IEEE Wireless Communications and Networking Conference (WCNC). San Francisco, USA: IEEE, 2017: 1 – 6. DOI: 10.1109/WCNC.2017.7925924
- [9] WEI Z Q, YUAN W J, LI S Y, et al. Orthogonal time-frequency space modulation: a promising next-generation waveform [J]. IEEE wireless communications, 2021, 28(4): 136 – 144. DOI: 10.1109/MWC.001.2000408
- [10] 3GPP TSG RA WG1. Overview of OTFS waveform for next generation RAT: Meeting #84-bis R1162929 [R]. Busan, South Korea: 3GPP, 2016
- [11] 3GPP TSG RA WG1. OTFS modulation waveform and reference signals for new RAT: Meeting #84-bis R1162930 [R]. Busan, South Korea: 3GPP, 2016
- [12] 3GPP TSG RA WG1. Performance results for OTFS modulation: Meeting #85 R1165620 [R]. Nanjing, China: 3GPP, 2016
- [13] MobileChina. White paper on 2030+ technology trends [R]. 2019
- [14] JING L Y, WANG H, HE C B, et al. Two dimensional adaptive multichannel decision feedback equalization for OTFS system [J]. IEEE communications letters, 2021, 25(3): 840 – 844. DOI: 10.1109/LCOMM.2020.3039982
- [15] THAJ T, VITERBO E. Low complexity iterative rake detector for orthogonal time frequency space modulation [C]/IEEE Wireless Communications and Networking Conference (WCNC). Seoul, Korea (South): IEEE, 2020: 1 – 6. DOI: 10.1109/WCNC45663.2020.9120526
- [16] THAJ T, VITERBO E. Low complexity iterative rake decision feedback equalizer for zero-padded OTFS systems [J]. IEEE transactions on vehicular technology, 2020, 69(12): 15606 – 15622. DOI: 10.1109/TVT.2020.3044276
- [17] JIN C X, BIE Z S, LIN X H, et al. A simple two-stage equalizer for OTFS with rectangular windows [J]. IEEE communications letters, 2021, 25(4): 1158 – 1162. DOI: 10.1109/LCOMM.2020.3043841
- [18] LI S Y, YUAN W J, WEI Z Q, et al. Cross domain iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, early access. DOI: 10.1109/TWC.2021.3110125
- [19] SHAN Y R, WANG F G. Low-complexity and low-overhead receiver for OTFS via large-scale antenna array [J]. IEEE transactions on vehicular technology, 2021, 70(6): 5703 – 5718. DOI: 10.1109/TVT.2021.3072667
- [20] XU X K, ZHAO M M, LEI M, et al. A damped GAMP detection algorithm for OTFS system based on deep learning [C]/IEEE 92nd Vehicular Technology Conference (VTC2020-Fall). Victoria, Canada: IEEE, 2020: 1 – 5. DOI: 10.1109/VTC2020-Fall49728.2020.9348493
- [21] ENKU Y K, BAI B M, WAN F, et al. Two-dimensional convolutional neural network based signal detection for OTFS systems [J]. IEEE wireless communications letters, 10(11): 2514 – 2518. DOI: 10.1109/LWC.2021.3106039
- [22] NAIKOTI A, CHOCKALINGAM A. Low-complexity delay-doppler symbol DNN for OTFS signal detection [C]/IEEE 93rd Vehicular Technology Conference (VTC2021-Spring). Helsinki, Finland: IEEE, 2021: 1 – 6. DOI: 10.1109/VTC2021-Spring51267.2021.9448630
- [23] ZHOU Z, LIU L J, XU J R, et al. Learning to equalize OTFS [EB/OL]. (2021-07-17)[2021-08-31]. <https://arxiv.org/abs/2107.08236>
- [24] PANDEY B C, MOHAMMED S K, RAVITEJA P, et al. Low complexity precoding and detection in multi-user massive MIMO OTFS downlink [J]. IEEE transactions on vehicular technology, 2021, 70(5): 4389 – 4405. DOI: 10.1109/TVT.2021.3061694
- [25] TIWARI S, DAS S S, RANGAMGARI V. Low complexity LMMSE Receiver for OTFS [J]. IEEE communications letters, 2019, 23(12): 2205 – 2209. DOI: 10.1109/LCOMM.2019.2945564
- [26] PFADLER A, JUNG P, SZOLLMANN T, et al. Pulse-shaped OTFS over doubly-dispersive channels: one-tap vs. full LMMSE equalizers [C]/IEEE International Conference on Communications Workshops (ICC Workshops). Montreal, Canada: IEEE, 2021: 1 – 6. DOI: 10.1109/ICCWorkshops50388.2021.9473535
- [27] SURABHI G D, CHOCKALINGAM A. Low-complexity linear equalization for OTFS modulation [J]. IEEE communications letters, 2020, 24(2): 330 – 334. DOI: 10.1109/LCOMM.2019.2956709
- [28] SINGH P, MISHRA H B, BUDHIRAJA R. Low-complexity linear MIMO-OTFS receivers [C]/IEEE International Conference on Communications Workshops (ICC Workshops). Montreal, Canada: IEEE, 2021: 1 – 6. DOI: 10.1109/ICCWorkshops50388.2021.9473839
- [29] ZOU T T, XU W J, GAO H, et al. Low-complexity linear equalization for OTFS systems with rectangular waveforms [C]/IEEE International Conference on

- Communications Workshops (ICC Workshops). Montreal, Canada: IEEE, 2021: 1 – 6. DOI: 10.1109/ICCWorkshops50388.2021.9473771
- [30] KOLLENGODE RAMACHANDRAN M, CHOCKALINGAM A. MIMO-OTFS in high-doppler fading channels: signal detection and channel estimation [C]/IEEE Global Communications Conference (GLOBECOM). Abu Dhabi, United Arab Emirates: IEEE, 2018: 206 – 212. DOI: 10.1109/GLOCOM.2018.8647394
- [31] RAVITEJA P, PHAN K T, JIN Q Y, et al. Low-complexity iterative detection for orthogonal time frequency space modulation [C]/IEEE Wireless Communications and Networking Conference (WCNC). Barcelona, Spain: IEEE, 2018: 1 – 6. DOI: 10.1109/WCNC.2018.8377159
- [32] RAVITEJA P, PHAN K T, HONG Y, et al. Interference cancellation and iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, 2018, 17(10): 6501 – 6515. DOI: 10.1109/TWC.2018.2860011
- [33] ZHANG H J, ZHANG T T. A low-complexity message passing detector for OTFS modulation with probability clipping [J]. IEEE wireless communications letters, 2021, 10(6): 1271 – 1275. DOI: 10.1109/LWC.2021.3063904
- [34] XIANG L P, LIU Y S, YANG L L, et al. Gaussian approximate message passing detection of orthogonal time frequency space modulation [J]. IEEE transactions on vehicular technology, 2021, 70(10): 10999 – 11004. DOI: 10.1109/TVT.2021.3102673
- [35] YUAN Z D, LIU F, YUAN W J, et al. Iterative detection for orthogonal time frequency space modulation with unitary approximate message passing [J]. IEEE transactions on wireless communications, early access. DOI: 10.1109/TWC.2021.3097173
- [36] GE Y, DENG Q W, CHING P C, et al. Receiver design for OTFS with a fractionally spaced sampling approach [J]. IEEE transactions on wireless communications, 2021, 20(7): 4072 – 4086. DOI: 10.1109/TWC.2021.3055585
- [37] CHENG J Q, JIA C L, GAO H, et al. OTFS based receiver scheme with multi-antennas in high-mobility V2X systems [C]/IEEE International Conference on Communications Workshops (ICC Workshops). Dublin, Ireland: IEEE, 2020: 1 – 6. DOI: 10.1109/ICCWorkshops49005.2020.9145313
- [38] LI S Y, YUAN W J, WEI Z Q, et al. Hybrid MAP and PIC detection for OTFS modulation [J]. IEEE transactions on vehicular technology, 2021, 70(7): 7193 – 7198. DOI: 10.1109/TVT.2021.3083181
- [39] LI H, DONG Y Y, GONG C H, et al. Low complexity receiver via expectation propagation for OTFS modulation [J]. IEEE communications letters, 2021, 25(10): 3180 – 3184. DOI: 10.1109/LCOMM.2021.3101827
- [40] YUAN W J, WEI Z Q, YUAN J H, et al. A simple variational Bayes detector for orthogonal time frequency space (OTFS) modulation [J]. IEEE transactions on vehicular technology, 2020, 69(7): 7976 – 7980. DOI: 10.1109/TVT.2020.2991443
- [41] QU H Y, LIU G H, ZHANG L, et al. Low-complexity symbol detection and interference cancellation for OTFS system [J]. IEEE transactions on communications, 2021, 69(3): 1524 – 1537. DOI: 10.1109/TCOMM.2020.3043007
- [42] THAJ T, VITERBO E. Low-complexity linear diversity-combining detector for

MIMO-OTFS [J]. IEEE wireless communications letters, 2021, early access. DOI: 10.1109/LWC.2021.3125986

Biographies

ZHANG Zhengquan (zhangzsqswjtu@163.com) received the Ph.D. degree in information and communication engineering from Southwest Jiaotong University, China in 2019. From 2008 to 2013, he was with ZTE Corporation as a communication engineer. From 2016 to 2017, he was a guest Ph.D. student with the KTH Royal Institute of Technology, Sweden. Since 2019, he has been with the Department of Communication Engineering, School of Information Science and Technology, Southwest Jiaotong University. His research interests include B5G and 6G wireless communication technologies. He is a member of IEEE.

LIU Heng received the Ph.D. degree in information and communication engineering from Southwest Jiaotong University, China in 2013. Since 2013, he has been with the Department of Communication Engineering, School of Information Science and Technology, Southwest Jiaotong University. His research interests include next-generation wireless communications, rail transit wireless communications, machine learning and intelligent wireless communications.

WANG Qianli received the B.S. and M.S. degrees in electronic and information engineering and the Ph.D. degree from the University of Electronic Science and Technology of China in 2013, 2016 and 2021, respectively. Since 2021, he has been with the Department of Communication Engineering, School of Information Science and Technology, Southwest Jiaotong University, China. His research interests include estimation theory, array signal processing, radar sensor network, and compressed sensing.

FAN Pingzhi is currently a distinguished professor of Southwest Jiaotong University, China, and a visiting professor of Leeds University, UK (1997 –). He is a recipient of the UK ORS Award (1992), the National Science Fund for Distinguished Young Scholars (1998, NSFC), IEEE VT Society Jack Neubauer Memorial Award (2018), IEEE SP Society SPL Best Paper Award (2018), IEEE WC-SP 10-Year Anniversary Excellent Paper Award (2009 – 2019), and IEEE/CIC ICC Best Paper Award (2020). He served as a chief scientist of the National “973” Plan Project between January 2012 and December 2016. He is an IEEE VTS Distinguished Speaker (2019 – 2022), a fellow of IEEE, IET, CIE and CIC. His research interests include high mobility wireless communications, massive random-access techniques, and signal design and coding.

Signal Detection and Channel Estimation in OTFS



Ashwitha NAIKOTI, Ananthanarayanan CHOCKALINGAM

(Department of ECE, Indian Institute of Science, Bangalore 560012, India)

Abstract: Orthogonal time frequency space (OTFS) modulation is a recently proposed modulation scheme that exhibits robust performance in high-Doppler environments. It is a two-dimensional modulation scheme where information symbols are multiplexed in the delay-Doppler (DD) domain. Also, the channel is viewed in the DD domain where the channel response is sparse and time-invariant for a long time. This simplifies channel estimation in the DD domain. This paper presents an overview of the state-of-the-art approaches in OTFS signal detection and DD channel estimation. We classify the signal detection approaches into three categories, namely, low-complexity linear detection, approximate maximum a posteriori (MAP) detection, and deep neural network (DNN) based detection. Similarly, we classify the DD channel estimation approaches into three categories, namely, separate pilot approach, embedded pilot approach, and superimposed pilot approach. We compile and present an overview of some of the key algorithms under these categories and illustrate their performance and complexity attributes.

Keywords: OTFS modulation; delay-Doppler domain; high-Doppler channels; signal detection; channel estimation

DOI: 10.12142/ZTECOM.202104003

<https://kns.cnki.net/kcms/detail/34.1294.TN.20211207.1647.002.html>, published online December 8, 2021

Manuscript received: 2021-10-18

Citation (IEEE Format): A. Naikoti and A. Chockalingam, "Signal detection and channel estimation in OTFS," *ZTE Communications*, vol. 19, no. 4, pp. 16 – 33, Dec. 2021. doi: 10.12142/ZTECOM.202104003.

1 Introduction

Next-generation wireless systems are expected to support a variety of use cases with a wide range of performance requirements. Interest in high-mobility use cases involving high-speed trains, unmanned vehicles/cars, drones, aeroplanes, etc., is on the rise. Also, in order to meet the growing bandwidth requirement, a wireless spectrum in the mmWave frequency band is preferred. Communication in such high-mobility and/or high carrier frequency scenarios has to deal with high Doppler shifts which are common in such environments. Orthogonal frequency division multiplexing (OFDM) is a widely used communication waveform in the current generation of wireless systems. De-

spite its popularity and adoption in current standards, OFDM suffers from severe performance degradation in high-Doppler scenarios. This is because of the increased loss of orthogonality among subcarriers and the resulting inter-carrier interference (ICI).

Orthogonal time frequency space (OTFS) modulation is a recently introduced 2-dimensional (2D) modulation^[1]. There has been growing interest in this modulation recently, because of its superior performance compared with OFDM in high-Doppler environments^[2-6]. In OTFS modulation, information symbols are multiplexed in the delay-Doppler (DD) domain. The symbols in the DD domain are converted to the time domain and transmitted. At the receiver, the received

signal in the time domain is converted back to the DD domain where the information symbols are recovered. The DD domain to time domain conversion and vice versa can be done using two approaches. In the first approach, symbols from the DD domain are mapped to the time domain in two steps: DD to time-frequency (TF) domain conversion using inverse symplectic finite Fourier transform (ISFFT), followed by the TF domain to time domain conversion using Heisenberg transform^[1]. The corresponding inverse transforms map the received time-domain signal to the TF domain and then to the DD domain (Wigner transform followed by SFFT). The second approach is a direct one-step approach, where DD domain to time domain mapping is done using inverse Zak transform^[7]. At the receiver, the Zak transform maps the signal from the time domain directly to the DD domain. While the first approach has been adopted in most of the studies reported in the literature so far, the second approach is also gaining popularity. While the first approach can be implemented as an overlay on existing TF modulation schemes (such as OFDM), the second approach has the benefit of reduced implementation complexity.

Since the introduction of OTFS in 2017, there has been a spurt of research activities in OTFS leading to an increasing volume of publications on OTFS^[5–51]. Some of the key areas of focus in these works include DD signal representation in OTFS, input-output relation in the DD domain in the form of a linear vector channel model, framework for DD signal processing, signal detection algorithms, techniques for DD channel estimation, characterization of the peak-to-average power ratio (PAPR), the effect of practical pulse shapes, diversity analysis of OTFS, the effect of oscillator phase noise and IQ imbalance, multi-antenna OTFS, space-time coding and precoding in OTFS, multiuser OTFS on the uplink and downlink, etc. Recognizing that efficient signal detection and channel estimation techniques are crucial for the successful realization of OTFS systems in practice, we focus on these two key receiver functions in this paper.

We classify the OTFS signal detection approaches into three broad categories. The first is the linear detection approach, where the focus is on exploiting the structure inherent in the effective channel matrix for reducing complexity. The second approach is based on approximations to maximum a posteriori (MAP) detection, which aim near-optimal performance at reduced complexity. The last one is a recent approach involving deep neural networks (DNN). We highlight some of the algorithms in these categories reported in the literature. In highlighting various detection algorithms, perfect DD channel knowledge will be assumed at the receiver.

Similarly, we classify the DD channel estimation approaches into three categories. In the first approach, separate pilot frames are employed for DD channel estimation. The channel estimates obtained during the pilot frames are used for detec-

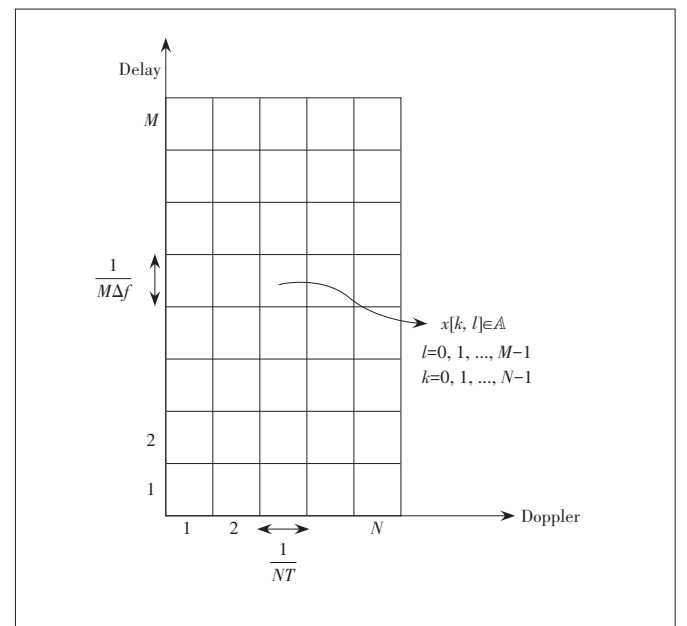
tion during data frames. The second approach involves embedding both pilot and data symbols in a frame. This further improves the throughput but the interference between pilot and data symbols needs to be taken into account by way of providing guard symbols around pilot symbols and/or interference cancellation. The last approach is the superimposed pilot approach, where pilot symbols are superimposed on data symbols. This further increases the throughput while demanding sophisticated signal processing (e.g., interference cancellation) to perform joint channel estimation and detection. We present algorithms reported in the literature under these categories.

The rest of this paper is organized as follows. Section 2 introduces the OTFS system model. Section 3 presents the state-of-the-art approaches and algorithms for OTFS signal detection. Section 4 presents the approaches and algorithms for DD channel estimation. Section 5 provides the conclusions.

2 OTFS Modulation and System Model

2.1 OTFS Modulation

In OTFS modulation, MN information symbols are multiplexed onto an $N \times M$ DD grid, where N is the number of Doppler bins and M is the number of delay bins, as shown in Fig. 1. The information symbols, denoted by $x[k, l]$, $k = 0, \dots, N-1$, $l = 0, \dots, M-1$, come from a modulation alphabet \mathbb{A} (e.g., QAM/PSK). The NM symbols are transmitted over a time duration of NT , occupying a bandwidth of $M\Delta f$, where $\Delta f = 1/T$. The Doppler resolution is $\frac{1}{NT}$ and the delay resolu-



▲ Figure 1. Multiplexing in delay-Doppler grid

tion is $\frac{1}{M\Delta f}$.

The symbols in the DD grid are mapped to a time-domain signal $x(t)$ for transmission. This can be done in two ways as shown in Fig. 2. In a two-step approach, the DD signal is first mapped to a time-frequency (TF) signal which is then mapped to a time-domain signal. The DD-to-TF domain mapping is done using ISFFT and the TF-to-time domain mapping is done using Heisenberg transform. In a one-step approach, the DD signal is directly mapped to a time-domain signal using inverse Zak transform. The corresponding inverse transforms are used at the receiver to demap the received time-domain signal to the DD domain. In this paper, we adopt the two-step approach which has been widely followed in the literature so far.

2.2 OTFS System Model

In this subsection, we present the OTFS system model for the two-step approach of DD-to-time domain conversion, as shown in Fig. 2(a). The symbols $x[k, l]$ in the DD domain are mapped to the TF domain using ISFFT, as

$$X[n, m] = \frac{1}{MN} \sum_{k=0}^{N-1} \sum_{l=0}^{M-1} x[k, l] e^{j2\pi \left(\frac{nk}{N} - \frac{ml}{M} \right)}. \quad (1)$$

This TF signal is transformed into a time-domain signal using Heisenberg transform, as

$$x(t) = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} X[n, m] g_{tx}(t - nT) e^{j2\pi m \Delta f (t - nT)}, \quad (2)$$

where $g_{tx}(t)$ defines the transmit pulse shape. The transmitted signal is passed through the channel whose response in the DD domain is given by

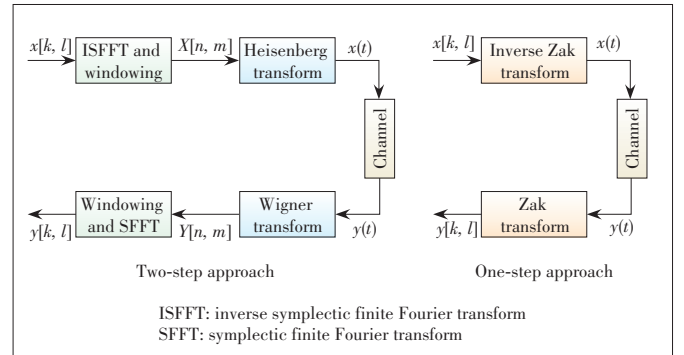
$$h(\tau, \nu) = \sum_{i=1}^P h_i \delta(\tau - \tau_i) \delta(\nu - \nu_i), \quad (3)$$

where h_i , τ_i , and ν_i are the channel gain, delay, and Doppler shift associated with the i -th path, respectively, and P is the number of resolvable paths in the DD domain.

The received time domain signal $y(t)$ at the receiver is given by

$$y(t) = \int_{\tau} \int_{\nu} h(\tau, \nu) x(t - \tau) e^{j2\pi \nu (t - \tau)} d\tau d\nu + v(t), \quad (4)$$

where $v(t)$ is the additive white Gaussian noise. Wigner transform is applied to $y(t)$ to transform it into a TF domain signal, as



▲ Figure 2. Orthogonal time frequency space (OTFS) modulation scheme

$$Y[n, m] = A_{g_{rx}, y}(t, f) \Big|_{t=nT, f=m\Delta f},$$

$$A_{g_{rx}, y}(t, f) = \int g_{rx}^*(t' - t) y(t) e^{-j2\pi f(t' - t)} dt', \quad (5)$$

where $g_{rx}(t)$ defines the receive pulse shape. If $g_{rx}(t)$ and $g_{tx}(t)$ satisfy the biorthogonality condition, the input-output relation in the TF domain is given by Ref. [11]:

$$Y[n, m] = H[n, m] X[n, m] + V[n, m], \quad (6)$$

where $V[n, m]$ is noise in TF domain and $H[n, m]$ is

$$H[n, m] = \int_{\tau} \int_{\nu} h(\tau, \nu) e^{j2\pi \nu nT} e^{-j2\pi (v + m\Delta f)\tau} d\nu d\tau. \quad (7)$$

The TF signal $Y[n, m]$ is transformed to the DD domain signal $y[k, l]$ using SFFT, as

$$y[k, l] = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} Y[n, m] e^{-j2\pi \left(\frac{nk}{N} - \frac{ml}{M} \right)}. \quad (8)$$

The above DD domain signal at the output of the SFFT can be derived to be of the form¹

$$y[k, l] = \sum_{i=1}^P h'_i x \left[(k - \beta_i)_N, (l - \alpha_i)_M \right] + v[k, l], \quad (9)$$

where $h'_i = h_i e^{-j2\pi \nu_i \tau_i}$, h_i s are i. i. d and are distributed as $\mathcal{CN}(0, 1/P)$ with uniform scattering profile, α_i and β_i are integers² corresponding to indices of delay and Doppler, respectively, for the i -th path, i.e., $\tau_i \triangleq \frac{\alpha_i}{M\Delta f}$ and $\nu_i \triangleq \frac{\beta_i}{NT}$, and $v[k, l]$ is the additive white Gaussian noise. By vectorizing the input-output relation in Eq. (9), we can write^[11]

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}, \quad (10)$$

where $\mathbf{x}, \mathbf{y}, \mathbf{v} \in \mathcal{C}^{MN \times 1}$, the $(k + Nl)$ -th entry of \mathbf{x} , $x_{k + Nl} =$

1. Refer to Ref. [11] for the detailed derivation.

2. For the purpose of exposition of detection and channel estimation algorithms, integer Dopplers are considered in this paper. Refer to Ref. [11] for a similar system model for fractional Dopplers. Also, refer to Ref. [13] for the MIMO-OTFS system model.

$x[k, l], k = 0, \dots, N-1, l = 0, \dots, M-1$ and $x[k, l] \in \mathbb{A}$. Similarly, $y_{k+NI} = y[k, l]$ and $v_{k+NI} = v[k, l], k = 0, \dots, N-1, l = 0, \dots, M-1$, and $\mathbf{H} \in \mathcal{C}^{MN \times MN}$ is the effective channel matrix, whose j -th row ($j = k + NI$), denoted by $\mathbf{H}[j]$, is given by $\mathbf{H}[j] = [\hat{h}((k-0)_N, (l-0)_M) \ \hat{h}((k-1)_N, (l-0)_M) \ \dots \hat{h}((k-N+1)_N, (l-M+1)_M)]$, where $\hat{h}(k, l)$ denotes the (k, l) -th element of the $N \times M$ DD channel matrix, given by

$$\hat{h}(k, l) = \begin{cases} h'_i, & \text{if } k = \beta_i, l = \alpha_i, i \in \{1, 2, \dots, P\} \\ 0, & \text{otherwise} \end{cases}. \quad (11)$$

It can be seen from the above that the effective channel matrix \mathbf{H} has only P non-zero entries in each row and column, i.e., there are only MNP non-zero elements in \mathbf{H} . The linear vector channel model in Eq. (10) is used for signal detection/equalization and channel estimation in OTFS.

3 OTFS Signal Detection

In this section, we present some of the signal detection algorithms proposed in the literature for OTFS modulation. Based on the approaches used, these algorithms are categorized into three groups, namely, 1) low-complexity linear detection, 2) approximate MAP detection, and 3) neural networks based detection, as shown in Fig. 3. We present algorithms under these categories in the following subsections, assuming perfect knowledge of the DD channel matrix. Later, in Section 4, we will present techniques/algorithms to estimate the channel matrix.

3.1 Low-Complexity Linear Detection

Linear equalizers detect the transmitted symbols by applying a linear transformation to the received vector \mathbf{y} followed by mapping to a symbol in the modulation alphabet \mathbb{A} which is closest in terms of euclidean distance. The linear transformation matrix is represented by \mathbf{G} and the mapping function is represented by $f(\cdot)$. Therefore, the estimate of the transmit vector \mathbf{x} is given by

$$\hat{\mathbf{x}} = f(\mathbf{G}\mathbf{y}). \quad (12)$$

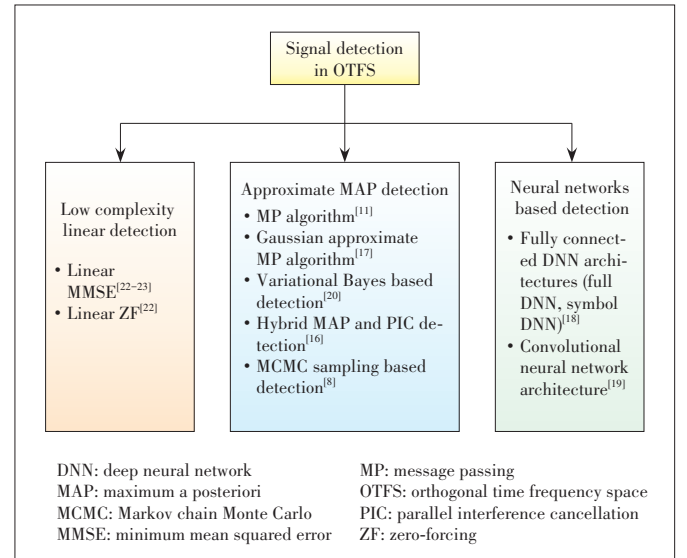
The transformation matrix for linear minimum mean squared error (LMMSE) equalization is given by

$$\mathbf{G}_{\text{lmmse}} = (\mathbf{H}^H \mathbf{H} + \sigma^2 \mathbf{I})^{-1} \mathbf{H}^H, \quad (13)$$

where σ^2 is the noise variance and the transformation matrix for zero-forcing (ZF) equalization is given by

$$\mathbf{G}_{\text{zf}} = (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H. \quad (14)$$

These equalizers in the case of OTFS have a computation



▲ Figure 3. Signal detection approaches in OTFS

complexity of $\mathcal{O}(M^3 N^3)$. However, the structure of the channel matrix in the DD domain can be exploited to reduce the complexity of these operations^[22-23].

3.1.1 Low-Complexity LMMSE Equalization

In the low-complexity linear minimum mean square error (LMMSE) equalization^[22], the channel matrix \mathbf{H} in Eq. (10) has a block circulant structure with M circulant blocks denoted by \mathbf{A}_i ($i = 0, 1, \dots, M-1$) of size $N \times N$. Using this property of the channel matrix, a low-complexity algorithm for implementing LMMSE equalization has been proposed in Ref. [22]. Let $\mathcal{C}_{M,N}$ denote the class of such block circulant matrices. These matrices have the following exploitable properties.

- Any matrix $\mathbf{H} \in \mathcal{C}_{M,N}$ can be unitarily diagonalizable as

$$\mathbf{H} = (\mathbf{F}_M \otimes \mathbf{F}_N)^H \mathbf{\Lambda} (\mathbf{F}_M \otimes \mathbf{F}_N), \quad (15)$$

where $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_{MN}\}$ such that λ_i is the i -th eigenvalue of \mathbf{H} , \mathbf{F}_M is the DFT matrix of size M , and \otimes is the Kronecker product operator.

- The matrix $\mathbf{\Lambda}$ can be written as

$$\mathbf{\Lambda} = \sum_{i=0}^{M-1} \mathbf{\Omega}_M^i \otimes \mathbf{A}_i, \quad (16)$$

where $\mathbf{\Omega}_M = \text{diag}\{1, \omega, \dots, \omega^{M-1}\}$ and $\omega = e^{j2\pi/M}$. \mathbf{A}_i is $N \times N$ diagonal matrix with eigenvalues of $N \times N$ circulant block \mathbf{A}_i on the diagonal.

- For $\mathbf{A}, \mathbf{B} \in \mathcal{C}_{M,N}$, the matrices $\mathbf{A}^T, \mathbf{A}^H, \mathbf{AB} = \mathbf{BA}$, $c_1 \mathbf{A} + c_2 \mathbf{B}$, $\sum_{r=0}^{R-1} c_r \mathbf{A}_r$ (c_r are all scalars) and \mathbf{A}^{-1} (if exists) are also block circulant and belong to $\mathcal{C}_{M,N}$.

As $\mathbf{H} \in \mathcal{C}_{M,N}$, using the above properties, we have the LMMSE transformation matrix $\mathbf{G}_{\text{lmmse}} \in \mathcal{C}_{M,N}$. Thus, by substituting Eq. (15) in Eq. (13), we get

$$\mathbf{G}_{\text{Immse}} = (\mathbf{F}_M \otimes \mathbf{F}_N)^H \mathbf{\Psi} (\mathbf{F}_M \otimes \mathbf{F}_N), \quad (17)$$

where $\mathbf{\Psi}$ is a diagonal matrix containing the eigenvalues of $\mathbf{G}_{\text{Immse}}$, given by

$$\mathbf{\Psi} = (\mathbf{\Lambda}^* \mathbf{\Lambda} + \sigma^2 \mathbf{I})^{-1} \mathbf{\Lambda}^*, \quad (18)$$

where $\mathbf{\Psi}_i = \frac{\lambda_i^*}{|\lambda_i|^2 + \sigma^2}$, $i = 1, 2, \dots, MN$. To reduce the complexity of computing $(\mathbf{F}_M \otimes \mathbf{F}_N) \mathbf{y}$, write \mathbf{y} as an $N \times M$ matrix \mathbf{Y} such that $\text{vec}(\mathbf{Y}) = \mathbf{y}$. This gives

$$\mathbf{z} = (\mathbf{F}_M \otimes \mathbf{F}_N) \mathbf{y} = \text{vec}(\mathbf{F}_N \mathbf{Y} \mathbf{F}_M). \quad (19)$$

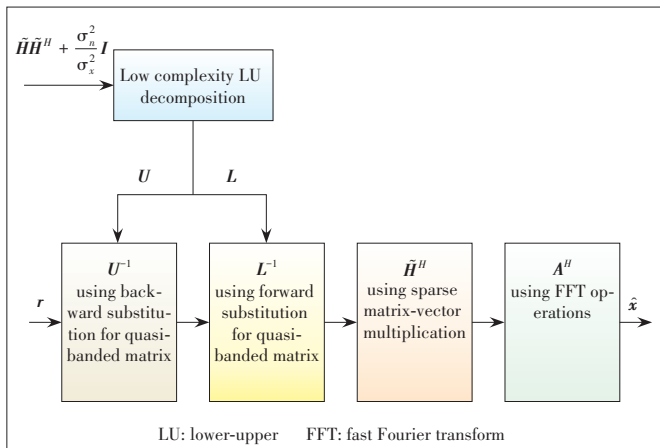
Now, compute $\mathbf{q} = \mathbf{\Psi} \mathbf{z}$ and write \mathbf{q} as a $N \times M$ matrix \mathbf{Q} such that $\text{vec}(\mathbf{Q}) = \mathbf{q}$. Finally, compute the estimated \mathbf{x} as

$$\hat{\mathbf{x}} = \mathbf{G}_{\text{Immse}} \mathbf{y} = \text{vec}(\mathbf{F}_N^H \mathbf{Q} \mathbf{F}_M), \quad (20)$$

which gives the exact LMMSE solution at a much less computational complexity. The complexity of computing \mathbf{z} involving N -point DFT and M -point IDFT operations is $\mathcal{O}(MN \log MN)$ and the complexity of computing \mathbf{q} is $\mathcal{O}(MN)$. Again, the computation of $\hat{\mathbf{x}}$ involves N -point IDFT and M -point DFT operations with complexity $\mathcal{O}(MN \log MN)$. Therefore, the overall complexity is $\mathcal{O}(2MN \log MN + MN)$ which is much small compared to the $\mathcal{O}(M^3 N^3)$ complexity of conventional LMMSE detection using matrix inversion.

3.1.2 Low-Complexity LMMSE Equalization

As shown in Fig. 4, a low-complexity LMMSE equalization method that takes advantage of the sparse and quasi-banded nature of the OTFS demodulation matrices has been pro-



▲ Figure 4. Low complexity linear minimum mean squared error (LMMSE) equalization^[23]

posed in Ref. [23]. Here, a different representation of the system is used. The transmit vector \mathbf{x} in the DD domain is written as an $M \times N$ matrix \mathbf{X} such that $\text{vec}(\mathbf{X}) = \mathbf{x}$. Assume that $E[x(k, l)x^*(k', l')] = \sigma_x^2 \delta(k - k', l - l')$. Using this representation, we obtain vector \mathbf{s} as

$$\mathbf{s} = \text{vec}(\mathbf{X} \mathbf{F}_N^H). \quad (21)$$

This vector \mathbf{s} can also be written in the form $\mathbf{s} = \mathbf{A} \mathbf{x}$, where $\mathbf{A} = \mathbf{F}_N^H \otimes \mathbf{I}_M$ is a unitary matrix. The received vector \mathbf{r} at the receiver is given by

$$\mathbf{r} = \tilde{\mathbf{H}} \mathbf{s} + \mathbf{n} = \tilde{\mathbf{H}} \mathbf{A} \mathbf{x} + \mathbf{n}, \quad (22)$$

where $\tilde{\mathbf{H}} = \sum_{i=1}^P h_i \mathbf{\Pi}^{\alpha_i} \Delta^{\beta_i}$, $\mathbf{\Pi} = \text{circ}\{[010\dots 0]_{MN \times 1}^T\}$ is a circulant delay matrix, $\Delta = \text{diag}\left(1, e^{j2\pi \frac{1}{MN}}, \dots, e^{j2\pi \frac{MN-1}{MN}}\right)$ is a diagonal Doppler matrix, and \mathbf{n} is i.i.d Gaussian noise vector with variance σ_n^2 . The detected symbol vector for this system model in Eq. (22) using LMMSE equalization is given by

$$\hat{\mathbf{x}} = (\tilde{\mathbf{H}} \mathbf{A})^H \left[(\tilde{\mathbf{H}} \mathbf{A}) (\tilde{\mathbf{H}} \mathbf{A})^H + \frac{\sigma_n^2}{\sigma_x^2} \mathbf{I} \right]^{-1} \mathbf{r}. \quad (23)$$

Due to the unitary nature of \mathbf{A} , the above equation reduces to

$$\hat{\mathbf{x}} = \mathbf{A}^H \underbrace{\tilde{\mathbf{H}}^H \left[\tilde{\mathbf{H}} \tilde{\mathbf{H}}^H + \frac{\sigma_n^2}{\sigma_x^2} \mathbf{I} \right]^{-1}}_{\mathbf{H}_{eq}} \mathbf{r}. \quad (24)$$

This detected vector $\hat{\mathbf{x}}$ is obtained in two steps. The first step involves a calculation of $\tilde{\mathbf{r}} = \mathbf{H}_{eq} \mathbf{r}$ and the second step involves the matched filter operation $\mathbf{A}^H \tilde{\mathbf{r}}$ to get $\hat{\mathbf{x}}$. Most complexity is in the first step as it involves an inverse of $\mathbf{\Psi} = \left[\tilde{\mathbf{H}} \tilde{\mathbf{H}}^H + \frac{\sigma_n^2}{\sigma_x^2} \mathbf{I} \right]$ to obtain \mathbf{H}_{eq} . This complexity is reduced by using a low-complexity LU decomposition of $\mathbf{\Psi}$ ^[23]. With LU decomposition of $\mathbf{\Psi}$, we can write Eq. (24) as

$$\hat{\mathbf{x}} = \mathbf{A}^H \tilde{\mathbf{H}}^H \underbrace{\mathbf{U}^{-1} \mathbf{L}^{-1} \mathbf{r}}_{\mathbf{r}_2}. \quad (25)$$

The computational complexity can be further reduced by using the quasi-banded nature of lower triangular matrix \mathbf{L} and upper triangular matrix \mathbf{U} . In Eq. (25), \mathbf{r}_1 is computed using the forward substitution method for quasi-banded lower triangular matrix and \mathbf{r}_2 is computed using backward substitution method for quasi-banded upper triangular matrix. The final computation of $\hat{\mathbf{x}} = \mathbf{A}^H \tilde{\mathbf{H}}^H \mathbf{r}_2$ is done by first obtaining $\tilde{\mathbf{r}} = \tilde{\mathbf{H}}^H \mathbf{r}_2$. This vector $\tilde{\mathbf{r}}$ is arranged in an $M \times N$ matrix as

$$\hat{\mathbf{Y}} = \begin{bmatrix} \tilde{r}(0) & \tilde{r}(M) & \cdots & \tilde{r}(MN - N) \\ \tilde{r}(1) & \tilde{r}(M + 1) & \cdots & \tilde{r}(MN - N + 1) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{r}(M - 1) & \tilde{r}(2M - 1) & \cdots & \tilde{r}(MN - 1) \end{bmatrix}. \quad (26)$$

Using the matrix $\hat{\mathbf{Y}}$, $\hat{\mathbf{x}}$ is obtained using DFT operation as

$$\hat{\mathbf{x}} = \text{vec}(\hat{\mathbf{Y}}\mathbf{F}_N). \quad (27)$$

The main reduction in the complexity is in the computation of \tilde{r} . All the steps involved in obtaining \mathbf{r}_2 together have a complexity of $\mathcal{O}(MN)$ and the final computation in Eq. (27) to obtain $\hat{\mathbf{x}}$ has a complexity of $\mathcal{O}\left(\frac{MN}{2}\log_2 N\right)$. On the whole, this LMMSE equalization method has a complexity of just $\mathcal{O}\left(\frac{MN}{2}\log_2 N + MN\right)$ compared to $\mathcal{O}(M^3N^3)$ of the conventional LMMSE equalization.

3.2 Approximate MAP Detection

From the vectorized representation $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}$ in Eq. (10), the MAP decision rule for detection of \mathbf{x} is given by

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x} \in \mathbb{A}^{MN}} \Pr(\mathbf{x}|\mathbf{y}, \mathbf{H}). \quad (28)$$

When the transmit symbol vectors are equally likely, the decision rule for maximum-likelihood (ML) detection is

$$\begin{aligned} \hat{\mathbf{x}} &= \arg \max_{\mathbf{a} \in \mathbb{A}^{MN}} \Pr(\mathbf{y}|\mathbf{x} = \mathbf{a}, \mathbf{H}) \Pr(\mathbf{x} = \mathbf{a}) = \\ &= \arg \max_{\mathbf{a} \in \mathbb{A}^{MN}} \frac{1}{|\mathbb{A}|} \Pr(\mathbf{y}|\mathbf{x} = \mathbf{a}, \mathbf{H}). \end{aligned} \quad (29)$$

Assuming the noise vector \mathbf{v} to be i.i.d Gaussian, the optimum decision rule is given by

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathbb{A}^{MN}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2. \quad (30)$$

The complexity of the optimum detector grows exponentially in MN because of the exhaustive enumeration/search involved. Therefore, several suboptimum detection algorithms have been proposed that are efficient with low complexity. In the following, some of the popular low-complexity approximate MAP algorithms for OTFS signal detection are presented.

3.2.1 Message Passing Algorithm

One of the popular detectors reported in the early OTFS literature is an approximate MAP detector based on low-complexity message passing^[11]. The key advantages of this detector are its linear complexity in MN and very good performance. The message passing algorithm involves the computation of approximate a posteriori probability of the modulation symbols by passing messages on a factor graph. The transmitted vector \mathbf{x} is represented by MN variable nodes and the received vector \mathbf{y} is represented by MN check nodes in the graph. As noted earlier in Section 2, \mathbf{H} is sparse with only L non-zero elements in any row or column (generally $L \ll MN$ and $L = P$ for non-fractional delays and Dopplers).

Let $\mathcal{I}(r)$ and $\mathcal{J}(c)$ denote the indices corresponding to non-zero elements in the r -th row and c -th column, respectively, such that $|\mathcal{I}(r)| = |\mathcal{J}(c)| = L$ for all rows and columns. In the factor graph, each variable node $x[c]$ has connections with L check nodes $y[r], r \in \mathcal{J}(c)$ and each check node $y[r]$ has connections with L variable nodes $x[c], c \in \mathcal{I}(r)$ as shown in Fig. 5. The symbol-by-symbol decision rule is given by

$$\begin{aligned} \hat{x}[c] &= \arg \max_{a_j \in \mathbb{A}} \frac{1}{|\mathbb{A}|} \Pr(\mathbf{y}|\mathbf{x}[c] = a_j, \mathbf{H}) \approx \\ &= \arg \max_{a_j \in \mathbb{A}} \prod_{r \in \mathcal{J}(c)} \Pr(y[r]|\mathbf{x}[c] = a_j, \mathbf{H}). \end{aligned} \quad (31)$$

This approximation is obtained assuming that the components of vector \mathbf{y} are independent for a given $\mathbf{x}[c]$ because of the sparsity of \mathbf{H} matrix. From the system model, we can write

$$y[r] = x[c]H[r, c] + \underbrace{\sum_{e \in \mathcal{I}(r), e \neq c} x[e]H[r, e]}_{\zeta_{r,c}^{(i)}} + v[r]. \quad (32)$$

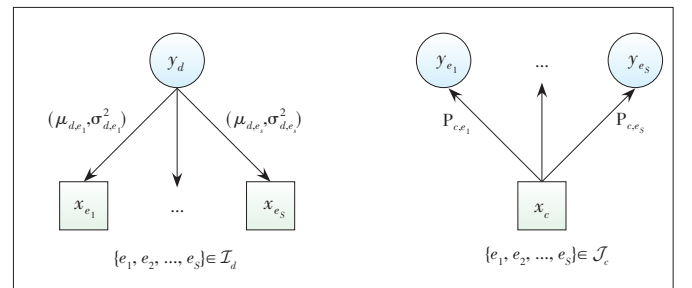
In the i -th iteration, the complete interference plus noise term in Eq. (32) is modelled as a single Gaussian random variable $\zeta_{r,c}^{(i)}$ with mean $\mu_{r,c}^{(i)}$ and variance $(\sigma_{r,c}^{(i)})^2$.

1) Message from check node $y[r]$ to variable node $x[c]$: The check nodes pass mean and variance information to the variable nodes, where

$$\mu_{r,c}^{(i)} = \sum_{e \in \mathcal{I}(r), e \neq c} \sum_{j=1}^{|\mathbb{A}|} p_{e,r}^{(i-1)}(a_j) a_j H[r, e], \quad (33)$$

$$\begin{aligned} (\sigma_{r,c}^{(i)})^2 &= \\ &= \sum_{e \in \mathcal{I}(r), e \neq c} \left(\sum_{j=1}^{|\mathbb{A}|} p_{e,r}^{(i-1)}(a_j) |a_j|^2 |H[r, e]|^2 - \left| \sum_{j=1}^{|\mathbb{A}|} p_{e,r}^{(i-1)}(a_j) a_j H[r, e] \right|^2 \right) + \sigma^2, \end{aligned} \quad (34)$$

and σ^2 is the variance of $v[r]$.



▲ Figure 5. Message passing between variable nodes and check nodes

2) Message from variable node $x[c]$ to check node $y[r]$: The variable nodes compute the probability mass function and pass it to the check nodes. Each component of the probability mass function $p_{(c,r)}^{(i)}$ is given by

$$p_{(c,r)}^{(i)}(a_j) = \Delta \tilde{p}_{(c,r)}^{(i)}(a_j) + (1 - \Delta) \tilde{p}_{(c,r)}^{(i-1)}(a_j), \quad (35)$$

where

$$\tilde{p}_{(c,r)}^{(i)}(a_j) = \prod_{e \in \mathcal{J}(c), e \neq r} \Pr(y[e] | x[c] = a_j, H) = \prod_{e \in \mathcal{J}(c), e \neq r} \frac{\xi_{(i)}(e, c, j)}{\sum_{t=1}^{|A|} \xi_{(i)}(e, c, t)}, \quad (36)$$

$$\xi_{(i)}(e, c, t) = \exp \left(\frac{-|y[e] - \mu_{e,c}^{(i)} - H_{e,c} a_t|^2}{(\sigma_{e,c}^{(i)})^2} \right). \quad (37)$$

3) Compute convergence parameter $\eta^{(i)}$ for some $\gamma > 0$, as

$$\eta^{(i)} = \frac{1}{MN} \sum_{c=1}^{MN} \mathbb{I} \left(\max_{a_j \in \mathbb{A}} p_c^{(i)}(a_j) \geq 1 - \gamma \right), \quad (38)$$

and

$$p_c^{(i)}(a_j) = \prod_{e \in \mathcal{J}(c)} \frac{\xi_{(i)}(e, c, j)}{\sum_{t=1}^{|A|} \xi_{(i)}(e, c, t)}. \quad (39)$$

4) If $\eta^{(i)} < \eta^{(i-1)}$ for $c = 1, 2, \dots, MN$, then update

$$\hat{x}[c] = \arg \max_{a_j \in \mathbb{A}} p_c^{(i)}(a_j). \quad (40)$$

5) Stop the iterations if any one of the following holds:

- The maximum limit set for the number of iterations has reached;

- $\eta^{(i)} = 1$;
- $\eta^{(i)} < \eta^{(s^*)} - \epsilon$ for some small ϵ , where $\eta^{(s^*)} = \max_{s < i} \eta^{(s)}$.

Modified variants of this message passing algorithm have also been proposed. For example, a low complexity variant that exploits channel hardening through match filtering operation on \mathbf{y} and message passing on the resulting system model is presented in Ref. [15]. Another variant in Ref. [17] is presented below.

3.2.2 Gaussian Approximate Message Passing Algorithm

In this variant of message passing, the a posteriori probability of each transmitted symbol is assumed to be Gaussian distributed instead of assuming the inter-symbol interference to be Gaussian as done earlier^[17]. We have

$$\Pr(\mathbf{y} | \mathbf{x}) = \prod_{r=1}^{MN} \Pr(y[r] | \mathbf{x}). \quad (41)$$

As the channel matrix is sparse, we get

$$\Pr(y[r] | \mathbf{x}) = \Pr(y[r] | \mathbf{x}_{\mathcal{I}(r)}), \quad (42)$$

where $\mathbf{x}_{\mathcal{I}(r)}$ contains the elements $x[c]$, $c \in \mathcal{I}(r)$. The Gaussian assumption is that

$$\Pr(y[r] | \mathbf{x}_{\mathcal{I}(r)}) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left(-\frac{1}{2\sigma^2} |y[r] - \mathbf{H}_{\mathcal{I}(r)} \mathbf{x}_{\mathcal{I}(r)}|^2 \right), \quad (43)$$

where $\mathbf{H}_{\mathcal{I}(r)}$ is a row vector containing the non-zero elements in r -th row of \mathbf{H} . This approach has been proposed in Ref. [17] and the results presented show that this approach has superior bit error rate (BER) performance with the same complexity order.

3.2.3 Variational Bayes Detection

An iterative algorithm that approximates the optimal MAP detection and has a faster convergence compared to the message passing algorithm has been proposed in Ref. [20]. An approximation of the a posteriori probability $p(\mathbf{x} | \mathbf{y})$ is obtained by using Kullback-Leibler (KL) divergence $\mathcal{D}(q || p)$ and the corresponding evidence lower bound (ELBO) is maximized iteratively using the variational Bayes approach. The convergence is guaranteed because the ELBO maximization problem is convexly resulting in a globally optimum solution. In this way, the marginal distribution for each symbol is obtained which is used for symbol-by-symbol MAP detection. The approximate distribution $q(\mathbf{x})$ is obtained by searching over a family of distributions \mathcal{Q} such that

$$q^*(\mathbf{x}) = \arg \max_{q \in \mathcal{Q}} \mathcal{D}(q || p) = \arg \max_{q \in \mathcal{Q}} \underbrace{\mathbb{E}_q[-\ln q(\mathbf{x}) + \ln p(\mathbf{x} | \mathbf{y})]}_{\mathcal{L}}. \quad (44)$$

The ELBO is given by \mathcal{L} which is the expectation over \mathbf{x} having distribution $q(\mathbf{x})$. In particular, when a family \mathcal{Q} with mutually independent variables is considered,

$$q(\mathbf{x}) = \prod_{i=1}^{MN} q_i(x[i]). \quad (45)$$

$q^*(x[i])$ is obtained iteratively $i = 1, 2, \dots, MN$, and the symbols are detected as

$$\hat{x}[i] = \arg \max_{s[i] \in \mathbb{A}} q_i^*(x[i]). \quad (46)$$

In addition to having complexity lower than that of MAP detection, this detection method has performance significantly close to the performance of MAP detection.

3.2.4 Hybrid MAP and PIC Detection

Another approximation to the MAP detection using a partitioning method based on path gains has been proposed in Ref. [16]. The hybrid MAP-PIC algorithm is a combination of both symbol-by-symbol MAP detection and message passing algorithm. The received symbols are partitioned into two subsets based on the path gains of the channel. On one part with good path gains MAP detection is used, and on the remaining part parallel interference cancellation (PIC) method is used. Define the following sets: $\mathbb{H}^{(i)} \triangleq \{h_j | 1 \leq j \leq P, j \neq i\}$, $\mathbb{Y}_{k,l} \triangleq \{y[(k + \beta_j)_N, (l + \alpha_j)_M] | 1 \leq j \leq P\}$, $\mathbb{X}_{k,l}^{(i)} \triangleq \{y[(k + \beta_i - \beta_j)_N, (l + \alpha_i - \alpha_j)_M] | 1 \leq j \leq P, j \neq i\}$, where P is the number of channel paths and i is the path index. The P received symbols that are associated with the transmitted symbol $x[k, l]$ are in set $\mathbb{Y}_{k,l}$. Similarly, the $P - 1$ transmitted symbols, other than $x[k, l]$, corresponding to the received symbol $\mathbb{Y}_{k,l}[i]$ are in set $\mathbb{X}_{k,l}^{(i)}$. The path gains in $\mathbb{H}^{(i)}$ are arranged in decreasing order such that $|h_m|^2 > |h_n|^2$ when $m < n$. The set $\mathbb{X}_{k,l}^{(i)}$ is partitioned into two subsets by enumerating different possible combinations of S (where S is the size of the first subset with good path gains) as follows:

$$\widetilde{\mathbb{X}}_{k,l}^{(i)} \triangleq \{\mathbb{X}_{k,l}^{(i)}[j] | 1 \leq j \leq S\}, \quad (47)$$

$$\bar{\mathbb{X}}_{k,l}^{(i)} \triangleq \{\mathbb{X}_{k,l}^{(i)}[j] | S + 1 \leq j \leq P - 1\}. \quad (48)$$

It is proposed to perform MAP detection on $\widetilde{\mathbb{X}}_{k,l}^{(i)}$ and PIC on $\bar{\mathbb{X}}_{k,l}^{(i)}$. The message passing algorithm that we discussed earlier is a special case of Hybrid-MAP-PIC detection when $S = 0$. Results have shown that choosing $S = P/2$ gives good error performance. A trade-off can be established between BER performance and computation complexity by choosing a suitable value for S .

3.2.5 MCMC Sampling Based Detection

This detection algorithm proposed in Ref. [8] uses the Markov chain Monte Carlo (MCMC) sampling method to obtain an approximate solution to Eq. (28). The joint probability distribution is given by

$$\Pr(\mathbf{x}|\mathbf{y}, \mathbf{H}) = \Pr(x_1, x_2, \dots, x_{MN}|\mathbf{y}, \mathbf{H}) \propto \exp\left(-\frac{\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2}{\sigma^2}\right). \quad (49)$$

The algorithm starts by initializing a random initial vector $\mathbf{x}^{(t=0)}$, where t denotes the iteration number. The MN coordinates of the \mathbf{x} vector are updated in each iteration based on the coordinates of the previous iteration as follows:

$$\mathbf{x}_1^{(t+1)} \sim \Pr(x_1|x_2^{(t)}, x_3^{(t)}, \dots, x_{MN}^{(t)}, \mathbf{y}, \mathbf{H}),$$

$$\mathbf{x}_2^{(t+1)} \sim \Pr(x_2|x_1^{(t)}, x_3^{(t)}, \dots, x_{MN}^{(t)}, \mathbf{y}, \mathbf{H}),$$

$$\mathbf{x}_3^{(t+1)} \sim \Pr(x_3|x_1^{(t)}, x_2^{(t)}, x_4^{(t)}, \dots, x_{MN}^{(t)}, \mathbf{y}, \mathbf{H}), \text{ and}$$

$$\mathbf{x}_{MN}^{(t+1)} \sim \Pr(x_{MN}|x_1^{(t)}, x_2^{(t)}, \dots, x_{MN-1}^{(t)}, \mathbf{y}, \mathbf{H}).$$

After updating over a certain number of iterations, the distribution obtained approximately converges to the distribution in Eq. (49). For a received vector, the symbol vector which has minimum ML cost $\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2$ in all the iterations is chosen as the detected symbol vector. A modification to this has also been proposed to reduce the number of iterations and also to overcome the phenomenon of stalling seen in the Gibbs sampling method at high SNRs which limits the BER performance. The modified joint distribution involves a temperature parameter α chosen based on the operating SNR and is given by

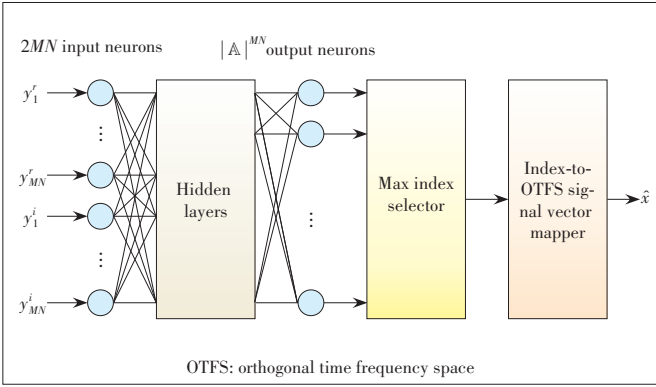
$$\Pr(\mathbf{x}|\mathbf{y}, \mathbf{H}) \propto \exp\left(-\frac{\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2}{\sigma^2 \alpha^2}\right). \quad (50)$$

Alternately, the sampling can be randomized by updating the parameters in each iteration using the conventional Gibbs sampling method with probability q (e.g., $q = \frac{1}{MN}$) and obtaining samples from a uniform distribution with probability $1 - q$. This randomized sampling has been shown to avoid stalling problems and achieve good BER performance.

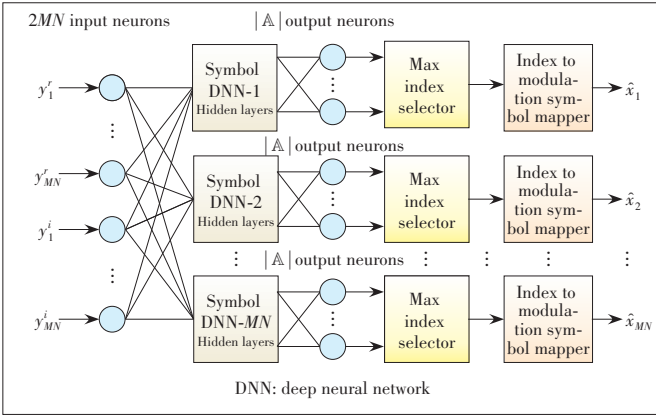
3.3 Neural Networks Based Detection

Apart from the detection methods based on conventional approaches, detectors based on DNN have been proposed recently. Two DNN approaches have been presented in Ref. [18]. One approach is to use a single fully-connected DNN to detect the signal vector. The detection problem is formulated as a multi-class classification problem where each class corresponds to each vector in the transmitted signal set, enabling joint detection of the transmitted symbol vector. The number of input neurons in the network is decided by the size of the received vector (MN) and the number of output neurons is decided by the size of the multi-dimensional modulation alphabet ($|\mathbb{A}|^{MN}$). This approach requires a large number of parameters to be learned and is computationally complex because of the exponential growth in the number of output neurons with the size of transmit symbol vector. The architecture of this fully connected DNN is shown in Fig. 6. The real and imaginary parts of the received vector \mathbf{y} are given as input to the DNN. The activation function used in the output layer is Softmax activation so that the output of each output neuron gives the probability of the corresponding transmitted signal vector and all these probabilities sum to one. The detected symbol vector is the one that has maximum probability.

Another architecture that uses multiple DNNs is shown in Fig. 7. In this approach, each symbol in the transmitted vector is detected by an individual DNN. In this way, each DNN



▲ Figure 6. Full deep neural network (DNN) architecture for detection



▲ Figure 7. Symbol DNN architecture for detection

has the number of output neurons growing linearly in the size of modulation alphabet ($|\mathbb{A}|$). Also, the number of DNNs grows linearly with the size of the transmit symbol vector (MN). This symbol-DNN architecture does symbol-by-symbol detection at lower complexity and achieves BER performance almost the same as that of full-DNN. Each DNN uses Softmax activation in the output layer and obtains probabilities corresponding to each symbol in \mathbb{A} .

The training of the DNNs is done by pseudo-randomly generating a set of training examples \mathbf{x}_T which are known both at the transmitter and the receiver. These training examples are sent to the receiver through the channel. The transmitted signal vector \mathbf{x}_T and the corresponding received signal vector \mathbf{y} form the training pair at the receiver. The real and imaginary parts of \mathbf{y} are given as input to the DNNs. The DNN trained in this manner learns the mapping from the received vector to the corresponding symbol in the transmitted vector. Another approach of using neural networks for signal detection has been reported in Ref. [19], where the two-dimensional structure of the OTFS frame with data augmentation based pre-processing is given as input to two-dimensional convolutional neural networks (CNN) for signal detection.

The benefits of using DNNs for signal detection can be predominantly seen when there are deviations in the noise model from the standard i.i.d. Gaussian model. In situations

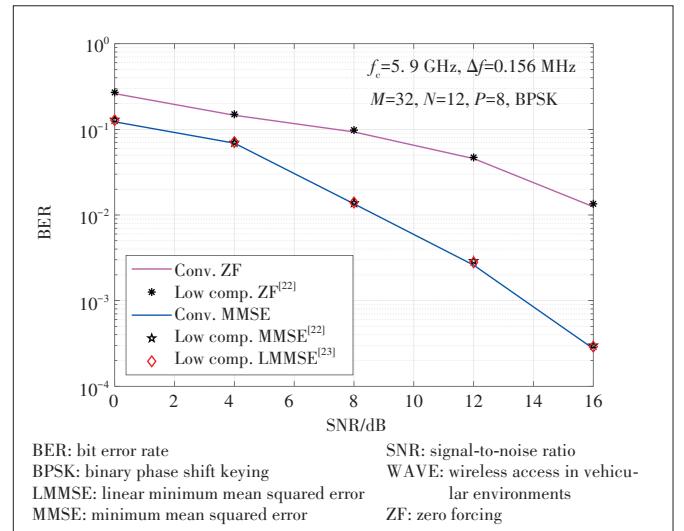
where there are deviations from the Gaussian as well as independence assumptions in the standard noise model, DNN based detection could outperform conventional ML detection. This is because ML detection is optimum only for the standard i.i.d. Gaussian noise model and the DNNs have the ability to learn the underlying deviations in the model.

3.4 Performance Results

In this subsection, we present the BER performance of some of the detectors presented in the previous subsections, assuming perfect DD channel knowledge at the receiver. The system parameters considered are according to the IEEE.802.11p standard for wireless access in vehicular environments (WAVE)^[52]. A carrier frequency of 5.9 GHz with a subcarrier spacing of 0.156 MHz, a maximum speed of 220 km/h, and a multipath channel with $P = 5$ paths are considered. BPSK modulation alphabet is used with a frame size of $M = 32$ and $N = 12$.

Fig. 8 shows the BER performance of the linear detectors including conventional MMSE/ZF detectors and low-complexity MMSE/ZF detectors. It can be observed from Fig. 8 that the performance of the conventional MMSE detector and the low-complexity MMSE detectors in Refs. [22] and [23] are the same. However, the detectors in Refs. [22] and [23] achieve this with significantly lower complexities compared to the conventional MMSE detector complexity. This is illustrated in Fig. 9 where the computational complexities (in the number of real operations) for these detectors are plotted.

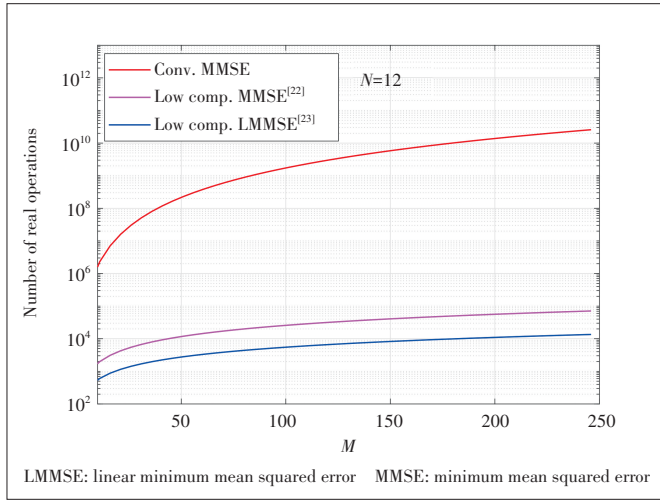
The BER performance of the message passing detector in Ref. [11] and the symbol-DNN based detector in Ref. [18] are shown in Fig. 10. MMSE detection performance is also shown for comparison. In this figure, a system with a carrier frequency of 4 GHz, subcarrier spacing of 15 kHz, OTFS frame size of $M = N = 16$, and a uniform power delay profile



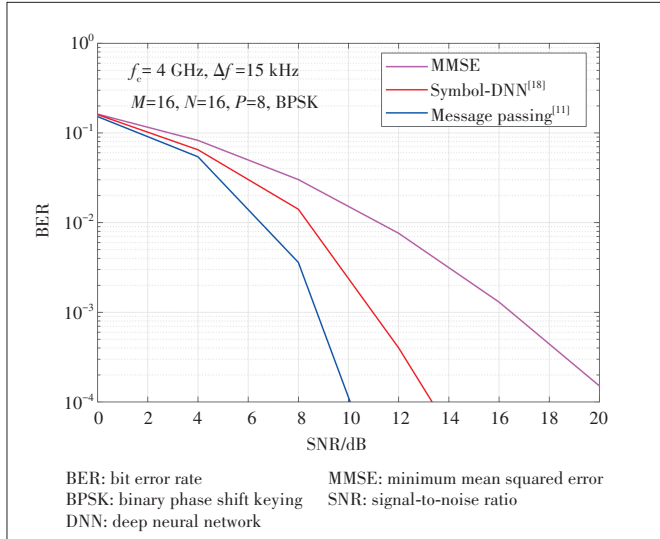
▲ Figure 8. BER performance of linear OTFS detectors for IEEE 802.11p WAVE channel model

channel with $P = 8$ are considered. The delay-Doppler profile considered is shown in Table 1. For the message passing algorithm, a damping factor of $\Delta = 0.6$ and maximum iterations of 30 are used. For the symbol-DNN based detection, the parameters of the neural network are shown in Table 2. It can be seen from Fig. 10 that the symbol-DNN performs better than the MMSE detector, and the message passing detector gives the best performance among them.

Next, the performance superiority of DNN-based detection compared to ML detection in correlated noise is illustrated in Fig. 11. This figure shows the BER performance of the ML



▲ Figure 9. Computational complexity of conventional MMSE and low-complexity MMSE detectors



▲ Figure 10. BER performance of MMSE, message passing, and symbol-DNN based detectors

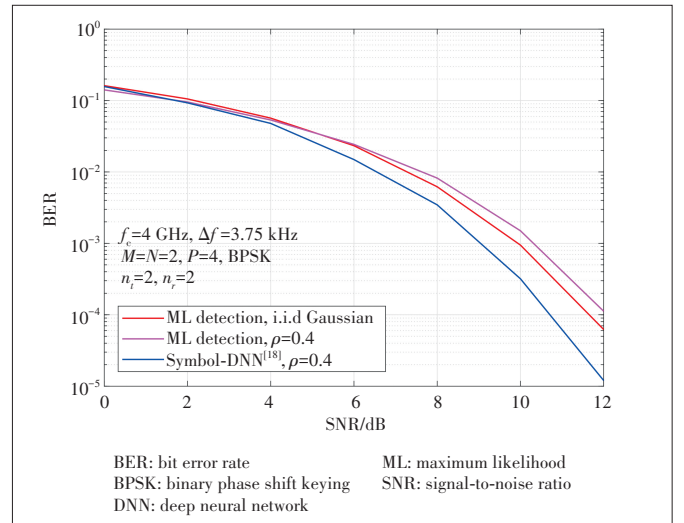
▼ Table 1. Delay-Doppler profile considered in Figure 10

Path (i)	1	2	3	4	5	6	7	8
τ_i (μ s)	0	4.16	8.32	12.48	16.64	20.8	24.96	29.12
ν_i (ms)	0	0	938.5	938.5	938.5	1875	1875	1875

▼ Table 2. Parameters of symbol-DNN detector in Figure 10

Parameters	Symbol-DNN
Number of input neurons	$2MN = 512$
Number of output neurons	$ A = 2$
Number of hidden layers	1
Number of neurons in hidden layers	256
Hidden layer activation	ReLU
Output layer activation	Softmax
Optimization	Adam
Loss function	Binary cross entropy
Training SNR	10 dB
Number of training examples	50 000
Number of epochs	50

DNN: deep neural network ReLU: Rectified Linear Unit SNR: signal-to-noise ratio



▲ Figure 11. BER performance of ML detection and symbol-DNN based detection in 2x2 MIMO-OTFS with correlated noise

detector and symbol-DNN detector for a MIMO-OTFS system when the noise is correlated. The correlated noise vector is taken to be $\mathbf{v}_c = \mathbf{N}_c \mathbf{v}$, where \mathbf{v} is the i.i.d Gaussian noise vector and \mathbf{N}_c is the correlation matrix $\mathbf{N}_c =$

$$\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n_r-1} \\ \rho & 1 & \rho & \cdots & \rho^{n_r-2} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \rho^{n_r-1} & \rho^{n_r-2} & \cdots & 1 \end{bmatrix}, \text{ given by Ref. [53], where } \rho$$

is the correlation coefficient ($0 \leq \rho \leq 1$), n_t is the number of transmit antennas, and n_r is the number of receive antennas. The following system parameters are considered in Fig. 11: carrier frequency of 4 GHz, subcarrier spacing of 3.75 kHz, frame size of $M = N = 2$, uniform power delay profile channel with $P = 4$, MIMO configuration with $n_t = n_r = 2$, and a correlation coefficient $\rho = 0.4$. The symbol-DNN architecture has an input layer with 16 nodes, one hidden layer with 32 nodes, and an output layer with 2 nodes. The hidden layer has ReLU activation and the output layer has Softmax activa-

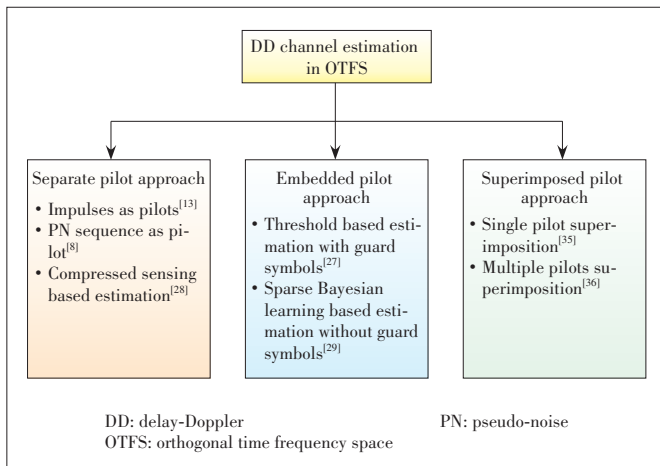
tion. The DNN is trained at an SNR of 8 dB for 60 epochs with 50 000 training examples. It can be seen from Fig. 11 that the symbol-DNN based detector outperforms the ML detector by almost 1 dB at BER of 10^{-4} . This is because the ML detector is optimal only when the noise is i.i.d Gaussian and is suboptimal in correlated noise. On the other hand, the symbol-DNN based detector performs well as it effectively learns the noise correlation leading to superior BER performance.

4 DD Channel Estimation

The task of channel estimation at the receiver is crucial as signal detection operation requires the knowledge of the channel state information. In OTFS, signal detection is carried out in the DD domain. In the system model in Eq. (10), knowledge of the DD channel matrix \mathbf{H} is needed for detection. In order to estimate \mathbf{H} , \mathbf{x} vector consisting of known pilot symbols is sent. Given the knowledge of the pilot symbol(s) in \mathbf{x} and the observation vector \mathbf{y} , channel estimation algorithms estimate \mathbf{H} . For the purpose of exposition, we classify the channel estimation approaches into three broad categories based on the OTFS frame pattern used to transmit the pilot and data symbols. Fig. 12 shows this classification consisting of 1) separate pilot approach, 2) embedded pilot approach, and 3) superimposed pilot approach. In the first approach, pilot frames consisting of only pilot symbols are used for channel estimation. The estimated channel matrix obtained during the pilot frame is used for detection during data frames. The second approach involves embedding both pilot and data symbols in a frame. In the third approach, pilot symbols are superimposed on data symbols. Some of the channel estimation techniques/algorithms employing these approaches are presented in the following subsections.

4.1 Separate Pilot Approach

As mentioned earlier, in this approach, separate frames



▲ Figure 12. DD channel estimation approaches in OTFS

are used for sending pilot symbols and data symbols. A pilot frame consists of only pilot symbol(s). One pilot frame per spatial coherence interval of the DD channel is sent. The channel estimated during the pilot frame is used for the detection of symbols in the data frames in that coherence interval. In the following, we present three-channel estimation methods using this approach.

4.1.1 Impulse Based Channel Estimation

In this method, impulses in the DD domain are sent as pilots^[13], i.e., the pilot symbol is given by

$$x_p[k, l] = \begin{cases} 1, & \text{if } (k, l) = (k_p, l_p) \\ 0, & \forall (k, l) \neq (k_p, l_p). \end{cases} \quad (51)$$

For this transmitted pilot, the received signal at the receiver is

$$\begin{aligned} y_p[k, l] &= \frac{1}{MN} \sum_{l'=0}^{M-1} \sum_{k'=0}^{N-1} x_p[k', l'] h_w \left(\frac{k-k'}{NT}, \frac{l-l'}{M\Delta f} \right) + \\ v[k, l] &= \frac{1}{MN} h_w \left(\frac{k-k_p}{NT}, \frac{l-l_p}{M\Delta f} \right) + v[k, l]. \end{aligned} \quad (52)$$

As the receiver knows the pilot locations k_p and l_p a priori, $h_w \left(\frac{k}{NT}, \frac{l}{M\Delta f} \right)$ can be estimated using Eq. (52) and the estimated channel matrix $\hat{\mathbf{H}}$ can be obtained.

4.1.2 PN Pilot Based Estimation

Instead of impulses as pilots, this method uses PN sequence as a pilot^[8]. The estimation is done in the discrete domain and the parameters to be estimated are delay tap τ_i , Doppler shift ν_i , and channel fade coefficient h'_i . The input-output relation for a P path channel in the time domain can be obtained as

$$\begin{aligned} y(t) &= \sum_{i=1}^P h_i x(t - \tau_i) e^{j2\pi\nu_i(t - \tau_i)} + v(t) = \\ &= \sum_{i=1}^P h'_i e^{j2\pi\nu_i t} x(t - \tau_i) + v(t), \end{aligned} \quad (53)$$

where $h'_i = e^{-j2\pi\nu_i\tau_i}$.

Let \mathcal{H} denote the vector space of complex-valued functions on the set of finite integers $\mathbb{Z}_{N_p} = \{0, 1, \dots, N_p - 1\}$ with an inner product defined as

$$\langle g_1, g_2 \rangle = \sum_{n \in \mathbb{Z}_{N_p}} g_1[n] g_2^*[n], \quad g_1, g_2 \in \mathcal{H}. \quad (54)$$

A signal is transmitted which is given by

$$S_A(t) = \sum_{n=0}^{M-1} S[n \bmod N_p] \text{sinc}(Wt - n), \quad (55)$$

where $S \in \mathcal{H}$, $M = N_p + \lceil W \max(\tau_i) \rceil \geq N_p$. For some $S \in \mathcal{H}$ that is transmitted, the received sequence $R[n]$ is

$$R[n] = \sum_{i=1}^P \alpha_i e(\omega_i n) S[n - \delta_i] + v[n], n \in \mathbb{Z}_{N_p}, \quad (56)$$

where $e(t) = e^{j \frac{2\pi}{N_p} t}$, $\delta_i, \omega_i \in \mathbb{Z}_{N_p}$, $\alpha_i \in \mathbb{C}$, and $v[n] \in \mathcal{H}$. Eq. (56) can be simplified as

$$R[n] = e(\omega_0 n) S[n - \delta_0] + v[n], \quad (57)$$

such that $(\delta_0, \omega_0) \in \mathbb{Z}_{N_p} \times \mathbb{Z}_{N_p}$. The (δ_0, ω_0) pairs are estimated using the time-frequency shift problem. A matched filter matrix for R and S is defined as

$$\mathcal{M}(R, S)[\delta, \omega] = \langle R[n], e(\omega n) S[n - \delta] \rangle = \begin{cases} 1 + \epsilon'_{N_p}, & \text{if } (\delta, \omega) = (\delta_0, \omega_0) \\ \epsilon'_{N_p}, & \text{if } (\delta, \omega) \neq (\delta_0, \omega_0), \end{cases} \quad (58)$$

where $|\epsilon'_{N_p}| \leq \frac{1}{\sqrt{N_p}}$ and $|\epsilon_{N_p}| \leq \frac{C+1}{\sqrt{N_p}}$ for some positive constant C . Thus, compute $\mathcal{M}(R, S)$ and choose (δ_0, ω_0) such that $\mathcal{M}(R, S)[\delta_0, \omega_0] \approx 1$. Once δ_0 and ω_0 are estimated, h'_i , τ_i and ν_i can be obtained.

4.1.3 Compressed Sensing Based Estimation

The channel estimation problem can be formulated as a sparse signal recovery problem using compressed sensing based methods like orthogonal matching pursuit (OMP) and modified sub-space pursuit (MSP)^[28]. The channel is estimated by sending a pilot matrix X_p in the DD domain with i.i.d Gaussian random sequences as pilots. The system model is rewritten as

$$\mathbf{y}_p = X_p \mathbf{h} + \mathbf{v}, \quad (59)$$

such that $X_p \in \mathbb{C}^{MN \times MN}$ and $\mathbf{h} \in \mathbb{C}^{MN \times 1}$ have P non-zero elements. The channel estimation problem as a sparse signal recovery problem is given by

$$\min \|\mathbf{h}\|_0 \quad \text{s.t.} \quad \mathbf{y}_p = X_p \mathbf{h} + \mathbf{v}. \quad (60)$$

OMP algorithm is used when the knowledge of the number of paths P is available. Initialize $\mathbf{h}^0 = 0$, $S^0 = \emptyset$, and $\mathbf{r}^0 = \mathbf{y}_p$. The following operations are performed in the i th iteration. The indices of the highest correlated columns are obtained as $T^i = \arg \max_j |\mathbf{X}_p^H \mathbf{r}^{i-1}|$, and the support is updated as $S^i = S^{i-1} \cup T^i$. The non-zero values corresponding to the support are $\mathbf{h}_{S^i} = (\mathbf{X}_p^{S^i})^\dagger \mathbf{y}_p$, where $(\cdot)^\dagger$ is the pseudo-inverse operator.

Finally, the residue is updated as $\mathbf{r}^i = \mathbf{y}_p - \mathbf{X}_p^{S^i} \mathbf{h}_{S^i}$. Stop the iterations when $\|\mathbf{r}^i\|_2$ is less than a threshold ϵ and obtain

the estimate as $\hat{\mathbf{h}}_{S^i} = (\mathbf{X}_p^{S^i})^\dagger \mathbf{y}_p$ and $\hat{\mathbf{h}}_{\bar{S}^i} = 0$.

When the knowledge of the number of channel paths P is not known, the subspace pursuit algorithm is modified to estimate the channel and the corresponding support using Algorithm 1. \mathbf{y}_p, X_p and ϵ are given as input and $\hat{\mathbf{h}}$ is obtained as output.

Algorithm 1. MSP based channel estimation^[28]

Inputs: \mathbf{y}, X, ϵ

Initialize: $i = 1, \mathbf{r}_1 = \mathbf{y}$

while $(\|\mathbf{r}_i\|_2 - \|\mathbf{r}_{i-1}\|_2 > \epsilon)$ **do**

Initialize: $t = 0, \mathbf{h}_i^0 = 0, S_i^0 = \{l_1^0, \dots, l_i^0\}$ are indices of i max. entries of $|\mathbf{X}^H \mathbf{y}|$, $\mathbf{b}_i^0 = \mathbf{X}_{S_i^0}^\dagger \mathbf{y}, \mathbf{r}_i^0 = \mathbf{y} - \mathbf{X}_{S_i^0} \mathbf{b}_i^0$

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

$t = t + 1$

$\tilde{S}_i^t = S_i^t \cup \Theta_i^t$, where Θ_i^t is a set of i indices corresponding to the i max. entries of $|\mathbf{X}^H \mathbf{r}_i^{t-1}|$

$\mathbf{u}_i^t = \mathbf{X}_{\tilde{S}_i^t}^\dagger \mathbf{y}$

$S_i^t = \{l_1^t, \dots, l_i^t\}$ are i entries from \tilde{S}_i^t which leads to i max. entries of $|\mathbf{u}_i^t|$

$\mathbf{b}_i^t = \mathbf{X}_{S_i^t}^\dagger \mathbf{y}$

$\mathbf{r}_i^t = \mathbf{y} - \mathbf{X}_{S_i^t} \mathbf{b}_i^t$

$\mathbf{r}_i = \mathbf{r}_i^{t_{\max}}$

$S_i = S_i^{t_{\max}}$

$i = i + 1$

Output: Estimated channel is $\hat{\mathbf{h}}_{S_i} = \mathbf{X}_{S_i}^\dagger \mathbf{y}$ and $\hat{\mathbf{h}}_{\bar{S}_i} = 0$

4.2 Embedded Pilot Approach

In this approach, instead of allocating an entire OTFS frame for pilot transmission, pilot symbols are transmitted in the same frame as data symbols with guard symbols around to prevent interference between pilot and data symbols. If no guard symbols are provided, then more sophisticated algorithms may be needed to handle the interference. In the following, we present two estimation algorithms for the embedded pilot approach. The first algorithm is applicable when guard symbols are provided. The second algorithm, based on sparse Bayesian learning, is applicable for embedded frames without guard symbols.

4.2.1 Embedded Pilot Based Estimation

In the embedded pilot based estimation^[27], let (k_p, l_p) be the pilot location such that $0 \leq k_p \leq N - 1$ and $0 \leq l_p \leq M - 1$. Define $\alpha = \max \{\alpha_i\}$ corresponding to the largest delay and $\beta = \max \{\beta_i\}$ corresponding to the largest Doppler. More

precisely, it is better to choose (k_p, l_p) such that $0 \leq l_p - \alpha \leq l_p \leq l_p + \alpha \leq M - 1$, and $0 \leq k_p - 2\beta \leq k_p \leq k_p + 2\beta \leq N - 1$.

The pilot, guard, and data symbols in an OTFS frame are arranged as follows (see Fig. 13 for an example):

$$x[k, l] = \begin{cases} x_p & k = k_p, l = l_p \\ 0 & k_p - 2\beta \leq k \leq k_p + 2\beta, l_p - \alpha \leq l \leq l_p + \alpha \\ x_d[k, l] & \text{otherwise.} \end{cases} \quad (61)$$

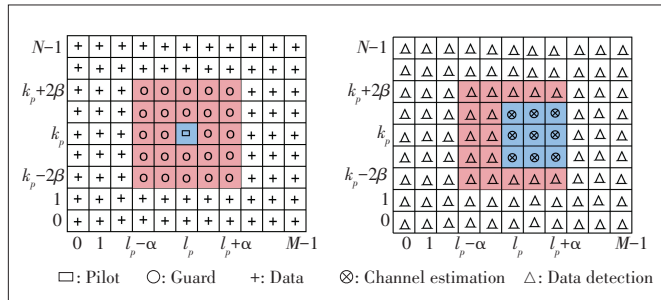
At the receiver, the subgrid in $y[k, l]$ used for channel estimation is given by $[k_p - \beta \leq k \leq k_p + \beta, l_p \leq l \leq l_p + \alpha]$. Within this subgrid, if $|y[k, l]| \geq \mathcal{T}$ for a detection threshold $\mathcal{T} > 0$, then $\hat{h}[k - k_p, l - l_p] = y[k, l]/x_p$ and $\hat{h}[k - k_p, l - l_p] = 0$ otherwise. If a path exists, then it must be seen in the received frame as a scaled version of the pilot plus Gaussian noise. It has been shown that choosing $\mathcal{T} = 3\sigma$ gives good estimation where σ^2 is the noise variance. An extension to this method has also been proposed in Ref. [27] considering the fractional Doppler scenario. A method to select a threshold value based on the receiver operating characteristics (ROC) curve has been demonstrated in Ref. [32].

4.2.2 Sparse Bayesian Learning Based Estimation

In this method, the problem of channel estimation is converted to a problem of sparse signal recovery by exploiting the sparsity of the channel in the DD domain^[29]. This method does not require guard symbols and uses pilot SNR to be the same as data SNR. This approach considers the case of fractional Dopplers as well. The structure of the OTFS frame is given by (see Fig. 14 for an example)

$$x[k, l] = \begin{cases} x_p, k_p - 2\beta - Q \leq k \leq k_p + 2\beta + Q, l_p - \alpha \leq l \leq l_p + \alpha \\ x_d[k, l], \text{ otherwise} \end{cases}, \quad (62)$$

where k_p and l_p are chosen to be $N/2$ and $M/2$, respectively, and Q is a parameter that approximates the effect of Doppler. Results in Ref. [27] show that the channel approximation is



▲ Figure 13. Transmit and receive symbol pattern for embedded pilot based channel estimation^[27]

good when $Q = 5$. Further, l_r is a parameter that obtains a trade-off between the error performance and pilot overhead. At the receiver, $y[k, l]$, $k \in [k_p - \beta, k_p + \beta]$, $l \in [l_p, l_p + l_r]$ are used for channel estimation. The system model for $L = (2\beta + 2Q + 1) \times (\alpha + 1)$ pilot symbols $(x_p[k, l])$ is modified as

$$\mathbf{y} = (\mathbf{X} \odot \mathbf{B}) \mathbf{h} + \mathbf{v} = \mathbf{\Phi} \mathbf{h} + \mathbf{v}, \quad (63)$$

where $\mathbf{X} \in \mathcal{C}^{MN \times L}$, $\mathbf{h} \in \mathcal{C}^{P \times 1}$, and $\mathbf{v} \in \mathcal{C}^{MN \times 1}$. \mathbf{B} is the phase compensation matrix which is a block diagonal matrix with the conjugate of the phase terms on the diagonal, and \odot is Hadamard product operator. The noise in Eq. (63) is assumed to be Gaussian with zero mean and variance $1/\nu_0$, $\Pr(\mathbf{v}|\nu_0) = \mathcal{N}(\mathbf{v}|\mathbf{0}, \nu_0^{-1} \mathbf{I})$. The precision parameter ν_0 is assumed to be Gamma distributed with parameters a and b . With this assumption, the distribution of the received vector is given as $\Pr(\mathbf{y}|\mathbf{\Phi} \mathbf{h}, \nu_0^{-1} \mathbf{I})$. By modelling the conjugate prior as per the Bayesian estimation framework, we get

$$\Pr(\mathbf{h}|\nu) = \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}), \quad (64)$$

$$\Pr(\nu; \lambda) = \prod_{i=1}^L \Gamma\left(\nu_i | 1, \frac{\lambda}{2}\right), \quad (65)$$

where $\mathbf{\Lambda} = \text{diag}(\nu)$ is the covariance matrix and ν_i is the variance of h_i . All the h_i 's are identical with a Laplace sparse prior distribution

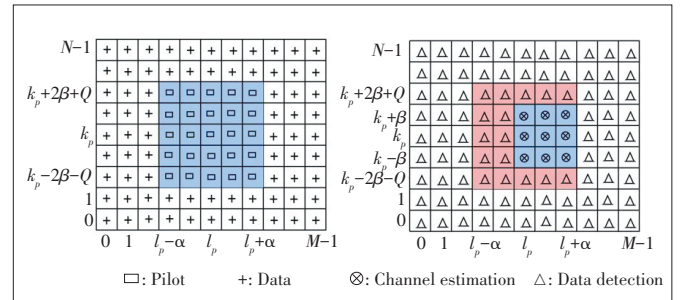
$$\Pr(h_i|\lambda) = \text{Laplace}\left(0, \frac{1}{\sqrt{\lambda}}\right), \quad i = 1, 2, \dots, L. \quad (66)$$

The joint probability distribution of this model is written as

$$\Pr(\mathbf{y}, \mathbf{h}, \nu, \nu_0) = \Pr(\mathbf{y}|\nu) \Pr(\mathbf{h}|\nu) \Pr(\nu) \Pr(\nu_0), \quad (67)$$

and

$$\Pr(\mathbf{h}|\mathbf{y}, \nu, \nu_0) = \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (68)$$



▲ Figure 14. Transmit and receive symbol pattern for sparse Bayesian learning based channel estimation^[29]

where $\boldsymbol{\mu} = \mathbf{v}_0 \boldsymbol{\Sigma} \boldsymbol{\Phi}^T \mathbf{y}$ and $\boldsymbol{\Sigma} = (\mathbf{v}_0 \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \boldsymbol{\Lambda}^{-1})^{-1}$. The values of \mathbf{v} and \mathbf{v}_0 are obtained by solving expectation maximization (EM) algorithm as

$$(\hat{\mathbf{v}}, \hat{\mathbf{v}}_0) = \arg \max_{\mathbf{v}, \mathbf{v}_0} E \{ \ln \Pr(\mathbf{y}, \mathbf{h}, \mathbf{v}, \mathbf{v}_0) \}, \quad (69)$$

which gives

$$\hat{v}_i = \frac{\sqrt{1 + 4\lambda(\Sigma_{ii} + \mu_i^2)} - 1}{2\lambda}, \quad i = 1, 2, \dots, L, \quad (70)$$

$$\hat{\mathbf{v}}_0 = \frac{2a - 2 + MN}{2b + \mathbf{v}_0^{-1} \sum_{i=1}^L (1 - v_i^{-1} \Sigma_{ii}) + \|\mathbf{y} - \boldsymbol{\Phi} \boldsymbol{\mu}\|_2^2}. \quad (71)$$

After obtaining $\boldsymbol{\mu}$, the first $P(2Q + 1)$ largest values in $\boldsymbol{\mu}$ are selected as \mathbf{h} .

4.3 Superimposed Pilot Approach

In this approach, pilot symbols are superimposed on data symbols in an OTFS frame. For example, each bin in the DD grid has a data symbol and a pilot symbol superimposed on it as shown in Fig. 15^[35]. Fig. 16 shows another example where all bins have data symbols and only one among them has a superimposed pilot symbol^[36].

4.3.1 Superimposed Pilots Based Estimation in Ref. [35]

In this method of estimation, low-powered pilot symbols are superimposed on the data symbols in the DD grid (Fig. 15). The mutual interference between the pilot and the data symbols is handled by optimum selection of pilot SNR and by adopting an iterative approach that iterates between channel

estimation and data detection. The system model considered is

$$\mathbf{y} = \mathbf{X} \mathbf{h} + \mathbf{v}, \quad (72)$$

where $\mathbf{X} \in \mathcal{C}^{MN \times P}$, $\mathbf{h} \in \mathcal{C}^{P \times 1}$, and $\mathbf{v} \in \mathcal{C}^{MN \times 1}$. When the pilot symbols are superimposed on the data symbols, the system model is given by

$$\mathbf{y} = \mathbf{X}_p \mathbf{h} + \underbrace{\mathbf{X}_d \mathbf{h} + \mathbf{v}}_{\mathbf{v}_d} = \mathbf{X}_p \mathbf{h} + \mathbf{v}_d, \quad (73)$$

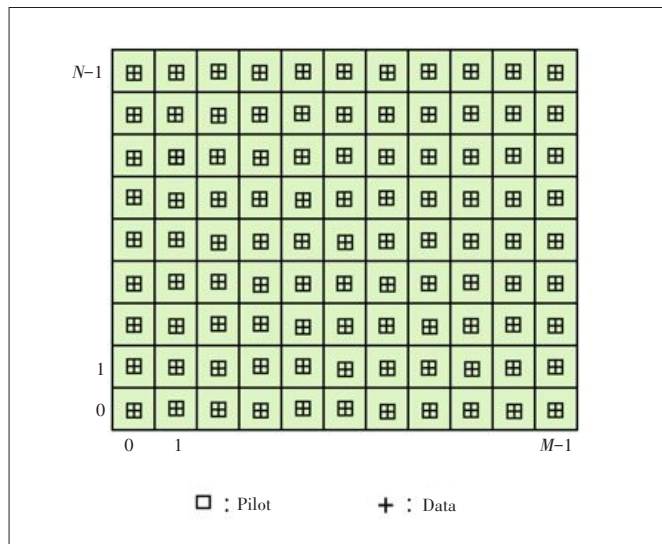
where \mathbf{X}_p corresponds to pilot symbols, \mathbf{X}_d corresponds to data symbols, and \mathbf{v}_d is the noise plus interference term having mean $\boldsymbol{\mu}_{\mathbf{v}_d} = \mathbf{0}_{MN \times 1}$ and covariance

$$\mathbf{C}_{\mathbf{v}_d} = \left(\left(\sum_{i=1}^P \sigma_{h_i}^2 \right) \sigma_d^2 + \sigma_v^2 \right) \mathbf{I}_{MN}, \quad (74)$$

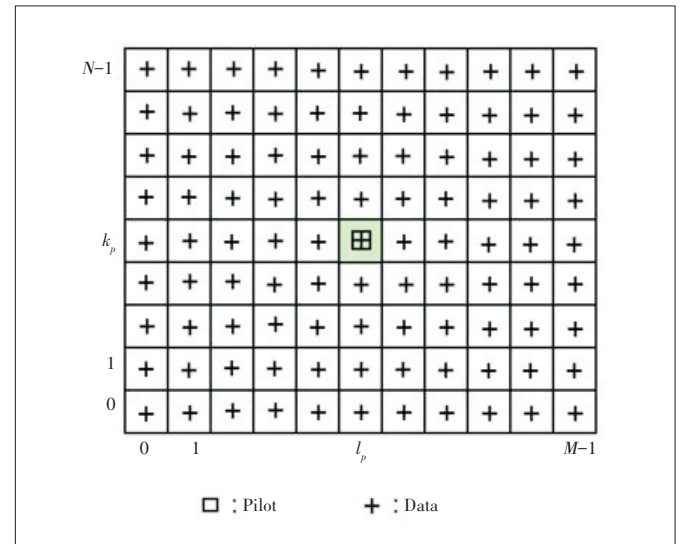
where $\sigma_d^2 = E \left[|x_d(k, l)|^2 \right]$, $\sigma_{h_i}^2$ is the variance of i th channel coefficient, and σ_v^2 is the noise variance. Under these assumptions, the MMSE estimate of the channel using superimposed pilots is given by

$$\hat{\mathbf{h}}_{sp} = \left(\mathbf{X}_p^H \mathbf{C}_{\mathbf{v}_d}^{-1} \mathbf{X}_p + \mathbf{C}_h^{-1} \right)^{-1} \mathbf{X}_p^H \mathbf{C}_{\mathbf{v}_d}^{-1} \mathbf{y}, \quad (75)$$

where $\mathbf{C}_h = \text{diag}(\sigma_{h_1}^2, \sigma_{h_2}^2, \dots, \sigma_{h_P}^2)$. Message passing algorithms along with the MMSE estimated channel are used to detect the data symbols $\hat{\mathbf{X}}_d^{(0)}$. This is used as initialization of \mathbf{X}_d for an iterative channel estimation algorithm which is more robust to the interference than the MMSE estimate. With this, the system model in Eq. (73) can be rewritten as



▲ Figure 15. Transmit symbol pattern in superimposed pilot scheme in Ref. [35]



▲ Figure 16. Transmit symbol pattern in superimposed pilot scheme in Ref. [36]

$$\mathbf{y} = (\mathbf{X}_p + \hat{\mathbf{X}}_d^{(0)})\mathbf{h} + (\mathbf{X}_d - \hat{\mathbf{X}}_d^{(0)})\mathbf{h} + \mathbf{v} = \mathbf{X}_{x_p, \hat{x}_d}^{(0)}\mathbf{h} + \xi_w^{(0)}. \quad (76)$$

Here, the data-aided pilot is given by $\mathbf{X}_{x_p, \hat{x}_d}^{(0)}$ and the interference plus noise term is given by $\xi_w^{(0)}$, where

$$\mathbf{X}_{x_p, \hat{x}_d}^{(0)} = \mathbf{X}_p + \hat{\mathbf{X}}_d^{(0)}, \quad \xi_w^{(0)} = (\mathbf{X}_d - \hat{\mathbf{X}}_d^{(0)})\mathbf{h} + \mathbf{v}. \quad (77)$$

For the system model in Eq. (76), the MMSE channel estimate in the n -th iteration $\hat{\mathbf{h}}^{(n)}$ is obtained using Eq. (75) as

$$\hat{\mathbf{h}}^{(n)} = \left(\left(\mathbf{X}_{x_p, \hat{x}_d}^{(n-1)} \right)^H \left(\mathbf{C}_{\xi_w}^{(n)} \right)^{-1} \mathbf{X}_{x_p, \hat{x}_d}^{(n-1)} + \mathbf{C}_h^{-1} \right)^{-1} \cdot \left(\mathbf{X}_{x_p, \hat{x}_d}^{(n-1)} \right)^H \left(\mathbf{C}_{\xi_w}^{(n)} \right)^{-1} \mathbf{y}, \quad (78)$$

where $\mathbf{C}_{\xi_w}^{(n)} = E[\xi_w^{(n)} \xi_w^{(n)H}]$ is given by

$$\mathbf{C}_{\xi_w}^{(n)} = 2 \left(\sum_{i=1}^P \sigma_{h_i}^2 \right) \sigma_d^2 \mathbf{I}_{MN} + \sigma_v^2 \mathbf{I}_{MN}. \quad (79)$$

Thus, the expression in Eq. (78) gives the channel estimation after n iterations.

4.3.2 Superimposed Pilot Based Estimation in Ref. [36]

A data-aided channel estimation that uses the whole OTFS frame for data transmission scheme with one pilot symbol superimposed on a data symbol in the (k_p, l_p) location of the grid is reported in Ref. [36]. The allocation of symbols in the DD grid is given by

$$x[k, l] = \begin{cases} x_p + x_d[k, l], & k = k_p, l = l_p \\ x_d[k, l], & \text{otherwise} \end{cases}. \quad (80)$$

The energy of the pilot symbol is denoted by $E_p = |x_p|^2$ and the average energy of the data symbol is denoted by $E_d = E[|x[k, l]|^2]$. With this frame structure, the received signal in the DD domain is given by

$$y[k, l] = x_p h_w \left[(k - k_p)_N, (l - l_p)_M \right] + \mathcal{I}_{k,l} + v[k, l], \quad (81)$$

where $\mathcal{I}_{k,l}$ is the interference due to data symbols, given by

$$\mathcal{I}_{k,l} = \sum_{k'=k-\beta}^{k+\beta} \sum_{l'=l-\alpha}^l x[k', l'] h_w \left[(k - k')_N, (l - l')_M \right]. \quad (82)$$

The channel is initially estimated using a modified threshold which incorporates the effect due to $\mathcal{I}_{k,l}$. Using this estimated channel, the data symbols are detected by a sum-product algorithm. The interference term can be simplified as

$$\mathcal{I}_{k,l} = \sum_{i \in \mathcal{Q}_{k,l}} h_i x \left[(k - \beta_i)_N, (l - \alpha_i)_M \right] e^{-j2\pi \frac{\alpha_i \beta_i}{MN}}, \quad (83)$$

where $\mathcal{Q}_{k,l}$ is the set of indices of the data symbols that contribute to $y[k, l]$ such that $|\mathcal{Q}_{k,l}| = P$. From this, the energy of the interference part is obtained as

$$E\{|\mathcal{I}_{k,l}|^2\} = \sum_{i \in \mathcal{Q}_{k,l}} E\{|h_i|^2\} E_s. \quad (84)$$

When $\sum_{i=1}^P E\{|h_i|^2\} = 1$, we get $E\{|\mathcal{I}_{k,l}|^2\} = E_s$. Using the interference energy, the threshold is obtained as

$$\gamma = 3 \left(\sqrt{N_0 + E_s} \right). \quad (85)$$

If $|y[k, l]| \geq \gamma$, the estimates of the channel coefficients can be obtained as

$$\hat{h}_w \left[(k - k_p)_N, (l - l_p)_M \right] = \frac{y[k, l]}{x_p}. \quad (86)$$

The data symbols are detected using the estimated \hat{h}_w by the decision rule:

$$\hat{x}[k, l] = \arg \min_{x[k, l] \in \mathbb{A}} \Pr(x[k, l] | y). \quad (87)$$

The marginal PDF is obtained by using a sum-product algorithm. After detecting the data symbols, the interference caused by them is cancelled and the resultant symbols are given by

$$\tilde{y}[k, l] = y[k, l] - \sum_{k'} \sum_{l'} \hat{x}[k', l'] \hat{h}_w \left[(k - k')_N, (l - l')_M \right]. \quad (88)$$

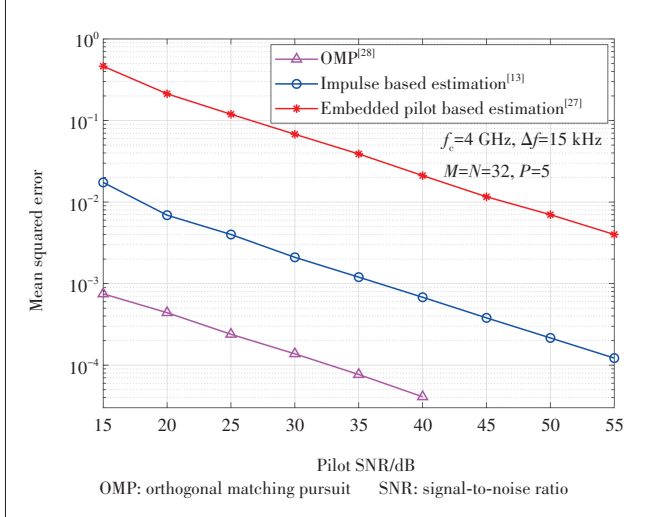
These symbols would contain only the pilot information if the interference was completely cancelled. However, due to imperfect estimates, this method of channel estimation followed by data detection and interference cancellation is performed iteratively to obtain better estimates.

4.4 Performance Results

In this subsection, we present the performance of some of the channel estimation methods presented in the previous subsections. The OTFS system considered has a carrier frequency of 4 GHz, a subcarrier spacing of 15 kHz, and a DD grid with $M = N = 32$. A multipath channel with $P = 5$ paths, exponential power delay profile, and delay-Doppler profile shown in Table 3 are considered. Fig. 17 shows the mean squared error (MSE) performance as a function of pilot SNR for 1) OMP, 2) impulse based estimation, and 3) embedded pilot based estimation. A pilot frame with i.i.d Gaussian random sequences occupying the entire DD grid is used for

▼ Table 3. Delay-Doppler profile for Figures 17 and 18

Path (i)	1	2	3	4	5
Delay τ_i	$\frac{1}{M\Delta f}$	$\frac{2}{M\Delta f}$	$\frac{3}{M\Delta f}$	$\frac{4}{M\Delta f}$	$\frac{5}{M\Delta f}$
Doppler ν_i	0	$\frac{1}{NT}$	$\frac{2}{NT}$	$\frac{3}{NT}$	$\frac{4}{NT}$



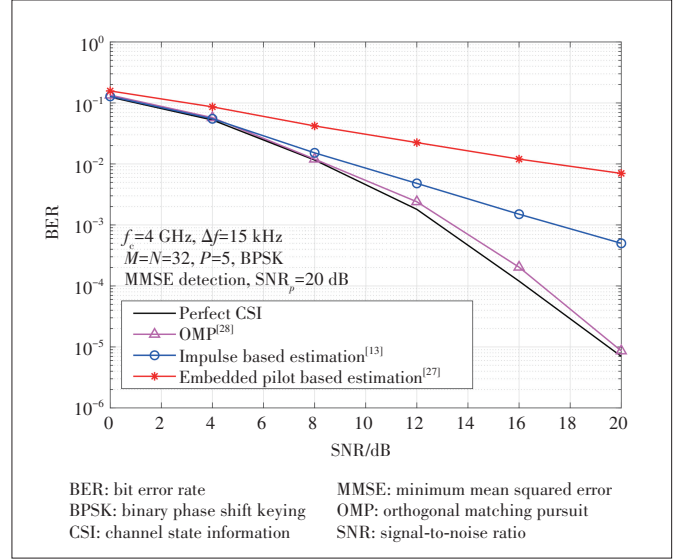
▲ Figure 17. Mean squared error performance of delay-Doppler (DD) channel estimation methods

OMP based channel estimation. A pilot frame with an impulse at location $(k_p, l_p) = (15, 15)$ and zeros elsewhere is considered for impulse based channel estimation. An embedded frame is used for embedded pilot based channel estimation, with impulse as a pilot at location $(k_p, l_p) = (15, 15)$, guard symbols in the locations $7 \leq k \leq 23, 10 \leq l \leq 20$, and data symbols elsewhere.

It can be seen from Fig. 17 that the OMP algorithm gives the channel estimate with small MSE, which is in the order of 10^{-3} for a pilot SNR of 15 dB. Impulse based channel estimation scheme is simpler but its MSE is high. The MSE of the embedded pilot based channel estimation is also high. Fig. 18 shows the BER performance of these estimation methods using MMSE detection as a function of data SNR for a pilot SNR of 20 dB. BPSK modulation is used. The BER performance of OMP based channel estimation is close to the performance using perfect channel knowledge. Impulse based estimation performance is inferior compared to OMP performance but is superior compared to that of embedded pilot based estimation. Embedded pilot based estimation can be used with high pilot SNRs to achieve good BER performance and higher throughput.

5 Conclusions

OTFS modulation is regarded as an attractive physical layer waveform for future wireless systems. It has demonstrated ro-



▲ Figure 18. BER performance with estimated delay-Doppler (DD) channel at pilot SNR of 20 dB

bust performance in high-Doppler scenarios which are expected in emerging standards. In this paper, we presented an overview of the state-of-the-art approaches in OTFS signal detection and DD channel estimation. We classified the detection approaches as low-complexity linear approach, approximate MAP approach, and DNN approach. Low complexities possible in the linear approach due to the structure of the channel matrix make it attractive for practical implementations. The iterative MAP approach (e.g., message passing) is known for its good performance at low complexities. DNN approach is emerging with good promise particularly when there are model deviations that are typical in practice. In the DD channel estimation space, we highlighted approaches based on exclusive pilot frames, embedded pilot frames, and superimposed pilot frames. More research in OTFS transceiver designs using the DNN approach can be pursued as future work.

References

- [1] HADANI R, RAKIB S, TSATSANIS M, et al. Orthogonal time frequency space modulation [C]/IEEE Wireless Communications and Networking Conference (WCNC). San Francisco, USA: IEEE, 2017: 1 - 6. DOI: 10.1109/WCNC.2017.7925924
- [2] HADANI R, MONK A. OTFS: a new generation of modulation addressing the challenges of 5G [EB/OL]. (2018-02-07) [2021-09-25]. <https://arxiv.org/ftp/arxiv/papers/1802/1802.02623.pdf>
- [3] HADANI R, RAKIB S, TSATSANIS M, et al. Orthogonal time frequency space modulation [EB/OL]. (2018-08-01) [2021-09-25]. <https://arxiv.org/abs/1808.00519v1>
- [4] HADANI R, RAKIB S, MOLISCH A F, et al. Orthogonal time frequency

- space (OTFS) modulation for millimeter-wave communications systems [C]/2017 IEEE MTT-S International Microwave Symposium (IMS). Honolulu, USA: IEEE, 2017: 681 – 683. DOI: 10.1109/MWSYM.2017.8058662
- [5] WIFFEN F, SAYER L, BOCUS M Z, et al. Comparison of OTFS and OFDM in ray launched sub-6 GHz and mmWave line-of-sight mobility channels [C]/IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). Bologna, Italy: IEEE, 2018: 73 – 79. DOI: 10.1109/PIMRC.2018.8580850
- [6] RAMACHANDRAN M K, SURABHI G D, CHOCKALINGAM A. OTFS: a new modulation scheme for high-mobility use cases [J]. Journal of the Indian institute of science, 2020, 100(2): 315 – 336. DOI: 10.1007/s41745-020-00167-4
- [7] MOHAMMED S K. Derivation of OTFS modulation from first principles [J]. IEEE transactions on vehicular technology, 2021, 70(8): 7619 – 7636. DOI: 10.1109/TVT.2021.3069913
- [8] MURALI K R, CHOCKALINGAM A. On OTFS modulation for high-Doppler fading channels [C]/Information Theory and Applications Workshop (ITA). San Diego, USA: IEEE, 2018: 1 – 10. DOI: 10.1109/ITA.2018.8503182
- [9] SURABHI G D, AUGUSTINE R M, CHOCKALINGAM A. On the diversity of uncoded OTFS modulation in doubly-dispersive channels [J]. IEEE transactions on wireless communications, 2019, 18(6): 3049 – 3063. DOI: 10.1109/TWC.2019.2909205
- [10] GUNTURU A, GODALA A R, SAHOO A K, et al. Performance analysis of OTFS waveform for 5G NR mmWave communication system [C]/IEEE Wireless Communications and Networking Conference (WCNC). Nanjing, China: IEEE, 2021: 1 – 6. DOI: 10.1109/WCNC49053.2021.9417346
- [11] RAVITEJA P, PHAN K T, HONG Y, et al. Interference cancellation and iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, 2018, 17(10): 6501 – 6515. DOI: 10.1109/TWC.2018.2860011
- [12] THAJ T, VITERBO E. Low complexity iterative rake detector for orthogonal time frequency space modulation [C]/IEEE Wireless Communications and Networking Conference (WCNC). Seoul, Korea (South): IEEE, 2020: 1 – 6. DOI: 10.1109/WCNC45663.2020.9120526
- [13] KOLLENGODE RAMACHANDRAN M, CHOCKALINGAM A. MIMO-OTFS in high-Doppler fading channels: signal detection and channel estimation [C]/IEEE Global Communications Conference (GLOBECOM). Abu Dhabi, United Arab Emirates: IEEE, 2018: 206 – 212. DOI: 10.1109/GLOBECOM.2018.8647394
- [14] LI L J, LIANG Y, FAN P Z, et al. Low complexity detection algorithms for OTFS under rapidly time-varying channel [C]/IEEE 89th Vehicular Technology Conference (VTC2019-Spring). Kuala Lumpur, Malaysia: IEEE, 2019: 1 – 5. DOI: 10.1109/VTCSpring.2019.8746420
- [15] ZHANG H J, ZHANG T T. A low-complexity message passing detector for OTFS modulation with probability clipping [J]. IEEE wireless communications letters, 2021, 10(6): 1271 – 1275. DOI: 10.1109/LWC.2021.3063904
- [16] LI S Y, YUAN W J, WEI Z Q, et al. Hybrid MAP and PIC detection for OTFS modulation [J]. IEEE transactions on vehicular technology, 2021, 70(7): 7193 – 7198. DOI: 10.1109/tvt.2021.3083181
- [17] XIANG L P, LIU Y S, YANG L L, et al. Gaussian approximate message passing detection of orthogonal time frequency space modulation [J]. IEEE transactions on vehicular technology, 2021, 70(10): 10999 – 11004. DOI: 10.1109/TVT.2021.3102673
- [18] NAIKOTI A, CHOCKALINGAM A. Low-complexity delay-Doppler symbol DNN for OTFS signal detection [C]/IEEE 93rd Vehicular Technology Conference (VTC2021-Spring). Helsinki, Finland: IEEE, 2021: 1 – 6. DOI: 10.1109/VTC2021-Spring51267.2021.9448630
- [19] ENKU Y K, BAI B M, WAN F, et al. Two-dimensional convolutional neural network-based signal detection for OTFS systems [J]. IEEE wireless communications letters, 2021, 10(11): 2514 – 2518. DOI: 10.1109/LWC.2021.3106039
- [20] YUAN W J, WEI Z Q, YUAN J H, et al. A simple variational Bayes detector for orthogonal time frequency space (OTFS) modulation [J]. IEEE transactions on vehicular technology, 2020, 69(7): 7976 – 7980. DOI: 10.1109/TVT.2020.2991443
- [21] SURABHI G D, CHOCKALINGAM A. Low-complexity linear equalization for OTFS modulation [J]. IEEE communications letters, 2020, 24(2): 330 – 334. DOI: 10.1109/LCOMM.2019.2956709
- [22] SURABHI G D, CHOCKALINGAM A. Low-complexity linear equalization for 2x2 MIMO-OTFS signals [C]/IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC). Atlanta, USA: IEEE, 2020: 1 – 5. DOI: 10.1109/SPAWC48557.2020.9154292
- [23] TIWARI S, DAS S S, RANGAMGARI V. Low complexity LMMSE Receiver for OTFS [J]. IEEE communications letters, 2019, 23(12): 2205 – 2209. DOI: 10.1109/LCOMM.2019.2945564
- [24] LONG F, NIU K, DONG C, et al. Low complexity iterative LMMSE-PIC equalizer for OTFS [C]/IEEE International Conference on Communications (ICC). Shanghai, China: IEEE, 2019: 1 – 6. DOI: 10.1109/ICC.2019.8761635
- [25] JING L Y, WANG H, HE C B, et al. Two dimensional adaptive multichannel decision feedback equalization for OTFS system [J]. IEEE communications letters, 2021, 25(3): 840 – 844. DOI: 10.1109/LCOMM.2020.3039982
- [26] PANDEY B C, MOHAMMED S K, RAVITEJA P, et al. Low complexity precoding and detection in multi-user massive MIMO OTFS downlink [J]. IEEE transactions on vehicular technology, 2021, 70(5): 4389 – 4405. DOI: 10.1109/TVT.2021.3061694
- [27] RAVITEJA P, PHAN K T, HONG Y. Embedded pilot-aided channel estimation for OTFS in delay – Doppler channels [J]. IEEE transactions on vehicular technology, 2019, 68(5): 4906 – 4917. DOI: 10.1109/TVT.2019.2906357
- [28] RASHEED O K, SURABHI G D, CHOCKALINGAM A. Sparse delay-Doppler channel estimation in rapidly time-varying channels for multiuser OTFS on the uplink [C]/IEEE 91st Vehicular Technology Conference (VTC2020-Spring). Antwerp, Belgium: IEEE, 2020: 1 – 5. DOI: 10.1109/VTC2020-Spring48590.2020.9128497
- [29] ZHAO L, GAO W J, GUO W B. Sparse Bayesian learning of delay-Doppler channel for OTFS system [J]. IEEE communications letters, 2020, 24(12): 2766 – 2769. DOI: 10.1109/LCOMM.2020.3021120
- [30] SRIVASTAVA S, SINGH R K, JAGANNATHAM A K, et al. Bayesian learning aided sparse channel estimation for orthogonal time frequency space modulated systems [J]. IEEE transactions on vehicular technology, 2020, 69(8): 8343 – 8348. DOI: 10.1109/TCOMM.2021.3123354
- [31] SHEN W Q, DAI L L, HAN S F, et al. Channel estimation for orthogonal time frequency space (OTFS) massive MIMO [C]/IEEE International Conference on Communications (ICC). Shanghai, China: IEEE, 2019: 1 – 6. DOI: 10.1109/ICC.2019.8761362
- [32] ZHAO H, KANG Z Q, WANG H. A novel channel estimation scheme for OTFS [C]/IEEE 20th International Conference on Communication Technology (ICCT). Nanning, China: IEEE, 2020: 12 – 16. DOI: 10.1109/ICCT50939.2020.9295699
- [33] BOMFIN R, CHAFII M, NIMR A, et al. Channel estimation for MIMO space time coded OTFS under doubly selective channels [C]/IEEE International Conference on Communications Workshops (ICC Workshops). Montreal, Canada: IEEE, 2021: 1 – 6. DOI: 10.1109/ICCWorkshops50388.2021.9473618
- [34] LIU F, YUAN Z D, GUO Q H, et al. Message passing based structured sparse signal recovery for estimation of OTFS channels with fractional Doppler shifts [J]. IEEE transactions on wireless communications, 2021, early access. DOI: 10.1109/TWC.2021.3087501
- [35] MISHRA H B, SINGH P, PRASAD A K, et al. OTFS channel estimation and data detection designs with superimposed pilots [J]. IEEE transactions on wireless communications, 2021, early access. DOI: 10.1109/TWC.2021.3110659
- [36] YUAN W J, LI S Y, WEI Z Q, et al. Data-aided channel estimation for OTFS systems with a superimposed pilot and data transmission scheme [J]. IEEE wireless communications letters, 2021, 10(9): 1954 – 1958. DOI: 10.1109/LWC.2021.3088836
- [37] YUAN Z D, LIU F, YUAN W J, et al. Iterative detection for orthogonal time frequency space modulation with unitary approximate message passing [EB/OL]. (2021-02-16)[2021-09-25]. <https://arxiv.org/abs/2008.06688v3>
- [38] LI L, WEI H, HUANG Y, et al. A simple two-stage equalizer with simplified orthogonal time frequency space modulation over rapidly time-varying channels [EB/OL]. (2017-09-08)[2021-09-25]. <https://arxiv.org/abs/1709.02505>
- [39] ZEMEN T, HOFER M, LOESCHENBRAND D. Low-complexity equalization for orthogonal time and frequency signaling (OTFS) [EB/OL]. (2017-10-26)[2021-09-25]. <https://arxiv.org/pdf/1710.09916v1.pdf>
- [40] THAJ T, VITERBO E. Low complexity iterative rake decision feedback equalizer for zero-padded OTFS systems [J]. IEEE transactions on vehicular technology, 2020, 69(12): 15606 – 15622. DOI: 10.1109/TVT.2020.3044276

- [41] LI S Y, YUAN W J, WEI Z Q, et al. Cross domain iterative detection for orthogonal time frequency space modulation [EB/OL]. (2021-01-11)[2021-09-25]. <https://arxiv.org/abs/2101.03822v1>
- [42] XU W J, ZOU T T, GAO H, et al. Low-complexity linear equalization for OTFS systems with rectangular waveforms [EB/OL]. (2019-11-19)[2021-09-25]. <https://arxiv.org/abs/1911.08133v1>
- [43] LIU Y S, ZHANG S, GAO F F, et al. Uplink-aided high mobility downlink channel estimation over massive MIMO-OTFS system [J]. *IEEE journal on selected areas in communications*, 2020, 38(9): 1994 – 2009. DOI: 10.1109/JSAC.2020.3000884
- [44] DAS S S, RANGAMGARI V, TIWARI S, et al. Time domain channel estimation and equalization of CP-OTFS under multiple fractional Dopplers and residual synchronization errors [J]. *IEEE Access*, 2021, 9: 10561 – 10576. DOI: 10.1109/ACCESS.2020.3046487
- [45] YAN H, WANG M. A low complexity channel estimation scheme for orthogonal time frequency space (OTFS) system with synchronization errors [C]// *IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*. Chengdu, China: IEEE, 2021: 576 – 581. DOI: 10.1109/ICCCS52626.2021.9449209
- [46] WU X D, MA S D, YANG X. Tensor-based low-complexity channel estimation for mmWave massive MIMO-OTFS systems [J]. *Journal of communications and information networks*, 2020, 5(3): 324 – 334. DOI: 10.23919/JCIN.2020.9200896
- [47] KUMAR SINGH V, FLANAGAN M F, CARDIFF B. Maximum likelihood channel path detection and MMSE channel estimation in OTFS systems [C]// *IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*. Victoria, Canada: IEEE, 2020: 1 – 5. DOI: 10.1109/VTC2020-Fall49728.2020.9348590
- [48] SHI D, WANG W J, YOU L, et al. Deterministic pilot design and channel estimation for downlink massive MIMO-OTFS systems in presence of the fractional Doppler [J]. *IEEE transactions on wireless communications*, 2021, early access. DOI: 10.1109/TWC.2021.3081164
- [49] ZHANG M C, WANG F G, YUAN X J, et al. 2D structured turbo compressed sensing for channel estimation in OTFS systems [C]// *IEEE International Conference on Communication Systems (ICCS)*. Chengdu, China: IEEE, 2018: 45 – 49. DOI: 10.1109/ICCS.2018.8689234
- [50] HASHIMOTO N, OSAWA N, YAMAZAKI K, et al. Channel estimation and equalization for CP-OFDM-based OTFS in fractional Doppler channels [C]// *IEEE International Conference on Communications Workshops (ICC Workshops)*. Montreal, Canada: IEEE, 2021: 1 – 7. DOI: 10.1109/ICCWorkshops50388.2021.9473532
- [51] QU H Y, LIU G H, ZHANG L, et al. Low-dimensional subspace estimation of continuous-Doppler-spread channel in OTFS systems [J]. *IEEE transactions on communications*, 2021, 69(7): 4717 – 4731. DOI: 10.1109/TCOMM.2021.3072744
- [52] ABDELGADER A M S, WU L N. The physical layer of the IEEE 802.11p WAVE communication standard: the specifications and challenges [C]// *World Congress on Engineering and Computer Science 2014*. San Francisco, USA: IAENG, 2014
- [53] KRUSEVAC S, RAPAJIC P, KENNEDY R A. Channel capacity estimation for MIMO systems with correlated noise [C]// *IEEE Global Telecommunications Conference*. St. Louis, USA: IEEE, 2005: 2812 – 2816. DOI: 10.1109/GLOCOM.2005.1578272

Biographies

Ashwitha NAIKOTI (ashwithan@iisc.ac.in) received the B.Tech. degree in electronics and communication engineering from the National Institute of Technology, Warangal, India in 2017. She is currently pursuing M. Tech (Research) degree with the Department of Electrical Communication Engineering, Indian Institute of Science (IISc), Bengaluru, India. She was with the Center for Development of Telematics, Bengaluru, as a Research Engineer from 2017 to 2019. Her current research interests include orthogonal time frequency space modulation and transceiver design using neural networks.

Ananthanarayanan CHOCKALINGAM received the Ph. D. degree in electrical communication engineering (ECE) from the Indian Institute of Science (IISc), Bangalore, India. He was a post-doctoral fellow and an assistant project scientist with the Department of Electrical and Computer Engineering, University of California, San Diego, USA. He was with Qualcomm, Inc., San Diego, USA as a Staff Engineer/Manager. Currently, he is a professor with the Department of ECE, IISc, Bangalore. He served as an associate editor for the *IEEE Transactions on Vehicular Technology*, an editor for the *IEEE Transactions on Wireless Communications*, and a guest editor for the *IEEE Journal on Selected Areas in Communications* and the *IEEE Journal of Selected Topics in Signal Processing*. He is an author of the book *Large MIMO Systems* published by Cambridge University Press.

Message Passing Based Detection for Orthogonal Time Frequency Space Modulation



YUAN Zhengdao¹, LIU Fei², GUO Qinghua³, WANG Zhongyong²

(1. The Open University of Henan, Zhengzhou 450000, China;

2. Zhengzhou University, Zhengzhou 450001, China;

3. University of Wollongong, Wollongong NSW 2522, Australia)

Abstract: The orthogonal time frequency space (OTFS) modulation has emerged as a promising modulation scheme for wireless communications in high-mobility scenarios. An efficient detector is of paramount importance to harvesting the time and frequency diversities promised by OTFS. Recently, some message passing based detectors have been developed by exploiting the features of the OTFS channel matrices. In this paper, we provide an overview of some recent message passing based OTFS detectors, compare their performance, and shed some light on potential research on the design of message passing based OTFS receivers.

Keywords: OTFS; detection; message passing; belief propagation; approximate message passing (AMP); unitary AMP (UAMP)

DOI: 10.12142/ZTECOM.202104004

<https://kns.cnki.net/kcms/detail/34.1294.TN.20211123.1116.002.html>, published online November 23, 2021

Manuscript received: 2021-10-10

Citation (IEEE Format): Z. D. Yuan, F. Liu, Q. H. Guo, et al., "Message passing based detection for orthogonal time frequency space modulation," *ZTE Communications*, vol. 19, no. 4, pp. 34 - 44, Dec. 2021. doi: 10.12142/ZTECOM.202104004.

1 Introduction

Recently the orthogonal time frequency space (OTFS) modulation has attracted much attention due to its capability of achieving reliable communications in high mobility scenarios^[1-6]. OTFS is a two-dimensional modulation scheme, and the information is modulated in the delay Doppler (DD) domain, which is in contrast to the time frequency (TF) domain modulation in the orthogonal frequency division multiplexing (OFDM). In OTFS, each symbol spreads over the time and frequency domains through the two dimensional (2D) inverse symplectic finite Fourier transform (SFFT), leading to both time and frequency diversities^[1-2]. It has been shown that OTFS can significantly outperform

OFDM in high mobility scenarios^[7].

To harvest the diversities promised by OTFS, the design of a powerful detector is paramount. The optimal maximum a posteriori (MAP) detector is impractical due to its complexity growing exponentially with the length of the OTFS block. Recently, significant efforts have been devoted to the design of more efficient detectors. In Ref. [8], an effective channel matrix in the DD domain was derived, based on which a low-complexity two-stage detector was proposed. The first-order Neumann series was used in Ref. [9] to approximately solve the matrix inverse problem involved in the linear minimum mean squared error (MMSE) estimation based detection. A detection scheme was developed in Ref. [10], where the MMSE equalization was used in the first iteration, followed by parallel interference cancellation with a soft-output sphere decoder in subsequent iterations. A rectangular waveform was considered in Ref. [11], where the sparsity and quasi banded structure of channel matrices without fractional Doppler shifts were exploited to reduce the detection complexity. The linear equaliz-

This work was supported by the National Natural Science Foundation of China (61901417, U1804152, 61801434) and Science and Technology Research Project of Henan Province (212102210556, 212102210566, 212400410179).

ers were extended to the multiple input and multiple output (MIMO)-OTFS systems in Ref. [12]. A cross-domain method was proposed in Ref. [13], where a conventional linear MMSE estimator is adopted for equalization in the time domain and a low-complexity symbol-by-symbol detection is utilized in the DD domain. A low complexity iterative rake decision feedback detector was proposed in Ref. [14], which extracts and coherently combines the multiple copies of the symbols (due to multipath propagation) in the DD grid using maximal ratio combining (MRC).

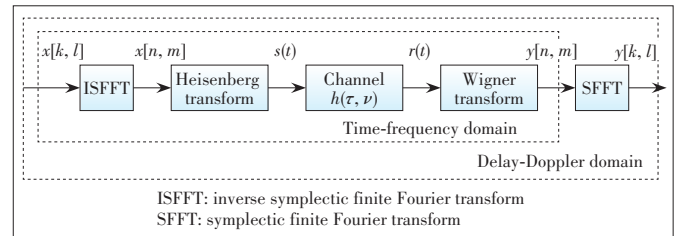
Another line of OTFS detector design is based on factor graphs and message passing techniques^[15, 23]. When the number of channel paths is small, the effective channel matrix in the DD domain is sparse, which allows efficient detection using the message passing algorithm (MPA)^[2]. An expectation propagation (EP) algorithm was proposed in Ref. [16], where EP is used for message update with Gaussian approximation. A variational Bayes (VB) based detector was proposed in Ref. [17] to achieve better convergence. Studying the matched filtering processing, the authors in Ref. [18] proposed a message passing detector, which is combined with a probability clipping solution. The detectors in Refs. [2, 17, 19] take advantage of the sparsity of the channel matrix in the DD domain, and their complexity depends on the number of nonzero elements in each row of the channel matrix, which is denoted by S . Without considering fractional Doppler shifts, S is equal to the number of channel paths. In general, a wideband system can provide sufficient delay resolution. The Doppler resolution depends on the time duration of the OTFS block. To fulfill the low latency requirement in wireless communications, the time duration of an OTFS block should be relatively small, where it is necessary to consider fractional Doppler shifts^[2, 20]. In this case, the value of S can be significantly larger than the number of channel paths. In the case of rich scattering environments, the complexity of these detectors can be a concern and the short loops in the corresponding system graph model may result in significant performance. To overcome the above issues, the design of OTFS detectors based approximate message passing (AMP)^[21 - 22] was investigated in Ref. [25]. AMP works well for independent and identically distributed (sub-) Gaussian system transfer matrix, but it suffers from performance loss or even diverges for a general system transfer matrix^[27 - 29]. Instead, the works in Refs. [25 - 26] resort to the unitary AMP (UAMP)^[27 - 29], which is a variant of AMP and was formerly called UTAMP^[27]. In UAMP, a unitary transformation of the original model is used, where the unitary matrix for transformation can be the conjugate transpose of the left singular matrix of the general system transfer matrix^[27] obtained through singular value decomposition (SVD). It is shown in Ref. [25] that UAMP is well suitable for OTFS due to the structure of block circulant matrix with circulant block (BCCB) of the DD domain channel matrices, where the 2D discrete Fourier transform is used for the unitary transformation,

leading to very efficient implementation using the 2D fast Fourier transform (FFT) algorithm. In addition, as the noise variance is normally unknown, the noise variance estimation is also incorporated into the UAMP-based detector in Ref. [25].

In this paper, we provide an overview of the message passing based detectors, provide some comparison results, and discuss potential research on the design of message passing based OTFS receivers. The notations used in this paper are as follows. Boldface lower-case and upper-case letters denote vectors and matrices, respectively. We use $(\cdot)^H$ and $(\cdot)^T$ to denote the conjugate transpose and the transpose, respectively. The superscript $*$ denotes the conjugate operation. We define $[\cdot]_M$ as the mod M operation. We use $N(x|\hat{x}, \nu_x)$ to denote the probability density function of a complex Gaussian variable with mean \hat{x} and variance ν_x . The notation $\langle f(\mathbf{x}) \rangle_{q(\mathbf{x})}$ denotes the expectation of the function $f(\mathbf{x})$ with respect to the distribution $q(\mathbf{x})$. The relation $f(x) = cg(x)$ for some positive constant c is written as $f(x) \propto g(x)$. The notation \otimes represents the Kronecker product, and $\mathbf{a} \cdot \mathbf{b}$ and $\mathbf{a} \oslash \mathbf{b}$ represent the component-wise product and the division between vectors \mathbf{a} and \mathbf{b} , respectively. We use $\mathbf{X} = \text{reshape}_M(\mathbf{x})$ to denote that the vector \mathbf{x} is reshaped as an $M \times N$ matrix \mathbf{X} column by column, where the length of \mathbf{x} is MN , and use $\mathbf{x} = \text{vec}(\mathbf{X})$ to represent vectorization of matrix \mathbf{X} column by column. The notation $\text{diag}(\mathbf{a})$ represents a diagonal matrix with the elements of \mathbf{a} as its diagonal. We use $|\mathbf{A}|^2$ to denote the element-wise magnitude squared operation for matrix \mathbf{A} . The notations $\mathbf{1}$ and $\mathbf{0}$ are used to denote an all-ones vector and an all-zeros vector with a proper length, respectively. The j -th entry of \mathbf{q} is denoted by q_j . The superscript t of \mathbf{s}' denotes the iteration index of the variable \mathbf{s} involved in an iterative algorithm.

2 System Model

The modulation and demodulation for OTFS are illustrated in Fig. 1, which are implemented with the 2D inverse SFFT (ISFFT) and SFFT at the transmitter and receiver, respectively^[1, 24]. Before the OTFS modulation, a (coded) bit sequence is mapped to symbols $x[k, l], k = 0, \dots, N-1, l = 0, \dots, M-1$ in the DD domain, where $x[k, l] \in \mathcal{A} = \{\alpha_1, \dots, \alpha_{|\mathcal{A}|}\}$, $|\mathcal{A}|$ is the cardinality of \mathcal{A} , l and k denote the indices of the delay and Doppler shifts, respectively, and N and M are the number of grids



▲ Figure 1. Modulation and demodulation in an orthogonal time frequency space (OTFS) system^[2]

of the DD plane. At the transmitter side, ISFFT is performed to convert the DD domain symbols to signals in the time-frequency (TF) domain.

$$X_{yf}[n, m] = \frac{1}{\sqrt{MN}} \sum_{k=0}^{N-1} \sum_{l=0}^{M-1} x[k, l] e^{j2\pi(\frac{nk}{N} - \frac{ml}{M})} \quad (1)$$

After that, the signals $X_{yf}[n, m]$ in the TF domain are converted to a continuous-time waveform $s(t)$ using the Heisenberg transform with a transmit waveform $g_{tx}(t)^{[2]}$, i.e.,

$$s(t) = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} X_{yf}[n, m] g_{tx}(t - nT) e^{j2\pi m \Delta f (t - nT)} \quad (2)$$

where Δf is subcarrier spacing and $T = 1/\Delta f$. Then the signal $s(t)$ is transmitted over a time-varying channel and the received signal in the time domain is given as:

$$r(t) = \iint h(\tau, \nu) s(t - \tau) e^{j2\pi \nu (t - \tau)} d\tau d\nu \quad (3)$$

where $h(\tau, \nu)$ is the channel impulse response in the continuous DD domain, and it can be expressed as^[1]:

$$h(\tau, \nu) = \sum_{i=0}^{P-1} h_i \delta(\tau - \tau_i) \delta(\nu - \nu_i) \quad (4)$$

with $\delta(\cdot)$ being the Dirac delta function, P being the number of channel paths, and h_i , τ_i and ν_i being the gain, delay and Doppler shift associated with the i -th path, respectively. The delay and Doppler-shift taps for the i -th path are given by

$$\tau_i = \frac{l_i}{M\Delta f}, \quad \nu_i = \frac{k_i + \kappa_i}{NT} \quad (5)$$

where l_i and k_i are the delay and Doppler indices of the i -th path, and $\kappa_i \in [-1/2, 1/2]$ is a fractional Doppler associated with the i -th path. In the above equation, $M\Delta f$ is the system bandwidth and NT is the duration of an OTFS block.

At the receiver, a receive waveform $g_{rx}(t)$ is used to transform the received signal $r(t)$ to the TF domain, i.e.,

$$Y(t, f) = \int g_{rx}^*(t' - t) r(t') e^{-j2\pi f(t' - t)} dt' \quad (6)$$

which is then sampled at $t = nT$ and $f = m\Delta f$, yielding $Y[n, m]$. Then SFFT is applied to $Y[n, m]$ to generate the DD domain signal $y[k, l]$, i.e.,

$$y[k, l] = \frac{1}{\sqrt{MN}} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} Y[n, m] e^{-j2\pi(\frac{nk}{N} - \frac{ml}{M})} \quad (7)$$

Assuming that the transmitted waveform and the received waveform satisfy the bi-orthogonal property^[1], in the DD domain we have the input-output relationship^[2].

$$y[k, l] = \sum_{i=0}^{P-1} \sum_{c=-N_i}^{N_i} h_i x([k - k_i + c]_N, [l - l_i]_M) \frac{1}{N} \frac{1 - e^{-j2\pi(-c - \kappa_i)}}{1 - e^{-j2\pi \frac{-c - \kappa_i}{N}}} e^{-j2\pi \frac{l_i(k_i + \kappa_i)}{MN}} + \omega[k, l] \quad (8)$$

where $N_i < N$ is an integer, and $\omega[k, l]$ is the noise in the DD domain. We can see that for each path, the transmitted signal is circularly shifted, and scaled by a corresponding channel gain. We arrange $\{x[k, l]\}$ as a vector $\mathbf{x} \in C^{MN \times 1}$, where the j -th element x_j is $x[k, l]$ with $j = kM + l$. Similarly, a vector $\mathbf{y} \in C^{MN \times 1}$ can also be constructed based on $y[k, l]$. Then Eq. (8) can be rewritten in a vector form as:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\omega} \quad (9)$$

where $\mathbf{H} \in C^{MN \times MN}$ is the effective channel matrix in the DD domain, and $\boldsymbol{\omega}$ denotes a white Gaussian noise with mean 0 and variance ϵ^{-1} (or precision ϵ). The channel matrix \mathbf{H} in Eq. (9) can be represented as^[25]:

$$\mathbf{H} = \sum_{i=0}^{P-1} \sum_{c=-N_i}^{N_i} \mathbf{I}_N(-[c - k_i]_N) \otimes [\mathbf{I}_M(l_i) h_i \times \frac{1 - e^{-j2\pi(-c - \kappa_i)}}{N - N e^{-j2\pi \frac{-c - \kappa_i}{N}}} e^{-j2\pi \frac{l_i(k_i + \kappa_i)}{MN}}] \quad (10)$$

where $\mathbf{I}_N(-[q - k_i]_N)$ denotes an $N \times N$ matrix obtained by circularly shifting the rows of the identity matrix by $-[q - k_i]_N$, and $\mathbf{I}_M(l_i)$ is obtained similarly. Without fractional Doppler, i.e., $\kappa_i = 0$, the channel matrix \mathbf{H} is reduced to

$$\mathbf{H} = \sum_{i=0}^{P-1} \mathbf{I}_N(k_i) \otimes [\mathbf{I}_M(l_i) h_i e^{-j2\pi \frac{l_i k_i}{MN}}] \quad (11)$$

3 Message Passing (MP) Based Detectors

Based on the model (9) in the DD domain, several detectors have been proposed using the message passing techniques.

3.1 MP Detector in Ref. [2]

In model (9), the $MN \times MN$ DD domain complex channel matrix \mathbf{H} is sparse (especially in the case without fractional Doppler shifts), which makes belief propagation suitable for implementing the OTFS detectors. In Eq. (2), \mathbf{y} and $\boldsymbol{\omega}$ are length- MN complex vectors with elements denoted by $y[d]$ and $\omega[d]$, $1 \leq d \leq MN$, the element of \mathbf{H} is denoted by $H[d, c]$, $1 \leq d, c \leq MN$, \mathbf{x} is a length- MN symbol vector with elements $x[c] \in \mathcal{A}$, $1 \leq c \leq MN$, and \mathcal{A} denotes the modulation alphabet.

Thanks to the sparsity of \mathbf{H} , the joint distribution of the random variables in model (9) can be represented with a sparsely-connected factor graph with MN variable nodes corresponding to \mathbf{x} and MN observation nodes corresponding to \mathbf{y} . As shown in

Fig. 2, each observation node $y[d]$ is connected to a set of variable nodes $\{x[e_s], e_s \in \mathcal{I}(d)\}$, and similarly, each variable node $x[c]$ is connected to a set of observation nodes $y[e_s], e_s \in \mathcal{J}(c)$, where $\mathcal{I}(d)$ and $\mathcal{J}(c)$ respectively denote the sets of indexes of non-zero elements in the d -th row and c -th columns of \mathbf{H} , $|\mathcal{I}(d)| = |\mathcal{J}(c)| = S$ and $1 \leq d \leq S$. The probability mass function (PMF) $p_{c,e_s} = \{p_{c,e_s}(a_j) | a_j \in \mathbf{A}\}$ represents the messages from variable nodes $x[c]$ to factor nodes $y[e_s]$.

Based on the factor graph in Fig. 2, a message passing algorithm was proposed in Ref. [2], and the detector is called MP detector in this paper. The following is a brief derivation of the message computations in the i -th iteration of the message computations.

1) Messages passing from observation node $y[d]$ to variable node $x[e_s]$

The message is approximated to be Gaussian, and the mean μ_{d,e_s}^i and variance $(\sigma_{d,e_s}^i)^2$ are computed as

$$\mu_{d,e_s}^i = \sum_{e \in \mathcal{J}(d), e \neq e_s} \sum_{j=1}^Q p_{e,d}^{i-1}(a_j) a_j H[d,e], \quad (12)$$

$$(\sigma_{d,e_s}^i)^2 = \sum_{e \in \mathcal{J}(d), e \neq e_s} \left(\sum_{j=1}^Q p_{e,d}^{i-1}(a_j) |a_j|^2 |H[d,e]|^2 - \left| \sum_{j=1}^Q p_{e,d}^{i-1}(a_j) a_j H[d,e] \right|^2 \right) + \epsilon^{-1}. \quad (13)$$

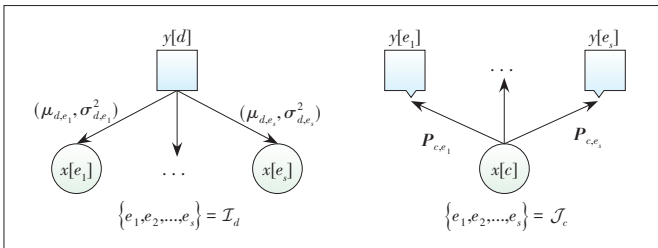
2) Messages passing from variable node $x[c]$ to observation node $y[e_s]$

The PMF p_{c,e_s}^i can be updated as

$$p_{c,e_s}^i(a_j) = \Delta \cdot \tilde{p}_{c,e_s}^i(a_j) + (1 - \Delta) \cdot \tilde{p}_{c,e_s}^{i-1}(a_j), \quad (14)$$

where $\Delta \in [0,1]$ is the damping factor and

$$\tilde{p}_{c,e_s}^i(a_j) \propto \prod_{e \in \mathcal{J}(c), e \neq e_s} \Pr(y[e] | x[c] = a_j, \mathbf{H}) = \prod_{e \in \mathcal{J}(c), e \neq e_s} \frac{s^i(e,c,j)}{\sum_{k=1}^Q s^i(e,c,k)}, \quad (15)$$



▲ Figure 2. Graph representation used to derive the message passing (MP) detector in Ref. [2]

with

$$s^i(e,c,k) = \exp \left(\frac{-|y[e] - \mu_{e,c}^i - H[e,c] a_k|^2}{(\sigma_{e,c}^i)^2} \right). \quad (16)$$

After a certain number of iterations by repeating 1) and 2), the decision on the transmitted symbol can be obtained, i.e.,

$$\hat{x}[c] = \arg \min_{a_j \in \mathbf{A}} p_c^i(a_j), \quad c = 1, \dots, MN, \quad (17)$$

where

$$p_c^i(a_j) = \prod_{e \in \mathcal{J}(c)} \frac{s^i(e,c,j)}{\sum_{k=1}^Q s^i(e,c,k)}. \quad (18)$$

The MP detector is summarized in Algorithm 1.

Algorithm 1. MPA detector in Ref. [2]

Input: y, \mathbf{H} , Initialize: $p_{c,e_s}^0 = 1/|\mathbf{A}|$, $c = 1, \dots, MN$, $e_s \in \mathcal{J}(c)$, $i = 1$

1: **Repeat**

2: $\forall d$: update μ_{d,e_s}^i and $(\sigma_{d,e_s}^i)^2$ with Eqs. (12) and (13)

3: $\forall c$: update $p_{c,d}^i$ with Eq. (14)

4: $i = i + 1$

5: **Until** terminate

Output: The decision on transmitted symbols $\hat{x}[c]$ using Eq. (17)

The MP algorithm shown above is an approximation to loopy belief propagation since it approximates the interference to be Gaussian to achieve lower complexity. The complexity of the algorithm is $\mathcal{O}(MNS|\mathbf{A}|)$ per iteration, which depends on the sparsity of the channel, i.e., the value of S . When S is small, the detector is very attractive because it has low complexity and the detector delivers a good performance as no short loops in the factor graph model. However, in the case of rich-scattering environments and fractional Doppler shifts, the value of S can be large, leading to a denser factor graph model, which can affect the performance of the MP detector and result in a significant increase in computational complexity.

3.2 VB Detector

The VB detector was proposed in Ref. [17] to guarantee the convergence of the iterative detector, which can be implemented with variational message passing. With model (9), the optimal MAP detection can be formulated as:

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x}} p(\mathbf{x} | \mathbf{y}). \quad (19)$$

However, the complexity of solving the above optimization problem increases exponentially with the size of \mathbf{x} . VB is ad-

opted to achieve low complexity approximate detection. In this method, a distribution $q(\mathbf{x})$ from a tractable distribution family \mathcal{Q} is found as an approximation to the a posteriori distribution $p(\mathbf{x}|\mathbf{y})$. The trial distribution $q(\mathbf{x})$ can be obtained by minimizing the Kullback-Leibler divergence $\mathcal{D}(q||p)$, i.e.,

$$q^*(\mathbf{x}) = \arg \max_{q \in \mathcal{Q}} \mathcal{D}(q||p) = \arg \max_{q \in \mathcal{Q}} \underbrace{\mathbb{E}_q[-\ln q(\mathbf{x}) + \ln p(\mathbf{x}|\mathbf{y})]}_{\mathcal{L}}, \quad (20)$$

where the expectation is taken over \mathbf{x} according to the trial distribution $q(\mathbf{x})$.

To simplify the optimization problem, $q(\mathbf{x})$ is assumed to be fully factorized, i.e.,

$$q(\mathbf{x}) = \prod_{k,l} q_{k,l}(x_{k,l}), \quad (21)$$

where $k \in [0, N-1]$, $M \in [0, M-1]$ and $x_{k,l}$ denotes the $(kM + l)$ -th entry of \mathbf{x} . With this assumption, $q(\mathbf{x})$ can be updated iteratively by maximizing \mathcal{L} . Since the noise sample $\omega_{k,l}$ and data symbol $x_{k,l}$, $\forall k, l$ are independent, and $\omega_{k,l} \sim \mathcal{CN}(\omega_{k,l}; 0, \epsilon^{-1})$, $p(\mathbf{x}|\mathbf{y})$ can be rewritten as:

$$p(\mathbf{x}|\mathbf{y}) \propto \prod_{k,l} p(x_{k,l}) p(y_{k,l}|\mathbf{y}), \quad (22)$$

where $y_{k,l} = \mathbf{h}_{k,l}^T \mathbf{x} + \omega_{k,l}$, $\mathbf{h}_{k,l}$ denotes the equivalent channel vector whose $(kM + l)$ -th entry is $h_{k,l}[k, l]$. Then the distribution $p(\mathbf{x}|\mathbf{y})$ can be further rewritten as:

$$p(\mathbf{x}|\mathbf{y}) \propto \prod_{k,l} \zeta_{k,l}(x_{k,l}) \prod_{k',l'} \psi_{k,l}(x_{k,l}, x_{k',l'}), \quad (23)$$

where

$$\zeta_{k,l}(x_{k,l}) = p(x_{k,l}) \exp\left(-\frac{\rho_{k,l}|x_{k,l}|^2 + \eta_{k,l}x_{k,l}}{\epsilon^{-2}}\right), \quad (24)$$

$$\psi_{k,l}(x_{k,l}, x_{k',l'}) = \exp\left(-\frac{\mathcal{Q}_{k,l,k',l'} x_{k,l} x_{k',l'}}{\epsilon^{-2}}\right), \quad (25)$$

with $\rho_{k,l} = \sum_{k',l'} |h_{k',l'}(k,l)|^2$, $\eta_{k,l} = 2 \sum_{k',l'} \mathcal{R}[h_{k',l'}[k,l] \cdot y_{k,l}^*]$, and $\mathcal{Q}_{k,l,k',l'} = 2\mathcal{R}[h_{k,l}[k,l] h_{k',l'}^*[k',l']]$. Substituting $p(\mathbf{x}|\mathbf{y})$ in Eq. (23) and $q(\mathbf{x})$ into \mathcal{L} yields

$$\mathcal{L} = \mathbb{E}_q \left[\sum_{k,l} \ln \psi_{k,l}(x_{k,l}, x_{k',l'}) - \sum_{k,l} \ln \frac{q_{k,l}(x_{k,l})}{\zeta_{k,l}(x_{k,l})} \right] = \mathbb{E}_q \left[-\frac{\sum_{k,l} \mathcal{Q}_{k,l,k',l'} x_{k,l} x_{k',l'}}{\epsilon^{-2}} - \sum_{k,l} \ln \frac{q_{k,l}(x_{k,l})}{\zeta_{k,l}(x_{k,l})} \right]. \quad (26)$$

To find a stationary point of \mathcal{L} , the partial derivations of \mathcal{L} with respect to all local functions $q_{k,l}(x_{k,l})$, $\forall k, l$ need to be zero. Take the latent variable $x_{k,l}$ as an example. Setting the partial derivation $\partial \mathcal{L} / \partial q_{k,l}(x_{k,l})$ to zero leads to:

$$\mathbb{E}_{q_{k,l}} \left[-\frac{\sum_{k',l'} \mathcal{Q}_{k,l,k',l'} x_{k,l} x_{k',l'}}{\epsilon^{-2}} \right] + \ln \zeta_{k,l}(x_{k,l}) - \ln q_{k,l}^{iter}(x_{k,l}) + C = 0, \quad (27)$$

where $q_{k,l} = \prod_{(k',l') \neq (k,l)} q_{k',l'}^{iter-1}(x_{k,l})$, $q_{k',l'}^{iter-1}(x_{k,l})$ is obtained in the $(iter - 1)$ -th iteration and C denotes a constant.

Then, solving Eq. (27) for $q_{k,l}(x_{k,l})$ results in the local distribution, which can be expressed as:

$$q_{k,l}^{iter}(x_{k,l}) \propto \zeta_{k,l}(x_{k,l}) \exp \left(\mathbb{E}_{q_{k,l}} \left[-\frac{\sum_{k',l'} \mathcal{Q}_{k,l,k',l'} x_{k,l} x_{k',l'}}{\epsilon^{-2}} \right] \right) \propto p(x_{k,l}) \exp \left(-\frac{\rho_{k,l}|x_{k,l}|^2 - m_{k,l}d_{k,l}}{\epsilon^{-2}} \right), \quad (28)$$

where $m_{k,l} = \eta_{k,l} - \sum_{(k',l') \neq (k,l)} \mathcal{Q}_{k,l,k',l'} \mathbb{E}_{q_{k',l'}^{iter-1}} x[k', l']$.

It is noted that the variance of $x_{k,l}$ is underestimated and only the noise variance is considered in Eq. (28). To fix the underestimation, a practical solution is to repeat the above procedure to approximate the a posteriori distribution for all the data symbols iteratively, resulting in the approximate marginal $q_{k,l}^*(x_{k,l})$, $\forall k, l$. Then, the decision on the symbols can be made by maximizing the approximate marginal distribution $q_{k,l}^*(x_{k,l})$, i.e.,

$$\hat{x}_{k,l} = \arg \max_{x_{k,l} \in \mathcal{A}} q_{k,l}^*(x_{k,l}). \quad (29)$$

The complexity of the algorithm per iteration is $\mathcal{O}(MNS|\mathcal{A}|)$.

3.3 UAMP Detector

Leveraging the UAMP algorithm, the UAMP detector was developed in Ref.[25], where the BCCB structure of the DD domain channel matrix is exploited, leading to a highly efficient OTFS detector with 2D FFT. It can be seen from Eqs. (10) and (11) that the DD domain channel matrix \mathbf{H} has a BCCB structure. A useful property of the BCCB matrix \mathbf{H} is that it can be diagonalized using 2D Discrete Fourier Transform matrix, i.e.,

$$\mathbf{H} = \mathbf{F}^H \mathbf{\Lambda} \mathbf{F}, \quad (30)$$

where $\mathbf{F} = \mathbf{F}_N \otimes \mathbf{F}_M$ with \mathbf{F}_N and \mathbf{F}_M being respectively the normalized N -point and M -point DFT matrices. In Eq. (30),

matrix \mathbf{A} is a diagonal matrix, i.e., $\mathbf{A} = \text{diag}(\mathbf{d})$, and \mathbf{d} is a length- MN vector that can be computed using 2D FFT.

$$\mathbf{d} = \text{vec}(\text{FFT2}(\mathbf{C})), \quad (31)$$

where $\text{FFT2}(\cdot)$ represents the 2D FFT operation, $\mathbf{C} = \text{reshape}_M(\mathbf{H}(:,1))$ is an $M \times N$ matrix, and $\mathbf{H}(:,1)$ with length- MN is the first column of matrix \mathbf{H} .

The above property is exploited in the design of the UAMP detector, leading to high computational efficiency while with outstanding performance compared with the existing detectors. Instead of using model (9) directly, the UAMP algorithm^[27-29] works with the unitary transform of the model. The channel matrix \mathbf{H} admits the diagonalization in Eq. (30), leading to the following unitary transform of the OTFS system model:

$$\mathbf{r} = \mathbf{A}\mathbf{F}\mathbf{x} + \mathbf{w}', \quad (32)$$

where $\mathbf{r} = \mathbf{F}\mathbf{y}$, $\mathbf{w}' = \mathbf{F}\mathbf{w}$, and the noise \mathbf{w}' has the same distribution with \mathbf{w} as \mathbf{F} is a unitary matrix. The precision of the noise is still denoted by ϵ , which needs to be estimated. Define $\Phi = \mathbf{A}\mathbf{F}$ and an auxiliary vector $\mathbf{z} = \Phi\mathbf{x}$. Then we can factorize the joint distribution of the unknown variables $\mathbf{x}, \mathbf{z}, \epsilon$ given \mathbf{r} as

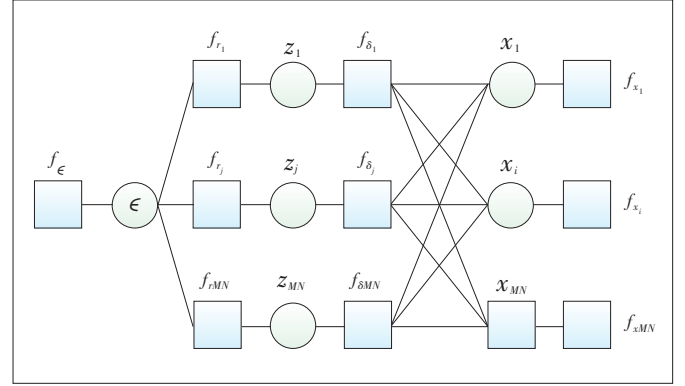
$$\begin{aligned} p(\mathbf{x}, \mathbf{z}, \epsilon | \mathbf{r}) &= p(\epsilon) p(\mathbf{r} | \mathbf{z}, \epsilon) p(\mathbf{z} | \mathbf{x}) p(\mathbf{x}) = \\ &= p(\epsilon) \prod_j p(r_j | z_j, \epsilon) p(z_j | \mathbf{x}) \prod_i p(x_i) = \\ &= f_\epsilon \prod_{j,j} f_{r_j}(z_j, \epsilon) f_{\delta_j}(z_j, \mathbf{x}) \prod_i f_{x_i}(x_i), \end{aligned} \quad (33)$$

where indices $i, j \in [1:MN]$. To facilitate the factor graph representation of the factorization in Eq. (33), the relevant notations are listed in Table 1, which shows the correspondence between the factor nodes and their associated distributions. The factor graph representation for the factorization in Eq. (33) is depicted in Fig. 3.

Following the UAMP algorithm, a UAMP based iterative detector can be designed, which is summarized in Algorithm 2. According to the derivation of (U)AMP using loopy belief propagation, UAMP provides the message from variable node z_j to function node f_{r_j} , which is Gaussian and denoted by $m_{z_j \rightarrow f_{r_j}}(z_j) = \mathcal{N}(z_j | p_j, \nu_{p_j})$. Here, the mean p_j and the variance ν_{p_j} are given in Lines 1 and 2 of the Algorithm in a vector

▼ Table 1. Factors, underlying distributions and functional forms associated with Eq. (31)

Factor	Distribution	Function Form
f_{r_j}	$p(r_j z_j, \epsilon)$	$\mathcal{N}(z_j; r_j, \epsilon^{-1})$
f_{δ_j}	$p(z_j \mathbf{x})$	$\delta(z_j - \Phi_j \mathbf{x})$
f_{x_i}	$p(x_i)$	$(1/ A) \sum_{a=1}^A \delta(x_i - \alpha_a)$
f_ϵ	$p(\epsilon)$	ϵ^{-1}



▲ Figure 3. Factor graph representation of Eq. (31)

form. With the mean field rule^[23] at the function node f_{r_j} , we can compute the message passed from function node f_{r_j} to variable node ϵ , i.e.,

$$\begin{aligned} m_{f_{r_j} \rightarrow \epsilon}(\epsilon) &\propto \exp \left\{ \left\langle \log f_{r_j}(r_j | z_j, \epsilon) \right\rangle_{b(z_j)} \right\} \propto \exp \left\{ -\epsilon (|r_j - \right. \\ &\quad \left. \hat{z}_j|^2 + \nu_{z_j}) \right\}, \end{aligned} \quad (34)$$

where $b(z_j)$ is the belief of z_j . It turns out that $b(z_j)$ is also Gaussian with its variance and mean given by

$$\nu_{z_j} = 1 / \left(1/\nu_{p_j} + \hat{\epsilon} \right), \quad \hat{z} = \nu_{z_j} (p_j / \nu_{p_j} + \hat{\epsilon} r_j), \quad (35)$$

respectively, where $\hat{\epsilon}$ is the estimate of ϵ in the last iteration. They can be expressed in a vector form shown in Lines 3 and 4 in Algorithm 2. The estimate of ϵ can be obtained based on the belief $b(\epsilon)$ at the variable node ϵ shown in Fig. 3, i.e.,

$$b(\epsilon) \propto f_\epsilon(\epsilon) \prod_{j=1}^{MN} m_{f_{r_j} \rightarrow \epsilon}(\epsilon). \quad (36)$$

And the estimate is given as

$$\hat{\epsilon} = \int_0^\infty \epsilon b(\epsilon) d\epsilon = MN / \sum_{j=1}^{MN} (|r_j - \hat{z}_j|^2 + \nu_{z_j}), \quad (37)$$

which can be rewritten in a vector form shown in Line 5 of the algorithm. With the mean field rule at the function node f_{r_j} again, the message passed from the function node f_{r_j} to the variable node z_j can be computed as:

$$m_{f_{r_j} \rightarrow z_j}(z_j) \propto \exp \left\{ \left\langle \log f_{r_j}(r_j | z_j, \hat{\epsilon}) \right\rangle_{b(\epsilon)} \right\} \propto \mathcal{N}(h_j | r_j, \hat{\epsilon}^{-1}). \quad (38)$$

Then the UAMP algorithm with known noise can be used as if the true noise precision is $\hat{\epsilon}$, leading to Lines 6 - 15 and

Lines 1 – 2 of the Algorithm 2. In Lines 10 – 13, the Gaussian message is combined with the discrete prior to obtain the MMSE estimates of the symbols in terms of their posterior means and variances. There is an extra operation in Line 14, which averages the variances of x_j . Thanks to the special form of the unitary matrix \mathbf{F} , 2D FFT is used in the implementations in Lines 2 and 9. It can be seen that the UAMP detector does not require any matrix-vector products, the algorithm requires only element-wise vector operations or scalar operations, except Lines 2 and 9, which are implemented with FFT. So the complexity of the UAMP detector is $\mathcal{O}(MN \log(MN)) + \mathcal{O}(MN|\mathcal{A}|)$ per OTFS block per iteration, which is independent of S .

Algorithm 2. UAMP detector for OTFS

Unitary transform: $\mathbf{r} = \mathbf{F}\mathbf{y} = \mathbf{A}\mathbf{F}\mathbf{x} + \boldsymbol{\omega}$ with $\mathbf{F} = \mathbf{F}_N \otimes \mathbf{F}_M$.

Calculated \mathbf{d} with Eq. (29), and define vector $\boldsymbol{\Lambda} = \mathbf{d} \cdot \mathbf{d}^*$.

Initialize $\mathbf{s}^{-1} = \mathbf{0}$, $\hat{\mathbf{x}} = \mathbf{0}$, $\hat{\epsilon}^{(0)} = 1$, $\nu_x^{(0)} = 1$, and $t = 0$.

Input: \mathbf{y}, H

Repeat

1: $\nu_p = \nu_x^t \boldsymbol{\Lambda}$

2: $\mathbf{p} = \mathbf{d} \cdot \text{vec}\left(\text{FFT2}\left(\text{reshape}_M(\hat{\mathbf{x}}^t)\right)\right) - \nu_p \cdot \mathbf{s}^{t-1}$

3: $\nu_z = 1./\left(1./\nu_p + \hat{\epsilon}^t\right)$

4: $\mathbf{z} = \nu_z \cdot \left(\mathbf{p}./\nu_p + \hat{\epsilon}^t \mathbf{r}\right)$

5: $\hat{\epsilon}^{t+1} = MN/\left(\|\mathbf{r} - \mathbf{z}\|_2^2 + 1^T \nu_z\right)$

6: $\nu_s = 1./\left(\nu_p + 1/\hat{\epsilon}^{t+1}\right)$

7: $\mathbf{s}^t = \nu_s \cdot \left(\mathbf{r} - \hat{\mathbf{p}}\right)$

8: $\nu_q = \boldsymbol{\Lambda}^T \nu_s / (MN)$

9: $\mathbf{q} = \hat{\mathbf{x}}^{(t)} + \nu_q \text{vec}\left(\text{IFFT2}\left(\text{reshape}_M(\mathbf{d} \cdot \mathbf{s}^t)\right)\right)$

10: $\forall j: \xi_{j,a} = \exp\left(-\nu_q^{-1} |\alpha_a - q_j|^2\right)$

11: $\forall j: \beta_{j,a} = \xi_{j,a} / \sum_{a=1}^{|\mathcal{A}|} \xi_{j,a}$

12: $\forall j: \hat{x}_j^{t+1} = \sum_{a=1}^{|\mathcal{A}|} \alpha_a \beta_{j,a}$

13: $\forall j: \nu_{x_j}^{t+1} = \sum_{a=1}^{|\mathcal{A}|} \beta_{j,a} |\alpha_a - \hat{x}_j^{t+1}|^2$

14: $\nu_x^{t+1} = \frac{1}{MN} \sum_{j=1}^{MN} \nu_{x_j}^{t+1}$

15: $t = t + 1$

Until terminated

Output: the estimate of \mathbf{x} i.e., $\hat{\mathbf{x}}$

Compared with the UAMP detector, the MP and VB detectors have a complexity of $\mathcal{O}(MNS|\mathcal{A}|)$ per OTFS block per iteration, which can be considerably higher than that of the UAMP detector in the case of rich scattering environments and when fractional Doppler shifts have to be considered (leading to a large S). Moreover, the UAMP detector can deliver much better performance when the number of paths is relatively large. In particular, the UAMP detector with estimated noise precision can significantly outperform other detectors with perfect noise precision. We note that,

the OTFS detector can be implemented directly with the AMP algorithm. However, due to the deviation of the channel matrix from the i.i.d. Gaussian matrix, the AMP detector may perform poorly.

4 Turbo Processing in Coded Systems

It is well known that joint decoding and detection can bring significant system performance improvement, and it can be realized in a way that the detector and decoder exchange information iteratively, i.e., the turbo processing^[30–31]. The OTFS detectors can be incorporated into a turbo receiver by endowing the OTFS detectors with the capabilities of taking the output log-likelihood ratios (LLRs) of the decoder as (soft) input and producing (soft) output in the form of extrinsic LLRs of the coded bits, i.e., the so-called soft input soft output (SISO) detector.

A typical turbo system is shown in Fig. 4, where Π and Π^{-1} represent interleaver and de-interleaver, respectively. The information bits are encoded and interleaved before symbol mapping, where each symbol $x_j \in \mathcal{A} = \{\alpha_1, \dots, \alpha_{|\mathcal{A}|}\}$ in the DD domain is mapped from a sub-sequence of the coded bit sequence, which is denoted by $\mathbf{c}_j = [c_j^1, \dots, c_j^{\log|\mathcal{A}|}]$. Each α_a corresponds to a length- $\mathcal{A} \log|\mathcal{A}|$ binary sequence, which is denoted by $\{\alpha_a^1, \dots, \alpha_a^{\log|\mathcal{A}|}\}$. Based on the LLRs provided by the SISO decoder and the output of the OTFS demodulator as shown in Fig. 4, the task of the SISO OTFS detector is to compute the extrinsic LLR for each coded bit, i.e.,

$$L^e(c_j^q) = \ln \frac{P(c_j^q = 0|\mathbf{r})}{P(c_j^q = 1|\mathbf{r})} - L^a(c_j^q), \quad (39)$$

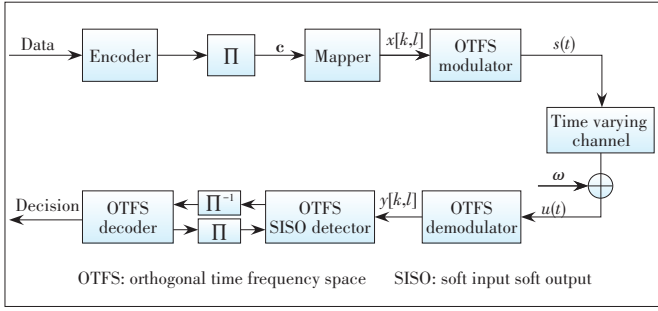
where $L^a(c_j^q)$ is the output extrinsic LLR of the decoder in the last iteration. The extrinsic LLR $L^e(c_j^q)$ is passed to the decoder. The extrinsic LLR $L^e(c_j^q)$ can be expressed in terms of extrinsic mean and variance of the symbols^[32], i.e.,

$$L^e(c_j^q) = \ln \frac{\sum_{\alpha_a \in \mathcal{A}_q^0} \exp\left(-\frac{|\alpha_a - m_j^e|^2}{v_j^e}\right) \prod_{q' \neq q} P(c_j^{q'} = \alpha_a^{q'})}{\sum_{\alpha_a \in \mathcal{A}_q^1} \exp\left(-\frac{|\alpha_a - m_j^e|^2}{v_j^e}\right) \prod_{q' \neq q} P(c_j^{q'} = \alpha_a^{q'})}, \quad (40)$$

where m_j^e and v_j^e are the extrinsic mean and variance of x_j , and \mathcal{A}_q^0 and \mathcal{A}_q^1 represent the subsets of all α_a corresponding to $c_j^q = 0$ and $c_j^q = 1$, respectively. The extrinsic variance and mean are defined in Ref. [32].

$$v_j^e = (1/v_j^p - 1/v_j)^{-1}, m_j^e = v_j^e (m_j^p/v_j^p - m_j/v_j), \quad (41)$$

where m_j and v_j are the a priori mean and variance of x_j calcu-



▲ Figure 4. Iterative joint detection and decoding in a coded OTFS system^[25]

lated based on the output LLRs of the SISO decoder^[30] and m_j^p and v_j^p are a posteriori mean and variance of x_j .

Taking the UAMP detector as an example, we show the incorporation of the OTFS detector into a turbo receiver. According to the derivation of the UAMP algorithm, we can find that \mathbf{q} and \mathbf{v}_q consist of the extrinsic means and variances of the symbols in \mathbf{x} as they are the messages passed from the observation side and do not contain the immediate a priori information about \mathbf{x} . Hence we have $m_j^e = q_j$ and $v_j^e = v_q$. Then Eq. (40) can be readily used to compute the extrinsic LLRs of the coded bits. With the LLRs provided by the SISO decoder, one can compute the probability $p(x_j = \alpha_a)$ for each x_j , which is no longer the “non-informative prior” in Algorithm 2. Therefore, $\xi_{j,a}$ in Line 7 of the algorithm is changed to

$$\xi_{j,a} = p(x_j = \alpha_a) \exp(-\nu_q^{-1} |\alpha_a - q_j|^2). \quad (42)$$

In addition, the iteration of the UAMP detector can be combined with the iteration between the SISO decoder and detector, which leads to a single loop iteration (i.e., inner iterations are not required).

The computational complexity of the detectors is summarized in Table 2. In the above discussion, we focus on the bi-orthogonal waveform. The detectors can be extended to OTFS systems with other waveforms, such as the simple rectangular waveform^[25].

5 Simulation Results

In this section, we compare the performance of the message passing based detectors. The low complexity MRC detector in Ref. [14] is also included. We set $M = 256$ and $N =$

▼ Table 2. Computational complexity of various detectors per iteration

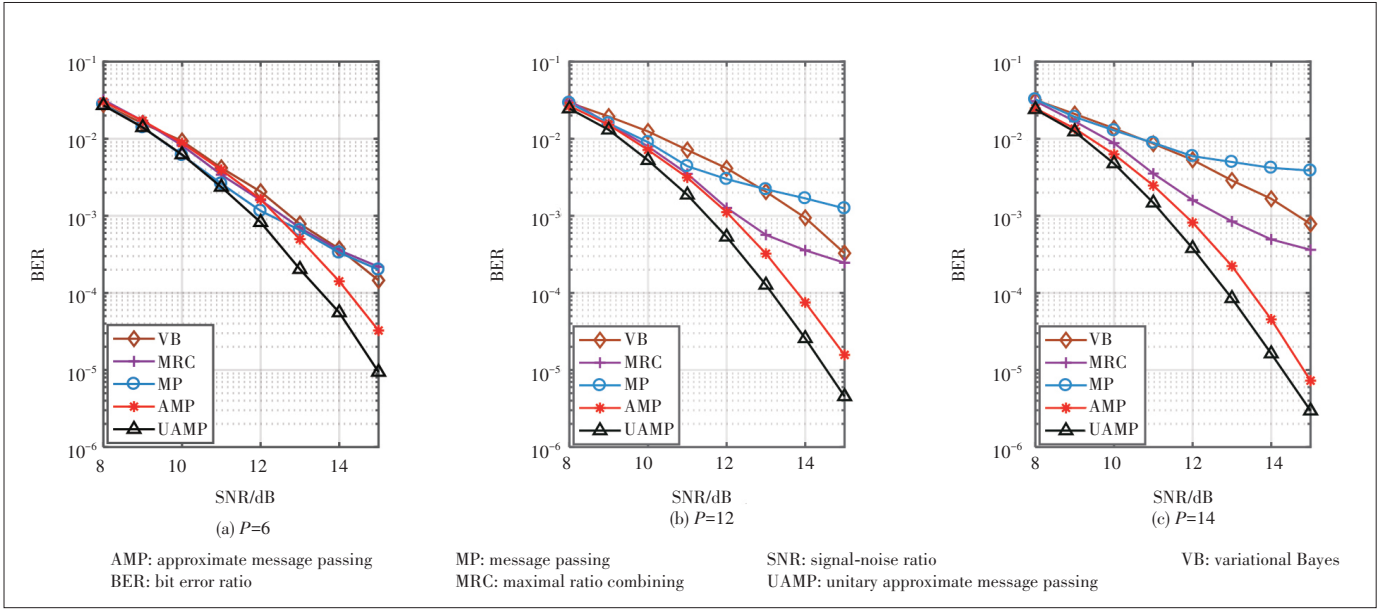
Detectors	Complexity
MP detector	$\mathcal{O}(MNSL\mathcal{A})$
VB detector	$\mathcal{O}(MNSL\mathcal{A})$
UAMP detector	$\mathcal{O}(MN \log(MN)) + \mathcal{O}(MNL\mathcal{A})$

MP: message passing UAMP: unitary approximate message passing
VB: variational Bayes

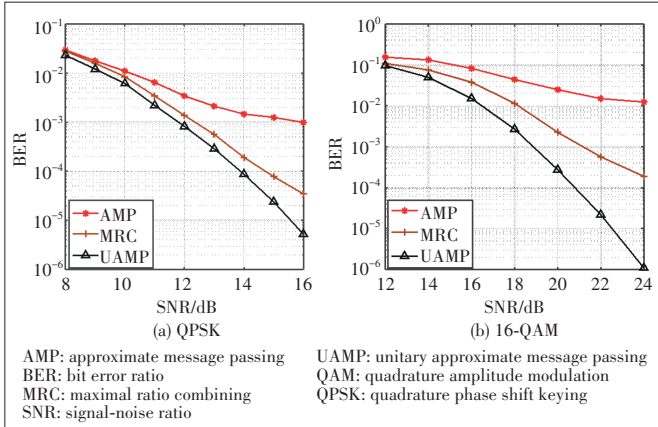
32, i.e., there are 32 time slots and 256 subcarriers in the TF domain. Both quadrature phase shift keying (QPSK) modulation and 16-quadrature amplitude modulation (QAM) are considered. The carrier frequency is 3 GHz, and the subcarrier spacing is 2 kHz. The speed of the mobile user is set to $v = 135$ km/h, leading to a maximum Doppler frequency shift index $k_{\max} = 6$. We assume that the maximum delay index is $l_{\max} = 14$. The Doppler index of the i -th path ($l_1 = 0$). We assume that the fractional Doppler κ_i is uniformly distributed within $[-1/2, 1/2]$, and the channel coefficients h_i are independently drawn from a complex Gaussian distribution with mean 0 and variance η^i , where the normalized power delay profile $\eta^i = \exp(-\alpha l_i) / \sum_i \exp(-\alpha l_i)$ with α being 0 or 0.1. The maximum number of iterations is set to 15 for all iterative detectors. We note that, all detectors except the MRC detector require the noise variance. The UAMP detector performs noise precision estimation, while the other detectors (except the MRC detector) including the AMP detector assume perfect noise precision. We evaluate the performance of the detectors in a variety of scenarios including the bi-orthogonal and rectangular waveforms with integer or fractional Doppler shifts, and QPSK or 16-QAM for modulations. In addition, both uncoded and coded systems are evaluated.

Fig. 5 shows the BER performance of various detectors in the case of the bi-orthogonal waveform with different numbers of paths, where we assume no fractional Doppler shifts, i.e., $S = P$. We also assume $\alpha = 0$, and QPSK is used. From this figure, we can see that, the MP detector performs well when $P = 6$, but with the increase of P , its performance becomes worse. The VB detector has a similar trend. The MRC detector performs similarly to the MP and VB detectors when $P=6$ and delivers better performance than the MP and VB detectors with larger P . The AMP and UAMP detectors perform well, where we can see that they enjoy the diversity gain and achieve better performance with the increase of P . In all cases, the UAMP based detector delivers the best performance and significantly outperforms other detectors.

With the rectangular waveform and fractional Doppler shifts, we compare the bit error ratio (BER) performance of the AMP, UAMP and MRC detectors in Fig. 6, where the number of paths $P = 9$ and $\alpha = 0.1$ is used for the power delay profile. Both QPSK and 16-QAM are considered. Due to the deviation of the channel matrix from the i.i.d. (sub-) Gaussian matrix, AMP exhibits performance loss, leading to significantly worse performance compared with the UAMP detector. Thanks to the robustness of UAMP against a general matrix, UAMP performs well. We can see that the MRC detector performs better than the AMP detector. The UAMP detector performs the best and the gaps between other detectors with the UAMP detector be-



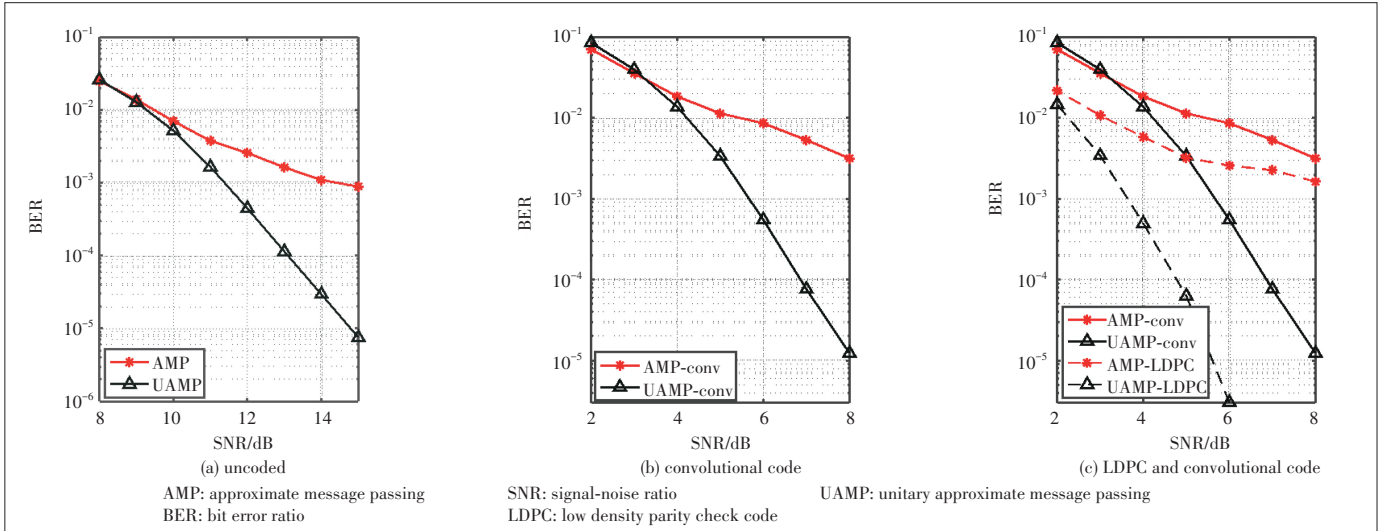
▲ Figure 5. BER performance of detectors with bi-orthogonal waveform and integer Doppler shifts (results are based on Ref. [25])



▲ Figure 6. BER performance of detectors with the rectangular waveform and fractional Doppler shifts (results are based on Ref. [25])

come larger in the case of higher order modulation 16-QAM, compared with QPSK.

We then evaluate the performance of the detectors in a coded OTFS system, where the turbo receiver in Fig. 4 is employed. The number of paths $P=14$, and a rectangular waveform is used. In Fig. 7(a), we show the performance of the uncoded system with the AMP and UAMP detectors. In Fig. 7(b), we use a rate-1/2 convolutional code with a generator $[5, 7]_8$ followed by a random interleaver and QPSK modulation. The length of the codeword is MN . The BCJR algorithm is used for the SISO decoder. We can find that the performance gaps between the AMP detector and the UAMP detector become larger in the coded system. The turbo receiver can achieve much better performance (about 3.5 – 4 dB at the BER of 10^{-4})



▲ Figure 7. BER performance comparison of coded and uncoded system with rectangular waveform (part of the results is based on Ref. [25])

thanks to the joint processing of decoding and detection. In Fig. 7(c), we investigate the performance of the system with a more powerful LDPC. The 8 192 information bits are coded at rate $R=1/2$ by an irregular LDPC code with an average column weight of 3, then the coded bits are randomly interleaved and mapped. As expected, the system performance is improved considerably when the LDPC is used. From Fig. 7(c), we can see that the use of the LDPC code can improve the performance of the UAMP based detector significantly and the performance gap between AMP and UAMP increases when the LDPC is used.

6 Conclusions and Potential Future Work

In this paper, we review and compare the recently proposed message passing based OTFS detectors, which exploit the structures of the OTFS channel matrices, such as sparsity and BCCB. According to the results, the MP and VB detectors are more suitable in the scenarios that the number of paths is relatively small and the modulation order is low, where they deliver good performance while with relatively low complexity. The UAMP detector seems very promising especially in the case of rich-scattering environments and/or when fractional Doppler shifts have to be considered, where the UAMP detector is attractive in both computational complexity and performance. The results also show that the OTFS system with a turbo receiver can provide significant performance gain.

The message passing techniques seem promising in the design of OTFS receivers. In this paper, we assume the OTFS channel matrix is known, which however has to be estimated for practical applications. Message passing based OTFS channel estimation has been investigated in the literature, such as the work in Ref. [33]. With the message passing techniques, channel estimation and detection can be integrated for joint channel estimation and detection, which is expected to lead to superior system performance and/or significant reduction of the training overhead. This is because the data symbols can be used to serve as a virtual training sequence and the guard band between the training symbols and data symbols is not necessary.

It has been shown that joint decoding and detection based on a turbo receiver can significantly improve the system performance. The system performance can be potentially further improved by optimizing the error control codes. This requires fast and accurate performance prediction of the iterative receiver, so that the error control codes, e.g., LDPC, can be optimized.

The message passing techniques could be used to implement sophisticated receivers in more complex systems, such as multi-user OTFS systems, grant-free multiple access with OTFS, multiple-output-multiple-input (MIMO)-OTFS, integrated sensing and communication with OTFS.

References

- [1] HADANI R, RAKIB S, TSATSANIS M, et al. Orthogonal time frequency space modulation [C]//2017 IEEE Wireless Communications and Networking Conference (WCNC). San Francisco, USA: IEEE, 2017: 1 - 6. DOI: 10.1109/WCNC.2017.7925924
- [2] RAVITEJA P, PHAN K T, HONG Y, et al. Interference cancellation and iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, 2018, 17(10): 6501 - 6515. DOI: 10.1109/TWC.2018.2860011
- [3] SURABHI G D, AUGUSTINE R M, CHOCKALINGAM A. On the diversity of uncoded OTFS modulation in doubly-dispersive channels [J]. IEEE transactions on wireless communications, 2019, 18(6): 3049 - 3063. DOI: 10.1109/TWC.2019.2909205
- [4] HADANI R, MONK A. OTFS: A new generation of modulation addressing the challenges of 5G [EB/OL]. [2021-10-01]. <https://arxiv.org/abs/1802.02623>
- [5] LI S Y, YUAN J H, YUAN W J, et al. Performance analysis of coded OTFS systems over high-mobility channels [J]. IEEE transactions on wireless communications, 2021, 20(9): 6033 - 6048. DOI: 10.1109/TWC.2021.3071493
- [6] WEI Z Q, YUAN W J, LI S Y, et al. Orthogonal time-frequency space modulation: a promising next-generation waveform [J]. IEEE wireless communications, 2021, 28(4): 136 - 144. DOI: 10.1109/MWC.001.2000408
- [7] FARHANG A, REZAZADEHREYHANI A, DOYLE L E, et al. Low complexity modem structure for OFDM-based orthogonal time frequency space modulation [J]. IEEE wireless communications letters, 2018, 7(3): 344 - 347. DOI: 10.1109/LWC.2017.2776942
- [8] LI L, WEI H, HUANG Y, et al. A simple two-stage equalizer with simplified orthogonal time frequency space modulation over rapidly time-varying channels [EB/OL]. [2021-10-10]. <https://arxiv.org/abs/1709.02505>
- [9] LONG F, NIU K, DONG C, et al. Low complexity iterative LMMSE-PIC equalizer for OTFS [C]//2019 IEEE International Conference on Communications (ICC). Shanghai, China: IEEE, 2019: 1 - 6. DOI: 10.1109/ICC.2019.8761635
- [10] ZEMEN T, HOFER M, LOESCHENBRAND D. Low-complexity equalization for orthogonal time and frequency signaling (OTFS) [EB/OL]. [2021-10-10]. <https://arxiv.org/abs/1710.09916v1>
- [11] SURABHI G D, CHOCKALINGAM A. Low-complexity linear equalization for OTFS modulation [J]. IEEE communications letters, 2020, 24(2): 330 - 334. DOI: 10.1109/LCOMM.2019.2956709
- [12] SINGH P, MISHRA H B, BUDHIRAJA R. Low-complexity linear MIMO-OTFS receivers [C]//2021 IEEE International Conference on Communications Workshops (ICC Workshops). Montreal, Canada: IEEE, 2021: 1 - 6. DOI: 10.1109/ICCWorkshops50388.2021.9473839
- [13] LI S Y, YUAN W J, WEI Z Q, et al. Cross domain iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, 2021, (99): 1. DOI: 10.1109/TWC.2021.3110125
- [14] THAJ T, VITERBO E. Low complexity iterative rake decision feedback equalizer for zero-padded OTFS systems [J]. IEEE transactions on vehicular technology, 2020, 69(12): 15606 - 15622. DOI: 10.1109/TVT.2020.3044276
- [15] KSCHISCHANG F R, FREY B J, LOELIGER H A. Factor graphs and the sum-product algorithm [J]. IEEE transactions on information theory, 2001, 47(2): 498 - 519. DOI: 10.1109/18.910572
- [16] LI H, DONG Y Y, GONG C H, et al. Low complexity receiver via expectation propagation for OTFS modulation [J]. IEEE communications letters, 2021, 25(10): 3180 - 3184. DOI: 10.1109/LCOMM.2021.3101827
- [17] YUAN W J, WEI Z Q, YUAN J H, et al. A simple variational Bayes detector for orthogonal time frequency space (OTFS) modulation [J]. IEEE transactions on vehicular technology, 2020, 69(7): 7976 - 7980. DOI: 10.1109/TVT.2020.2991443
- [18] ZHANG H J, ZHANG T T. A low-complexity message passing detector for OTFS modulation with probability clipping [J]. IEEE wireless communications letters, 2021, 10(6): 1271 - 1275. DOI: 10.1109/LWC.2021.3063904
- [19] TIWARI S, DAS S S, RANGAMGARI V. Low complexity LMMSE Receiver for OTFS [J]. IEEE communications letters, 2019, 23(12): 2205 - 2209. DOI: 10.1109/LCOMM.2019.2945564
- [20] RAVITEJA P, VITERBO E, HONG Y. OTFS performance on static multipath channels [J]. IEEE wireless communications letters, 2019, 8(3): 745 - 748. DOI: 10.1109/LWC.2018.2890643

- [21] DONOHO D L, MALEKI A, MONTANARI A. Message passing algorithms for compressed sensing: motivation and construction [C]//2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo). Cairo, Egypt: IEEE, 2010: 1 – 5. DOI: 10.1109/ITWKSPPS.2010.5503193
- [22] DONOHO D L, MALEKI A, MONTANARI A. Message passing algorithms for compressed sensing: analysis and validation [C]//2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo). Cairo, Egypt: IEEE, 2010: 1 – 5. DOI: 10.1109/ITWKSPPS.2010.5503228
- [23] WINN J, BISHOP C M. Variational message passing [J]. Journal of machine learning research, 2005, 6(4): 661 – 694.
- [24] MONK A, HADANI R, TSATSANIS M, et al. OTFS - Orthogonal Time Frequency Space [EB-OL]. [2021-10-10]. <https://arxiv.org/abs/1608.02993v1>.
- [25] YUAN Z D, LIU F, YUAN W J, et al. Iterative detection for orthogonal time frequency space modulation with unitary approximate message passing [J]. IEEE transactions on wireless communications, 2021. DOI: 10.1109/TWC.2021.3097173
- [26] LIU F, YUAN Z D, GUO Q H, WANG Z Y, et al. Multi-block UAMP based detection for OTFS with rectangular waveform [J]. IEEE wireless communications letters, 2021. DOI: 10.1109/LWC.2021.3126871
- [27] GUO Q H, XI J T. Approximate message passing with unitary transformation [EB/OL]. [2021-10-10]. <https://arxiv.org/abs/1504.04799>
- [28] YUAN Z D, GUO Q H, LUO M. Approximate message passing with unitary transformation for robust bilinear recovery [J]. IEEE transactions on signal processing, 2021, 69: 617 – 630. DOI: 10.1109/TSP.2020.3044847
- [29] LUO M, GUO Q H, JIN M, et al. Unitary approximate message passing for sparse Bayesian learning [J]. IEEE transactions on signal processing, 2021, 69: 6023 – 6039. DOI: 10.1109/TSP.2021.3114985
- [30] TUCHLER M, SINGER A C, KOETTER R. Minimum mean squared error equalization using a priori information [J]. IEEE transactions on signal processing, 2002, 50(3): 673 – 683. DOI: 10.1109/78.984761
- [31] GUO Q H, PING L. LMMSE turbo equalization based on factor graphs [J]. IEEE journal on selected areas in communications, 2008, 26(2): 311 – 319. DOI: 10.1109/JSAC.2008.080208
- [32] GUO Q H, HUANG D D. A concise representation for the soft-in soft-out LMMSE detector [J]. IEEE communications letters, 2011, 15(5): 566 – 568. DOI: 10.1109/LCOMM.2011.032811.102073
- [33] LIU F, YUAN Z D, GUO Q H, et al. Message passing based structured sparse signal recovery for estimation of OTFS channels with fractional Doppler shifts [J]. IEEE transactions on wireless communications, 2021. DOI: 10.1109/TWC.2021.3087501

Biographies

YUAN Zhengdao received the B.E. degree in communication and information system from Henan University of Science and Technology, China in 2006, the M.E. degree in communication engineering from Soochow University, China in 2009, and the Ph.D. degree in information and communication engineering from the National Digital Switching System Engineering and Technological Research Center, China in 2018. He is currently an associate professor with the Open University of Henan. He was a visiting scholar with the University of Wollongong, Australia in 2019. His research interests are mainly in massive MIMO, sparse channel estimation, message passing algorithm, and iterative receiver.

LIU Fei received the B.E. and M.E. degrees in information and communication engineering from Zhengzhou University, China in 2015 and 2017, respectively. He is currently working toward the Ph.D. degree with the School of Information and Engineering, Zhengzhou University, China. His research interests are message passing algorithm, sparse signal recovery, and OTFS.

GUO Qinghua (qguo@uow.edu.au) received the B.E. degree in electronic engineering and the M.E. degree in signal and information processing from Xidian University, China in 2001 and 2004, respectively, and the Ph.D. degree in electronic engineering from the City University of Hong Kong, China in 2008. He is currently an associate professor with the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Australia and an adjunct associate professor with the School of Engineering, The University of Western Australia. His research interests include signal processing, machine learning and telecommunications. He was a recipient of the Australian Research Council's inaugural Discovery Early Career Researcher Award in 2012.

WANG Zhongyong received the B.S. and M.S. degrees in automatic control from Harbin Shipbuilding Engineering Institute, China in 1986 and 1988, respectively, and the Ph.D. degree in automatic control theory and application from Xi'an Jiaotong University, China in 1998. From 1988 to 2002, he has been with the Department of Electronics, Zhengzhou University, China. Now he is a professor with the Department of Communication Engineering, Zhengzhou University. His research interests include numerous aspects within embedded systems, signal processing, and communication theory.



Performance of LDPC Coded OTFS Systems over High Mobility Channels

ZHANG Chong^{1,2,3,4}, XING Wang^{1,2,3,4}, YUAN Jinhong⁵, ZHOU Yiqing^{1,2,3,4}

(1. Beijing Key Laboratory of Mobile Computing and Pervasive Device, Beijing 100190, China;

2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;

3. University of Chinese Academy of Sciences, Beijing 100049, China;

4. Zhongke Nanjing Mobile Communication & Computing Innovation Institute, Nanjing 211135, China;

5. School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney NSW 2052, Australia)

Abstract: The upcoming 6G wireless networks have to provide reliable communications in high-mobility scenarios at high carrier frequencies. However, high-mobility or high carrier frequencies will bring severe inter-carrier interference (ICI) to conventional orthogonal frequency-division multiplexing (OFDM) modulation. Orthogonal time frequency space (OTFS) modulation is a recently developing multi-carrier transmission scheme for wireless communications in high-mobility environments. This paper evaluates the performance of coded OTFS systems. In particular, we consider 5G low density parity check (LDPC) codes for OTFS systems based on 5G OFDM frame structures over high mobility channels. We show the performance of the OTFS systems with 5G LDPC codes when sum-product detection algorithm and iterative detection and decoding are employed. We also illustrate the effect of channel estimation error on the performance of the LDPC coded OTFS systems.

Keywords: OTFS; LDPC codes; OFDM

DOI: 10.12142/ZTECOM.202104005

<https://kns.cnki.net/kcms/detail/34.1294.TN.20211210.1815.002.html>, published online December 14, 2021

Manuscript received: 2021-10-29

Citation (IEEE Format): C. Zhang, W. Xing, J. H. Yuan, et al., "Performance of LDPC coded OTFS systems over high mobility channels," *ZTE Communications*, vol. 19, no. 4, pp. 45 – 53, Dec. 2021. doi: 10.12142/ZTECOM.202104005.

1 Introduction

Future wireless networks are expected to provide high-speed and ultra-reliable communication^[1-6] for a number of emerging wireless applications, such as the millimeter wave (mmWave)^[7], low-earth-orbit satellites (LEOSs)^[8], high-speed trains^[9], and unmanned aerial vehicles (UAVs)^[10]. In practice, one of the challenges for these systems is that wireless communication for high mobility is always accompanied by severe Doppler spread. While orthogonal frequency division multiplexing (OFDM) currently deployed in Long Term Evolution (LTE) and 5G systems can achieve high spectral efficiency for time-varying frequency selective channels, it suffers from heavy performance degradation in high Doppler conditions due to severe Doppler spread. Therefore, new modulation and signal processing techniques that are

more robust to time-varying channels are required to meet the challenging requirements of high mobility communications.

Recently, the orthogonal time frequency space (OTFS) modulation proposed in Refs. [11 – 12] has drawn a lot of attention due to its advantages over OFDM in high Doppler channels where each transmitted data symbol can exploit the channel's delay and Doppler variations. OTFS is a two-dimensional (2D) modulation scheme, where the information symbols are modulated into the 2D delay-Doppler (DD) domain rather than into the time-frequency (TF) domain as the classic OFDM modulation. Thus, the time-varying channel in the TF domain can be transformed into a 2D quasi-time-invariant channel in the DD domain, where attractive properties, such as separability, stability, and compactness, can be exploited^[12]. With the application of inverse symplectic finite Fourier transform (ISFFT) in OTFS, symbols in the DD domain can be transformed to the TF domain, in which each symbol in the DD domain spans the whole bandwidth and time duration of the transmission frame. Therefore, the OTFS modulation can offer the potential of ex-

This paper was partially supported by National Key R&D Program of China (No. 2020YFB1807802) the National Science Fund for Distinguished Young Scholars (No. 61901453), and Jiangsu Provincial Key Research and Development Program (No. BE2021013-2). ZHOU Yiqing is the corresponding author.

exploiting the full diversity. However, modulating information symbols into the DD domain makes conventional receiver technologies cannot be applied directly, which in turn requires advanced channel coding, signal estimation and detection methods for OTFS to achieve the potential full diversity and reliable error performance.

Accurately estimating the channel parameters in OTFS systems is a challenging but vital requirement for reliable detection. A simple channel estimation algorithm for OTFS is to embed the pilot signals in the DD domain^[13-14] to obtain DD channel responses of each transmitting path. This method has low complexity and signal overhead. But it suffers from performance degradation for channels with fractional Doppler. In Ref. [15], the authors proposed a channel estimation method based on pseudo-noise (PN) to estimate the Doppler frequency of each transmitting path, which has high computational complexity to estimate the fractional Doppler. Using deep learning algorithms to assist in the channel estimation has recently been proposed as a potential technology to get more accurate estimation parameters. For example, in Ref. [16], two deep learning assisted algorithms were proposed to facilitate channel estimation for massive machine-type communication, which could potentially be applied to OTFS channel estimation to deal with fractional Doppler.

Many existing studies focus on signal detection for OTFS modulation. In Ref. [17], the authors gave an iterative detection based on the message passing (MP) algorithm. However, the MP detection treats the interference from other information symbols as a Gaussian variable to reduce the detection complexity, which may fail to converge and result in performance degradation. To solve this problem, the authors in Refs. [18-19] explored the approximate message passing algorithm. They proposed a detection algorithm by covariance processing to obtain better BER performance in Ref. [18] and a convergence guaranteed receiver based on the variational Bayes framework in Ref. [19]. An OTFS detection approach based on approximate message passing (AMP) with a unitary AMP transformation (UAMP) was developed in Ref. [20], which enjoys the structure of the channel matrix and allows efficient implementation. Inspired by the connection between the orthogonality and message passing, the authors in Ref. [21] proposed a novel cross domain iterative detection algorithm for OTFS modulation, where the extrinsic information is passed between the time domain and DD domain via the corresponding unitary transformations. This algorithm has low computational complexity and can achieve almost the same error performance as the maximum-likelihood sequence detection even in the presence of fractional Doppler shifts.

Most papers focus on the basic principle or key algorithms for the estimation or detection of OTFS systems, while little attention has been paid to coded OTFS systems. The authors in Ref. [22] analyzed coded OTFS performance, but they only considered classical codes, ignoring modern coding technologies like LDPC, Turbo,

or Polar codes for OTFS systems, which are important to achieve high reliability and channel capacity. Therefore, we are going to evaluate modern LDPC coded OTFS performance in this work.

In this paper, we first provide a brief overview of the fundamental concepts of OTFS. Then we consider 5G LDPC codes for OTFS systems over high mobility channels. As OTFS can be built on the top of the OFDM frame structure, we also consider the 5G OFDM frame structure in our evaluations. The effect of channel estimation errors on the LDPC coded OTFS system performance is evaluated as well. The rest of this article is organized as follows. Section 2 introduces related knowledge of OTFS including the basic conceptions, system model, OTFS modulation/demodulation, and the channel input-output relationship. In Section 3, we provide an overview of the error performance analysis for OTFS systems. In Section 4, we present the details of the performance evaluation of OTFS systems with 5G LDPC codes. Section 5 concludes the paper.

2 OTFS System

In this section, we first present an OTFS system model. Then, we briefly review the input-output relations for the OTFS systems^[11-12,17].

2.1 OTFS System Model

An OTFS modulation is a 2D modulation, which modulates information in the DD domain before transforming signals to TF and time domains. The time-frequency signal plane is discretized to a grid by sampling time and frequency axes at intervals T (s) and Δf (Hz), respectively, i.e.,

$$\Lambda = \left\{ (nT, m\Delta f), n = 0, \dots, N-1, m = 0, \dots, M-1 \right\}. \quad (1)$$

For some integers $N, M > 0$, where $T = 1/\Delta f$, N and M are the numbers of time and frequency grids, respectively. Therefore, we can obtain that an OTFS frame is transmitted with duration $T_{\text{frame}} = NT$ and occupies a bandwidth $B_{\text{frame}} = M\Delta f$.

Accordingly, let us define the delay-Doppler plane as,

$$\Gamma = \left\{ \left(\frac{k}{NT}, \frac{l}{M\Delta f} \right), k = 0, \dots, N-1, l = 0, \dots, M-1 \right\}, \quad (2)$$

where $\frac{1}{M\Delta f}$ and $\frac{1}{NT}$ represent the quantization steps of the delay and Doppler frequency, respectively. They are also called delay and Doppler resolution.

Let a set $\mathbb{A} = \{a_1, \dots, a_Q\}$ denote the constellation alphabet with sized Q and a set of NM modulated symbols $x_0, \dots, x_{NM-1} \in \mathbb{A}$ to be transmitted by OTFS. The OTFS modulation firstly maps symbols x_0, \dots, x_{NM-1} to DD plane Γ . Here, we use $x[k, l]$, $k = 0, \dots, N-1, l = 0, \dots, M-1$ to represent the baseband modulated symbols to be transmitted in DD plane, where k

denotes Doppler index and l denotes delay index.

Then, the signal in DD plane Γ is transformed to TF plane Λ , whereby the symbols in DD domain $x[k, l], k = 0, \dots, N-1, l = 0, \dots, M-1$ are mapped into TF domain $X[n, m], n = 0, \dots, N-1, m = 0, \dots, M-1$, where n denotes the time index and m denotes the frequency index. Such mapping of $x[k, l]$ to $X[n, m]$ can be realized by ISFFT^[11].

$$X[n, m] = \text{ISFFT}(x[k, l]) = \frac{1}{\sqrt{NM}} \sum_{k=0}^{N-1} \sum_{l=0}^{M-1} x[k, l] e^{j2\pi \left(\frac{nk}{N} - \frac{ml}{M} \right)}. \quad (3)$$

Next, a multi-carrier modulator, such as OFDM, is used to transform the samples $X[n, m]$ at each time slot to a continuous time waveform $s(t)$ with a pulse shaping $g_{tx}(t)$ as the transmitted pulse. Such a transformation can be realized by discrete Heisenberg transform^[11],

$$s(t) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X[n, m] e^{j2\pi m \Delta f (t - nT)} g_{tx}(t - nT). \quad (4)$$

The cross-ambiguity function between the transmitted pulse shaping $g_{tx}(t)$ and the received pulse shaping $g_{rx}(t)$ is given by^[11]

$$A_{g_{rx}, g_{tx}}(t, f) = \int g_{rx}^*(t' - t) g_{tx}(t') e^{-j2\pi f(t' - t)} dt'. \quad (5)$$

Here, we only discuss OTFS modulation/demodulation principles without channel effects. Let $r(t)$ denote the received signal. At the receiver, a match filter is applied to compute the cross-ambiguity function $A_{g_{rx}, r}(t, f)$ as

$$Y(t, f) = A_{g_{rx}, r}(t, f) \triangleq \int g_{rx}^*(t' - t) r(t') e^{-j2\pi f(t' - t)} dt'. \quad (6)$$

By sampling $Y(t, f)$ as $Y[n, m] = Y(t, f)|_{t=nT, f=m\Delta f}$, $Y[n, m], n = 0, \dots, N-1, m = 0, \dots, M-1$ can be obtained.

Then, by transforming TF plane Λ signal $Y[n, m]$ back to DD plane Γ signal $y[k, l]$ with symplectic finite Fourier transform (SFFT), symbols in DD domain can be recovered as^[11]

$$y[k, l] = \text{SFFT}(Y[n, m]) = \frac{1}{\sqrt{NM}} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} Y[n, m] e^{-j2\pi \left(\frac{nk}{N} - \frac{ml}{M} \right)}. \quad (7)$$

2.2 Input-Output Relations of OTFS Signal

In this subsection, we present the channel model and the input-output relations between the transmitter and the receiver^[17]. We assume that only a few reflectors are moving within one OTFS frame duration in practice, and only a small number of channel taps are associated with Doppler shift. Therefore, the channel response of the DD domain is sparse compared

with the whole DD plane. Considering a channel with P independent distinguishable paths, the channel response can be represented as

$$h(\tau, \nu) = \sum_{i=1}^P h_i \delta(\tau - \tau_i) \delta(\nu - \nu_i), \quad (8)$$

where h_i , τ_i and ν_i denote the channel coefficient, delay and Doppler shift associated with the i -th path, respectively. Without considering additive white Gaussian noise (AWGN), the relation of $s(t)$ and $r(t)$ is given by^[17]

$$r(t) = \iint h(\tau, \nu) e^{j2\pi \nu(t - \tau)} s(t - \tau) d\tau d\nu. \quad (9)$$

The relation of $X[n, m]$ and $Y[n, m]$ in TF domain is given by^[17]

$$Y[n, m] = \sum_{n'=n-1}^n \sum_{m'=0}^{M-1} H_{n,m}[n', m'] X[n', m'] = H_{n,m}[n, m] X[n, m] + \sum_{m'=0, m' \neq m}^{M-1} H_{n,m}[n, m'] X[n, m'] + \sum_{m'=0}^{M-1} H_{n,m}[n-1, m'] X[n-1, m'], \quad (10)$$

in which

$$H_{n,m}[n', m'] = \iint h(\tau, \nu) A_{g_{rx}, g_{tx}}((n - n')T - \tau, (m - m')\Delta f - \nu) e^{j2\pi \nu n' T} \times e^{j2\pi (\nu + m'\Delta f)((n - n')T - \tau)} d\tau d\nu. \quad (11)$$

The second term in Eq. (10) is the samples $X[n, m']$ at different frequencies $m' \neq m$, which can be seen as the interference to the current sample $X[n, m]$ in the same time slot n . On the other hand, the third term in Eq. (10) accumulates the interference from the samples $X[n-1, m']$ in the previous time slot $n-1$. Hence, we call the second and third terms as the inter-carrier interference (ICI) and inter-symbol interference (ISI), respectively. It is clear that since the delay spread and Doppler spread, there are severe ICI and ISI in the TF domain.

The relation of $x[k, l]$ and $y[k, l]$ in DD domain is given by^[17]

$$y[k, l] \approx \sum_{i=1}^P \sum_{q=-N_i}^{N_i} h_i e^{j2\pi \left(\frac{l - l_{\tau_i}}{M} \right) \left(\frac{k_{\nu_i} + \kappa_{\nu_i}}{N} \right)} \alpha_i(k, l, q) x \left[\left[k - k_{\nu_i} + q \right]_N, \left[l - l_{\tau_i} \right]_M \right], \quad (12)$$

$$\alpha_i(k, l, q) = \begin{cases} \frac{1}{N} \beta_i(q), & l_{\tau_i} \leq l < M \\ \frac{1}{N} (\beta_i(q) - 1) e^{-j2\pi \frac{[k - k_{\nu_i} + q]_N}{N}}, & 0 \leq l < l_{\tau_i}, \end{cases}$$

where

$$\beta_i(q) = \frac{e^{-j2\pi(-q-\kappa_{\nu_i})} - 1}{e^{-j\frac{2\pi}{N}(-q-\kappa_{\nu_i})} - 1}$$

and

In Eq. (12), h_i denotes the channel coefficient and N_i is the number of neighboring transmitted signals in the Doppler domain, where $0 < N_i \ll N$. The l_{τ_i} and k_{ν_i} are the delay and Doppler indices corresponding to the i -th path, respectively, and we have

$$\tau_i = \frac{l_{\tau_i}}{M\Delta f}, \quad \nu_i = \frac{k_{\nu_i} + \kappa_{\nu_i}}{NT}. \quad (13)$$

Note that the term $-\frac{1}{2} < \kappa_{\nu_i} \leq \frac{1}{2}$ denotes the fractional Doppler shifts which correspond to the fractional shifts from the nearest Doppler indices^[17]. For wideband systems, the typical value of the sampling time in the delay domain is usually sufficiently small. Therefore, the impact of fractional delays in typical wide-band systems can be neglected^[24]. Here, we have $-k_{\max} \leq k_i + \kappa_i \leq k_{\max}$, where k_{\max} is the maximum Doppler index satisfying $k_{\max} = \lceil NT\nu_{\max} \rceil$ ^[17].

From Eq. (12), we can see that the received signal $y[k, l]$ is a linear combination of $S = \sum_{i=1}^P 2N_i + 1$ transmitted signals.

The signal corresponding to $q = 0$, $x \left[[k - k_{\nu_i}]_N, [l - l_{\tau_i}]_M \right]$ contributes the most, and all the other $2N_i$ signals can be seen as interference. Such interference is due to the transmitted signals neighboring $x \left[[k - k_{\nu_i}]_N, [l - l_{\tau_i}]_M \right]$ in the Doppler domain caused by fractional Doppler and we refer to this interference as inter-Doppler interference (IDI). In the TF domain, the delay spread and Doppler spread cause severe ICI and ISI as the second and the third terms in Eq. (10), which are hard to distinguish in the received signals. However, in the DD domain, they mainly affect the phase shifts as $e^{j2\pi \left(\frac{l - l_{\tau_i}}{M} \right) \left(\frac{k_{\nu_i} + \kappa_{\nu_i}}{N} \right)}$.

Now, let us represent the DD domain input-output relation in a vector form. Let $\mathbf{x} \triangleq \text{vec}(\mathbf{X}) \in \mathbb{A}^{MN}$ and $\mathbf{y} \triangleq \text{vec}(\mathbf{Y}) \in \mathbb{A}^{MN}$ denote the vector forms of the transmitted symbols X and the received symbols Y in DD domain, respectively. According to Eq. (9), we have

$$\mathbf{y} = \mathbf{H}_{\text{eff}} \mathbf{x} + \mathbf{w}, \quad (14)$$

where \mathbf{w} is the corresponding noise vector and \mathbf{H}_{eff} of size $MN \times MN$ is the effective channel matrix in the DD domain. Assuming that both $g_{tx}(t)$ and $g_{rx}(t)$ are rectangular pulses, with a reduced CP frame format, the effective channel matrix \mathbf{H}_{eff} is given by^[23]

$$\mathbf{H}_{\text{eff}} = \sum_{i=1}^P h_i (\mathbf{F}_N \otimes \mathbf{I}_M) \mathbf{\Pi}^{l_{\tau_i}} \Delta^{(k_{\nu_i} + \kappa_{\nu_i})} (\mathbf{F}_N^H \otimes \mathbf{I}_M), \quad (15)$$

where $\mathbf{\Pi}$ is the permutation matrix (forward cyclic shift), i.e.,

$$\mathbf{\Pi} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & \ddots & \ddots & \ddots & 0 \\ \vdots & 1 & 0 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{MN \times MN}, \quad (16)$$

and Δ is a diagonal matrix,

$$\Delta = \text{diag} \{ e^{j2\pi \frac{0}{NM}}, e^{j2\pi \frac{1}{NM}}, \dots, e^{j2\pi \frac{NM-1}{NM}} \}. \quad (17)$$

3 Error Performance Analysis

In this section, we discuss the error performance of the uncoded and coded OTFS systems, respectively. We assume that ideal channel state information (CSI) is available at the receiver.

3.1 Uncoded OTFS System Performance

According to Ref. [26], Eq. (14) can be rewritten as

$$\mathbf{y} = \mathbf{\Phi}_{\tau,\nu}(\mathbf{x}) \mathbf{h} + \mathbf{w}, \quad (18)$$

where $\mathbf{\Phi}_{\tau,\nu}(\mathbf{x})$ is referred to as the equivalent code-word matrix and it is a concatenated matrix of size $MN \times P$ constructed by the column vector $\mathbf{\Xi}_i \mathbf{x}$, i.e.,

$$\mathbf{\Phi}_{\tau,\nu}(\mathbf{x}) = [\mathbf{\Xi}_1 \mathbf{x}, \mathbf{\Xi}_2 \mathbf{x}, \dots, \mathbf{\Xi}_P \mathbf{x}], \quad (19)$$

and $\mathbf{\Xi}_i$ is given by

$$\mathbf{\Xi}_i \triangleq (\mathbf{F}_N \otimes \mathbf{I}_M) \mathbf{\Pi}^{l_{\tau_i}} \Delta^{(k_{\nu_i} + \kappa_{\nu_i})} (\mathbf{F}_N^H \otimes \mathbf{I}_M), 1 \leq i \leq P. \quad (20)$$

In Eq. (18), \mathbf{h} is the channel coefficient vector of size $P \times 1$, i.e., $\mathbf{h} = [h_1, h_2, \dots, h_P]^T$, where the elements in \mathbf{h} are assumed to be independently and identically distributed complex Gaussian random variables. Besides, we assume a uniform power delay and Doppler profile of the channel so that the channel coefficient h_i has mean μ and variance $1/2P$ per real dimension for $1 \leq i \leq P$ and is independent from the delay and Doppler indices^[25]. In particular, we note that if $\mu = 0$, h_i follows the Rayleigh distribution, which will be considered as a special case in the error performance analysis and code design.

Based on Eq. (18), for a given channel realization, we define the conditional Euclidean distance $d_{h,\tau,\nu}^2(\mathbf{x}, \mathbf{x}')$ between a pair of code-words \mathbf{x} and $\mathbf{x}' (\mathbf{x} \neq \mathbf{x}')$ as

$$d_{h,\tau,\nu}^2(\mathbf{x}, \mathbf{x}') = d_{h,\tau,\nu}^2(\mathbf{e}) \triangleq \|\mathbf{\Phi}_{\tau,\nu}(\mathbf{e}) \mathbf{h}\|^2 = \mathbf{h}^H \mathbf{\Omega}_{\tau,\nu}(\mathbf{e}) \mathbf{h}, \quad (21)$$

where $\mathbf{e} = \mathbf{x} - \mathbf{x}'$ is the corresponding code-word difference (error) sequence and $\mathbf{\Omega}_{\tau,\nu}(\mathbf{e}) = (\mathbf{\Phi}_{\tau,\nu}(\mathbf{e}))^H (\mathbf{\Phi}_{\tau,\nu}(\mathbf{e}))$ is referred to as the code-word difference matrix. Here we have

$$\mathbf{\Omega}_{\tau,\nu}(\mathbf{e}) = \begin{bmatrix} \mathbf{e}^H \Xi_1^H \Xi_1 \mathbf{e} & \mathbf{e}^H \Xi_1^H \Xi_2 \mathbf{e} & \cdots & \mathbf{e}^H \Xi_1^H \Xi_P \mathbf{e} \\ \mathbf{e}^H \Xi_2^H \Xi_1 \mathbf{e} & \mathbf{e}^H \Xi_2^H \Xi_2 \mathbf{e} & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{e}^H \Xi_P^H \Xi_1 \mathbf{e} & \cdots & \cdots & \mathbf{e}^H \Xi_P^H \Xi_P \mathbf{e} \end{bmatrix}. \quad (22)$$

Note that the code-word difference matrix $\mathbf{\Omega}(\mathbf{e})$ is positive semidefinite Hermitian with a rank r , where $r \leq P$. Let us denote by $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_P\}$ the eigenvectors of $\mathbf{\Omega}(\mathbf{e})$ and $\{\lambda_1, \lambda_2, \dots, \lambda_P\}$ the corresponding nonnegative real eigenvalues sorted in the descending order, where $\lambda_i > 0$ for $1 \leq i \leq r$ and $\lambda_i = 0$ for $r+1 \leq i \leq P$. Thus, the conditional pairwise-error probability (PEP)^[27-28] is upper-bounded by

$$\Pr(\mathbf{x}, \mathbf{x}' | \mathbf{h}, \tau, \nu) \leq \exp\left(-\frac{E_s}{4N_0} \sum_{i=1}^r \lambda_i |\tilde{h}_i|^2\right), \quad (23)$$

where E_s is the average symbol energy and $\tilde{h}_i = \mathbf{h}_i \cdot \mathbf{v}_i$, for $1 \leq i \leq r$. It can be shown that $\{\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_r\}$ are independent complex Gaussian random variables with mean $\mu_{\tilde{h}_i} = \mathbb{E}[\mathbf{h}] \cdot \mathbf{v}_i$ and $1/2P$ variance per real dimension. It has been defined in the previous work^[26, 29-30] that the rank of $\mathbf{\Omega}(\mathbf{e})$ is the diversity gain of the uncoded OTFS system. Specifically, it has been shown in Ref. [30] that the diversity gain of uncoded OTFS modulation systems can be one but the full diversity can be obtained by suitable precoding schemes. Furthermore, Ref. [26] has shown that the full diversity can be achieved almost surely for the case of $P = 2$ when the frame size is sufficiently large, even for uncoded OTFS modulation systems.

3.2 Coded OTFS System Performance

Based on the previous analysis, Ref. [22] gave the error performance of the coded OTFS systems. With the assumptions of the wide-sense stationary-uncorrelated scattering (WSSUS) channel and Rayleigh fading, Eq. (23) can be further simplified as

$$\Pr(\mathbf{x}, \mathbf{x}' | \tau, \nu) \leq \left(\prod_{i=1}^r \lambda_i / P\right)^{-1} \left(\frac{E_s}{4N_0}\right)^{-r} = \frac{1}{\prod_{i=1}^r \lambda_i} \left(\frac{E_s}{4N_0 P}\right)^{-r}, \quad (24)$$

and it is approximated by

$$\Pr(\mathbf{x}, \mathbf{x}') \approx \left(\frac{d_E^2(\mathbf{e})}{P}\right)^{-r} \left(\frac{E_s}{4N_0}\right)^{-r}. \quad (25)$$

Based on Eq. (25), we note that the unconditional PEP for OTFS modulation only depends on $d_E^2(\mathbf{e})$, the rank of $\mathbf{\Omega}(\mathbf{e})$, and the number of independent resolvable paths P , and is inde-

pendent of the specific distribution of delay and Doppler indices. The power of the signal-to noise ratio (SNR) is referred to as the diversity gain, and the term $d_E^2(\mathbf{e})/P$ is referred to as the coding gain, which characterizes the approximate improvement of coded OTFS systems over the uncoded counterpart with the same diversity gain, i.e., the same exponent $-r$ ^[27]. Considering the total diversity, there exists a fundamental trade-off between the diversity gain and the coding gain. Based on the previous work^[26, 29-30], we can notice that the diversity gain depends on the number of independent resolvable paths P . When P (the rank of $\mathbf{\Omega}(\mathbf{e})$) is small, the diversity gain is small. It is crucial for OTFS systems that using an optimized channel code can greatly improve the error performance. For a large value of P and a reasonably high SNR, the unconditional PEP of OTFS systems can be approximately upper-bounded^[22] by

$$\Pr(\mathbf{x}, \mathbf{x}') \leq \exp\left(-\frac{E_s}{16N_0} d_E^2(\mathbf{e})\right). \quad (26)$$

Form Eq. (26), we can notice that the channel with a large number of diversity paths approaches an AWGN model^[35], which indicates that it is reasonable to use coding gain approximation for AWGN channels for evaluating the coding gain of OTFS systems with a large P . A preliminary guideline for the code design of the OTFS systems is to maximize the minimum value of $d_E^2(\mathbf{e})$ among all pairs of code-words of the code. Note that, the error performance of coded OTFS systems still depends on the channel parameters, which is widely observed in the system design for fading channels^[36].

4 Evaluation of Coded OTFS with 5G LDPC Codes

In this section, we provide the performance evaluation of the coded OTFS with 5G LDPC codes. Without loss of generality, we consider the sum-product algorithm (SPA)^[31] or message passing algorithm for detection, where the details can be found in Refs. [17, 32]. In specific, we consider 5 MHz channel bandwidth¹ and 15 kHz sub-carrier for a time slot. Followed by a 5G frame structure, 5 MHz channel bandwidth contains 4.5 MHz efficient bandwidth with 300 sub-carriers and a time slot contains 14 OFDM symbols^[33], where $N = 14$ and $M = 300$. In all simulations, we consider the Rayleigh fading case. If not otherwise specified, we only consider the integer delay and Doppler case and set the maximum delay index as $l_{\max} = 4$ and the maximum Doppler index as $k_{\max} = 4$. For each channel realization, we randomly select the delay and Doppler indices according to the uniform distribution, so that we have $-k_{\max} \leq k_{\nu_i} \leq k_{\max}$ and $0 \leq l_{\tau_i} \leq l_{\max}$. We first present the performance for uncoded OTFS systems with SPA detec-

¹ It contains efficient bandwidth and guard band in the 5G frame structure.

tion. Then, we discuss the performance of coded OTFS systems with 5G LDPC codes.

4.1 Uncoded OTFS System via SPA Detection

In Fig. 1, it shows the receiver of the uncoded OTFS system via the iterative SPA detection. The SPA detector needs to iterate the soft information between the detector and the demodulator.

The block-error-rate (BLER) performance of the uncoded OTFS systems with $P = 4$ is shown in Fig. 2. Here, we compare the BLER performance under different iterations. In Fig. 2, "ite $\times n$ " means n iterations in the iterative detections. It can be observed that SPA detection can achieve better performance with iterations. For BPSK modulation, increasing the iteration number over two does not provide much improvement. The same phenomenon for QPSK modulation is observed for over four iterations. We notice that there is a trade-off between the numbers of iterations and BLER performance. For higher order modulation, we need more iterations for better performance.

4.2 Coded OTFS System via SPA Detection

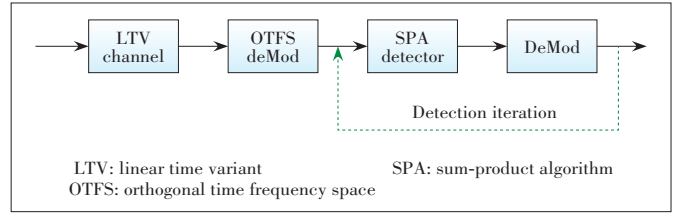
Here, we provide performance evaluation of coded OTFS based on 5G LDPC codes^[34]. The decoder uses the sum-product decoding algorithm with a floating value, and the maximum decoding iteration number is 50. We show the coded OTFS performance for various modulation and different code lengths.

Fig. 3 shows the receiver of the coded OTFS system via iterative SPA detection. We firstly only consider iterative detection in the coded OTFS system. The received symbols after the OTFS demodulation will be put into the SPA detector. The SPA detector gives soft bit messages corresponding to the received symbols to the demodulator. Soft bit messages after demodulation will back to the SPA detector via the detection iteration loop. After several iterations, soft bit messages will be sent to the decoder for decoding. Compared with Fig. 1, Fig. 3 adds the function of coding and decoding.

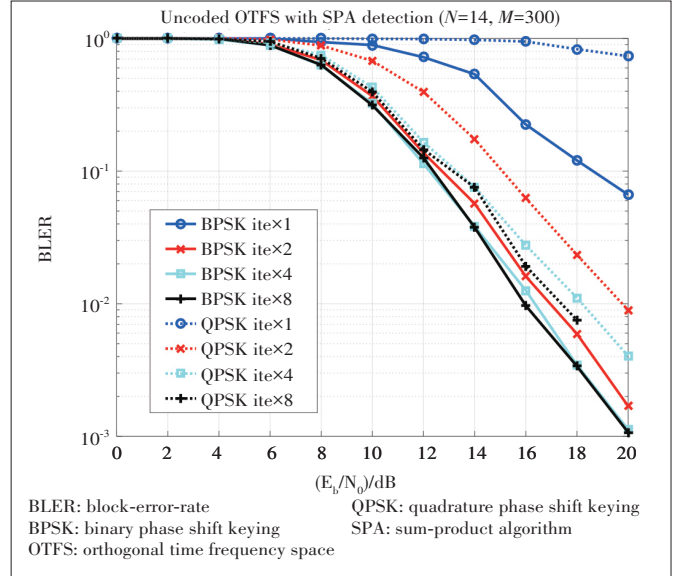
Fig. 4 shows the BLER performance of the OTFS systems with $P = 4$. We list the parameters of the codes we use in Table 1. Here, K is the information bit length and R is the code rate. We can observe that for the BPSK modulation, increasing the iteration number does not provide much improvement. For the QPSK modulation, two iterations can obtain better performance. For 16QAM modulation, four iterations can obtain larger performance gain. We notice that there exists an optimal iteration number for different modulations.

Next, we only consider the decoding iteration shown in Fig. 5, where we can notice that the detection iteration loop is replaced by decoding iteration loop. Here, the soft bit messages after demodulation will be straightly sent to the decoder. After decoding, the soft bit messages will be passed to the SPA detector via the decoding iteration loop.

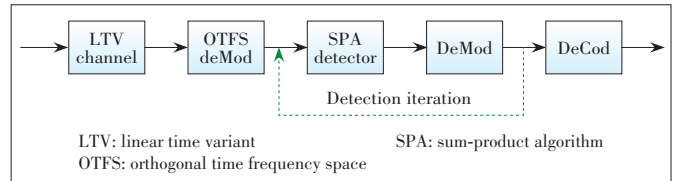
We compare the BLER performance with iterative detection



▲ Figure 1. An illustration of uncoded OTFS system via iterative SPA detection



▲ Figure 2. BLER performance of the uncoded OTFS systems with $P=4$



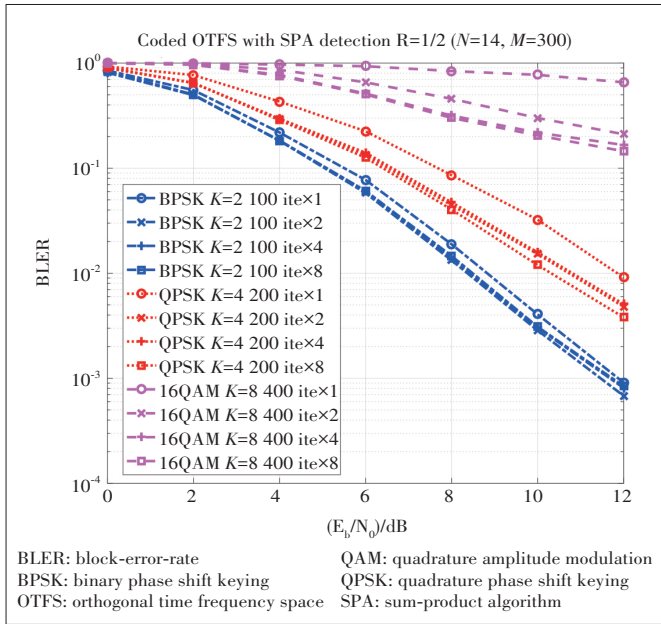
▲ Figure 3. An illustration of coded OTFS system via iterative SPA detection

only and with iterative detection-decoding in Figs. 6 and 7. For the BPSK modulation in Fig. 6, the performances are similar.

For the QPSK modulation in Fig. 7, we notice that the performance of the iterative detection is better than that of the iterative detection-decoding.

Here, conducting two and four iterative detections outperforms the iterative decoding at BLER 10^{-2} , with about 0.8 dB and 0.6 dB more gain, respectively. We can notice here that conducting iterative detections can improve the performance more than only iterating between decoding and detection.

Now, we consider the hybrid iterative decoding and detection. In Fig. 8, we have two iteration links. For detection iteration loop (we call it the inner iterative layer), we do SPA detection and demodulation as shown in Fig. 3. For decoding iteration loop (we call it the outer iterative layer), we do decoding and SPA detection as shown in Fig. 5. After several inner iterations, the demodulator sends the soft bit messages to the decoder for decoding, and after decoding, the soft bit messages

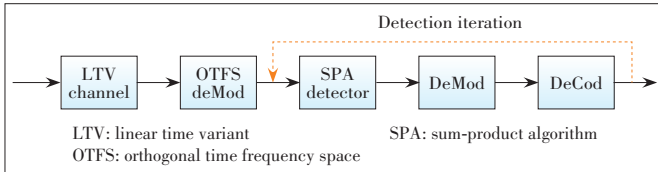


▲ Figure 4. BLER performance of the coded OTFS systems

▼ Table 1. Parameters of LDPC codes

Information Bits K	Code Rate R	Modulation
2 100	1/2	BPSK
4 200	1/2	QPSK
8 400	1/2	16QAM

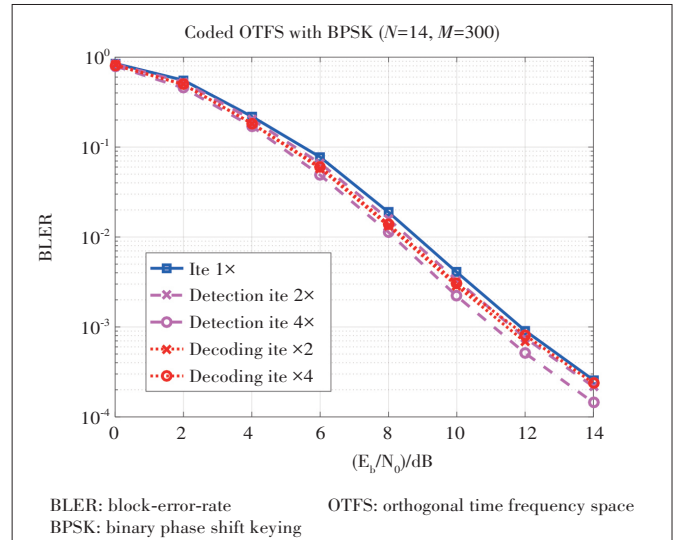
BPSK: binary phase shift keying
QPSK: quadrature phase shift keying
QAM: quadrature amplitude modulation



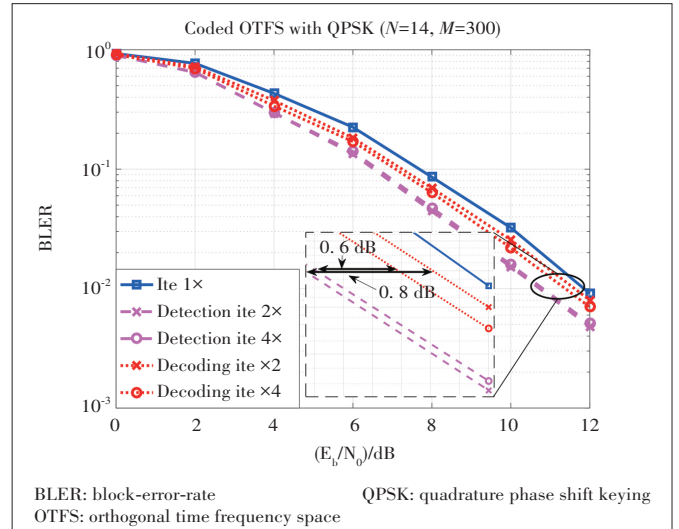
▲ Figure 5. Illustration of coded OTFS system via decoding iteration

es will be sent back to the SPA detector via the outer iterative layer. Here, “ite 2×2 ” means doing two decoding iterations and in each decoding iteration there are two detection iterations. In Fig. 9, we can observe that for the BPSK modulation, doing hybrid iterative decoding and detection cannot obtain much gain. For the QPSK modulation, the performance of “ite 2×2 ” hybrid iteration can approach the performance of doing 8 iterative detections. For 16QAM modulation, the performance of “ite 2×2 ” hybrid iterative decoding and detection is similar to the performance of doing 8 iterative detections.

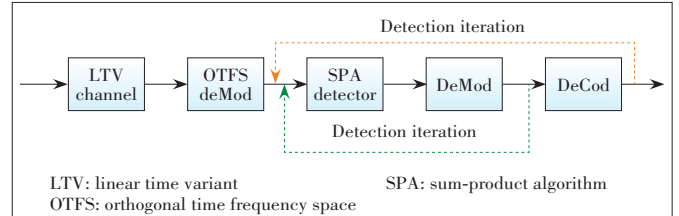
Consequently, we have some observations for coded OTFS systems as follows. For the iterative SPA detection, the optimal iterative numbers depend on the modulation order. Higher order modulation needs a higher number of iterations. Only iterating between decoding and detection does not improve the performance much. Hybrid iterative decoding and detection can achieve better performance, especially for higher order



▲ Figure 6. BLER performance of the coded OTFS systems with BPSK modulation



▲ Figure 7. BLER performance of the coded OTFS systems with QPSK modulation



▲ Figure 8. An illustration of coded OTFS system via hybrid iteration

modulations.

4.3 Effect of Channel Estimation on Coded OTFS System

Furthermore, we consider the OTFS performance with the channel estimation, where the DD domain channel estimation is employed^[14]. In Fig. 10, we compare the performances of uncoded OTFS and coded OTFS system with the channel estimation. Here, we only consider the QPSK modulation with SPA

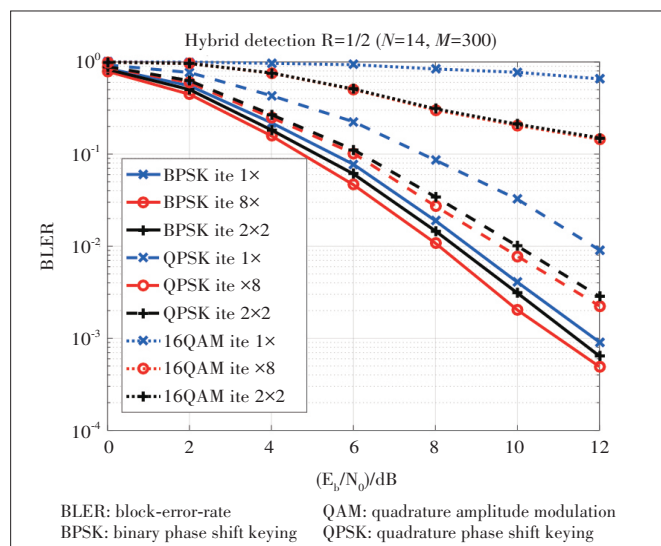
detection, where the iteration number is 2. “Pilot 30 dB” means the SNR of the pilot is 30 dB. We can see from the figure that the channel estimation has more impact on the performance of the uncoded system than that of the coded OTFS system. For example, the performance gap between “Pilot 30 dB” and “Pilot 40 dB” of the uncoded OTFS system is about 2 dB. For the coded OTFS system, the performance gap between “pilot 30 dB” and “pilot 40 dB” is about 1 dB. In other words, the channel codes reduce the effect of channel estimation on the system. However, the performance of the coded OTFS system shows an error floor below BLER at 10^{-2} , particularly at high SNR due to the channel estimation error.

5 Conclusions

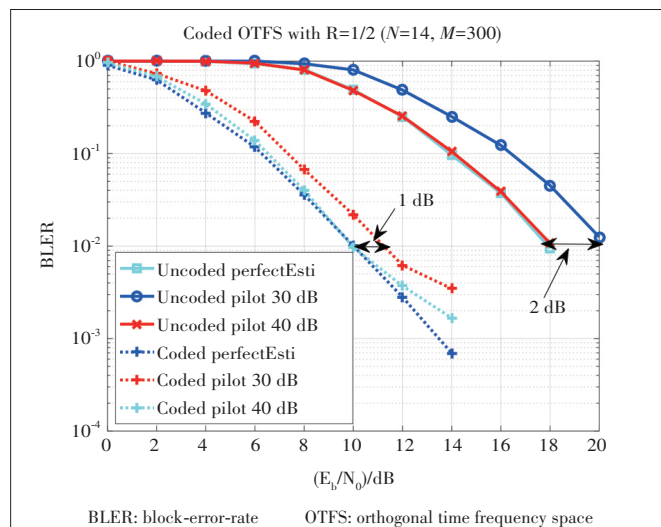
OTFS has great potential in providing reliable communications for next generation wireless communication systems. In this paper, we gave a brief overview of the fundamental concept of OTFS and presented the coded OTFS performance based on 5G LDPC codes. We showed the BLER performance of uncoded OTFS system and coded OTFS system over high mobility channels, and discussed three different coded OTFS schemes with iterative detection and decoding. We also evaluated the impact of channel estimation on the performance of coded OTFS systems.

References

- [1] ZHOU Y Q, LIU L, WANG L, et al. Service-aware 6G: an intelligent and open network based on the convergence of communication, computing and caching [J]. Digital communications and networks, 2020, 6(3): 253 – 260. DOI: 10.1016/j.dcan.2020.05.003
- [2] ZHOU Y Q, TIAN L, LIU L, et al. Fog computing enabled future mobile communication networks: a convergence of communication and computing [J]. IEEE communications magazine, 2019, 57(5): 20 – 27. DOI: 10.1109/MCOM.2019.1800235
- [3] LIU L, ZHOU Y Q, YUAN J H, et al. Economically optimal MS association for multimedia content delivery in cache-enabled heterogeneous cloud radio access networks [J]. IEEE journal on selected areas in communications, 2019, 37(7): 1584 – 1593. DOI: 10.1109/JSAC.2019.2916280
- [4] LIU L, ZHOU Y Q, ZHUANG W H, et al. Tractable coverage analysis for hexagonal macrocell-based heterogeneous UDNs with adaptive interference-aware CoMP [J]. IEEE transactions on wireless communications, 2019, 18(1): 503 – 517. DOI: 10.1109/TWC.2018.2882434
- [5] LIU L, ZHOU Y Q, GARCIA V, et al. Load aware joint CoMP clustering and inter-cell resource scheduling in heterogeneous ultra-dense cellular networks [J]. IEEE transactions on vehicular technology, 2018, 67(3): 2741 – 2755. DOI: 10.1109/TVT.2017.2773640
- [6] GARCIA V, ZHOU Y Q, SHI J L. Coordinated multipoint transmission in dense cellular networks with user-centric adaptive clustering [J]. IEEE transactions on wireless communications, 2014, 13(8): 4297 – 4308. DOI: 10.1109/TWC.2014.2316500
- [7] JAMEEL F, WYNE S, NAWAZ S J, et al. Propagation channels for mmWave vehicular communications: state-of-the-art and future research directions [J]. IEEE wireless communications, 2019, 26(1): 144 – 150. DOI: 10.1109/MWC.2018.1800174
- [8] SU Y T, LIU Y Q, ZHOU Y Q, et al. Broadband LEO satellite communications: architectures and key technologies [J]. IEEE wireless communications, 2019, 26



▲ Figure 9. BLER performance of the coded OTFS systems via hybrid detection and decoding OTFS



▲ Figure 10. BLER performance of the coded OTFS systems with channel estimation

- [2]: 55 – 61. DOI: 10.1109/MWC.2019.1800299
- [9] NOH G, HUI B, KIM I. High speed train communications in 5G: Design elements to mitigate the impact of very high mobility [J]. IEEE wireless communications, 2020, 27(6): 98 – 106. DOI: 10.1109/MWC.001.2000034
- [10] BAI L, HAN R, LIU J W, et al. Air-to-ground wireless links for high-speed UAVs [J]. IEEE journal on selected areas in communications, 2020, 38(12): 2918 – 2930. DOI: 10.1109/JSAC.2020.3005471
- [11] HADANI R, RAKIB S, TSATSANIS M, et al. Orthogonal time frequency space modulation[C]//2017 IEEE Wireless Communications and Networking Conference (WCNC). San Francisco, USA: IEEE, 2017: 1 – 6. DOI: 10.1109/WCNC.2017.7925924
- [12] WEI Z Q, YUAN W J, LI S Y, et al. Orthogonal time-frequency space modulation: a promising next-generation waveform [J]. IEEE wireless communications, 2021, 28(4): 136 – 144. DOI: 10.1109/MWC.001.2000408
- [13] HEBRON Y, RAKIB S, HADANI R, et al. Channel acquisition using orthogonal time frequency space modulated pilot signals: PCTUS20 17/025 166 [P]. 2016
- [14] RAVITEJA P, PHAN K T, HONG Y. Embedded pilot-aided channel estimation for OTFS in delay – Doppler channels [J]. IEEE transactions on vehicular technology, 2019, 68(5): 4906 – 4917. DOI: 10.1109/TVT.2019.2906357
- [15] MURALI K R, CHOCKALINGAM A. On OTFS modulation for high-Doppler

- fading channels [J]. 2018 information theory and applications workshop (ITA), 2018: 1 – 10. DOI: 10.1109/ITA.2018.8503182
- [16] LIU B, WEI Z Q, YUAN W J, et al. Channel estimation and user identification with deep learning for massive machine-type communications [J]. IEEE transactions on vehicular technology, 2021, 70(10): 10709 – 10722. DOI: 10.1109/TVT.2021.3111081
- [17] RAVITEJA P, PHAN K T, HONG Y, et al. Interference cancellation and iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, 2018, 17(10): 6501 – 6515. DOI: 10.1109/TWC.2018.2860011
- [18] LI L J, LIANG Y, FAN P Z, et al. Low complexity detection algorithms for OTFS under rapidly time-varying channel [C]//2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring). Kuala Lumpur, Malaysia: IEEE, 2019: 1 – 5. DOI: 10.1109/VTCSpring.2019.8746420
- [19] YUAN W J, WEI Z Q, YUAN J H, et al. A simple variational Bayes detector for orthogonal time frequency space (OTFS) modulation [J]. IEEE transactions on vehicular technology, 2020, 69(7): 7976 – 7980. DOI: 10.1109/TVT.2020.2991443
- [20] YUAN Z D, LIU F, YUAN W J, et al. Iterative detection for orthogonal time frequency space modulation with unitary approximate message passing [J]. IEEE transactions on wireless communications, 2021, 71(7): 1. DOI: 10.1109/TWC.2021.3097173
- [21] LI S Y, YUAN W J, WEI Z Q, et al. Cross domain iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, 2021, 70(9): 1. DOI: 10.1109/TWC.2021.3110125
- [22] LI S Y, YUAN J H, YUAN W J, et al. Performance analysis of coded OTFS systems over high-mobility channels [J]. IEEE transactions on wireless communications, 2021, 70(9): 6033 – 6048. DOI: 10.1109/TWC.2021.3071493
- [23] RAVITEJA P, HONG Y, VITERBO E, et al. Practical pulse-shaping waveforms for reduced-cyclic-prefix OTFS [J]. IEEE transactions on vehicular technology, 2019, 68(1): 957 – 961. DOI: 10.1109/TVT.2018.2878891
- [24] TSE D, VISWANATH P. Fundamentals of wireless communication [M]. Cambridge: Cambridge University Press, 2005. DOI: 10.1017/cbo9780511807213
- [25] MOLISCH A F. Wireless communications [M]. Hoboken, USA: John Wiley & Sons, 2012
- [26] RAVITEJA P, HONG Y, VITERBO E, et al. Effective diversity of OTFS modulation [J]. IEEE wireless communications letters, 2020, 9(2): 249 – 253. DOI: 10.1109/LWC.2019.2951758
- [27] TAROKH V, SESHADRI N, CALDERBANK A R. Space-time codes for high data rate wireless communication: Performance criterion and code construction [J]. IEEE transactions on information theory, 1998, 44(2): 744 – 765. DOI: 10.1109/18.661517
- [28] VUCETIC B, YUAN J H. Space-time coding [M]. Chichester, UK: John Wiley & Sons, Ltd, 2003. DOI: 10.1002/047001413x
- [29] BIGLIERI E, RAVITEJA P, HONG Y. Error performance of orthogonal time frequency space (OTFS) modulation [C]//IEEE International Conference on Communications Workshops. Shanghai, China: IEEE, 2019: 1 – 6. DOI: 10.1109/ICCWork.2019.8756831
- [30] SURABHI G D, AUGUSTINE R M, CHOCKALINGAM A. On the diversity of uncoded OTFS modulation in doubly-dispersive channels [J]. IEEE transactions on wireless communications, 2019, 18(6): 3049 – 3063. DOI: 10.1109/TWC.2019.2909205
- [31] KSCHISCHANG F R, FREY B J, LOELIGER H A. Factor graphs and the sum-product algorithm [J]. IEEE transactions on information theory, 2001, 47(2): 498 – 519. DOI: 10.1109/18.910572
- [32] LI S Y, YUAN W J, WEI Z Q, et al. Hybrid MAP and PIC detection for OTFS modulation [J]. IEEE transactions on vehicular technology, 2021, 70(7): 7193 – 7198. DOI: 10.1109/TVT.2021.3083181
- [33] 3GPP. Technical specification group radio access network: 3GPP TS 38.214 V15.0.0 [S]. 2017
- [34] 3GPP. Technical specification group radio access network: 3GPP TS 38.212 V15.0.0 [S]. 2017
- [35] VENTURA-TRAVESET J, CAIRE G, BIGLIERI E, et al. Impact of diversity reception on fading channels with coded modulation. Part I: coherent detection [J]. IEEE transactions on communications, 1997, 45(5): 563 – 572. DOI: 10.1109/26.592556
- [36] BIGLIERI E, PROAKIS J, SHAMAI S. Fading channels: Informationtheoretic and communications aspects [J]. IEEE transactions on information theory, 1998, 44(6): 2619 – 2692

Biographies

ZHANG Chong (zhangchong@ict.ac.cn) received the B.S. degree in electronic & information engineering from Huaqiao University, China in 2017, and the M.S. degree in electronic and communication engineering from University of Chinese Academy of Science, China in 2020. He is currently pursuing the Ph.D. degree in the Institute of Computing Technology, Chinese Academy of Science, China. His research interests include channel coding, modulation, and wireless communications.

XING Wang received the B.S. degree from University of Science and Technology Beijing, China in 2019. He is currently pursuing the Ph.D. degree at the Wireless Communication Research Center, Institute of Computing Technology, Chinese Academy of Sciences. His current research interests include orthogonal time frequency space modulation and the convergence of communication, computation, and caching.

YUAN Jinhong received the B.E. and Ph.D. degrees in electronics engineering from the Beijing Institute of Technology, China, in 1991 and 1997, respectively. In 2000, he joined the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia, where he is currently a professor and Head of Telecommunication Group with the School. He has published two books, five book chapters, over 300 papers in telecommunications journals and conference proceedings, and 50 industrial reports. He is a co-inventor of one patent on MIMO systems and two patents on low-density-parity-check codes. He has co-authored four Best Paper Awards and one Best Poster Award. He is an IEEE Fellow and currently serving as an Associate Editor for the *IEEE Transactions on Wireless Communications*. He served as the IEEE NSW Chapter Chair of Joint Communications/Signal Processions/Ocean Engineering Chapter during 2011-2014 and served as an Associate Editor for the *IEEE Transactions on Communications* during 2012-2017. His current research interests include error control coding and information theory, communication theory, and wireless communications.

ZHOU Yiqing received the B.S. degree in communication and information engineering and the M.S. degree in signal and information processing from Southeast University, China in 1997 and 2000, respectively, and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong, China in 2004. She is currently a professor with the Wireless Communication Research Center, Chinese Academy of Sciences. She has published over 150 articles and four books/book chapters. She received the best paper awards from WCSP2019, IEEE ICC2018, ISCIT2016, PIMRC2015, ICCS2014, and WCNC2013. She also received the 2014 Top 15 Editor Award from *IEEE TVT* and the 2016 – 2017 Top Editors of *ETT*. She is also the TPC Co-Chair of ChinaCom2012, an Executive Co-Chair of IEEE ICC2019, a Symposia Co-Chair of ICC2015, a Symposium Co-Chair of GLOBECOM2016 and ICC2014, a Tutorial Co-Chair of ICC2014 and WCNC2013, and the Workshop Co-Chair of Smart-GridComm2012 and GlobeCom2011. She is also the Associate/Guest Editor of some renowned journals.

Coded Orthogonal Time Frequency Space Modulation



LIU Mengmeng¹, LI Shuangyang², ZHANG Chunqiong¹, WANG Boyu¹, BAI Baoming¹

(1. State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China;

2. School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney 2032, Australia)

Abstract: Orthogonal time frequency space (OTFS) modulation is a novel two-dimensional modulation scheme for high-Doppler fading scenarios, which is implemented in the delay-Doppler (DD) domain. In time and frequency selective channels, OTFS modulation is more robust than the popular orthogonal frequency division multiplexing (OFDM) modulation technique. To further improve transmission reliability, some channel coding schemes are used in the OTFS modulation system. In this paper, the coded OTFS modulation system is considered and introduced in detail. Furthermore, the performance of the uncoded/coded OTFS system and OFDM system is analyzed with different relative speeds, modulation schemes, and iterations. Simulation results show that the OTFS system has the potential of full diversity gain and better robustness under high mobility scenarios.

Keywords: OTFS modulation; OFDM; channel coding; fading channel

DOI: 10.12142/ZTECOM.202104006

<https://kns.cnki.net/kcms/detail/34.1294.TN.20211123.1849.004.html>, published online November 24, 2021

Manuscript received: 2021-11-01

Citation (IEEE Format): M. M. Liu, S. Y. Li, C. Q. Zhang, et al., "Coded orthogonal time frequency space modulation," *ZTE Communications*, vol. 19, no. 4, pp. 54 – 62, Dec. 2021. doi: 10.12142/ZTECOM.202104006.

1 Introduction

The 5G network has achieved the peak rate of 10 – 20 Gbit/s, which is more than ten times that of 4G Long Term Evolution (LTE) cellular networks. Some new scenarios with high mobility have emerged in 5G/B5G, such as V2X (vehicle-to-vehicle—V2V and vehicle-to-infrastructure—V2I) with the terminal speed up to 300 km/h, high speed train (HST) with the maximum speed up to 500 km/h and unmanned aerial vehicle (UAV). In these cases, the higher Doppler spread will be induced. In addition, a higher data rate is required, which is considered to be solved by using a higher frequency band, such as a millimeter wave band or even a terahertz band. Both high mobility and high frequency

will lead to large Doppler shifts, yielding the large frequency dispersion. Although orthogonal frequency division multiplexing (OFDM) modulation is used in 4G and 5G, it has good robustness only in time-invariant channels and is very sensitive to carrier frequency offsets. However, the channel is time-varying in high Doppler scenarios. The orthogonality of sub-carriers in an OFDM symbol is seriously damaged so that the channel estimation is no longer accurate, which will lead to severe inter-carrier interference (ICI) and the disappearance of the near-capacity advantage.

To deal with communication scenarios with high Doppler shifts, a novel two-dimensional (2D) modulation scheme called orthogonal time frequency space (OTFS) modulation was proposed by R. HADANI et al. in 2017, whose pioneering works^[1-4] introduced the principle of OTFS modulation and demonstrated its significant performance on OFDM modulation in channels with high Doppler or at high frequencies.

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61771364, and by the National Key R&D Program of China under Grant No. 2020YEB1807104.

Compared with the OFDM modulation, OTFS modulation has the potential of full diversity gain and better robustness, which can effectively deal with the impact of high Doppler shifts. One more advantage of OTFS is that it can be implemented as pre- and post-processing blocks applied to a time-frequency signaling scheme, such as OFDM^[5]. Furthermore, an implementation scheme of OTFS modulation based on the OFDM has been proposed in Ref. [6], which greatly reduces the complexity of implementation. In Ref. [7], the vector form of concise and elegant input-output relationship of the OFDM-based OTFS system has also been derived by utilizing the properties of the Kronecker product in matrices and vectors, which is also suitable for general time-varying channels with arbitrary Doppler and windowing functions. It is worth mentioning that this representation is very popular in subsequent research work.

As for the significant advantage of achieving the full diversity, the detailed formal analysis on the diversity order of OTFS in doubly-dispersive channels has been presented in Ref. [8], which points out that the full diversity in the delay-Doppler (DD) domain can be extracted by using the phase rotation method. In addition, when the OTFS frame is long enough, even the uncoded OTFS modulation system can obtain almost full diversity in the case of path number $P = 2^9$. In order to make full use of full diversity, effective equalization is needed, which depends on the accurate channel estimation. A well-known channel estimation scheme for OTFS has been proposed in Ref. [10], in which pilots, protection symbols, and data symbols are cleverly arranged on the delay Doppler grid plane to effectively avoid the interference between pilots and data symbols at the receiver and enable the channel estimation and data detection to be performed in the same OTFS frame with the minimum overhead. However, the performance of such algorithms^[10-11] is very sensitive to the availability of protection space. In fact, more advanced channel estimation methods based on compressed sensing^[12], orthogonal matching pursuit (OMP)^[13-14] or sparse Bayesian learning^[15] algorithms have been proposed, which take advantage of the channel sparsity in the DD domain. However, the channel in the DD domain may not always be sparse, especially in the case of fractional Doppler^[5]. An effective solution is to enhance channel sparsity by applying time-frequency (TF) domain windows, such as Dolph-Chebyshev (DC) window^[16]. In addition, advanced detection algorithms are also an important part of OTFS to achieve potential full diversity gain^[17]. A message passing algorithm (MPA) based on the maximum a posteriori probability (MAP) detection criterion has been introduced in Ref. [5], which processes the interference from other information symbols as Gaussian variables to reduce the detection complexity. However, due to the short period of the probabilistic graphical model, the proposed MPA may not converge, resulting in performance degradation. In order to solve this problem, a convergence protection receiver based on variable Bayes (VB) framework has been presented in Ref. [18], which utilizes the relative entropy to approximate the optimal detection of the cor-

responding a posteriori distribution to realize the MPA on a simple graphical model. In addition, a hybrid detection scheme has been demonstrated in Ref. [19], which takes the MAP and the parallel interference cancellation (PIC) into account and achieves a good trade-off between the error performance and detection complexity.

Channel coding is also one of the effective methods to ensure diversity gain and achieve reliable communications. However, most of the research work analyzes the performance of the uncoded OTFS modulation system. In Ref. [20], the BER performance of the coded OTFS system has been analyzed in detail. The derivation of pairwise-error probability (PEP) and its upper bounds have demonstrated a very interesting trade-off between the coding gain and diversity gain of the coded OTFS system. Moreover, a channel coding design criterion is derived, that is, maximizing the minimum Euclidean distance between all code-word pairs. This criterion is very similar to the channel coding design in the additive white Gaussian noise (AWGN) channel. However, the OTFS modulation experiences a time-varying channel with both time dispersion and frequency dispersion. Thus, both the inter-symbol interference (ISI) and ICI will be generated, which depend on the delay τ , Doppler ν of the channel and the cross-ambiguity function of pulses at the transmitter and receiver. The assumed ideal pulse-shaping waveforms^[1,3] that satisfy the bi-orthogonality condition in both time and frequency do not exist in practical applications. Therefore, ISI and ICI are inevitable in the actual OTFS system. Obviously, it is necessary to consider the properties of OTFS and characteristics of the DD domain channel to design the channel code suitable for the OTFS system.

In this paper, we consider the coded OTFS modulation and describe it in detail. Then, recent work about the coded OTFS system analysis is summarized. On this basis, the upper bound on the unconditional PEP of the coded OTFS system is further supplemented. In addition, the joint iterative strategy of detection and decoding is also considered to improve the system performance. According to the considered iterative system, the coding design scheme of the OTFS system is discussed. Finally, we analyze the performance of the coded/uncoded OTFS and OFDM systems with different relative speeds, modulation schemes and iterations. Simulation results show that OTFS systems have significant robustness compared with OFDM.

2 Principle and System Model

2.1 Relationship Between Time-Frequency Domain and Delay-Doppler Domain

In Ref. [4], the authors proposed that the OTFS modulation can be viewed as a time-frequency spreading scheme, which was based on the Fourier duality relation between the time-frequency plane and the delay-Doppler plane, resulting in a sim-

ple pre-processing step over an arbitrary multicarrier modulation (such as OFDM). In view of the importance of the transform between the DD domain and the TF domain, we will review this relationship in this subsection.

The grid in the TF domain and the corresponding reciprocal grid in the DD domain are shown in Fig. 1, with the size of $M \times N$. According to Fig. 1, the TF grid can be represented as:

$$\Lambda = \{(nT, m\Delta f), n = 0, \dots, N-1, m = 0, \dots, M-1\}, \quad (1)$$

where T is the sampling interval along the time axis, Δf is the sampling interval along the frequency axis, and N and M are the corresponding numbers of sampling points on the TF plane. According to the principle of the time-frequency modulation explicated in Ref. [1], the transmitted packet can be regarded as a burst one with a total duration of NT seconds and a total bandwidth of $M\Delta f$ Hz. Then, the modulated symbols $X_{TF}[m, n]$, $m = 0, 1, \dots, M-1$, $n = 0, 1, \dots, N-1$ are transmitted over the burst packet in the TF domain.

The reciprocal delay-Doppler grid is represented as:

$$\Lambda^\perp = \left\{ \left(\frac{k}{NT}, \frac{l}{M\Delta f} \right), k = 0, \dots, N-1, l = 0, \dots, M-1 \right\}, \quad (2)$$

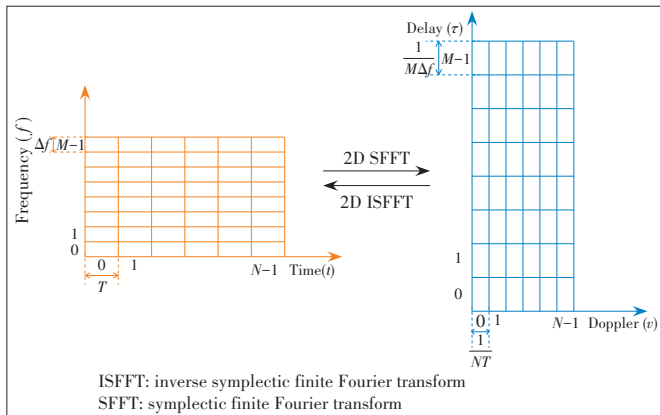
where $\frac{1}{NT}$ and $\frac{1}{M\Delta f}$ represent the sampling intervals along the Doppler axis and the delay axis, respectively.

The mapping between signals in the TF domain and DD domain depends on two-dimensional symplectic finite Fourier transform (2D SFFT) pairs, which can be exemplified as:

$$h(\tau, \nu) = \iint H(t, f) e^{-j2\pi(\nu t - f\tau)} dt df, \quad (3)$$

$$H(t, f) = \iint h(\tau, \nu) e^{j2\pi(\nu t - f\tau)} d\tau d\nu, \quad (4)$$

where $h(\tau, \nu)$ and $H(t, f)$ are the responses of linear time-varying (LTV) wireless channels in the DD domain and TF domain,



▲ Figure 1. Grids in time frequency (TF) plane and delay-Doppler (DD) plane

respectively. Without loss of generality, the DD domain representation of an LTV wireless channel can be expressed as:

$$h(\tau, \nu) = \sum_{i=1}^P h_i \delta(\tau - \tau_i) \delta(\nu - \nu_i), \quad (5)$$

where $\delta(\cdot)$ denotes the Dirac delta function, P is the number of resolvable paths, and h_i , τ_i and ν_i are the channel coefficient, delay and Doppler shift of the i -th path respectively. Here, τ_i and ν_i are defined as:

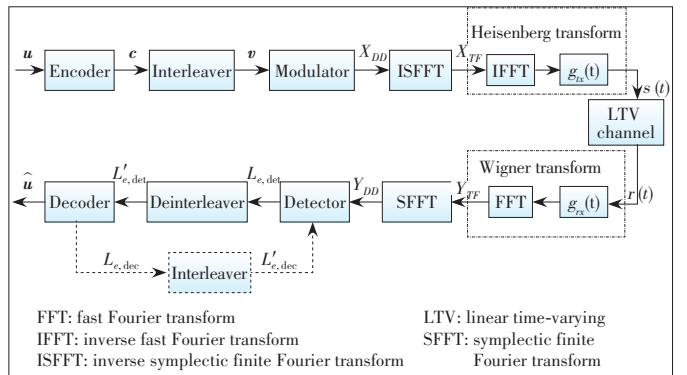
$$\tau_i = \frac{l_i}{M\Delta f}, \nu_i = \frac{k_i + \kappa_i}{NT}. \quad (6)$$

In Eq. (6), l_i represents the index of the delay with integer values, k_i represents the index of the Doppler shift with integer values, and $\kappa_i \in (-0.5, 0.5]$ is the real number, indicating the fractional shift from the nearest Doppler index k_i , which is also called fractional Doppler^[5].

Specifically, the 2D SFFT pairs can be realized by simple discrete Fourier transform (DFT) pairs or fast Fourier transform (FFT) pairs. For example, in view of the DFT, the inverse symplectic finite Fourier transform (ISFFT) can be regarded as the M -point DFT along the delay axis and the N -point inverse DFT (IDFT) along the Doppler axis for the two-dimensional signal with the size of $M \times N$ in the DD domain, resulting in the corresponding TF domain signal.

2.2 Coded OTFS System Model

Fig. 2 shows the proposed coded OTFS system model in this paper. Suppose that the information bit sequence u of length K is encoded using a forward error correction (FEC) code, resulting in the codeword c of length N_c . After interleaving, the interleaved sequence v is then mapped to an M -ary signal constellation \mathbb{A} , such as M -ary phase shift keying (MPSK) or M -ary quadrature amplitude modulation (MQAM), and the modulated symbol vector \mathbf{x}_{DD} of length MN is arranged as a two-dimensional signal matrix $\mathbf{x}_{DD} \in \mathbb{A}^{M \times N}$ in the DD domain, where M is the number of the sub-carriers, and N is the number of time slots for each OTFS symbol. The element $x[l, k]$ of the \mathbf{x}_{DD} represents the modulated signals in the l -th Delay and k -



▲ Figure 2. Proposed model of coded OTFS system

th Doppler grid, for $l \in \{0, 1, \dots, M-1\}$ and $k \in \{0, 1, \dots, N-1\}$. Then the symbol $X[m, n]$ in the m -th frequency and n -th time grid is obtained by the ISFFT, which is given as:

$$X[m, n] = \text{ISFFT}(x[l, k]) = \frac{1}{\sqrt{NM}} \sum_{l=0}^{M-1} \sum_{k=0}^{N-1} x[l, k] e^{j2\pi \left(\frac{nk}{N} - \frac{ml}{M} \right)}, \quad (7)$$

for $m = 0, 1, \dots, M-1, n = 0, 1, \dots, N-1$. The two-dimensional signal matrix in the TF domain is denoted by $\mathbf{X}_{TF} \in \mathbb{C}^{M \times N}$. The early literature^[1-3] explained that the composition of the ISFFT and the windowing function in the TF domain are referred to as the OTFS transform. The window operations of the transmitter and receiver affect the cross-symbol interference of the effective impulse response^[1]. The window design has the potential to increase the effective channel sparsity in the DD domain, which is conducive to the channel estimation, as described in Refs. [6] and [7]. Furthermore, the influence of the design of the TF domain window on improving the performance of the channel estimation and data detection is discussed in Ref. [16]. Here, the rectangular window is considered.

The transmitted signal in the time domain is obtained from the TF domain symbols $X[m, n]$ using the Heisenberg transform parameterized by the pulse shaping filter $g_{tx}(t)$, which can be written as:

$$s(t) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X[m, n] g_{tx}(t - nT) e^{j2\pi m \Delta f (t - nT)}. \quad (8)$$

This can be regarded as a general form of the OFDM modulation^[3]. Moreover, OTFS modulation can be implemented as a cascade of a pre-coder (ISFFT) and a traditional OFDM modulator^[6], as shown in Fig. 2.

Assume that the channel is the LTV channel described in Eq. (5), the received signal can be expressed as:

$$r(t) = \iint h(\tau, \nu) s(t - \tau) e^{j2\pi \nu (t - \tau)} d\tau d\nu + w(t), \quad (9)$$

where $w(t)$ is the additive white Gaussian noise with zero mean and one-sided power spectral density of N_0 .

At the receiver, $r(t)$ is subject to the Wigner transform to obtain the received symbols $Y[m, n]$ in the TF domain, given by

$$Y[m, n] = \int r(t) g_{rx}^*(t - nT) e^{-j2\pi m \Delta f (t - nT)} dt, \quad (10)$$

where $g_{rx}(t)$ is the pulse shaping filter at the receiver. Eq. (10) can be further written as^[15]:

$$Y[m, n] = \sum_{m'=0}^{M-1} \sum_{n'=0}^{N-1} H_{m,n}[m', n'] X[m', n'] + \bar{w}[m, n], \quad (11)$$

where $\bar{w}[m, n]$ is the noise sample in the TF domain, and $H_{m,n}[m', n']$ is the channel impulse response in the TF domain, i.e.,

$$H_{m,n}[m', n'] = \iint h(\tau, \nu) A_{g_{rx}, g_{tx}}((m - m')\Delta f - \nu, (n - n')T - \tau) e^{j2\pi(\nu + m'\Delta f)((n - n')T - \tau)} e^{j2\pi m' n' T} d\tau d\nu. \quad (12)$$

In Eq. (12), $A_{g_{rx}, g_{tx}}(\tau, \nu)$ is referred to as the cross-ambiguity function, which represents the interference between symbols in the DD domain caused by the channel dispersion^[20], and can be expressed as:

$$A_{g_{rx}, g_{tx}}(\tau, \nu) = \int g_{rx}^*(t - \tau) g_{tx}(t) e^{j2\pi \nu \Delta f t} dt. \quad (13)$$

The received symbols in the DD domain are given as:

$$y[l, k] = \text{SFFT}(Y[m, n]) = \frac{1}{\sqrt{NM}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} Y[m, n] e^{-j2\pi \left(\frac{nk}{N} - \frac{ml}{M} \right)} + w[l, k], \quad (14)$$

for $l = 0, 1, \dots, M-1, k = 0, 1, \dots, N-1$, where $w[l, k]$ is the noise sample in the DD domain. Upon the received symbols $y[l, k]$, a signal detection algorithm is then performed. In addition, a joint iterative strategy between detection and decoding can also be considered.

It is generally known that the MAP detection is optimum for OTFS systems. However, the complexity of the MAP detection increases exponentially with the block size of each OTFS frame. As a compromise of the MAP detection, a lot of literature has studied the message passing detection algorithm based on the factor graph, which can effectively reduce the detection complexity, such as Ref. [5]. For the coded system, the iterative signal processing of the detector and the decoder is usually considered at the receiver, as shown in Fig. 2. Correspondingly, soft decision detection algorithms should be adopted, such as MP, unitary approximate MP (UAMP), vector AMP (VAMP), sum-product algorithm (SPA) and other message passing algorithms. With Log-Likelihood Ratios (LLRs), the message $L_{e, \text{det}}$ passed from the detector to the decoder is calculated as:

$$L_{e, \text{det}}(x_{l,k} = a_j) = L_{\text{app}, \text{det}}(x_{l,k} = a_j) - L_{a, \text{det}}(x_{l,k} = a_j) = \ln \frac{P(x_{l,k} = a_j | Y_{DD})}{P(x_{l,k} = 0 | Y_{DD})} - L_{a, \text{det}}(x_{l,k} = a_j), \quad (15)$$

where $L_{\text{app}, \text{det}}$ and $L_{a, \text{det}}$ represent the a posteriori LLRs and the priori LLRs of the detector, respectively; $P(\cdot)$ denotes the priori symbol probabilities, $a_j \in \mathbb{A}, j = 1, \dots, |\mathbb{A}|$, and the binary vector (signal label) corresponding to a_j can be expressed as \mathbf{v}_j . Similarly, the extrinsic LLRs of the decoder $L_{e, \text{dec}}$ are also obtained by subtracting priori LLRs $L_{a, \text{dec}}$ from the a posteriori LLRs $L_{\text{app}, \text{dec}}$, and $L_{a, \text{dec}}$ is updated by the extrinsic LLRs of the detector. The iterative process between the detector and the decoder can be described as follows.

Algorithm 1. Algorithm of the iterative process between the detector and the decoder

- 1: **Initialization:** Set the number of joint iterations $\eta = 0$, the maximum number of joint iterations η_{\max} , the maximum number of detection iterations η_{\max}^{\det} , the maximum number of decoding iterations η_{\max}^{dec} , and the priori LLRs $L_{a,\det} = 0$
- 2: **while** $\eta \leq \eta_{\max}$ **do**
- 3: Perform detection until η_{\max}^{\det} is satisfied.
- 4: Calculate the a posteriori LLRs $L_{\text{app},\det}$ and the extrinsic LLRs of the detector $L_{e,\det}(x_{l,k} = a_j)$ for all $j = 1, \dots, |\mathbb{A}|$, $k = 0, \dots, N-1, l = 0, \dots, M-1$
- 5: Convert the symbol form of $L_{e,\det}(x_{l,k} = a_j)$ into the bit form $L_{e,\det}(c_i)$, $i = 0, 1, \dots, N_c - 1$
- 6: Deinterleave the extrinsic LLRs $L_{e,\det}(c_i)$, resulting in $L'_{e,\det}(c_i)$
- 7: Set $L_{a,\text{dec}}(c_i) = L'_{e,\det}(c_i)$, $i = 0, 1, \dots, N_c - 1$
- 8: Perform decoding until η_{\max}^{dec} is satisfied
- 9: Calculate the a posteriori LLRs $L_{\text{app},\text{dec}}$ and the extrinsic LLRs of the decoder $L_{e,\text{dec}}(c_i)$ for $i = 0, 1, \dots, N_c - 1$
- 10: **if** $\eta < \eta_{\max}$ **then**
- 11: Interleave the extrinsic LLRs $L_{e,\text{dec}}(c_i)$, resulting in $L'_{e,\text{dec}}(c_i)$
- 12: Convert the bit form of $L'_{e,\text{dec}}(c_i)$ into the symbol form $L'_{e,\text{dec}}(x_{l,k} = a_j)$
- 13: Set $L_{a,\det}(x_{l,k} = a_j) = L'_{e,\text{dec}}(x_{l,k} = a_j)$
- 14: **end if**
- 15: $\eta = \eta + 1$
- 16: **end while**
- 17: Make decisions of c_i according to the a posteriori LLRs $L_{\text{app},\text{dec}}$ for $i = 0, 1, \dots, K-1$

2.3 Vectorization Representation of the System

With respect to the vectorization, the following definitions are given, $\mathbf{x}_{DD} = \text{vec}(\mathbf{x}_{DD}) \in \mathbb{A}^{MN \times 1}$, $\mathbf{x}_{TF} = \text{vec}(\mathbf{X}_{TF}) \in \mathbb{C}^{MN \times 1}$, $\mathbf{y}_{TF} = \text{vec}(\mathbf{Y}_{TF}) \in \mathbb{A}^{MN \times 1}$, and $\mathbf{y}_{DD} = \text{vec}(\mathbf{Y}_{DD}) \in \mathbb{C}^{MN \times 1}$, where $\text{vec}(\cdot)$ denotes the vectorized version of the 2D matrix formed by stacking the columns of the one into a single column vector. Besides, the N -point DFT matrix and its inverse are represented by \mathbf{F}_N and \mathbf{F}_N^H respectively and are assumed to be normalized so that $\mathbf{F}_N \mathbf{F}_N^H = \mathbf{I}_N$. According to the introduction of the coded OTFS system in the above subsection, the relationship between the symbol matrix $\mathbf{X}_{TF} \in \mathbb{C}^{M \times N}$ in the TF domain and $\mathbf{x}_{DD} \in \mathbb{C}^{M \times N}$ in the DD domain can be described as:

$$\mathbf{X}_{TF} = \mathbf{F}_M \mathbf{x}_{DD} \mathbf{F}_N^H. \quad (16)$$

The vectorized form can be expressed as:

$$\mathbf{x}_{TF} = \text{vec}(\mathbf{X}_{TF}) = (\mathbf{F}_N^H \otimes \mathbf{F}_M) \mathbf{x}_{DD}. \quad (17)$$

Considering that the pulse shaping filter $g_{tx}(t)$ is the rectangular form, the output of the Heisenberg transform is given by

$$\mathbf{S} = \mathbf{F}_M^H \mathbf{X}_{TF} = \mathbf{x}_{DD} \mathbf{F}_N^H, \quad (18)$$

where $\mathbf{S} \in \mathbb{C}^{M \times N}$ represents the transmitted signal matrix in the time domain. Vectorize \mathbf{S} by stacking each column of \mathbf{S} into a vector, we have

$$\mathbf{s} = \text{vec}(\mathbf{S}) = (\mathbf{F}_N^H \otimes \mathbf{I}_M) \mathbf{x}_{DD}. \quad (19)$$

At the receiver, the received signal expressed by Eq. (9) in discrete form is

$$r(n) = \sum_{i=1}^P h_i e^{\frac{j2\pi k_i(n-l_i)}{MN}} s\left(\left[n-l_i\right]_{MN}\right) + w(n), \quad (20)$$

where $[\cdot]_{MN}$ indicates mod MN operation, $r(n)$, $s(n)$ and $w(n)$ are the corresponding discrete forms of $r(t)$, $s(t)$ and $w(t)$, respectively. Thus, the received signal can be written in the vector form as

$$\mathbf{r} = \mathbf{H}_T \mathbf{s} + \mathbf{w}. \quad (21)$$

In the above formula, \mathbf{H}_T is an $MN \times MN$ matrix, given by Ref. [10],

$$\mathbf{H}_T = \sum_{i=1}^P h_i \mathbf{\Pi}^{l_i} \Delta^{(k_i)}, \quad (22)$$

where $\mathbf{\Pi}$ is the permutation matrix (forward cyclic shift), and $\Delta^{(k_i)}$ is the diagonal matrix, as shown below.

$$\mathbf{\Pi} = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}_{MN \times MN}, \quad (23)$$

$$\Delta^{(k_i)} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & e^{\frac{j2\pi k_i}{MN}} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\frac{j2\pi k_i(MN-1)}{MN}} \end{bmatrix}_{MN \times MN}. \quad (24)$$

The received signal vector \mathbf{r} is devectorized into an $M \times N$ matrix \mathbf{R} . Then, the Wigner transform and SFFT can be successive to obtain the received signal matrix \mathbf{Y}_{DD} with the size $M \times N$ in the DD domain as follows:

$$\mathbf{Y}_{DD} = \mathbf{F}_M^H (\mathbf{F}_M \mathbf{R}) \mathbf{F}_N = \mathbf{R} \mathbf{F}_N. \quad (25)$$

The vector form can be obtained by

$$\mathbf{y} = \text{vec}(\mathbf{Y}_{DD}) = (\mathbf{F}_N \otimes \mathbf{I}_M) \mathbf{r}. \quad (26)$$

Substituting Eqs. (19) and (21) with Eq. (26), we can get the vector form of the input-output relation in the DD domain as follows:

$$\mathbf{y} = (\mathbf{F}_N \otimes \mathbf{I}_M) \mathbf{H}_T (\mathbf{F}_N^H \otimes \mathbf{I}_M) \mathbf{x}_{DD} + \tilde{\mathbf{w}} = \mathbf{H}_{eff} \mathbf{x}_{DD} + \tilde{\mathbf{w}}. \quad (27)$$

The vectorized forms of each operation are simple and more vivid, which contribute to understanding the OTFS modulation more clearly and they are widely used in the research of the OTFS modulation.

3 Error Performance of the Coded OTFS System

3.1 Error Performance Analysis

PEP is commonly used in communication systems for analyzing the error performance of the system. In Ref. [8], the achievable diversity order of the OTFS system is analyzed based on the PEP under the maximum likelihood (ML) detection. Similarly, the PEP under the ML detection is also used to analyze the error performance of OTFS modulation in Ref. [21]. On this basis, the effective diversity (ED) is introduced from the perspective of PEP^[9]. In Ref. [20], the conditional PEP and the unconditional PEP are utilized to analyze the error performance of coded OTFS systems. And an approximate upper bound on the unconditional PEP for small P is derived by:

$$\Pr(\mathbf{x}_{DD} \rightarrow \hat{\mathbf{x}}_{DD}) \leq (d_E^2(\mathbf{e})/P)^{-r} \left(\frac{E_s}{4N_0} \right)^{-r}, \quad (28)$$

where $\mathbf{e} \triangleq \mathbf{x}_{DD} - \hat{\mathbf{x}}_{DD}$ is the codeword difference vector, $d_E^2(\mathbf{e}) = \mathbf{e}^H \mathbf{e}$ is the squared Euclidean distance between \mathbf{x}_{DD} and $\hat{\mathbf{x}}_{DD}$, and r is the rank of the positive semidefinite Hermite matrix $\mathbf{\Omega}(\mathbf{e})$ given by Eq. (18)^[20]. In Eq. (28), the exponent r and the term $d_E^2(\mathbf{e})/P$ are regarded as the diversity gain and the coding gain, respectively. According to the early works, e.g. Refs. [1] and [8], OTFS can achieve full diversity, whose order is the number of the separable multipath P . When the channel code is given, the term $d_E^2(\mathbf{e})$ is also fixed. Thus, as described in Corollary 1 in Ref. [20], the diversity gain increases and the coding gain decreases with the increase of P , which reveals an interesting trade-off between them. In addition, an approximate upper bound on the unconditional PEP for large P is also given by

$$\Pr(\mathbf{x}_{DD} \rightarrow \hat{\mathbf{x}}_{DD}) \leq \exp\left(-\frac{E_s}{16N_0} d_E^2(\mathbf{e})\right), \quad (29)$$

which only depends on the signal to noise ratio (SNR) and $d_E^2(\mathbf{e})$, and demonstrates that channels with a large number of resolvable paths approach an AWGN model.

More detailed derivation and illustration can be found in Ref. [20]. It should be noted that the upper bound on the unconditional PEP shown in Eqs. (28) and (29) are approximate. Referring to the appendix A in Ref. [22], the upper bound on the unconditional PEP has a more accurate display.

Note that $\mathbf{\Omega}(\mathbf{e})$ is also a Gram matrix^[23] corresponding to vectors $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_P\}$, where $\mathbf{u}_i \triangleq \mathbf{\Xi}_i \mathbf{e}$, and $\mathbf{\Xi}_i \triangleq (\mathbf{F}_N \otimes \mathbf{I}_M) \mathbf{H}^{l_i} \Delta^{(k_i)} (\mathbf{F}_N^H \otimes \mathbf{I}_M)$. According to the appendix A in Ref. [22], the determinant of Gram matrix $\mathbf{\Omega}(\mathbf{e})$ can be calculated by

$$\begin{aligned} \det(\mathbf{\Omega}(\mathbf{e})) &= GD\left(\left\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{P-1}\right\}\right) \left\|\tilde{\mathbf{u}}_P\right\|^2 \leq \\ &GD\left(\left\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{P-1}\right\}\right) \left\|\mathbf{u}_P\right\|^2 \\ &\vdots \\ &\leq \prod_{j=1}^P \left\|\mathbf{u}_j\right\|^2 = \prod_{j=1}^P \mathbf{e}^H \mathbf{e} = (d_E^2(\mathbf{e}))^P, \end{aligned} \quad (30)$$

where $\tilde{\mathbf{u}}_j$ is the orthogonal projection of \mathbf{u}_j onto the orthogonal complement of $\text{span}(\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{j-1})$. Besides, the maximum value of the rank of the matrix $\mathbf{\Omega}(\mathbf{e})$ is the number of resolvable paths P . In particular, when matrix $\mathbf{\Omega}(\mathbf{e})$ is full-rank, we have $r = P$. Then the upper bound on the unconditional PEP of the coded OTFS system can be rewritten as

$$\Pr(\mathbf{x}_{DD} \rightarrow \hat{\mathbf{x}}_{DD}) \leq \left(\frac{d_E^2(\mathbf{e})}{P}\right)^{-P} \left(\frac{E_s}{4N_0}\right)^{-P}, \quad (31)$$

Furthermore, the equality holds if $\mathbf{\Omega}(\mathbf{e})$ is a diagonal matrix.

3.2 Design Issues of Channel Codes for OTFS Systems

In Ref. [20], the code design criterion for the coded OTFS is given based on the PEP analysis, which is to maximize the minimum squared Euclidean distance of all possible codeword pairs. Simulation results show the performance of the coded OTFS system under convolutional codes with different minimum squared Euclidean distances and verify the proposed code design criterion. At present, most channel codes are designed for AWGN channels. In Ref. [20], authors also reveal that the channel with a large number of diversity paths approaches an AWGN channel when the number of resolvable paths P is large enough. In this case, some good channel codes can be used in the OTFS system. However, the increase of P will bring about large ISI, making signal detection more complicated. Therefore, it is necessary to design channel codes according to the characteristics of the OTFS modulation. In particular, the joint iteration between decoding and detection is needed, when the channel conditions are poor. In a word, the design of the channel coding scheme is still an inter-

esting challenge.

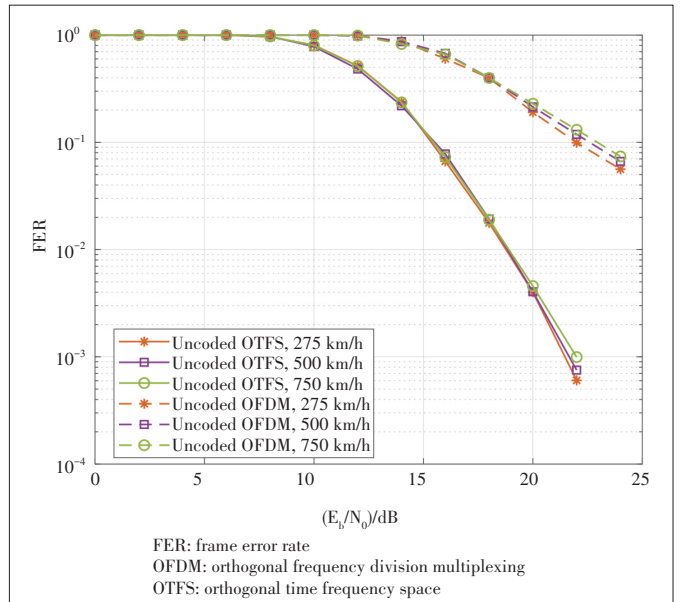
A simple and direct method to analyze coded OTFS systems is to use the extrinsic information transfer (EXIT) chart^[24], which is commonly used to aid the construction of good iteratively-decoded error-correcting codes. EXIT charts are especially popular in the analysis of low-density parity-check (LDPC) codes and Turbo codes. In the most works of coded OFDM systems, the tool, EXIT chart, is also commonly utilized to optimize the performance of iterative decoding, and parameters of the corresponding channel coding scheme and detection, such as Refs. [25 – 26].

Another possible method is to learn from the code construction method under ISI channels. In general, ISI channels can be conveniently represented by a trellis^[27] or a factor graph^[28]. Codes such as Turbo codes and LDPC codes can also be represented by a trellis or a factor graph. Note this, the channel factor graph and the code factor graph are considered together to obtain the joint channel/code graph in Ref. [29]. The limits of the performance of LDPC codes over binary linear ISI channels are also studied in Ref. [29]. With the use of density evolution, the noise tolerance threshold is calculated. This may provide some reference for the design of coded OTFS system, because the received signals with ISI can also be represented by a trellis or a factor graph.

4 Numerical Results

Numerical results of the considered coded OTFS system are provided in this section. The 5G LDPC code is used, whose code rate and length of information sequence are $R = 1/2$ and $K = 1\,024$, respectively. Without loss of generality, quadrature phase shift keying (QPSK) and 16QAM are chosen as the traditional modulation schemes, whose corresponding OTFS frame sizes are $M \times N = 64 \times 16$ and $M \times N = 32 \times 16$, respectively. In all simulations, the LTV channel with path number $P = 4$ is used, where the path gain follows the Rayleigh distribution with respect to the exponential power delay profile. For the DD domain channel, the indices of delay and Doppler shifts are integers. Moreover, according to 4G LTE and 5G NR, the carrier frequency and subcarrier interval are selected as 4 GHz and 15 kHz, respectively. Thus, we consider the maximum delay index $l_{\max} = 5$ and the maximum Doppler shift index $k_{\max} = 1, 2, 3$, corresponding to the cases in which relative speeds are around 275 km/h, 500 km/h, and 750 km/h, respectively. It should be noted that the delay and Doppler shift indices are generated uniformly at random. At the receiver, the near-optimal symbol-by-symbol MAP detection algorithm^[19] is used, unless otherwise specified. In order to accelerate the iterative convergence between the detector and the decoder, the offset min-sum algorithm (MSA) with an offset factor of 0.5 is adopted by the decoder. The maximum iteration number of the detection is 10, while that of the decoding is 50.

Fig. 3 shows the frame error rate (FER) performances of the uncoded OTFS and OFDM systems with 16QAM and different relative speeds, such as 275 km/h, 500 km/h and 750 km/h. For a fair comparison, we also apply the near-optimal symbol-by-symbol MAP detection^[19] for OFDM systems, which is designed to exploit all the interference (including both ISI and ICI). Unless otherwise specified, this detection algorithm will be used in subsequent simulations of uncoded/coded OFDM systems. As shown in Fig. 3, we first observe that both uncoded OTFS and OFDM systems have good robustness at different relative speeds. This is because the channel coherence time ($T_c = 1/f_d = 0.981, 0.539, 0.360$ ms corresponding to relative speeds of 275 km/h, 500 km/h and 750 km/h, respectively) is longer than the OFDM symbol time ($T_s = 1/\Delta f = 0.067$ ms). The channel variation is slow at considered relative speeds, and the interference between adjacent subcarriers demonstrates similar property. It is also assumed that the channel state information is perfectly known to the receiver. Thus, with the use of the near-optimal symbol-by-symbol MAP detection, all the ISI and ICI can be effectively cancelled. Furthermore, the error performances for OFDM transmission with considered relative speeds are similar. On the other hand, for OTFS transmission, different Doppler shifts caused by different relative speeds do not change the 2D convolution nature of the signal-channel interaction in the DD domain. Therefore, OTFS is insensitive to Doppler effects. In addition, we notice that the OTFS system has a better error performance than the corresponding OFDM system. Moreover, the slope of the FER curve for the OTFS system is greatly higher than that for the OFDM system, which indicates that OTFS enjoys a larger diversity ad-

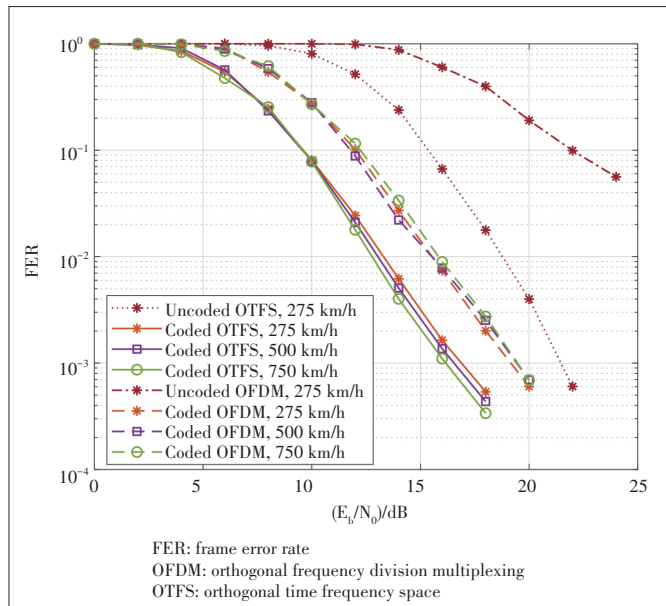


▲ Figure 3. FER performance of uncoded OTFS and OFDM systems with 16QAM, where relative speeds are 275 km/h, 500 km/h, and 750 km/h respectively

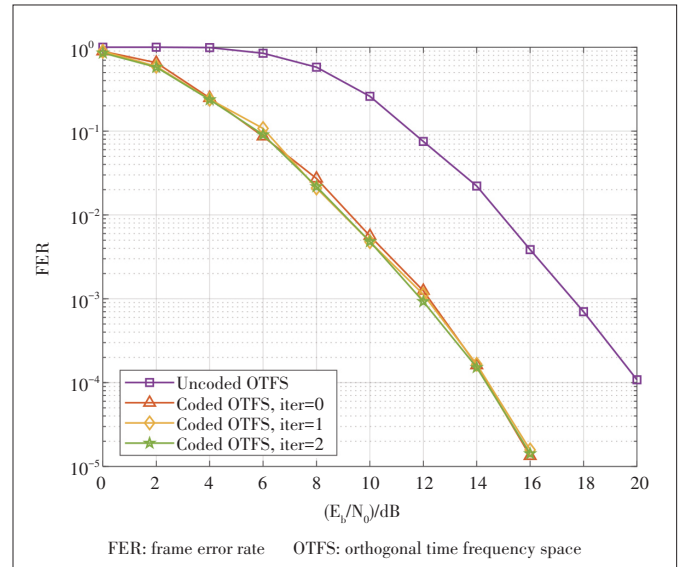
vantage. Those observations align with the findings in Refs. [5] and [20].

The FER performances of the coded OTFS and OFDM systems without the joint iteration are also shown in Fig. 4, where the relative speeds are 275 km/h, 500 km/h, and 750 km/h, respectively. The 16QAM modulated symbols are considered in the simulation. Similar to Fig. 3, we observe that the FER performances of both coded OTFS and OFDM systems do not change much with different relative speeds, thanks to the near-optimal MAP detection. Furthermore, compared with uncoded cases, both coded OTFS and OFDM systems enjoy an improved error performance. In addition, we also notice that the error performance of the coded OTFS is much better than that of the coded OFDM. However, it can be noticed that the coding improvement for the OFDM system is more significant compared with that of the OTFS system. Moreover, the FER curve of the coded OFDM system shares almost the same slope as that of the coded OTFS system. This is because OTFS has the potential to achieve the full channel diversity and consequently, channel coding cannot improve the diversity performance very much for OTFS systems. In contrast, OFDM systems rely deeply on the channel coding to achieve the larger diversity gain. Those observations are also consistent with the analysis in Ref. [20].

The FER performance of the coded OTFS system with different joint detection and decoding iterations is compared in Fig. 5, as well as the uncoded OTFS system. The modulation type is QPSK and the relative speed is 500 km/h in the simulation. We can obviously observe that the channel coding significantly improves the error performance. In addition, we notice that the iterations between the detector and decoder do not im-



▲ Figure 4. FER performance of coded OTFS and OFDM systems with 16QAM, where relative speeds are 275 km/h, 500 km/h, and 750 km/h respectively



▲ Figure 5. FER performance of coded OTFS system with different iterations between detection and decoding, where QPSK and relative speed 500 km/h are considered

prove the error performance very much. This is because the near-optimal symbol-by-symbol MAP detection algorithm used in the coded OTFS system exploits all possible interference patterns, and therefore the a priori information from channel decoding cannot improve the extrinsic information from the near-optimal detection for decoding. Consequently, the iterations between the detector and decoder cannot improve the error performance very much.

5 Conclusions

In this paper, the coded OTFS system is introduced and the existing research work of the coded OTFS is summarized. Based on this, the upper bound on unconditional PEP for the coded OTFS system is supplemented, and the design issues of the channel coding scheme are discussed. Furthermore, the performance of the uncoded OTFS and OFDM systems is analyzed, as well as that of 5G LDPC coded systems. Simulation results show that the performance of the OTFS system significantly outperforms that of OFDM system under different speeds, modulation schemes and iterations, whether coded or uncoded.

References

- [1] HADANI R, RAKIB S, TSATSANIS M, et al. Orthogonal time frequency space modulation [C]//IEEE Wireless Communications and Networking Conference. San Francisco, USA: IEEE, 2017: 1 – 6. DOI: 10.1109/WCNC.2017.7925924
- [2] HADANI R, RAKIB S, MOLISCH A F, et al. Orthogonal Time Frequency Space (OTFS) modulation for millimeter-wave communications systems [C]//IEEE MTT-S International Microwave Symposium. Honolulu, USA: IEEE, 2017: 681 – 683. DOI: 10.1109/MWSYM.2017.8058662
- [3] HADANI R, RAKIB S, TSATSANIS M, et al. Orthogonal time frequency space

- modulation [J]. IEEE wireless communications and networking conference, 2017: 1 – 6. DOI: 10.1109/WCNC.2017.7925924
- [4] HADANI R, MONK A. OTFS: A new generation of modulation addressing the challenges of 5G [EB/OL]. (2018-02-07) [2021-11-01]. <https://arxiv.org/abs/1802.02623>
- [5] RAVITEJA P, PHAN K T, HONG Y, et al. Interference cancellation and iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, 2018, 17(10): 6501 – 6515. DOI: 10.1109/TWC.2018.2860011
- [6] FARHANG A, REZAADEHREYHANI A, DOYLE L E, et al. Low complexity modem structure for OFDM-based orthogonal time frequency space modulation [J]. IEEE wireless communications letters, 2018, 7(3): 344 – 347. DOI: 10.1109/LWC.2017.2776942
- [7] REZAADEHREYHANI A, FARHANG A, JI M Y, et al. Analysis of discrete-time MIMO OFDM-based orthogonal time frequency space modulation [C]// 2018 IEEE International Conference on Communications (ICC). Kansas City, USA: IEEE, 2018: 1 – 6. DOI: 10.1109/ICC.2018.8422467
- [8] SURABHI G D, AUGUSTINE R M, CHOCKALINGAM A. On the diversity of uncoded OTFS modulation in doubly-dispersive channels [J]. IEEE transactions on wireless communications, 2019, 18(6): 3049 – 3063. DOI: 10.1109/TWC.2019.2909205
- [9] RAVITEJA P, HONG Y, VITERBO E, et al. Effective diversity of OTFS modulation [J]. IEEE wireless communications letters, 2020, 9(2): 249 – 253. DOI: 10.1109/LWC.2019.2951758
- [10] RAVITEJA P, HONG Y, VITERBO E, et al. Practical pulse-shaping waveforms for reduced-cyclic-prefix OTFS [J]. IEEE transactions on vehicular technology, 2019, 68(1): 957–961. DOI: 10.1109/TVT.2018.2878891
- [11] KOLLENGODE RAMACHANDRAN M, CHOCKALINGAM A. MIMO-OTFS in high-Doppler fading channels: signal detection and channel estimation [C]// IEEE Global Communications Conference. Abu Dhabi, United Arab Emirates: IEEE, 2018: 206 – 212. DOI: 10.1109/GLOCOM.2018.8647394
- [12] ZHANG M C, WANG F G, YUAN X J, et al. 2D structured turbo compressed sensing for channel estimation in OTFS systems [C]//2018 IEEE International Conference on Communication System. Chengdu, China: IEEE, 2018: 45 – 49. DOI: 10.1109/ICCS.2018.8689234
- [13] SHEN W Q, DAI L L, AN J P, et al. Channel estimation for orthogonal time frequency space (OTFS) massive MIMO [J]. IEEE transactions on signal processing, 2019, 67(16): 4204 – 4217. DOI: 10.1109/TSP.2019.2919411
- [14] LI M Y, ZHANG S, GAO F F, et al. A new path division multiple access for the massive MIMO-OTFS networks [J]. IEEE journal on selected areas in communications, 2021, 39(4): 903 – 918. DOI: 10.1109/JSAC.2020.3018826
- [15] ZHAO L, GAO W J, GUO W B. Sparse Bayesian learning of delay-Doppler channel for OTFS system [J]. IEEE communications letters, 2020, 24(12): 2766 – 2769. DOI: 10.1109/LCOMM.2020.3021120
- [16] WEI Z Q, YUAN W J, LI S Y, et al. Transmitter and receiver window designs for orthogonal time-frequency space modulation [J]. IEEE transactions on communications, 2021, 69(4): 2207 – 2223. DOI: 10.1109/TCOMM.2021.3051386
- [17] WEI Z Q, YUAN W J, LI S Y, et al. Orthogonal time-frequency space modulation: a promising next-generation waveform [J]. IEEE wireless communications, 2021, 28(4): 136 – 144. DOI: 10.1109/MWC.001.2000408
- [18] YUAN W J, WEI Z Q, YUAN J H, et al. A simple variational Bayes detector for orthogonal time frequency space (OTFS) modulation [J]. IEEE transactions on vehicular technology, 2020, 69(7): 7976 – 7980. DOI: 10.1109/TVT.2020.2991443
- [19] LI S Y, YUAN W J, WEI Z Q, et al. Hybrid MAP and PIC detection for OTFS modulation [J]. IEEE transactions on vehicular technology, 2021, 70(7): 7193 – 7198. DOI: 10.1109/TVT.2021.3083181
- [20] LI S Y, YUAN J H, YUAN W J, et al. Performance analysis of coded OTFS systems over high-mobility channels [J]. IEEE transactions on wireless communications, 2021, 20(9): 6033 – 6048. DOI: 10.1109/TWC.2021.3071493
- [21] BIGLIERI E, RAVITEJA P, HONG Y. Error performance of orthogonal time frequency space (OTFS) modulation [C]//IEEE International Conference on Communications Workshops (ICC Workshops). Shanghai, China: IEEE, 2019: 1 – 6. DOI: 10.1109/ICC.2019.8756831
- [22] LI S Y W, YUAN W J, LIU C, et al. ISAC transmission framework based on spatially-spread orthogonal time frequency space modulation [EB/OL]. (2021-09-01) [2021-11-01]. <https://arxiv.org/abs/2109.00440>
- [23] ROTHSTEIN M. The Gram Matrix, ProjectionOrthogonal, and Volume [EB/OL]. [2021-10-10]. <https://www.presentica.com/doc/10443352/the-gram-matrix-orthogonal-projection-and-volume-pdf-document>
- [24] BRINK STEN. Convergence of iterative decoding [J]. Electronics letters, 1999, 35(10): 806. DOI:10.1049/el: 19990555
- [25] SAND S, PLASS S, DAMMANN A. EXIT chart analysis of iterative receivers for space-time-frequency coded OFDM systems [C]//2007 IEEE 66th Vehicular Technology Conference. Baltimore, USA: IEEE, 2007: 725 – 729. DOI: 10.1109/VETECF.2007.161
- [26] HE Y, JIANG M, LING X T, et al. Protograph-based EXIT analysis and optimization of LDPC coded DCO-OFDM in VLC systems [J]. IEEE photonics technology letters, 2018, 30(21): 1898 – 1901. DOI: 10.1109/LPT.2018.2871836
- [27] FORNEY G. Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference [J]. IEEE transactions on information theory, 1972, 18(3): 363 – 378. DOI:10.1109/TIT.1972.1054829
- [28] TANNER R. A recursive approach to low complexity codes [J]. IEEE transactions on information theory, 1981, 27(5): 533 – 547. DOI: 10.1109/TIT.1981.1056404
- [29] KAVCIC A, MA X, MITZENMACHER M. Binary intersymbol interference channels: gallager codes, density evolution, and code performance bounds [J]. IEEE transactions on information theory, 2003, 49(7): 1636 – 1652. DOI: 10.1109/TIT.2003.813563

Biographies

LIU Mengmeng received her B.S. degree in communication engineering from Xidian University, China in 2017. She is currently pursuing her Ph.D. degree with Xidian University. Her research interests include signal processing and channel coding for wireless communications.

LI Shuangyang received his B.S. and M.S. degrees from Xidian University, China in 2013 and 2016, respectively. He is currently pursuing his Ph.D. degree in Xidian University and University of New South Wales, Australia. His research interests include signal processing, channel coding and their applications to communication systems.

ZHANG Chunqiong received her B.S. degree in communication engineering from Xidian University, China in 2019. She is currently pursuing the M.S. degree with Xidian University. Her research interests include signal processing and channel coding for wireless communications.

WANG Boyu received his B.S. degree in mathematics from China University of Mining and Technology-Beijing, China in 2017, and M.S. degree in mathematics from University of Sheffield, UK in 2019. He is currently studying at Xidian University, China. His research interests include coding theory, and algorithm design and analysis via convex optimization in LDPC decoding.

BAI Baoming (bmbai@mail.xidian.edu.cn) received his B.S. degree from the Northwest Telecommunications Engineering Institute, China in 1987, and the M.S. and Ph.D. degrees in communication engineering from Xidian University, China in 1990 and 2000, respectively. From 2000 to 2003, he was a senior research assistant in the Department of Electronic Engineering, City University of Hong Kong, China. Since April 2003, he has been with the State Key Laboratory of Integrated Services Networks (ISN), School of Telecommunication Engineering, Xidian University, where he is currently a professor. In 2005, he was with the University of California, USA as a visiting scholar. His research interests include information theory and channel coding, wireless communication, and quantum communication.

OTFS Enabled NOMA for mMTC Systems over LEO Satellite



MA Yiyan¹, MA Guoyu¹, WANG Ning², ZHONG Zhangdui¹, AI Bo^{1,3}

(1. Beijing Jiaotong University, Beijing 100044, China;

2. Zhengzhou University, Zhengzhou 450001, China;

3. Peng Cheng Laboratory, Shenzhen 518000, China)

Abstract: As a complement of terrestrial networks, non-terrestrial networks (NTN) have advantages of wide-area coverage and service continuity. The NTN is potential to play an important role in the 5G new radio (NR) and beyond. To enable the massive machine type communications (mMTC), the low earth orbit (LEO) satellite is preferred due to its lower transmission delay and path loss. However, the LEO satellite may generate notable Doppler shifts to degrade the system performance. Recently, orthogonal time frequency space (OTFS) modulation has been proposed. It provides the opportunity to allocate delay Doppler (DD) domain resources, which is promising for mitigating the effect of high mobility. Besides, as the LEO satellite constellation systems such as Starlink are thriving, the space spectrum resources have become increasingly scarce. Therefore, non-orthogonal multiple access (NOMA) is considered as a candidate technology to realize mMTC with limited spectrum resources. In this paper, the application of OTFS enabled NOMA for mMTC over the LEO satellite is investigated. The LEO satellite based mMTC system and the OTFS-NOMA schemes are described. Subsequently, the challenges of applying OTFS and NOMA into LEO satellite mMTC systems are discussed. Finally, the potential technologies for the systems are investigated.

Keywords: mMTC; LEO satellite; OTFS; NOMA

DOI: 10.12142/ZTECOM.202104007

<https://kns.cnki.net/kcms/detail/34.1294.TN.20211126.0933.002.html>, published online November 26, 2021

Manuscript received: 2021-10-09

Citation (IEEE Format): Y. Y. Ma, G. Y. Ma, N. Wang, et al., "OTFS enabled NOMA for mMTC systems over LEO satellite," *ZTE Communications*, vol. 19, no. 4, pp. 63 – 70, Dec. 2021. doi: 10.12142/ZTECOM.202104007.

1 Introduction

1.1 Non-Terrestrial Networks

Nowadays, non-terrestrial networks (NTN) are playing an important role in human society, especially in navigation,

ground monitoring and communications services. It regains attention recently in both academia and industry, advertised by several companies' plans of launching thousands of satellites in the low earth orbit (LEO), such as OneWeb and SpaceX. The NTN refers to the segment of network that uses air-borne or space-borne vehicles as relay nodes or base stations (BS) for transmission^[1]. In a typical NTN, it features the following elements: the satellite-gateway that connects the NTN to the public network, satellites or unmanned aerial systems (UAS), and user equipment.

There are kinds of available orbits for space-borne vehicles, including the high elliptical orbit (HEO) (400 – 50 000 km), geostationary earth orbit (GEO) (35 786 km), medium earth orbit (MEO) (7 000 – 25 000 km), and LEO (300 – 1 500 km). Different orbits exhibit different coverage and transmission characteristics, which determine the types of services on them.

This work is supported by the Fundamental Research Funds for the Central Universities under Grant Nos. 2021YJS202, 2020JBZD005 and 2021RC205, the National Key Research and Development Program under Grant Nos. 2016YFE0200900 and 2016YFB1200102-04, NSFC under Grant Nos 61725101 and U1834210, the Royal Society Newton Advanced Fellowship under Grant Nos. 61961130391 and NA191006, Beijing Natural Haidian Joint Fund under Grant No. L172020, Major Projects of Beijing Municipal Science and Technology Commission under Grant No. Z181100003218010, State Key Lab of Rail Traffic Control and Safety under Grant Nos. RCS2021ZQ002, RCS2018ZZ007 and RCS2020ZT010, Teaching Reform Project under Grant No. 134496522, the Open Research Fund from Shenzhen Research Institute of Big Data under Grant No. 2019ORF01006, and the PCL Future Greater-Bay Area Network Facilities for Large-scale Experiments and Applications Project under Grant No. LZ0019.

The non-GEO (NGSO) satellites refer to LEO and MEO satellites, whose orbital periods vary between 1.5 and 10 h. The UAS has an altitude range of 8 – 50 km and keeps a fixed position in terms of elevation with respect to a given earth point^[1].

1.2 mMTC System over LEO Satellite

In the 5G new radio (NR) system, massive machine type communications (mMTC) are introduced for Internet of Things (IoT). 5G NR considers the IoT communications with low energy consumption, low data rate, burst transmissions and massive connections. In the beyond 5G (B5G) and 6G networks, IoT communications will take a larger proportion^[2]. However, it is unavoidable that the required large-scale deployment of 5G or 6G BSs in the future could generate greater overhead compared with short-term revenue. Furthermore, terrestrial networks (TN) are hard to be deployed due to the geographical environment and cost in unmanned areas whereas IoT communication services are required. However, the NTN is not restricted by geographical environment and can still work in unmanned and geological disaster areas. Therefore, the NTN is considered to be involved as a complement of TN to construct an air-space-ground integrated network for the sake of the superiority of cost, coverage and massive connections^[1–3]. Ref. [3] demonstrates that Release 17 will study the feasibility of adapting narrowband NB-IoT to support NTN. Additionally, potential modifications of NB-IoT at physical and higher layer aspects to support NTN are investigated^[3]. For example, the transmission bandwidth can be reduced to improve uplink signal to noise ratio (SNR) and to realize transmission under limited power of devices, based on which it is possible to support low data rate GEO satellite communication using the 23 dBm device power class. Besides, directional antenna can also be equipped by mMTC devices to obtain transmission gain. Among the candidate orbits, path loss and transmission delay of LEO communications are smaller than those of MEO and GEO, so LEO is more competitive for connection of terrestrial power-constrained mMTC devices.

The 3GPP^[4] suggests that the connectivity of mMTC is 10^6 devices/km². Since the typical beam footprint size of an LEO satellite is 100 – 1 000 km^[1], there could be 10^8 – 10^9 devices inside. Though the TN could coordinate with NTN for mMTC connection, heavy control signal overhead in the grant-based multiple access (MA) technologies will be generated due to the massive concurrent devices. Besides, the scarce frequency resources are different to be allocated to the massive IoT devices orthogonally^[5]. Researchers have tried to introduce non-orthogonal multiple access (NOMA) into satellite framework^[6–8]. The related works focus on improving system spectrum efficiency and show the feasibility of applying NOMA into NTN. Moreover, the reliable access of massive devices affected by power limitation and Doppler shift remains to be a challenging topic.

On the other hand, the LEO satellite rotates at a high speed

and the Doppler shift effect is obvious. The Doppler characterization of LEO satellites was investigated in Ref. [9], as a function of the maximum elevation angular and satellite position. For instance, the normalized Doppler of the satellite with circular orbit altitude 1 000 km and inclination 53° is around 10^{-5} for a terminal located at (39° N, 77° W), which varies at the rate 0.1 ppm/s approximately. Since the severe Doppler shift could introduce inter carrier interference (ICI) to degrade the transmission performance, various Doppler estimation schemes have been proposed^[10–14]. In general, the Doppler estimators are based on either geometric information or preamble data. For the device capable of tracking satellites based on ephemeris, such as the ground gateway, the Doppler shift can be estimated and compensated according to the obtained Doppler shift curve and its own position information. However, Doppler estimation at mMTC devices could raise complexity and consume extra energy^[15]. Thus, the coexistence of Doppler shift at the user side needs to be considered, which could be handled at the gateway.

1.3 Combining OTFS and NOMA for mMTC System over LEO Satellite

Taking the effect of scarce spectrum and notable Doppler shift into account, the mMTC over LEO satellite prefers the NOMA technologies which are robust to multiple Doppler shift. Under such vision, the classic orthogonal frequency division multiple (OFDM) framework is no longer suitable for the system due to the vulnerability to Doppler shift. In Ref. [15], the adaptability of NB-IoT technologies for NTN communications was investigated, which mitigated the effect of Doppler shift by user grouping geographically. Then the maximum differential Doppler shift of users in the same group is limited to a tolerable range of NB-IoT, which is 950 Hz. Subsequently, the gateway compensates the common Doppler shift component of users in each user group. By doing so, NB-IoT is practical for IoT communications over the LEO satellite. However, this scheme is not general for the NOMA technologies, since the inter carrier interference (ICI) remains to be unprocessed. It will limit the system user capacity and improves the complexity of the ground gateway, especially for the schemes where concurrency is larger than NB-IoT. Recently, orthogonal time frequency space (OTFS) modulation has been proposed. OTFS is a chance to utilize the Doppler diversity to realize the NOMA system robustness against multiple users' Doppler shift. In what follows, the application of OTFS enabled NOMA for mMTC over LEO satellite is investigated. Specifically, the system architecture is described. Adaptability research of OTFS-NOMA for mMTC over LEO satellite is discussed. Moreover, the potential technologies for the system are listed.

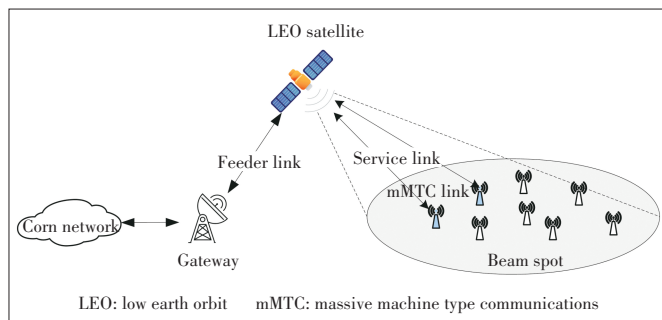
2 System Description

According to the 3GPP NTN standard^[1], there are three

parts in the mMTC system over the LEO satellite, including ground massive devices, the satellite and the ground gateway (Fig. 1). The link between the LEO satellite and mMTC devices refers to the service link while that between the LEO satellite and ground gateway refers to feeder link. In the system, the satellite could work in a transparent or regenerative mode. In the transparent mode, the satellite performs frequency carrier changing, filtering and amplifying merely. For the regenerative payload, the satellite performs signal digital processing additionally, including demodulation/re-modulation, decoding/re-encoding, etc. Due to the difficulty of air interface protocol application on the satellite, the transparent mode of the satellite is supposed in this paper. Besides, this paper concentrates on the uplink access procedure, which refers to the link from the devices to the ground gateway. The interaction between the LEO satellite and mMTC devices during uplink access depends on the utilized MA technologies. For example, four-step random access channel (RACH) enhancements for NTN are investigated in Ref. [1] considering the long transmission delay. After receiving mMTC devices' data, the LEO satellite would forward it to the ground gateway, where user identification and data recovery are performed.

As introduced before, LEO could generate a notable Doppler shift to degrade the symbol performance. In the feeder link, it is assumed that the link is available all the time. Thus, the gateway could always track the satellite and compensate the Doppler shift inside. In the service link, the Doppler shift is determined by the LEO satellite altitude, the maximum and minimum elevation angles of the devices, etc. Hence, the OTFS enabled NOMA for mMTC over the LEO satellite could be designed based on devices' geographic locations.

Current MA technologies for NTN enabled communications can be divided into the grant-based and grant-free. The grant-based MA technologies include time division multiple access (TDMA), frequency division multiple access (FDMA), code division multiple access (CDMA), physical random access channel (PRACH) in 5G-NR, etc. For example, FDMA and TDMA are applied in the Thuraya, AceS and Iridium satellite systems; CDMA is adopted by the Odyssey and Glonass systems. In the 3GPP 5G NR-NTN R16, PRACH formats and preamble sequences in Rel-15 are reused for random access^[1]. Besides,



▲ Figure 1. mMTC system over LEO satellite

several options are provided for pre-compensating the timing and frequency offset at the device side. In general, the grant-based MA have several rounds of interactions before data transmission between devices and the satellite, which are designed for device access and contention resolution. In addition, each device occupies certain specific resources independently. In the grant-free category, there are ALOHA, contention resolution diversity slotted ALOHA (CRDSA), etc. Devices in such grant-free MA systems do not require the satellite to perform dynamic scheduling authorization, but transmit data independently and obtain resources by competition. To reduce the control signal overhead, this paper adopts the grant-free access mode for the mMTC over LEO satellite.

3 OTFS-NOMA for LEO Satellite MMTC System

3.1 NOMA Schemes for LEO Satellite MMTC System

Up to now, multiple NOMA schemes have been proposed for the mMTC system over the LEO satellite^[6, 7, 16]. Simulation results demonstrate that the satellite communication systems can benefit from the application of non-orthogonal transmission schemes in terms of capacity and user fairness. However, the related studies rarely considered the Doppler effect and directly assumed perfect/imperfect channel satellite information (CSI) in the system. As introduced before, the coexistence of multiple Doppler shifts during the uplink transmission needs to be considered in the system due to the users' power limitation. Thus, OTFS is considered to be involved in such systems.

OTFS-NOMA was first proposed in Ref. [17], which considered a scenario where there was only one user with high mobility. In the mMTC system over the LEO satellite, the model in Ref. [17] could suffer more severe interference and the corresponding performance need to be evaluated. Meanwhile, there are OTFS enabled code or space domain NOMA schemes, such as OTFS-SCMA in Ref. [18] and OTFS-tandem spreading multiple access (TSMA) in Ref. [19]. The applicability and improvement of different OTFS-NOMA schemes in the LEO satellite system requires further research.

3.2 OTFS Modulation

OTFS modulation designs the transceiver in the delay Doppler (DD) domain^[20]. Note that the channel impulse response (CIR) of the double-selective channel is time varying in the time frequency (TF) domain. In the DD domain, the CIR reflects the scattering environment, which can be regarded as time-invariant during a transmission frame. Besides, CIR in the DD domain is sparse due to the limited number of scatters. The channel equalization is simplified as a result. In the TF domain, different Doppler shifts of users occupying the same TF resource block are difficult to be equalized in the grant-free NOMA mode. Meanwhile, the effect of Doppler shifts is

considered as interference in the TF domain, and its information entropy is not utilized. Now in the OTFS, transmission can be scheduled considering the users' Doppler shift characteristic, and the Doppler diversity could bring additional performance gain. In the DD domain, the transceiver input-output relationship depends on the DD resource granularity and adopted waveform types^[21]. If the waveforms ideally satisfy the bio-orthogonal property and the DD resource granularity is able to locate the CIR exactly, the output will be expressed as:

$$y[k, l] = \sum_{i=1}^P h_i e^{-j2\pi\nu_i\tau_i} x\left[\left[k - k_{\nu_i}\right]_N, \left[l - l_{\tau_i}\right]_M\right], \quad (1)$$

where x and y refer to the input and output in the DD domain respectively, P denotes the number of multipath, h_i refers to the channel fading, τ_i and ν_i are the values of delay and Doppler shift of the i -th path, $[\cdot]_N = \text{mod}(\cdot, N)$. In the TF domain, it is assumed that the system occupies the time domain resource with NT and the frequency domain resource with $M\Delta f$, where N denotes the number of time intervals T and M denotes the number of subcarrier intervals $\Delta f = 1/T$. Then the scattered Doppler domain resource is obtained as $\frac{l}{M\Delta f} \in \left[0, \frac{M-1}{M\Delta f}\right]$, $l = 0, 1, \dots, M-1$, and the scattered delay

domain resource is $\frac{k}{NT} \in \left[0, \frac{N-1}{NT}\right]$, $k = 0, 1, \dots, N-1$ ^[20].

Therein, the maximum Doppler and delay values are assumed to be less than Δf and T respectively. Additionally, suppose that τ_i and ν_i are dividable by $\frac{1}{NT}$ and $\frac{1}{M\Delta f}$, then $k_{\nu_i} = \nu_i NT$ and $l_{\tau_i} = \tau_i M\Delta f$ are obtained. Eq. (1) illustrates that based on the bio-orthogonal transceiver waveforms, y is the superposition of the faded x , which is two-dimension cyclically shifted by the corresponding DD CIR. Under the non-delicate DD resource granularity or the practical waveforms such as the rectangular waveform, there are additional ICI and inter symbol interference (ISI) in the output.

As for the OTFS receiver, the data detection and channel estimation scheme should be able to reverse the two-dimensional cyclic shift in Eq. (2). In terms of data detection, a bespoken optimal maximum a posteriori (MAP) detector was proposed in Ref. [21]. Such kind of detector is designed based on the sparsity of the DD domain channel matrix $\mathbf{H} \in \mathbb{C}^{NM \times NM}$, and is with excessive complexity. Therefore, researchers have focused on complexity reduction of the MAP detector^[22]. A variational Bayes (VB) OTFS detector was proposed in Ref. [23]. In such a detector, the distributions of OTFS symbols are constructed adaptively according to corresponding interference patterns to make the OTFS detection converge rapidly. Besides, Ref. [24] proposed a detector named cross-domain iterative detection (CDID). In this detector, a linear minimum mean squared error (L-MMSE) estimator is adopted for equal-

ization in the time domain and low-complexity symbol-by-symbol detection is utilized in the DD domain. Both VB and CDID based detectors are shown to be with close-to-optimal performance, where the computational complexity is reduced^[22]. There are other potent schemes such as the combination of hybrid MAP and PIC detection proposed in Ref. [25].

In terms of channel estimation, OTFS modulation outperforms the schemes designed in the TF domain, due to the DD domain channel sparsity and quasi-stationarity^[22]. When the channel sparsity is damaged (e.g., there are fractional Doppler shifts), the requirement of larger guard space would result in heavy training overhead. The solution to enhancing the channel sparsity has been proposed by TF domain window designing^[26]. Specifically, a Dolph-Chebyshev (DC) window is applied in the transceiver of OTFS to suppress the channel spreading. It is demonstrated that better channel estimation accuracy is obtained by the DC windowing compared with the conventional rectangular window. Moreover, coded OTFS can be used to improve the system performance^[27].

In the mMTC system over the LEO satellite, the satellite-land channel is different from that in terrestrial communications. In terms of large-scale fading, there are obvious atmospheric absorption, rain attenuation, cloud attenuation and scintillation in the satellite-land channel, which should be considered in the link budget. In terms of the small-scale fading, the LOS probability of the satellite-land channel is higher in the urban, suburban and rural scenarios^[1]. Additionally, the additive white Gaussian noise (AWGN) channel can be assumed in the open area, such as the devices on boats or aircrafts. Therefore, the DD domain CIR is sparser than that in the terrestrial communication, and the DD domain input/output relationship under the ideal pulses is written as:

$$y[k, l] = h e^{-j2\pi\nu\tau} x\left[\left[k - k_{\nu}\right]_N, \left[l - l_{\tau}\right]_M\right]. \quad (2)$$

It should be noted that k_{ν} and l_{τ} are mainly related to the devices' geographical position. Thus, the unique Doppler shift and delay values of difficult users can be regarded as a novel orthogonal space. It provides opportunity to schedule the system transmission from the perspective of devices' geography to mitigate multi-user interference (MUI). As a result, the pre-compensation of Doppler shift at devices could be omitted.

3.3 Combining NOMA with OTFS for mMTC over LEO Satellite

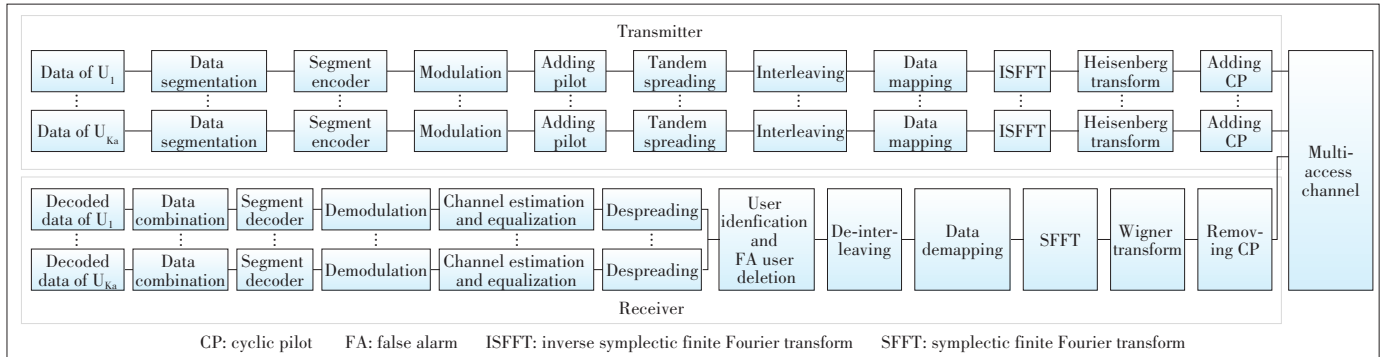
To embed NOMA into OTFS framework for mMTC over the LEO satellite, a novel transceiver could be designed. In terms of the transmitter, the main concern is to design the resource allocation scheme adapting to the DD domain channel characteristics. In terms of the receiver, equalization technologies need to be considered for the grant-free transmission. Furthermore, strategies to scale up the system capacity, such as massive multiple input multiple output (MIMO) and access point

(AP) assistant communications, could be involved in. For the code domain grant-free OTFS-NOMA schemes, data spreading and interleaving are able to reverse two-dimension cyclic shift of the DD domain, like those shown in OTFS-TSMA^[19].

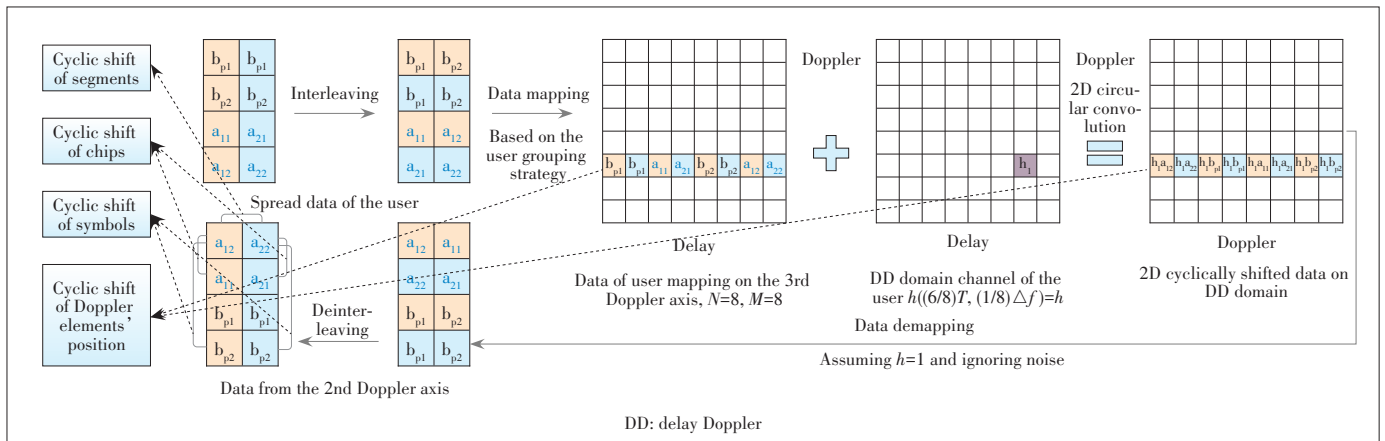
TSMA is a code domain NOMA scheme, which innovates on transceiver design and data settings. At the TSMA transmitter, users' data are segmented and encoded to generate redundant segments. Subsequently, each user is allocated with a unique tandem spreading codeword, considering the indexes of the spreading sequences utilized by each segment. Then the segments are tandemly spread according to the codeword. Therein, the tandem spreading codebook C is designed according to the maximum distance separable (MDS) code, which maximizes the user capacity under the limitation of the number of colliding segments. At the TSMA receiver, the superposition of active users' data is separated by orthogonal correlation detection on each segment according to C . Then the active users set and corresponding data are obtained. Subsequently, the colliding segments inside are deleted and the redundant segments are used for segment decoding. Therefore, TSMA scales up the user capacity by tandem spreading, ensures transmission reliability by orthogonal spreading sequences and sacrifices data rate.

As shown in Fig. 2, novel data interleaving and correspond-

ing data recovery strategies are proposed for OTFS-TSMA. The chip-level data interleaving/de-interleaving procedures are demonstrated in Fig. 3. It shows that based on the interleaving/de-interleaving strategies, the two-dimension cyclic shift of DD domain resources is transformed into cyclic shifts of Doppler domain elements, segments, systems and chips. Additionally, OTFS-TSMA adopts the cyclic orthogonal spreading sequences, namely discrete Fourier transform (DFT) sequences, whose orthogonality is maintained by the interleaving/de-interleaving strategies. At the OTFS-TSMA receiver, the segment-level cyclic shift is recovered during user identification, the symbol-level cyclic shift is recovered after de-spreading, phase rotation caused by the chip level cyclic shift is recovered before channel equalization, and Doppler diversity brought by the Doppler elements level cyclic shift is utilized during data combination. Therefore, two-dimension cyclic convolution of the DD domain is de-solved from the perspective of data design. Besides, OTFS-TSMA cleverly combines the multiple access characteristics of TSMA and the robustness of OTFS to dual selective channels. It could be involved into OTFS-NOMA enabled mMTC over the LEO satellite and inspire other OTFS-NOMA schemes of the code domain. Additionally, to make full use of DD domain resources, the schemes need to consider users' geometry positions.



▲ Figure 2. Orthogonal time frequency space (OTFS)-tandem spreading multiple access (TSMA) transmission diagram in Ref. [19]



▲ Figure 3. Resource allocation and data interleaving demonstration of orthogonal time frequency space (OTFS)-tandem spreading multiple access (TSMA) in Ref. [19]

Therefore, users with certain longitude and latitude values are scheduled to transmit during certain time slots according to their DD domain CIR characteristics.

For the power domain OTFS-NOMA, MUI brought by the DD domain channel could be mitigated by resource, rate and power allocation policies. In Ref. [17], users with high mobility are served in the DD domain and those with low mobility are serviced in the TF domain. Inspired by Ref. [17], both the TF and DD domains could be utilized for mMTC over the LEO satellite. For example, two users with the same Doppler shift and delay values (symmetrical with the sub satellite trajectory) could be serviced in two domains respectively. Besides, the users could be scheduled according to the channel conditions, whereas adaptive-rate transmission or fixed-rate transmission could be considered depending on the NOMA users' ability. Furthermore, power allocation schemes could be designed to facilitate the implementation of successful success interference cancellation (SIC), according to the users' QoS requirements and channel conditions. Note that the power domain OTFS-NOMA schemes rarely consider the uplink channel estimation, which always assume perfect CSI at transceiver. They perform better in system design in terms of communication capacity and transmission fairness. On the fast-varying satellite-ground channel, the performance of power domain OTFS-NOMA under imperfect CSI remains to be studied. Furthermore, in terms of the grant-free OTFS-NOMA design for mMTC systems over the LEO satellite, data colliding caused by non-cooperative transmission would effect user identification and data recovery. A classic grant-free OTFS-NOMA scheme utilizes the sparsity of devices^[28], which is the idea from compressive sensing based multi-user detection (CSMUD). Therein, each user is provided with a preamble sequence known to the receiver, and the sequences are used for user identification and channel equalization based on compressed sensing. Next, the data could be divided by SIC. There is also another way to design the grant-free OTFS-NOMA in the time domain using the idea of ALOHA. Overall, the performance degradation brought by non-cooperative transmission in the grant-free OTFS-NOMA mode can be compensated by additional design in space, code, power, time or frequency domains.

3.4 Challenges of Combining OTFS and NOMA for mMTC over LEO Satellite

Firstly, the effect of a notable Doppler shift and heterogeneous delay is challenging for the system design. In the OTFS modulation, the DD domain resource plane requires that the subcarrier interval should be larger than the maximum Doppler shift, and the time interval should be larger than the maximum delay. Therefore, the required time and frequency resources are dozens of times that of similar services on ground communications, due to the normalized Doppler shift at 24 ppm and the propagation delay at 25.77 ms. It seems difficult

to realize the system design with the rare space spectrum at sub-6G which is preferred for mMTC. Therefore, the Doppler and delay processing schemes need to be designed for OTFS-NOMA in the grant-free mode. For example, the system could concentrate on the differential values of the Doppler shift and delay, rather than the absolute values. In addition, the flexible configuration and mobility of the satellite system will bring different delay and Doppler characteristics. It also puts forward requirements for the flexibility of Doppler and delay processing schemes.

Secondly, the performance of OTFS-NOMA under rectangular pulses and fractional Doppler shift needs to be investigated. The DD domain input/output relationship in this scenario is written as:

$$y[k, l] \approx \sum_{q=-N'}^{N'} h e^{j2\pi \left(\frac{l-l_r}{M} \right) \left(\frac{k-k_r+\kappa_r}{N} \right)} \alpha(k, l, q) x \left[\left[k - k_r + q \right]_N, \left[l - l_r \right]_M \right], \quad (3)$$

where

$$\alpha(k, l, q) = \begin{cases} \frac{1}{N} \beta(q) & l_r \leq l < M \\ \frac{1}{N} (\beta(q) - 1) e^{-j2\pi \frac{[k - k_r + q]_N}{N}} & 0 \leq l < l_r \end{cases}, \quad (4)$$

and

$$\beta(q) = \frac{e^{-j2\pi(-q-\kappa_r)} - 1}{e^{-j\frac{2\pi}{N}(-q-\kappa_r)} - 1}. \quad (5)$$

Eqs. (3) - (5) illustrate that ICI and ISI are introduced. Therefore, the system capacity and transmission reliability performance of OTFS-NOMA remain to be investigated. Besides, the OTFS-NOMA transmitter can be re-designed considering the geometry position of signal x and the LEO satellite, which determines k_r and l_r , where the unknown quantity is only h . The novel resource allocation scheme can be proposed correspondingly.

Thirdly, the retransmission, beamforming and handover strategies remain to be studied. Due to propagation fading, the retransmission strategy for massive devices needs to be designed. It is constrained by the short transmission window, limited devices power and overall throughput. In addition, since the massive terrestrial devices need to be served by the earth-fixed beams, beamforming of the LEO satellite should be investigated. It could expand the system user capacity in the meantime. Besides, owing to the high mobility of the LEO satellite, a handover proposal for reliable communications is required.

3.5 Potential Technologies for OTFS Enabled NOMA for mMTC over LEO Satellite

In addition to the technologies for OTFS modulation and NOMA, some additional technologies are still needed by the mMTC over LEO satellite to achieve reliable access and communications.

Since the cost of cyclic pilot (CP) could be huge under the large delay and the Doppler shift is notable, frequency and time advance at devices can be adopted. Therein, the propagation delay and Doppler shift could be formulated by the location information and the LEO ephemeris. If the devices do not support such kind of calculation, certain value of Doppler shift and time advance could be initialized in the devices. Thereafter, the LEO satellite concentrates on the differential Doppler shift and delay, which is much smaller than the absolute value.

To tackle with the differential Doppler shift and delay, user grouping based on geography is a potential solution. It is able to limit the differential values of the users' Doppler shift and delay in a certain interval. Subsequently, the interference could be mitigated and the resources required by the OTFS could be reduced. Besides, resource allocation and data mapping schemes of OTFS could be designed correspondingly. In order to realize grouping and seamless convergence of the massive devices in the earth-fixed beams, beamforming strategies are required. They can also be combined with spectrum sharing technologies, such as the cognitive radio, to increase the spectrum resources efficiency.

In the wide area application of mMTC over the LEO satellite, handover strategies need to be considered to enable the constant service. Firstly, the spot beam handover, which switches the connection to a difficult spot-beam, could be researched to achieve the maximum system throughput. Secondly, fast inter-satellite handover could be investigated to connect the worldwide mMTC devices, under the high mobility of the LEO satellite. Therein, the design of LEO satellite cancellations is required.

In addition to mMTC over the LEO satellite, the air network consisting of the high-altitude platforms (HAPs) and unmanned aerial vehicles (UAVs) could complement the NTN. In detail, the HAPs are with lower mobility and smaller communication coverage properties. Therefore, heterogeneous architecture can be adopted according to the service requirement. Additionally, the mobility management and routing algorithms need to be considered.

4 Conclusions

In this paper, we investigate the OTFS enabled NOMA for mMTC systems over the LEO satellite. The architecture and scenarios of the LEO satellite enabled mMTC systems are described. The MA schemes in the current NTN are also listed. Under the contradiction between the massive access and scarce

spectrum, the NOMA technology is introduced in the system. Moreover, in the grant-free access system, OTFS is introduced to tackle with the notable Doppler shift by Doppler diversity, rather than pre-compensation. The designs and challenges of applying OTFS and NOMA into mMTC systems over the LEO satellite are then analyzed. In the end, the potential technologies and further studies of the system are suggested.

References

- [1] 3GPP. Solutions for NR to support non-terrestrial networks (NTN) (Release 16): TR 38.821 v16.0.0 [S]. 2019
- [2] SHEN X M S, CHENG N, ZHOU H B, et al. Air-space-ground integrated network technology: exploration and prospects [J]. Chinese journal on Internet of Things, 2020, 4(3): 3 – 19. DOI: 10.11959/j.issn.2096-3750.2020.00142
- [3] LIBERG O, LÖWENMARK S E, EULER S, et al. Narrowband Internet of Things for non-terrestrial networks [J]. IEEE communications standards magazine, 2020, 4(4): 49 – 55. DOI: 10.1109/MCOMSTD.001.2000004
- [4] YAN X J, AN K, LIANG T, et al. The application of power-domain non-orthogonal multiple access in satellite communication networks [J]. IEEE access, 2019, 7: 63531 – 63539. DOI: 10.1109/ACCESS.2019.2917060
- [5] CHAE S H, JEONG C, LEE K. Cooperative communication for cognitive satellite networks [J]. IEEE transactions on communications, 2018, 66(11): 5140 – 5154. DOI: 10.1109/TCOMM.2018.2850813
- [6] PEREZ-NEIRA A I, CAUS M, VAZQUEZ M A. Non-orthogonal transmission techniques for multibeam satellite systems [J]. IEEE communications magazine, 2019, 57(12): 58 – 63. DOI: 10.1109/MCOM.001.1900249
- [7] CHU J H, CHEN X M, ZHONG C J, et al. Robust design for NOMA-based multibeam LEO satellite Internet of Things [J]. IEEE Internet of Things journal, 2021, 8(3): 1959 – 1970. DOI: 10.1109/JIOT.2020.3015995
- [8] LIU X, ZHAI X B, LU W D, et al. QoS-guarantee resource allocation for multibeam satellite industrial Internet of Things with NOMA [J]. IEEE transactions on industrial informatics, 2021, 17(3): 2052 – 2061. DOI: 10.1109/TII.2019.2951728
- [9] ALI I, AL-DHAHIR N, HERSHEY J E. Doppler characterization for LEO satellites [J]. IEEE transactions on communications, 1998, 46(3): 309 – 313. DOI: 10.1109/26.662636
- [10] YOU M H, LEE S P, HAN Y. Adaptive compensation method using the prediction algorithm for the doppler frequency shift in the LEO mobile satellite communication system [J]. ETRI journal, 2000, 22(4): 32 – 39. DOI: 10.4218/etri.00.0100.0404
- [11] LIN J N, HOU Z W, ZHOU Y Q, et al. Map estimation based on Doppler characterization in broadband and mobile LEO satellite communications [C]//83rd Vehicular Technology Conference (VTC Spring). Nanjing, China: IEEE, 2016: 1 – 5. DOI: 10.1109/VTCSpring.2016.7504336
- [12] LIU Y J, ZHU X, LIM E G, et al. High-robustness and low-complexity joint estimation of TOAs and CFOs for multiuser SIMO OFDM systems [J]. IEEE transactions on vehicular technology, 2018, 67(8): 7739 – 7743. DOI: 10.1109/TVT.2018.2821152
- [13] TIAN D, ZHAO Y, TONG J F, et al. Frequency offset estimation for 5G based LEO satellite communication systems [C]//IEEE/CIC International Conference on Communications in China (ICCC). Changchun, China: IEEE, 2019: 647 – 652. DOI: 10.1109/ICCCChina.2019.8855824
- [14] PAN M G, HU J L, YUAN J H, et al. An efficient blind Doppler shift estimation and compensation method for LEO satellite communications [C]//20th International Conference on Communication Technology (ICCT). Nanning, China: IEEE, 2020: 643 – 648. DOI: 10.1109/ICCT50939.2020.9295821
- [15] KODHELI O, ANDRENACCI S, MATURO N, et al. An uplink UE group-based scheduling technique for 5G mMTC systems over LEO satellite [J].

- IEEE access, 2019, 7: 67413 – 67427. DOI: 10.1109/ACCESS.2019.2918581
- [16] ZHANG Z J, LI Y, HUANG C W, et al. User activity detection and channel estimation for grant-free random access in LEO satellite-enabled Internet of Things [J]. IEEE Internet of Things journal, 2020, 7(9): 8811 – 8825. DOI: 10.1109/IIOT.2020.2997336
- [17] DING Z G, SCHÖBER R, FAN P Z, et al. OTFS-NOMA: an efficient approach for exploiting heterogeneous user mobility profiles [J]. IEEE transactions on communications, 2019, 67(11): 7950 – 7965. DOI: 10.1109/TCOMM.2019.2932934
- [18] DEKA K, THOMAS A, SHARMA S. OTFS-SCMA: A code-domain NOMA approach for orthogonal time frequency space modulation [J]. IEEE transactions on communications, 2021, 69(8): 5043 – 5058. DOI: 10.1109/TCOMM.2021.3075237
- [19] MA Y Y, MA G Y, WANG N, et al. OTFS-TSMA for massive Internet of Things in high-speed railway [J]. IEEE transactions on wireless communications, Early access, 2021. DOI: 10.1109/TWC.2021.3098033
- [20] HADANI R, RAKIB S, TSATSANIS M, et al. Orthogonal time frequency space modulation [C]/Wireless Communications and Networking Conference (WCNC). San Francisco, USA: IEEE, 2017: 1 – 6. DOI: 10.1109/WCNC.2017.7925924
- [21] RAVITEJA P, PHAN K T, HONG Y, et al. Interference cancellation and iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, 2018, 17(10): 6501 – 6515. DOI: 10.1109/TWC.2018.2860011
- [22] WEI Z, YUAN W, LI S, et al. Orthogonal time-frequency space modulation: a promising next generation waveform [J]. IEEE wireless communications, 2021, 28(4): 136 – 144. DOI: 10.1109/MWC.001.2000408
- [23] YUAN W, WEI Z, YUAN J, et al. A simple variational bayes detector for orthogonal time frequency space (OTFS) modulation [J]. IEEE transactions on vehicular technology, 2020, 69(7): 7976 – 7980. DOI: 10.1109/TVT.2020.2991443
- [24] LI S, YUAN W, WEI Z, et al. Cross domain iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, 2021, early access. DOI: 10.1109/TWC.2021.3110125
- [25] LI S, YUAN W, WEI Z, et al. Hybrid MAP and PIC detection for OTFS modulation [J]. IEEE transactions on vehicular technology, 2021, 70(7): 7193 – 7198. DOI: 10.1109/TVT.2021.3083181
- [26] WEI Z, YUAN W, LI S, et al. Transmitter and receiver window designs for orthogonal time-frequency space modulation [J]. IEEE transactions on communications, 2021, 69(4): 2207 – 2223. DOI: 10.1109/TCOMM.2021.3051386
- [27] LI S, YUAN J, YUAN W, et al. Performance analysis of coded OTFS systems over high-mobility channels [J]. IEEE transactions on wireless communications, 2021, 20(9): 6033 – 6048. DOI: 10.1109/TWC.2021.3071493
- [28] WANG F G, MA G Y. Massive machine type communications: multiple access schemes [M]. Heidelberg, Germany: Springer, 2019

Biographies

MA Yiyao received the B.S. degree in applied physics from Beijing Jiaotong University, China in 2019, and is currently working toward the Ph.D. degree at the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. His current research interests include the field of Internet of Things and massive machine type communications.

MA Guoyu (magy@bjtu.edu.cn) received the B.S. and Ph.D. degrees in electrical engineering from Beijing Jiaotong University, China in 2012 and 2019, re-

spectively. Currently he is an associate professor at the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. His current research interests include machine-type communications and random access.

WANG Ning received the B.E. degree in communication engineering from Tianjin University, China in 2004, the M.A.Sc. degree in electrical engineering from The University of British Columbia, Canada in 2010, and the Ph.D. degree in electrical engineering from the University of Victoria, Canada in 2013. He was on the Finalist of the Governor General's Gold Medal for Outstanding Graduating Doctoral Student with the University of Victoria in 2013. From 2004 to 2008, he was with the China Information Technology Design and Consulting Institute as a mobile communication system engineer, specializing in planning and design of commercial mobile communication networks, network traffic analysis, and radio network optimization. He was a postdoctoral research fellow with the Department of Electrical and Computer Engineering, The University of British Columbia, from 2013 to 2015. Since 2015, he has been with the School of Information Engineering, Zhengzhou University, China, where he is currently an associate professor. He also holds adjunct appointments with the Department of Electrical and Computer Engineering, McMaster University, Canada, and the Department of Electrical and Computer Engineering, University of Victoria. He has served on the technical program committees of international conferences, including the IEEE GLOBECOM, IEEE ICC, IEEE WCNC, and CyberC. His research interests include resource allocation and security designs of future cellular networks, channel modeling for wireless communications, statistical signal processing, and cooperative wireless communications.

ZHONG Zhangdui is currently a professor and advisor of Ph.D. candidates with Beijing Jiaotong University, China. He is also a chief scientist with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. He is a director of the Innovative Research Team of Ministry of Education and a chief scientist of Ministry of Railways in China. He is an executive council member of Radio Association of China and a deputy director of Radio Association of Beijing. He has authored/coauthored seven books, five invention patents, and more than 200 scientific research papers in his research area. His interests include wireless communications for railways, control theory and techniques for railways, and GSM-R system. His research results have been widely used in railway engineering, including the Qinghai-Xizang railway, Datong-Qinhuangdao heavy haul railway and many high-speed railway lines in China. Prof. ZHONG was the recipient of a Maoyisheng Scientific Award of China, Zhantianyou Railway Honorary Award of China, and Top Ten Science/Technology Achievements Award of Chinese Universities. He is a fellow of IEEE.

AI Bo (boai@bjtu.edu.cn) graduated from Tsinghua University, China, with the honor of Excellent Postdoctoral Research Fellow in 2007. He received the master's and Ph.D. degrees from Xidian University, China in 2002 and 2004, respectively. He is currently working as a full professor and a Ph.D. advisor with Beijing Jiaotong University, China, where he is also the deputy director of the State Key Laboratory of Rail Traffic Control and Safety and the International Joint Research Center. He has authored or coauthored eight books and published over 300 academic research papers in his research area. He holds 26 invention patents. He has been the research team leader for 26 national projects and has won some important scientific research prizes. Five of his papers have been the ESI highly cited papers. He has been notified by the Council of Canadian Academies (CCA) that, based on Scopus database, he has been listed as one of the Top 1% authors in his field all over the world. His research interests include the research and applications of channel measurement and channel modeling and dedicated mobile communications for rail traffic systems. He is a fellow of IEEE and IET.



Orthogonal Time Frequency Space Modulation in Multiple-Antenna Systems

WANG Dong¹, WANG Fanggang¹, LI Xiran¹, YUAN Pu², JIANG Dajie²

(1. State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China;

2. Vivo Mobile Communication Co., Ltd., Beijing 100016, China)

Abstract: The application of the orthogonal time frequency space (OTFS) modulation in multiple-antenna systems is investigated. The diversity and/or the multiplexing gain can be achieved by deploying various multiple-antenna techniques, and thus the reliability and/or the spectral efficiency are improved correspondingly. We provide two classes of OTFS-based multiple-antenna approaches for both the open-loop and the closed-loop systems. Specifically, in the open-loop system, a transmitting diversity approach, which resembles the space-time coding technique, is proposed by allocating the information symbols appropriately in the delay-Doppler domain. In the closed-loop system, we adopt the Tomlinson-Harashima precoding in our derived delay-Doppler equivalent transmission model. Numerical evaluations demonstrate the advantages of applying the multiple-antenna techniques to the OTFS. At last, several challenges and opportunities are presented.

Keywords: OTFS; space-time coding; Tomlinson-Harashima precoding

DOI: 10.12142/ZTECOM.202104008

<https://kns.cnki.net/kcms/detail/34.1294.TN.20211102.1456.002.html>, published online December 5, 2021

Manuscript received: 2021-10-13

Citation (IEEE Format): D. Wang, F. G. Wang, X. R. Li, et al., "Orthogonal time frequency space modulation in multiple-antenna systems," *ZTE Communications*, vol. 19, no. 4, pp. 71 - 78, Dec. 2021. doi: 10.12142/ZTECOM.202104008.

1 Introduction

Recently, the orthogonal time frequency space (OTFS) modulation has been proposed to deal with the severe inter-carrier interference problem caused by the Doppler effect in the high-speed mobility scenario^[1]. In the OTFS system, the modulation symbols are multiplexed in the delay-Doppler domain instead of the time-frequency domain, which provides an alternative representation of a time-varying channel geometry modeling of the transmission paths^[2]. However, the two-dimensional convolution of the input-output relation in the OTFS system makes the equalization involved. The OTFS equalization schemes have been ex-

tensively studied^[3-5]. In Ref. [3], a message-passing (MP) algorithm was developed for joint interference cancellation and symbol detection. In Ref. [4], the authors proposed a cross domain iterative detection algorithm to enhance the error performance of OTFS modulation.

By utilizing the mathematical least-square minimum residual algorithm, the authors in Ref. [5] proposed a low-complexity equalizer and a block-wise interference eliminator.

The aforementioned OTFS schemes focus on the single-input and single-output system. By exploiting the multiple antenna techniques, the reliability and/or the spectral efficiency can be improved in the OTFS multiple-antenna system. Similar to the orthogonal frequency division multiplexing (OFDM) multiple-antenna system, the transmission schemes in the OTFS multiple-antenna system can be categorized into two classes: open-loop and closed-loop. The main difference is that the channel state information (CSI) reported to the transmitter is

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant Nos. 2020JBM081 and 2020JBZD005, in part by the National Key R&D Program of China under Grant 2020YFB1806903, in part by the National Natural Science Foundation under Grant No. U1834210.

required in the closed-loop system whereas it is not necessary for the open-loop system. In the literature of the open-loop OTFS, the authors in Ref. [6] proposed a low-complexity linear equalizer by utilizing the block circulant property of the equivalent channel. In Ref. [7], the quasi-banded sparse structure of the equalization matrix was used to design a low-complexity linear equalization scheme in the delay-Doppler domain. However, the transmission schemes in Refs. [6] and [7] only consider the equalization at the receiver without designing the transmitter, the spatial diversity of which is not fully exploited. Then, several papers studied the spatial diversity schemes in OTFS multiple-antenna systems, which can be harnessed by the space-time coding scheme^[8-10]. In Ref. [11], a space-time coding scheme with cyclic delay diversity was presented. By assuming that the delay-Doppler channel is invariant in the two consecutive OTFS frames, a space-time code using the Alamouti code structure is proposed in the open-loop system^[12]. However, the channel varies drastically in the two consecutive OTFS frames under the high-speed scenarios, which degrades the bit error ratio (BER) performance. In the closed-loop transmission, the feedback of the channel information is required, which further improves the throughput of the OTFS multiple-antenna system. The authors in Ref. [13] proposed a Tomlinson-Harashima precoding (THP) scheme in the delay-time domain. In Ref. [14], an uplink-aided high mobility downlink channel estimation scheme for the massive multiple-input and multiple-output (MIMO)-OTFS systems was proposed. The expectation-maximization-based variational Bayesian framework was adopted to recover the uplink channel parameters including the angle, the delay, the Doppler frequency, and the channel gain for each physical scattering path. Then, in Ref. [15], the authors designed an effective path scheduling algorithm to map different users to the delay-Doppler domain grids without inter-user interference in the same OTFS block. The authors in Ref. [16] proposed a maximal ratio combining (MRC) precoding scheme in the multi-user massive MIMO-OTFS system. However, the previous work mainly focuses on the ideal pulse shaping for the transmission design whereas the problem of its application with the practical pulses, such as the rectangular pulse shaping, is not fully addressed.

In this paper, we investigate the transmission schemes in the open-loop and closed-loop OTFS multiple-antenna systems. We preclude the assumption that the channel states in the two consecutive frames are invariant. Both the ideal pulse shaping and the rectangular pulse shaping are studied. The main contributions of this paper are summarized as follows:

- We design a transmit diversity approach within one OTFS frame duration in the open-loop systems. The guards are carefully allocated to avoid sub-frames interfering with each other and to ensure the two sub-frames can be regarded as passing the identical channel in the delay-Doppler domain.

- We then design a scheme with the rectangular pulse shaping. In addition to the guards in the ideal pulse shaping, we

place the guards at the last l_{\max} symbols along the delay domain in the rectangular pulse shaping. Then, the channel of each sub-frame in the equivalent transmission model is identical.

- We further indicate that the linear precoding in the time-frequency domain can be alternatively expressed in the delay-Doppler domain. By deriving the equivalent transmission model in the OTFS system, the THP-based schemes in the delay-Doppler domain under the zero-forcing (ZF) and minimum mean square error (MMSE) criterion are designed.

2 Open-Loop System

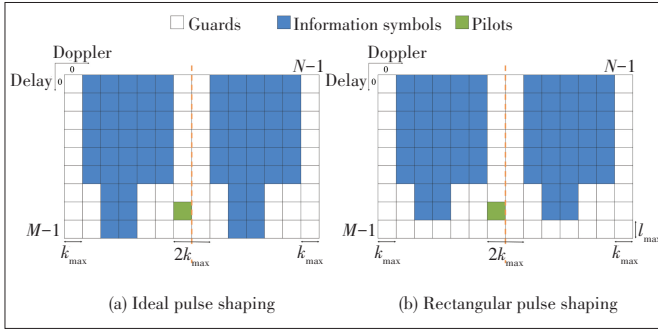
In the open-loop system, the channel information is not required at the transmitter, which is suitable for the transmission in the high-speed scenarios. In contrast to the two-frame block fading assumption^[12], we introduce the transmit diversity approach within one OTFS frame duration by appropriately allocating information symbols, guards, and pilots with the ideal pulse shaping and rectangular pulse shaping.

2.1 Spatial Diversity Approach

In this paper, we consider the wide-sense non-stationary channel with the Jakes' formula, different from the wide-sense stationary channel. The amplitude and the phase shift of each path are different in the two consecutive frames, which causes the delay-Doppler channel changes. The traditional space-time coding is applied on two-time slots by assuming the invariant channel, which is not practical in the non-stationary channel. Then, how to design a transmit diversity within one frame duration is a challenge in the OTFS system. Considering a scenario that the transmitter is equipped with two antennas and the receiver is equipped with a single antenna, we propose a spatial diversity scheme within one frame by allocating information symbols, guards, and pilots. The positions of information symbols and guards are the same for each transmit antenna, which is shown in Fig. 1. Information symbols, guards, and pilots are appropriately allocated in one OTFS frame. In addition to the guards allocated in the ideal pulse shaping, we place the guards at the last l_{\max} symbols along the delay domain in the rectangular pulse shaping. Then, the channel of each sub-frame in the equivalent transmission model is identical.

Firstly, we divide the resource units of a frame into two sub-frames along the Doppler domain. Symbols in the first sub-frame and the second sub-frame are expressed as $x[l, k_1], k_1 \in \{0, 1, \dots, \frac{N}{2} - 1\}$ and $x[l, k_2], k_2 \in \{\frac{N}{2}, \frac{N}{2} + 1, \dots, N - 1\}$, respectively. Then, the received signal of each sub-frame can be obtained as:

$$y[l, k_t] = \sum_{p=0}^{P-1} h_p e^{-j \frac{2\pi l k_t p}{MN}} x[l - l_p, k - k_p] + v[l, k_t], t \in \{1, 2\}, \quad (1)$$



▲ Figure 1. Ideal and rectangular pulse shaping

where $y[l, k_1]$ and $y[l, k_2]$ represent the received signal of the first and the second sub-frame, respectively; h_p denotes the channel response of the p -th path; l_p and k_p are the delay and Doppler indices of the p -th path. For the first sub-frame signal, since $k_1 \in \{0, 1, \dots, \frac{N}{2} - 1\}$, and $k_p \in \{-k_{\max}, \dots, 0, 1, \dots, k_{\max}\}$, we can obtain $[k - k_p]_N \in [0, \frac{N}{2} - 1 + k_{\max}] \cup [N - k_{\max}, N - 1]$. From Eq. (1), we can see that the received signal of the first sub-frame is interfered by symbols of the second sub-frame when $[k - k_p]_N > \frac{N}{2} - 1$. In order to avoid the interference with the first sub-frame, the guards are placed at the interference symbols in the second sub-frame, i. e., $x[l, k] = 0, l = \{0, 1, \dots, M - 1\}, k \in \mathcal{G}_1$, where $\mathcal{G}_1 = \{[\frac{N}{2}, \frac{N}{2} - 1 + k_{\max}] \cup [N - k_{\max}, N - 1]\}$. Moreover, interference symbols in the first frame are similar to the above analysis. Overall, the allocation of the information symbols can be obtained as $x[l, k], l = \{0, 1, \dots, M - 1\}, k \in \mathcal{G}_1 \cup \mathcal{G}_2$, where $\mathcal{G}_2 = \{[0, k_{\max} - 1] \cup [\frac{N}{2} - k_{\max}, \frac{N}{2} - 1]\}$. Therefore, two sub-frames are not interfered with each other, and the equivalent channel of each sub-frame is identical. Furthermore, by considering the channel estimation in the delay-Doppler domain, pilots and guards should be carefully placed to protect the pilots from interference with other information symbols at the receiver. In order to utilize guard patterns introduced in this section, we allocate the pilots and the guards to the last $3l_{\max} + 2$ symbol positions along the delay domain. The allocation of the pilots and the corresponding guards for the i -th antenna is given by:

$$x_i[l, k] = \begin{cases} x_0, & k = k_0, l = M - il_{\max} - 1; \\ z^{k_0 l}, & k_0 - 2k_{\max} \leq k \leq k_0 + 2k_{\max}, M - il_{\max} - 1 \leq l \leq M - 1, \end{cases} \quad (2)$$

where $k_0 = \frac{N}{2} - 1$ and $i \in \{1, 2\}$.

Next, we introduce the transmit pattern. By deploying the Alamouti code structure, the transmit vectors at the first antenna and the second antenna are obtained as $\tilde{\mathbf{x}}_1 = [\mathbf{x}_1^T, \hat{\mathbf{x}}_1^T]^T \in \mathbb{C}^{NM}$ and $\tilde{\mathbf{x}}_2 = [\mathbf{x}_2^T, \hat{\mathbf{x}}_1^T]^T \in \mathbb{C}^{NM}$, where $\hat{\mathbf{x}}_1 = \mathbf{P}\mathbf{x}_1^*$, $\hat{\mathbf{x}}_2 = -\mathbf{P}\mathbf{x}_2^*$, and \mathbf{P} is the permutation matrix. Then, the received signal can be obtained as

$$\mathbf{y}_1 = \mathbf{H}_1 \mathbf{x}_1 + \mathbf{H}_2 \mathbf{x}_2 + \mathbf{v}_1, \quad (3)$$

$$\mathbf{y}_2 = -\mathbf{H}_1 \hat{\mathbf{x}}_2^* + \mathbf{H}_2 \hat{\mathbf{x}}_1^* + \mathbf{v}_2, \quad (4)$$

where \mathbf{y}_1 and \mathbf{y}_2 are received signals at the first and the second sub-frame, respectively; \mathbf{H}_1 and \mathbf{H}_2 are the equivalent channels for the first and second sub-frame, respectively. From the block circulant property of \mathbf{H}_1 and \mathbf{H}_2 , an alternative representation of Eqs. (3) and (4) are expressed as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & -\mathbf{X}_2^H \\ \mathbf{X}_2 & \mathbf{X}_1^H \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}, \quad (5)$$

where $\mathbf{h}_1 \in \mathbb{C}^{MN/2}$ and $\mathbf{h}_2 \in \mathbb{C}^{MN/2}$ denote the first column of \mathbf{H}_1 and \mathbf{H}_2 ; \mathbf{X}_1 and \mathbf{X}_2 are the equivalent code words for the transmit vectors \mathbf{x}_1 and \mathbf{x}_2 , respectively. By applying the maximal ratio combining receiver, the received signal of the two frames can be split apart as $\tilde{\mathbf{y}}_1 = \tilde{\mathbf{H}} \mathbf{x}_1 + \tilde{\mathbf{v}}_1$ and $\tilde{\mathbf{y}}_2 = \tilde{\mathbf{H}} \mathbf{x}_2 + \tilde{\mathbf{v}}_2$, where $\tilde{\mathbf{H}} = \mathbf{H}_1^H \mathbf{H}_1 + \mathbf{H}_2^H \mathbf{H}_2$. Therefore, \mathbf{x}_1 and \mathbf{x}_2 are not interfered with each other, then the maximal likelihood (ML) detector or MMSE detector can be applied to decode \mathbf{x}_1 and \mathbf{x}_2 , respectively. However, the computational complexity of the detectors is high since the non-linear iterations is involved in the ML receiver and the inverse operation of a high-dimensional matrix is applied in the MMSE receiver. By exploiting the matrix transformation and matrix decomposition, the computational complexity of the detector can be reduced to $O(l_{\max}^2 MN^3)$, which is lower than the conventional MMSE receiver $O(M^3 N^3)$.

2.2 Modified Approach with Rectangular Pulse Shaping

In this section, the transmission scheme in the rectangular pulse shaping is considered. We introduce the allocation of information symbols with the rectangular pulse shaping and design the corresponding transmitting and receiving structure.

The allocation of the information symbols is based on the input-output relation with rectangular pulse shaping, which is given by

$$\bar{y}[l, k_t] = \sum_{p=0}^{P-1} h_p e^{-j \frac{2\pi l k_p}{MN}} \alpha_p(l, k) x[l - l_p]_M, [k - k_p]_N + v[l, k_t], t \in \{1, 2\}, \quad (6)$$

where

$$\alpha_p(l, k) = \begin{cases} e^{-j \frac{2\pi k}{N} z_p^{k_p(l+M)}}, & \text{if } l < l_p \\ z_p^{k_p l}, & \text{if } l \geq l_p. \end{cases} \quad (7)$$

Eq. (7) is the phase shift caused by the rectangular pulse shaping. The two sub-frames do not interfere with each other by allocating the guards introduced in Section 2.1. However, the equivalent channel of each sub-frame is different due to

the phase shift. From Eq. (7), we can see that the phase shift of $x[l, k]$ is related to l, k when $l < l_p$ but only related to l when $l \geq l_p$. Then, the guards are placed at the last l_{\max} symbols along the delay domain, i. e., $x[l, k] = 0, l = [M - l_{\max}, M - 1], k \in \{0, 1, \dots, N\}$ to keep the phase shift of each sub-frame identical. The allocation of information symbols, guards, and pilots in the OTFS frame is shown in Fig. 1. For the channel estimation, the position of pilots and guards is similar to that of the ideal pulse shaping in Eq. (2). Then, we also adopt the transmit code word of the Alamouti code structure in Section 2.1. However, the permutation matrix \bar{P} should be redesigned to keep the guards not being permuted. The transmit vectors at the first antenna and the second antenna are obtained as $\bar{\mathbf{x}}_1 = [\mathbf{x}_{11}^T, \mathbf{x}_{12}^T]^T \in \mathbb{C}^{NM}$ and $\bar{\mathbf{x}}_2 = [\mathbf{x}_{21}^T, \mathbf{x}_{22}^T]^T \in \mathbb{C}^{NM}$, respectively, where $\mathbf{x}_{12} = -\bar{P}\mathbf{x}_{21}^*$ and $\mathbf{x}_{22} = -\bar{P}\mathbf{x}_{21}^*$. Then, the received signal can be expressed as:

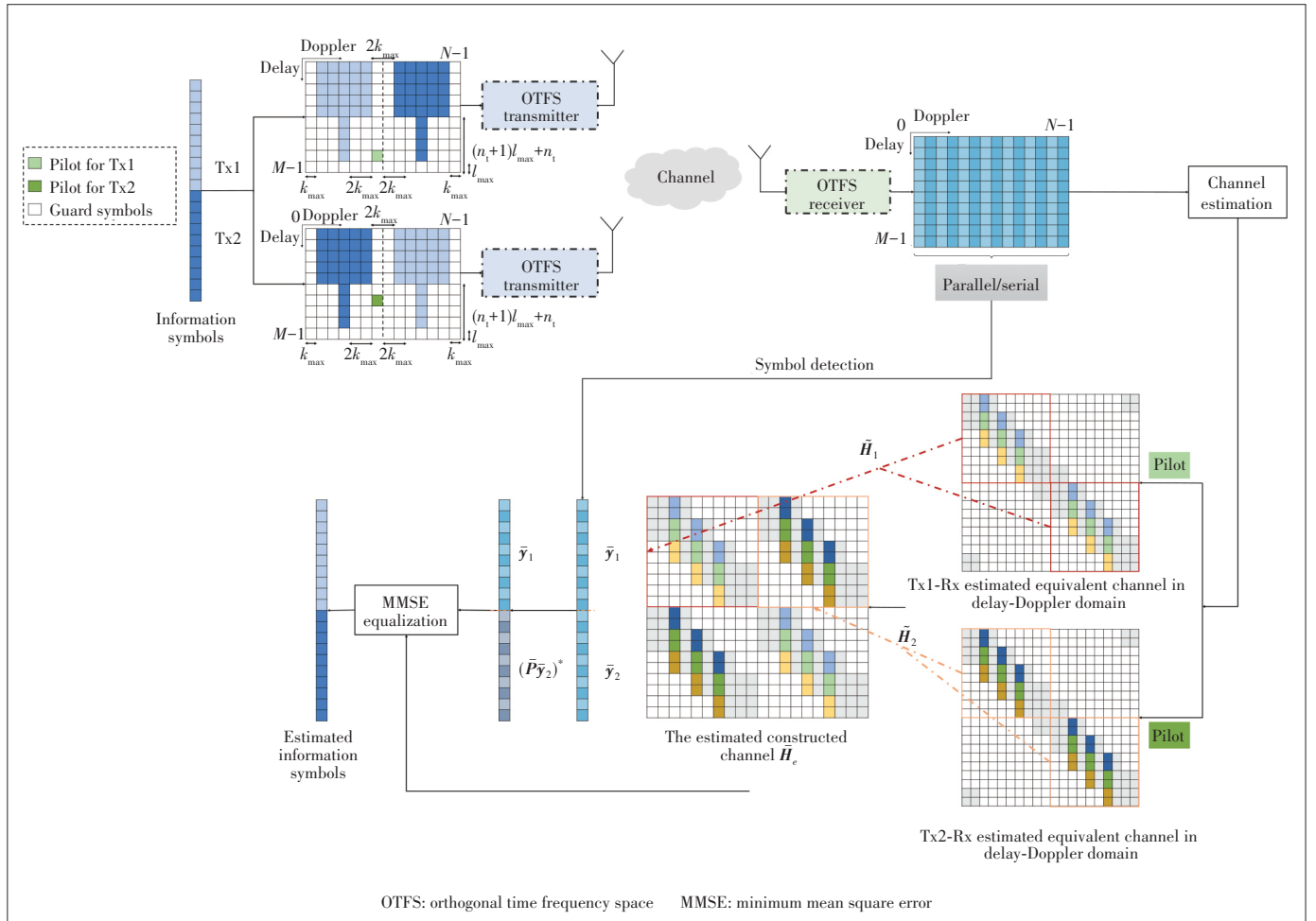
$$\begin{bmatrix} \bar{\mathbf{y}}_1 \\ \bar{\mathbf{y}}_2^* \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{H}}_1 & \bar{\mathbf{H}}_2 \\ \mathbf{P}\bar{\mathbf{H}}_2^* \mathbf{P} & -\mathbf{P}\bar{\mathbf{H}}_1^* \mathbf{P} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{11} \\ \mathbf{x}_{21} \end{bmatrix} + \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2^* \end{bmatrix}, \quad (8)$$

where $\hat{\mathbf{y}}_2 = \mathbf{P}\bar{\mathbf{y}}_2$; $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ are received signals of the first sub-

frame and the second sub-frame, respectively; $\bar{\mathbf{H}}_1$ and $\bar{\mathbf{H}}_2$ are equivalent channels for \mathbf{x}_{11} and \mathbf{x}_{21} , respectively. Then, the MMSE receiver or the MP receiver can be adopted to decode \mathbf{x}_{11} and \mathbf{x}_{21} . In summary, the transmitting and receiving process with rectangular pulse shaping is shown in Fig. 2. Moreover, the proposed diversity scheme in this section can be extended to the multiple-antenna scenarios by employing the structure similar to Jafarkhani code, which is a quasi-orthogonal space-time block coding introduced in Ref. [19].

3 Closed-Loop System

In the closed-loop system, the transmitter dynamically varies the precoding matrix based on the CSI report such as the precoding matrix index, the rank indicator, and the channel quality indicator. However, the precoding problem in the closed-loop OTFS multiple-antenna system is not fully addressed. In this section, we observe the equivalence of the linear precoding between the delay-Doppler and the time-frequency domain. Then, delay-Doppler THP (DD-THP) schemes are introduced under the ZF or the MMSE criterion.



▲ Figure 2. Proposed diversity approach with rectangular pulse shaping

Finally, we provide an overview of other closed-loop transmission schemes in the OTFS multiple-antenna systems.

3.1 Linear Precoding in Delay-Doppler Domain

In this section, the linear precoding in the time-frequency and the equivalent representation in the delay-Doppler domain are provided. We recall that in the MIMO-OFDM system, the linear codebook is selected to map symbols in the time-frequency domain to the transmit antennas. In contrast, symbols in the OTFS system are multiplexed in the delay-Doppler domain, then the precoding can be deployed in the delay-Doppler domain or the time-frequency domain. We observe that the precoding in the time-frequency domain can be alternatively represented in the delay-Doppler domain, which is shown as an example in the following. Considering a scenario that the transmitter is equipped with N_t antennas and the receiver is equipped with N_r antennas, we denote $\mathbf{W}_{\text{TF}} = \text{diag}\{\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_{MN}\}$ as the codebook in the time-frequency domain, where $\mathbf{W}_i, i = 1, 2, \dots, MN$ is the codebook on each subcarrier. Then, the equivalent precoding in the delay-Doppler domain \mathbf{W}_{DD} is given by

$$\mathbf{W}_{\text{DD}} = \left((\mathbf{F}_N^* \otimes \mathbf{F}_M) \otimes \mathbf{I}_{N_t} \right)^{-1} \mathbf{W}_{\text{TF}} \left((\mathbf{F}_N^* \otimes \mathbf{F}_M) \otimes \mathbf{I}_{N_t} \right), \quad (9)$$

where \mathbf{F}_N and \mathbf{F}_M are the discrete Fourier transform matrices with N -point and M -point, respectively; The operator \otimes denotes the Kronecker product. By the equivalent representation in Eq. (9), the precoding process can be carried out in the delay-Doppler domain instead of the time-frequency domain since the channel estimation in the delay-Doppler domain is more stable than that in the time-frequency domain, especially in high-speed scenarios.

3.2 DD-THP Approach

THP is a well-known non-linear precoding scheme, which can also be adopted in the closed-loop OTFS multiple-antenna system. In this subsection, we introduce the DD-THP schemes in the delay-Doppler domain.

In the OTFS multiple-antenna system, the vector form of input-output relationship in the delay-Doppler domain is expressed as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}, \quad (10)$$

where \mathbf{H} is the equivalent channel matrix in the delay-Doppler domain; $\mathbf{x} \in \mathbb{C}^{N_t MN}$ is the vector form of the transmit symbols in the delay-Doppler domain; $\mathbf{y} \in \mathbb{C}^{N_r MN}$ is the received symbol in the delay-Doppler domain; $\mathbf{v} \in \mathbb{C}^{N_r MN}$ is the vector form of the zero mean circularly symmetric complex Gaussian noise at the receiver.

The equivalent channel matrix \mathbf{H} is fed back to the transmitter for the precoding. A block diagram of applying THP in the OTFS multiple-antenna system is shown in Fig. 3. For the ZF-

DD-THP approach, the conjugate transpose of the equivalent channel matrix \mathbf{H}^H is decomposed into a unitary matrix and an upper triangular matrix, i.e., $\mathbf{H}^H = \mathbf{Q}\mathbf{R}$. Then, the forward and feedback filters are given by $\mathbf{F} = \mathbf{Q}\mathbf{G}$, $\mathbf{B} = \mathbf{R}^H\mathbf{G}$, respectively, where $\mathbf{G} = \text{diag}\{r_1^{-1}, r_2^{-1}, \dots, r_{N_r MN}^{-1}\}$ and $r_1, r_2, \dots, r_{N_r MN}$ are the diagonal elements of \mathbf{R}^H . In addition, β is introduced to normalize the power of transmitted signals. The received signal is given by

$$\mathbf{y} = \mathbf{H}\mathbf{F}\mathbf{B}^{-1}(\mathbf{s} + \mathbf{a}) + \mathbf{v}. \quad (11)$$

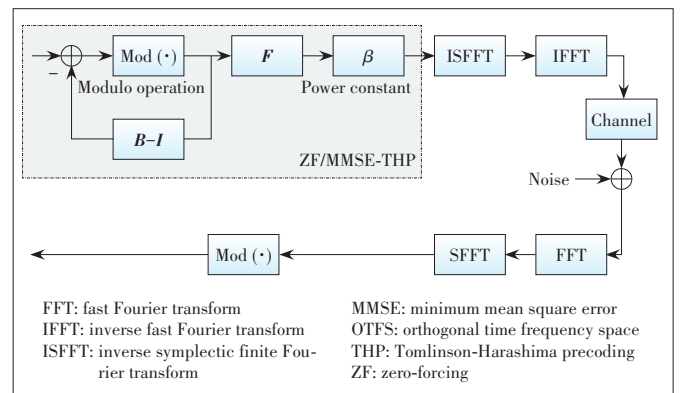
Alternatively, the MMSE criterion can also be employed in the THP approach, which is called MMSE-DD-THP. We define

$$\bar{\mathbf{H}}_e = \mathbf{H}^{-1} \left(\mathbf{H}\mathbf{H}^H + \frac{\sigma_n^2}{\sigma_v^2} \mathbf{I} \right) \quad (12)$$

Similarly, $\bar{\mathbf{H}}_e$ is decomposed into a unitary matrix and an upper triangular matrix, i.e., $\bar{\mathbf{H}}_e = \mathbf{Q}\mathbf{R}$. We can obtain $\mathbf{F} = \mathbf{Q}\mathbf{G}$ and $\mathbf{B} = \mathbf{R}^H\mathbf{G}$. Moreover, there are several approaches to further improvement of the MMSE-DD-THP. In Ref. [16], the authors rearranged the precoding order of symbols by considering the channel conditions. Furthermore, the linear Wiener transmit filter was adopted to obtain the optimizations after ordering the symbols^[17]. In Ref. [18], a block-wise fashion and the Tx-Rx matrices were jointly optimized to minimize the MSE. However, it is noted that the computation complexity of DD-THP schemes is high due to the non-linear processing, which may limit the implementation of the DD-THP schemes in practice.

3.3 Other Precoding Approaches

The linear precoding is a widely used precoding technique in the cellular system. The advantage of the linear precoding over the THP is that the former has low computational complexity. The linear precoding is studied in the MIMO-OTFS system^[14-16]. By exploiting the reciprocity of the uplink chan-



▲ Figure 3. Block diagram of applying THP in OTFS multiple-antenna system

nel and downlink channel, the authors in Ref. [14] proposed an uplink-aided transmission scheme in the MIMO-OTFS system. Since there are few scatterers between the transmitter and the user, and the angular spread of transmitted signals is small in the high-speed scenarios where the received signals occupy only a small part of the channel in the whole angle domain. The channel of the MIMO-OTFS system in the angle domain is sparse. By exploiting the reciprocity of wireless channels, the angle-delay-Doppler channel is estimated through the uplink channel estimation^[15]. Then, based on the estimated angular direction of each user, each beamforming vector is formed to avoid multi-user interference in the downlink. In Ref. [16], an MRC precoder was proposed in the multi-user massive MIMO-OTFS system. The transmitter precodes the symbols by multiplying the Hermitian of the equivalent channels. Although the computational complexity of the precoding is low in Ref. [16], the interference among users is not completely eliminated, which degrades the BER performance.

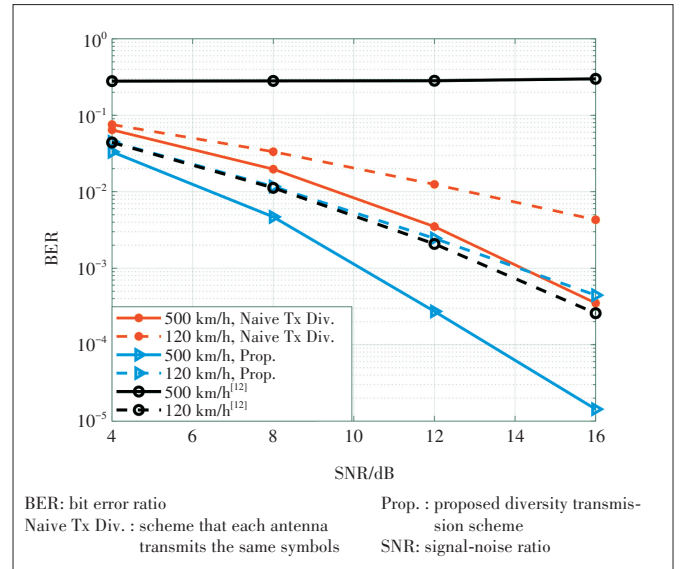
4 Simulation Results

In this section, we evaluate the BER performance of the transmission schemes in the open-loop and the closed-loop MIMO-OTFS systems. The rectangular pulse shaping and the MMSE detector are employed in the simulation. The default setup of the simulation is listed in Table 1. We adopt the tapped-delay-line-A (TDL-A) channel model. The Doppler shift corresponding to the i -th tap is generated by Jakes' formula.

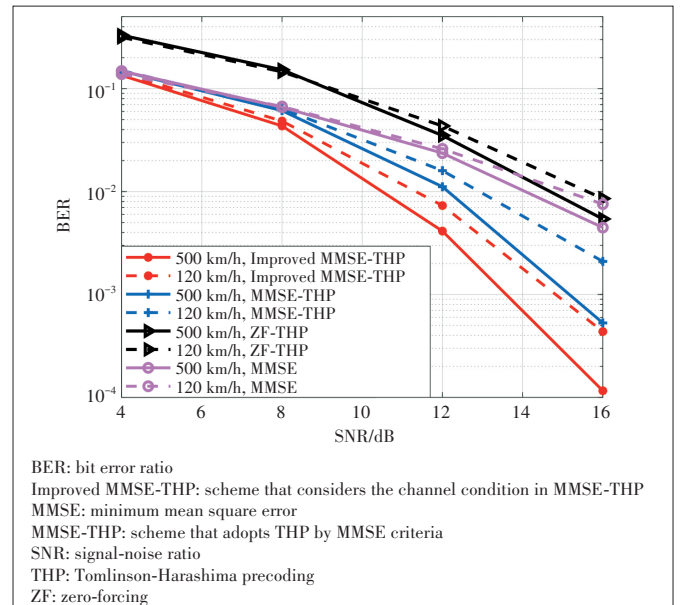
We evaluate the BER performance of our proposed open-loop scheme and the existing works in Fig. 4. We set the number of the receive antennas as one, and the number of the carriers is $M = 32$. In Fig. 4, the proposed diversity transmission scheme outperforms the Naive Tx Div. and the scheme in Ref. [12] at the speed of 500 km/h. The fundamental reason can be summarized into the two aspects: 1) the diversity gain and the coding gain can be obtained by using the proposed code word structure. That is the reason why the diversity scheme outperforms the Naive Tx Div. scheme; 2) The proposed scheme is achieved within one frame duration by precluding the two-frame block fading assumption. However, the scheme in Ref. [12] is implemented over the duration of the two consecutive frames where the channel state varies rapidly in the high-speed scenario, which leads to the poor BER performance. In

comparison, the BER performance of the scheme in Ref. [12] becomes slightly better than the proposed scheme when the velocity reduces to 120 km/h. This is because: 1) the channel varies slowly at the speed of 120 km/h, which can be approximately regarded as the same; 2) the equivalent code word in Ref. [12] can achieve much coding gain since the orthogonality of the code word. In addition, the larger Doppler diversity can be achieved at the speed of 500 km/h over 120 km/h, which results in the better BER performance than that of 120 km/h.

In Fig. 5, we compare the BER performance of the ZF-DD-THP, the MMSE-DD-THP and the MMSE-THP with an ordering matrix. The number of the receive antennas is 2, and the



▲ Figure 4. Evaluated BER performance for diversity approaches in OTFS system



▲ Figure 5. Evaluated BER performance for the delay-Doppler (DD)-THP approaches in OTFS system

▼ Table 1. Setup of simulation parameters

Parameters	Values
Number of OTFS symbols	14
Number of subcarriers	16, 64
Carrier frequency	4 GHz
Subcarrier spacing	15 KHz
Number of transmit antennas	2
Pulse shaping	Rectangular

OTFS: orthogonal time frequency space

number of the carriers is set as $M = 16$. Benchmarks are shown as follows: 1) MMSE: each antenna transmits different symbols and uses MMSE equalization in the OTFS system; 2) ZF-THP: the transmission scheme based on ZF-DD-THP; 3) MMSE-THP: the transmission scheme based on MMSE-DD-THP; 4) Improved MMSE-THP: the transmission scheme based on the MMSE-THP considering the channel condition. We can see that compared with the ZF-DD-THP, the MMSE-DD-THP has a performance gain of about 4 – 5 dB at the same BER level. By considering the ordering matrix to encode the symbols in an optimized precoding order, the improved MMSE-THP can further improve the performance of MMSE-DD-THP.

5 Challenges and Opportunities

OTFS confronts many challenges when combined with the multiple-antenna system in the high-speed scenarios. In this section, we introduce challenges and opportunities in the open-loop and the closed-loop MIMO-OTFS transmission system.

5.1 Feedback in Closed-Loop Transmission

The equivalent channel with the dimension of $N_r MN \times N_t MN$ in the delay-Doppler domain is required to feed back to the transmitter in the DD-THP OTFS system, which causes a high-feedback overhead. Since the fast time-varying of the high-speed scenarios, the feedback of channel information may not match the current channel state, which results in the precoder mismatch to the current channel state. Therefore, how to reduce the feedback overhead and improve the real-time property of the precoder to match the current channel is inevitable in the closed-loop transmission systems. By observing the law of channel changes, one possible solution is to adopt a reliable channel prediction method by the previous channel estimation. Another approach is to design a codebook set to match the delay-Doppler channel, which is similar to the feedback pattern to OFDM closed-loop networks. In this way, the receiver selects a codebook to match the estimated channel and feeds back to the transmitter by a specific indicator. Then, each feedback become less.

5.2 Low-Complexity Precoding and Receiving Approaches

Due to the two-dimensional convolution in the delay-Doppler domain, the size of the precoding and the equalization matrices are large which results in the high computational complexity. Moreover, the linear receiver such as the MMSE receiver also leads to a high complexity due to the inverse operation of a high-dimensional matrix. Therefore, how to reduce the complexity of the precoding and equalization schemes is a big challenge in the OTFS multiple-antenna system. The study of the variational Bayes can be utilized to be the detector in the OTFS multiple-antenna systems.

5.3 Channel Estimation in MIMO-OTFS

The channel estimation with the fractional Doppler and the

fractional delay is a more practical problem in the future research. Furthermore, in the aspect of the Doppler spectrum, most of the existing works focus on the channel estimation of the discrete-Doppler-spread (also known as limit-Doppler-spread) scenarios^[20]. However, when the propagation environment involves many scattering objects, the Doppler shift of transmit paths are infinite and the Doppler spectrum is continuous, which is treated as the continuous-Doppler-spread channel. How to reduce the overhead of the channel estimation in the meantime improve the accuracy in these practical channel models is a big challenge. In Ref. [21], a low-dimensional subspace is constructed to characterize the variation of the equivalent channel responses in the continuous-Doppler-spread channel, which is modeled by a sum of the projection coefficients and the basis functions.

6 Conclusions

In this paper, we introduce both the open-loop and the closed-loop multi-antenna approaches for the OTFS with the ideal and the rectangular pulse shaping. In the open-loop design, the main contribution is that the transmit diversity approaches resembling space-time coding are provided, which take the practical issues into account, i. e., the rectangular pulse shaping and the rapidly time-varying channel. For the closed-loop design, we suggest to adopt the Tomlinson-Harashima precoding in the delay-Doppler domain since we have developed the relation between the precoding matrix in the time-frequency domain and that in the delay-Doppler domain. The reason why we recommend precoding in the delay-Doppler domain is that the channel in the delay-Doppler domain varies more slowly within a frame in contrast to that in the time-frequency domain. In the end, we discuss challenges and opportunities of the OTFS multiple-antenna system, which can be further investigated in future.

References

- [1] LI S Y, YUAN W J, WEI Z Q, et al. A tutorial to orthogonal time frequency space modulation for future wireless communications [C]//2021 IEEE/CIC International Conference on Communications in China (ICCC Workshops). Xiamen, China, 2021: 439 – 443. DOI: 10.1109/ICCCWorkshops52231.2021.9538891
- [2] WEI Z Q, YUAN W J, LI S Y, et al. Orthogonal time-frequency space modulation: a promising next-generation waveform [J]. IEEE wireless communications, 2021, 4(28): 136 – 144. DOI: 10.1109/MWC.001.2000408
- [3] RAVITEJA P, PHAN K T, HONG Y, et al. Interference cancellation and iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, 2018, 17(10): 6501 – 6515. DOI: 10.1109/TWC.2018.2860011
- [4] LI S Y, YUAN W J, WEI Z Q, et al. Cross domain iterative detection for orthogonal time frequency space modulation [J]. IEEE transactions on wireless communications, 2021. DOI: 10.1109/TWC.2021.3110125

- [5] QU H Y, LIU G H, ZHANG L, et al. Low-complexity symbol detection and interference cancellation for OTFS system [J]. IEEE transactions on communications, 2021, 69(3): 1524 – 1537. DOI: 10.1109/TCOMM.2020.3043007
- [6] SURABHI G D, CHOCKALINGAM A. Low-complexity linear equalization for 2×2 MIMO-OTFS signals [C]//IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications. Atlanta, USA: IEEE, 2020: 1 – 5. DOI: 10.1109/SPAWC48557.2020.9154292
- [7] SINGH P, MISHRA H B, BUDHIRAJA R. Low-complexity linear MIMO-OTFS receivers [C]//IEEE International Conference on Communications Workshops. Montreal, Canada: IEEE, 2021: 1 – 6. DOI: 10.1109/ICCWorkshops50388.2021.9473839
- [8] VUCETIC B, YUAN J H. Space-time coding [M]. Chichester, UK: John Wiley & Sons, 2003. DOI: 10.1002/047001413x
- [9] ALAMOUTI S M. A simple transmit diversity technique for wireless communications [J]. IEEE journal on selected areas in communications, 1998, 16(8): 1451 – 1458. DOI: 10.1109/49.730453
- [10] TAROKH V, SESHADRI N, CALDERBANK A R. Space-time codes for high data rate wireless communication: Performance criterion and code construction [J]. IEEE transactions on information theory, 1998, 44(2): 744 – 765. DOI: 10.1109/18.661517
- [11] BOMFIN R, CHAFII M, NIMR A, et al. Channel estimation for MIMO space time coded OTFS under doubly selective channels [C]//2021 IEEE International Conference on Communications Workshops. Montreal, Canada: IEEE, 2021: 1 – 6. DOI: 10.1109/ICCWorkshops50388.2021.9473618
- [12] AUGUSTINE R M, SURABHI G D, CHOCKALINGAM A. Space-time coded OTFS modulation in high-Doppler channels [C]//IEEE 89th Vehicular Technology Conference. Kuala Lumpur, Malaysia: IEEE, 2019: 1 – 6. DOI: 10.1109/VTCSpring.2019.8746394
- [13] DELFELD J, RAKIB S S. Tomlinson-harashima precoding in an OTFS communication system: US11018731 [P]. 2021
- [14] LIU Y, ZHANG S, GAO F F, et al. Uplink-aided high mobility downlink channel estimation over massive MIMO-OTFS system [EB/OL]. (2020-03-16) [2021-09-18]. <https://arxiv.org/abs/2003.07045>
- [15] LI M Y, ZHANG S, GAO F F, et al. A new path division multiple access for the massive MIMO-OTFS networks [J]. IEEE journal on selected areas in communications, 2021, 39(4): 903 – 918. DOI: 10.1109/JSAC.2020.3018826
- [16] PANDEY B C, MOHAMMED S K, RAVITEJA P, et al. Low complexity precoding and detection in multi-user massive MIMO OTFS downlink [J]. IEEE transactions on vehicular technology, 2021, 70(5): 4389 – 4405. DOI: 10.1109/TVT.2021.3061694
- [17] HABENDORF R, IRMER R, RAVE W, et al. Nonlinear multiuser precoding for non-connected decision regions [C]//IEEE Workshop on Signal Processing Advances in Wireless Communications. New York, USA: IEEE, 2005., 535 – 539. DOI: 10.1109/SPAWC.2005.1506197
- [18] KUSUME K, JOHAM M, UTSCHICK W, et al. Efficient Tomlinson-Harashima precoding for spatial multiplexing on flat MIMO channel [C]//IEEE International Conference on Communication. Seoul, Korea (South): IEEE, 2005: 2021 – 2025. DOI: 10.1109/ICC.2005.1494693
- [19] JAFARKHANI H. A quasi-orthogonal space-time block code [J]. IEEE transactions on communications, 2001, 1(49): 1 – 4. DOI: 10.1109/26.898239
- [20] QU H Y, LIU G H, ZHANG L, et al. Low-dimensional subspace estimation of continuous-Doppler-spread channel in OTFS systems [J]. IEEE transactions on communications, 2021, 69(7): 4717 – 4731. DOI: 10.1109/TCOMM.2021.3072744
- [21] RAVITEJA P, PHAN K T, HONG Y. Embedded pilot-aided channel estimation for OTFS in delay – Doppler channels [J]. IEEE transactions on vehicular technology, 2019, 68(5): 4906 – 4917. DOI: 10.1109/TVT.2019.2906357

Biographies

WANG Dong received the B.Eng. degree from the School of Electronic and Information Engineering, Hebei University, China in 2016. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, China. His current research interests include multiway relaying communications and MIMO communications.

WANG Fanggang (wangfg@bjtu.edu.cn) received the B.Eng. and Ph.D. degrees from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China in 2005 and 2010, respectively. He was a Post-Doctoral Fellow with the Institute of Network Coding, The Chinese University of Hong Kong, China from 2010 to 2012. He was a visiting scholar with the Massachusetts Institute of Technology, USA from 2015 to 2016 and the Singapore University of Technology and Design, Singapore in 2014. He is currently a professor with the State Key Laboratory of Rail Traffic Control and Safety, School of Electronic and Information Engineering, Beijing Jiaotong University, China. His research interests are in wireless communications, signal processing, and information theory. He served as an editor for the *IEEE Communications Letters* and a technical program committee member for several conferences.

LI Xiran received the B.Eng degree from the school of Information and Communication Engineering, Beijing Jiaotong University, China in 2021. She is currently pursuing the M.A. degree with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. Her current research interests include MIMO communications and orthogonal time frequency space.

YUAN Pu received the B.Eng. degree from Tianjin University, China in 2007 and the M.S. and Ph.D. degrees from Nanyang Technological University, Singapore in 2010 and 2015, respectively. He is currently with vivo Mobile Communication Co., Ltd., China. From 2016 to 2019, he was a research engineer with the 2012 Laboratories, Huawei Technologies Company, Ltd., China. His research interests include signal processing in communication and information theory.

JIANG Dajie received the B.S. degree in communication engineering and M.S. degree in digital signal processing from Beijing University of Posts and Telecommunications, China in 2005 and 2008, respectively. From 2008 to 2017, he was a research engineer for 4G and 5G wireless research & standardization with China Mobile Research Institute. He is currently with vivo Mobile Communication Co., Ltd., China. His research interests include potential technologies for 6G including RIS and joint communication and sensing.



Study on Security of 5G and Satellite Converged Communication Network

YAN Xincheng^{1,2}, TENG Huiyun²,
PING Li², JIANG Zhihong²,
ZHOU Na^{1,2}

(1. State Key Laboratory of Mobile Network and
Mobile Multimedia Technology, Shenzhen
518055, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202104009

[http://kns.cnki.net/kcms/detail/34.1294.
TN.20211022.1252.002.html](http://kns.cnki.net/kcms/detail/34.1294.TN.20211022.1252.002.html), published online
October 22, 2021

Manuscript received: 2021-08-09

Abstract: The 5G and satellite converged communication network (5G SCCN) is an important component of the integration of satellite-terrestrial networks, the national science, and technology major projects towards 2030. Security is the key to ensuring its operation, but at present, the research in this area has just started in our country. Based on the network characteristics and security risks, we propose the security architecture of the 5G SCCN and systematically sort out the key protection technologies and improvement directions. In particular, unique thinking on the security of lightweight data communication and design reference for the 5G SCCN network architecture is presented. It is expected to provide a piece of reference for the follow-up 5G SCCN security technology research, standard evolution, and industrialization.

Keywords: 5G SCCN; non-terrestrial networks; 5G security; satellite security; integration of satellite-terrestrial networks

Citation (IEEE Format): X. C. Yan, H. Y. Teng, L. Ping, et al., "Study on security of 5G and satellite converged communication network," *ZTE Communications*, vol. 19, no. 4, pp. 79 – 89, Dec. 2021. doi: 10.12142/ZTECOM.202104009.

1 Introduction

The development of mobile communication technology has greatly improved the informatization level of all industries in the whole society. However, due to factors such as space and quantity, 5G communication networks are currently deployed in limited areas. Satellites are an ideal choice for wide coverage communications^[1], especially for areas where ground transmission towers cannot be deployed (oceans, mountains, islands, etc.) and for scenarios of disaster relief and emergency response. Building 5G and satellite converged communication network (5G SCCN) has become an important direction for future network development, deeply combining the excellent access capabilities and mobility of 5G networks with the extensive coverage capabilities of satellite networks, and giving play to the respective advantages of the networks to achieve global wide-area full coverage and seamless high-speed interconnection. However, the satellite network has the characteristics of environmental openness, time-varying topology, and limited computing resources, bringing 5G and satellite networks more complex security challenges; meanwhile, 5G SCCN will carry more critical and urgent communication services for industries, individuals, and public affairs, which makes it particularly important to ensure

the security of 5G SCCN.

A wave of satellite Internet constellation construction is underway around the world. At present, at least 15 companies around the world have announced low-orbit communication satellite plans, and many have carried out research and practice related to 5G and satellite converged networks^[2]. In 2017, the European Union funded the Satellite and Terrestrial Network for 5G (SaT5G) alliance to promote solutions that integrate satellite communications with 5G, software-defined networking/network functions virtualization SDN/NFV, and other technologies^[3-4]. In 2018, the European Space Agency (ESA) launched the ALIX project to promote the standardization of 5G satellite components and its interfaces with other networks^[5]. In 2019, Telesat verified that low-orbit satellites provided effective solutions to 5G base station relays^[6]. In April 2021, the China Satellite Network Group was established. It plans to provide satellite communication services including 5G satellite converged networks to ground and air terminals. In July 2021, Beijing University of Posts and Telecommunications completed a low-orbit satellite and 5G private network integration test between two cities.

The academia has researched on early development of satellite communications^[7-9]. In recent years, institutions and univer-

sities have gradually carried out technical research on satellite-terrestrial converged communications^[10-15] and its security, such as dual access through satellite and ground base stations, 5G New Radio (NR) and satellite network convergence, and satellite networks and 5G core network heterogeneous convergence. In the security aspect, the National Digital Switching System Engineering & Technological R&D Center has researched on satellite communication security^[16-17]. The Ph.D. thesis “Research on Security Protocol of Broadband Satellite Network” improves security protocols such as IP Security (IPSec) and Internet key exchange (IKE) for satellite communications^[18].

Standard organizations such as International Telecommunication Union (ITU) and the 3rd Generation Partnership Project (3GPP) proposed that satellite networks can be used as extensions of terrestrial networks^[19-24], and research in this area has been carried out. Among them, ITU-R M.[NGAT_SAT]^[19] defines and discusses the key technical issues, service characteristics, network structure, and deployment scenarios regarding satellite networks integration into 5G networks. 3GPP’s research on 5G and satellite converged networks is mainly carried out in two projects, TR 38.811^[20] and TR 22.822^[21]. Among them, “Study on Using Satellite Access in 5G” (TR22.822) analyzes the functional requirement of 5G satellites and introduces 12 functional requirements and their corresponding usage scenarios. While “Study on NR to Support Non-Terrestrial Networks” (TR 38.811) proposes three functions of satellite communications for 5G networks. It serves as a supplementary coverage for terrestrial 5G networks, provides continuous communications for high-speed mobile carriers, and uses new services such as satellite multicast and broadcast. The project also introduces service characteristics, network structures, deployment scenarios, and non-ground-based network channel models of 5G and satellite converged networks and proposes a variety of non-terrestrial network architecture options.

However, although 3GPP defines the network form of 5G SCCN, its security issues have not been considered. At present, there is also a lack of technical requirements in this area, and standards are also absent. This paper intends to analyze and discuss the security requirements and key security technologies of 5G SCCN. On the one hand, it is a reference for future research on 5G and satellite converged network security technology, standard promotion, and industrialization; On the other hand, based on the concept of “security-synchronized design”, it is hoped that the security design can provide a reference for 5G SCCN design.

2 Security Challenges and Requirements

2.1 Security Challenges

5G SCCN has different network characteristics from terrestrial 5G networks. These characteristics are mainly derived

from satellite networks. Meanwhile, the cross-network and cross-domain integration of 5G and satellite networks, and the introduction of 5G diversified services will jointly constitute new features of the converged network, making 5G SCCN face new security challenges.

1) Borderless security issues are caused by the open network environment.

Different from the terrestrial network, the satellite network nodes are exposed and channels are open, and the satellite node runs in the exposed space orbit for a long time. Thus, new threats emerge. For example, the inter-satellite and satellite-to-ground wireless communication links are more susceptible to the adverse natural environment and malicious users; network nodes are more susceptible to forgery and hijacking; communication links are more susceptible to human interference, eavesdropping, replay attack, and wireless resource occupation. Therefore, higher risks of confidentiality, integrity, availability and reliability of the network are posed.

2) Dynamic changes in network topology lead to changes in security policies.

The 5G SCCN includes satellite nodes and ground nodes. Satellite nodes are always in high-speed operation and may frequently join or exit the network. This makes the network topology change dynamically, and the communication objects change as a consequence, which leads to network security function switching and security strategy migration, such as the update and synchronization of the original authentication policy, or the renegotiation of the original IPSec/transport layer security (TLS) tunnel.

3) Heterogeneous interconnection causes security applicability issues.

5G and satellite converged communications are based on different forms of physical resources and present a “chimney-like” development model. Different satellite systems are relatively independent and dedicated, lacking a unified network protocol specification. This may also make the mature security protocols applicable to terrestrial networks while inapplicable to satellite networks. In addition, the heterogeneous interconnection and long-distance communication of the 5G SCCN make it more difficult to protect data in transmission, and the risk of user data being stolen, tampered with, and damaged increases.

4) Insufficient security computing power is caused by low on-board processing capability.

Satellites usually use aerospace-grade chips to cope with the complexity and harshness of the space environment. In order to improve the reliability of the chip, it is necessary to reduce the density of computing units on the chip, and strictly control the amount of computing of the software carried by the satellite, which makes the computing power of the satellite far lower than that of the ground communication node. Therefore, satellite nodes are more susceptible to availability attacks from asymmetric computing power, such as distributed denial

of service (DDoS) attacks. Meanwhile, some traditional computationally intensive encryption algorithms cannot run on satellites. All these make the security of satellite nodes face greater challenges, and therefore, it is necessary to research on a new type of 5G security architecture and security technology suitable for satellite networks.

2.2 Security Requirements

Based on the above analysis of the 5G SCCN characteristics and security challenges, four security requirements can be summarized: identity authentication, lightweight communication security, enhanced availability protection, and fine-grained resource sharing and isolation.

1) Universal identity authenticity needs

Terrestrial communication networks usually adopt physical isolation or physical dedicated lines. Network nodes are usually in the same physical or logical trust domain, and there is a default trust relationship among network nodes. However, the 5G SCCN conducts ultra-long distance communication in an open space environment, and the satellite network has a time-varying topology, which makes the communication objects highly dynamic. Therefore, ensuring the authenticity of communication nodes, especially the authenticity of network equipment, is a key requirement for 5G SCCN. Through the access authentication of the terminal and the authentication between the network nodes, a communication system with an open external environment and trustworthy internal communication can be established.

2) Lightweight communication security requirements

In the 5G SCCN, the service link, inter-satellite link, and feeder link all use wireless links for communication, making it more vulnerable to eavesdropping, tampering, and replay attacks. Therefore, the confidentiality and integrity protection of the transmitted data is especially necessary. On the other

hand, due to the limited processing resources on the satellite, it is necessary to avoid running computationally intensive encryption algorithms on the satellite as much as possible. The 5G SCCN consequently needs to design and adopt lightweight communication security architecture and technology to ensure the security of communication data while avoiding excessive computational burden on the satellite network.

3) Enhanced availability protection requirements

Availability attacks on existing networks will still exist in 5G SCCN, such as DDoS attacks and signaling storms. Considering the openness of the satellite network environment, lower processing capacity, and high value of the services, the security risks are severer. Therefore, more systematic and efficient technical measures need to be adopted to ensure the availability of functions and services on satellite nodes.

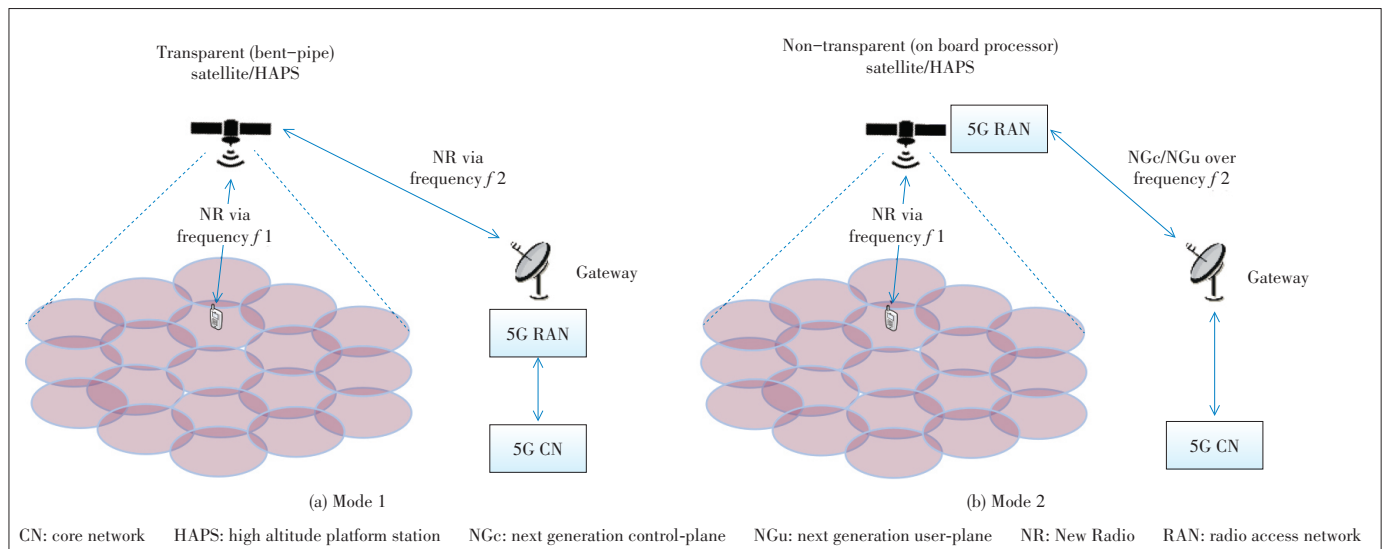
4) Fine-grained resource sharing and isolation requirements

The 5G SCCN will provide differentiated network services for public users, industry users and special users on shared network infrastructure. Therefore, it is necessary to isolate shared resources securely and effectively to prevent side-channel attacks and threats from spreading. Limited satellite network resources put forward higher requirements on the granularity of resource sharing, and more refined network resource management technologies are required.

3 Network Security Architecture

3.1 Service Architecture

3GPP TR 38.811 defines two typical 5G and satellite converged network modes (excluding relay nodes), as shown in Fig. 1. In mode 1, 5G RAN is still deployed on the ground, and the satellite network is used as a transparent forwarding channel for the 5G access network; In mode 2, 5G radio access network (RAN) is deployed on the satellite and connected



▲ Figure 1. Typical access network mode of 5G and satellite integration (Source: 3GPP TR 38.811)

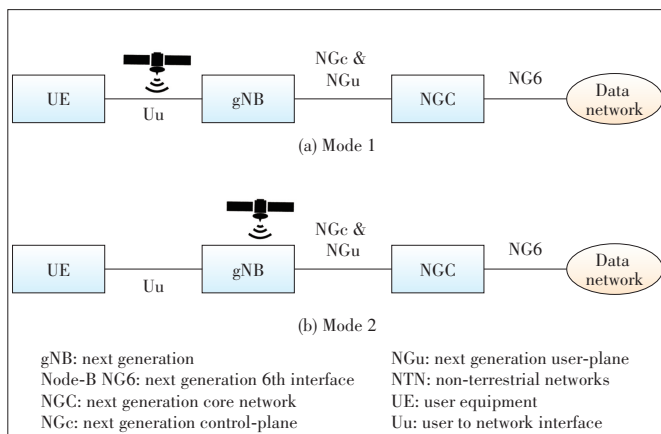
to the ground core network through non-terrestrial networks (NTN) gateway.

Fig. 2(a) shows NTN featuring access network serving user equipment (UE), based on a satellite/aerial with bent pipe payload and gNB on the ground (satellite hub or gateway level). Fig. 2(b) shows NTN featuring an access network serving UE, based on a satellite/aerial with gNB on board.

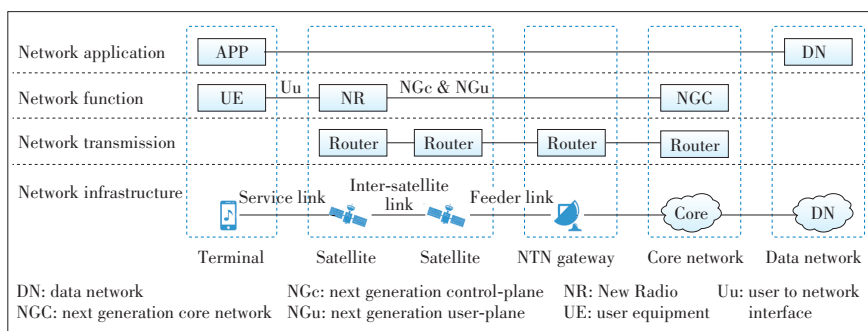
In contrast, Mode 2 is easier to inherit the existing 5G access technologies, including air interface scheduling technology, mobile handover technology, terminal secure access technology, etc., and it is also easier to achieve the goal of mobile terminal access everywhere with one device, which has better industrialization foundation and better serviceability. Therefore, the follow-up technical research herein mainly focuses on the second service model, which is about the gNB on board the satellite network.

Fig. 3 shows the service architecture in Mode 2 of 3GPP TR 38.811. A mobile phone terminal accesses Internet services through a 5G SCCN. From left to right, the terminal UE located on the ground or in the air communicates with the base station NR on the low-orbit satellite. The inter-satellite link is routed to the ground satellite gateway station, then reaches the core network, and finally accesses Internet services.

In the vertical dimension, the entire service model can be abstracted into four levels, from bottom to top including the network infrastructure layer, the network transmission layer, the network function layer, and the network application layer.



▲ Figure 2. Network logical view of two modes



▲ Figure 3. Typical network logic view of the integration of base stations and satellites

the network function layer, and the network application layer. The network transmission layer and the network function layer realize respectively the forwarding of IP packets and the communication of the mobile network. At the network transmission layer, assuming that satellites and satellite gateways have basic routing functions and follow the basic IP routing protocol to realize the transmission and forwarding of messages on inter-satellite links and the satellite-to-ground links, and at the network function layer, assuming that the base station is on the satellite, a 5G communication network is therefore formed consisting of network functions such as the terminal, the on-board base station and the terrestrial core network, which has the basic features and capabilities of a 5G network.

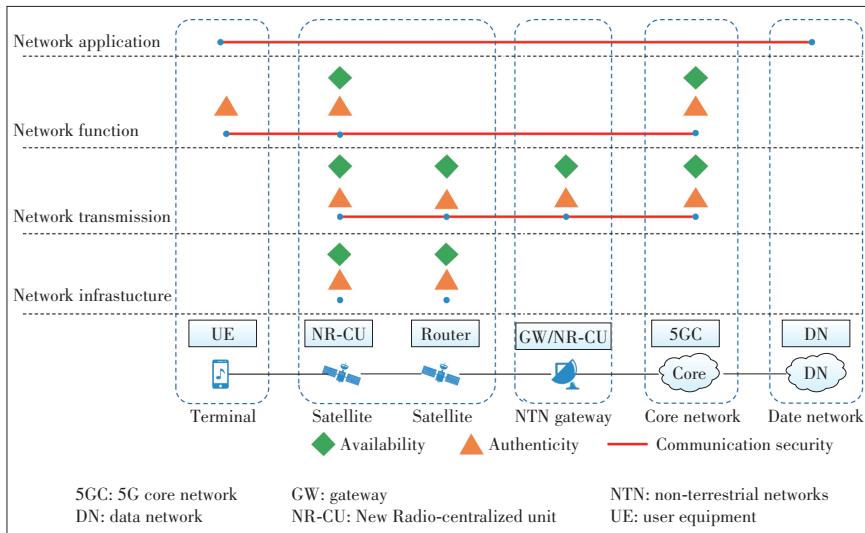
Through the service architecture, we can decompose the capabilities of each communication facility, and thus map more clearly the requirements and capabilities of the traditional 5G network and bearer network to the NTN network. For example, the NGc&NGu port in the 5G network is composed of inter-satellite links, feeder links, and the ground bearer network between satellite gateways and the core network in the NTN network. As another example, because the satellite has the capability of a base station at the network functional layer, the UE's access guarantee can be enhanced with the help of existing 5G access technologies to a large extent.

3.2 Security Architecture

Due to the openness, mobility, and low power consumption of satellite links, the offensive and defensive situations of 5G SCCN and terrestrial 5G networks are quite different. But many similar technologies can be used and referenced. As shown in Fig. 4, based on the service architecture, the security of the 5G SCCN is analyzed layer by layer, focusing on the three types of security attributes of availability, authenticity, and communication security (confidentiality, integrity, and communication isolation) for key communication nodes and interfaces.

In the 5G SCCN, in addition to the general security attributes and technologies of the 5G network, it is important to consider the security issues caused by the characteristics of the satellite network and its integration with the 5G network. Given the different service characteristics and security attributes of each layer, the required security technologies are also different.

At the infrastructure layer, considering the difficulty of upgrading and maintaining the equipment on the satellite, it is necessary to establish an active immune mechanism for the satellite node through the trusted boot to resist attacks from unknown threats to ensure the authenticity of a single node. The infrastructure protection method for other nodes is similar to that of the 5G network. In the radio frequency part, con-



▲ Figure 4. 5G satellite converged communication network (SCCN) security architecture

sidering the power asymmetry between the satellite and the ground attacker, the availability of satellite nodes needs to be paid more attention.

At the network transmission layer, in order to ensure the confidentiality and integrity of the information transmission of the inter-satellite link and the back-haul network, the bearer network communication security technology is needed to be in place. Considering the limited resources of satellite networks, lightweight attack detection methods can be used to defend against DDoS attacks on satellite networks. In addition, with the slice isolation technology of the bearer network, the satellite network-related traffic is isolated from other 5G network traffic to avoid mutual influence.

At the network function layer, to reduce the burden on the on-board base station, on-demand air interface signaling encryption and decryption and user plane confidentiality and integrity protection functions are considered. To ensure that the satellite access network and related core network resources are not affected by other 5G services, end-to-end slice isolation from the RAN to the core network should be adopted. Regarding the malicious directed call attack that may exist in the case of the asymmetry of space and ground resources, the anti-UE random access technology should be adopted to ensure the regular access of legitimate users. In addition, in order to ensure the legitimate access of mass terminals and dynamic nodes, it is necessary to perform two-way authentication on the terminals and the network nodes.

At the network application layer, in order to protect the confidentiality and integrity of application layer data, end-to-end encryption technology is used between the terminal and the application service. To reduce the processing burden of the satellite, the satellite nodes can implement transparent transmission.

The specific security technologies are referred to in Table 1.

4 Key Protection Technology

Combining the analysis result in Table 1, this section analyzes the key technologies involved in aspects of identity authentication, data communication security, network availability, and network resource sharing. Due to space limitations, we only put forward directional suggestions for each technical requirement, and elaborate the key technical features needed to meet the requirements, the current technologies that can be inherited or learned from, and the improvement suggestions when applying this technology to 5G SCCN. The technical details will not be discussed.

4.1 Identity Authentication

Due to the openness and the time-varying topology of the 5G SCCN, it is necessary to verify the authenticity of the communication node. In addition to the terminal access authorization in the traditional 5G network, the 5G SCCN also needs to authenticate the network nodes. Communications among network nodes should take identity authentication as a prerequisite, and ensure the security and independence of the 5G SCCN system by enabling the identity authentication process. Meanwhile, technologies such as trusted boot and trusted environment can also be used to further ensure the authenticity of hardware devices and their running software.

4.1.1 Terminal Access Authentication

In the 5G SCCN, due to the openness of the service link and the diversity of access terminals, the trusted communication of the service link has received much attention. 3GPP defines a complete user access system for 5G networks. Based on the 5G unified authentication architecture (5G-AKA or EAP-AKA'), the terminal and the service network are mutually authenticated to ensure mutual trust between users and the network. The 5G SCCN can follow this set of access authentication frameworks to ensure the trusted access of wireless terminals and solve the problem of pseudo base stations, pseudo terminals, and pseudo networks. Meanwhile, 5G SCCN also needs to enhance 5G access authentication for new network features. For example, for the weak processing capabilities of satellite nodes and new multicast services, group authentication and lightweight authentication methods^[25] need to be considered; The topology of the satellite network is time-varying, and it is necessary to enhance the switching capability of the Xn ports between the base stations.

4.1.2 Authentication of Network Nodes

Authentication of network nodes is a key feature that distinguishes 5G SCCN from terrestrial 5G networks. In the commu-

▼Table 1. Catalogue of security protection technology

Network Layer	Security Attributes	Categories of Security Technology	Sub-Categories of Security Technology
Network application layer	Communication security	Data communication security	User data communication security
Network function layer	Authenticity	Identity authenticity	Terminal access authentication
			Authentication between network nodes
	Availability	Network availability	Anti-DDoS attack
			Anti-UE random access attack
	Communication security	Network resource sharing	RAN slice isolation
			Core network slice isolation
Network transmission layer	Communication security	Data communication security	Network function communication security
		Network resource sharing	Bearer network slice isolation
	Availability	Data communication security	Bearer network communication security
		Network availability	Anti-DDoS attack
Infrastructure layer	Authenticity	Identity authenticity	Authentication between network nodes
	Availability	Network availability	Anti-wireless communication interference
	Authenticity	Identity authenticity	Trusted boot of satellite nodes

DDoS: distributed denial of service UE: user equipment RAN: radio access network

nication of each layer of the 5G SCCN, whether the communication is between network functions (such as between NR and 5G core network elements) or between transmission nodes (such as on-board routing and forwarding), the authentication of the network node is required as a precondition to prevent attackers from impersonating legitimate network functions to access the 5G network, or impersonating legitimate transmission nodes to establish routing adjacencies with legitimate satellites, thereby stealing or tampering with user data and routing information in the network.

There are two typical authentication methods. One is to use SDN-like technology. The communication forwarding node uniformly authenticates the management node, and the token for network communication is obtained after the authentication is passed. Since there may be blind spots in communication, this method has higher requirements on the topology of the management network. The other way is to carry out mutual authentication between communication nodes. This way has a relatively high technical maturity. For example, two-way authentication between network function nodes can be performed based on the IKE protocol, and the authentication function in the dynamic routing protocol can be enabled to implement identity authentication between routing and transmission nodes. However, in this way, the overhead brought to the satellite node and the system complexity introduced by the dynamic switching of communication objects need to be considered.

4.1.3 Trusted Boot of Satellite Nodes

Satellite nodes run in the space orbit for a long time, making upgrades and maintenance difficult. Software and hardware vulnerabilities are difficult to update in time, and the nodes are more vulnerable to attacks from unknown threats. Therefore, the satellite nodes need stronger self-immunity. Satellite nodes need to be reinforced under the principle of the

least privilege, such as shutting down unnecessary processes and ports. Meanwhile, with trusted computing technology, there forms a trusted chain of level-by-level verification to ensure the operation of satellite nodes through digital signature and integrity verification technology. Based on trusted execution environment (TEE) storage device identity fingerprints, combined with technologies such as authentication and remote certification, the authenticity of the 5G SCCN communication system can be further assured.

4.2 Data Communication Security

Due to the openness of the 5G SCCN, the confidentiality and integrity protection of the transmitted data has become particularly important. However, the cryptographic computing used for data confidentiality and integrity protection requires a large amount of computing power, which contradicts the low processing capabilities of satellites. Therefore, the confidentiality and integrity protection of the 5G SCCN needs to be considered systematically, especially to avoid enhancing the overhead of satellite nodes greatly. This section discusses the communication security protection technology of critical data such as 5G signaling, IP routing, and user data in the 5G SCCN, and proposes a framework solution that can effectively avoid the impact on satellite computing resources caused by confidentiality and integrity requirements.

4.2.1 Bearer Network Communication Security

On the ground network, if a section of the bearer network is in an insecure or untrustworthy environment, it is usually recommended to protect its confidentiality and integrity. On the satellite network, although the inter-satellite and satellite-to-ground links use wireless communications, considering the limited processing capabilities of the satellite nodes, we do not recommend this kind of protection. Confidentiality and integrity protection in 5G SCCN communication requires refined

design. For the data transmitted in the network, including 5G signaling and user data, it is recommended that the network function layer and application layer be resolved. This part will be discussed in subsequent sections. The network transmission layer should focus on ensuring the confidentiality and integrity of the routing information exchanged in the network. If each network node in the 5G SCCN has undergone strict authentication, whether it is necessary to protect the confidentiality and integrity of routing information requires further research. If considering the bit error rate of wireless transmission, cyclic redundancy check (CRC) may be more suitable for satellite communications than MD5.

4.2.2 Network Function Communication Security

This section focuses on the security of signaling communications between 5G network functions, and the security of user data communication carried by the 5G network is discussed in the next section. In order to ensure the security of 5G air interface communication and UE access signaling, 3GPP has standardized the confidentiality and integrity protection of radio resource control (RRC) signaling between UE and NR, and non-access stratum (NAS) signaling between UE and 5GC. In order to enhance the communication security of the 5G SCCN, it is recommended to enable transmission protection for RRC and NAS signaling. However, the opening of the confidentiality and integrity of RRC signaling means that the load on the on-board base station will increase significantly.

In view of the limited computing resources of satellites, confidentiality and integrity computing on the satellite should be avoided. As shown in Fig. 5, we recommend the Control and User Plane Separation technology, to adopt the deployment method of distributed unit-centralized unit (DU-CU) separation. Among them, the DU is deployed on the satellite, and the CU is deployed on the ground. The physical layer, media access control (MAC) layer, and radio link control (RLC) layer with high real-time requirements are placed in the DU for processing, while the packet data convergence protocol (PDCP) and RRC layers with relatively low real-time requirements are placed in the CU for processing. Since the confidentiality and integrity of the RRC signaling are completed at the PDCP layer by the DU located on the ground, it is possible to effectively avoid heavy-duty cryptographic computing on the satellite nodes.

4.2.3 User Data Communication Security

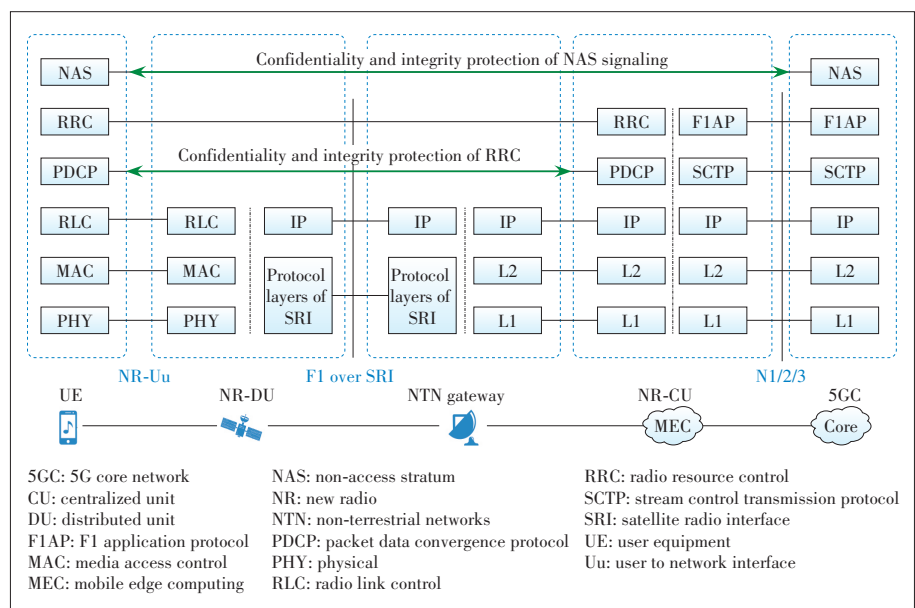
In Fig. 6, there are requirements for communication security assurance at the network transmission layer, network

function layer, and network application layer. In traditional network protection, the idea of in-depth multi-level protection is usually adopted, that is, each system and each protocol layer is protected independently. When the internal system is protected, the assumption that the external system has been protected cannot be made. This may result in the network transmission layer, network function layer, and network application layer protecting the confidentiality and integrity of user data on their own. However, in satellite communications, this idea cannot be fully applied, instead, the principles of minimalism and optimality should be adopted.

To solve the contradiction between the security protection of data communication and the weak processing capability of satellite nodes, we can use the idea of “transmitting on-satellite, processing off-satellite” and place the encryption, decryption and integrity check of high computing costs on the ground node for processing. The satellite is mainly responsible for the forwarding of encrypted user plane data to achieve a balance between performance and security. The higher layer the encryption is applied to, the closer to the end-to-end encryption and the higher level of the security is achieved. Based on this idea, we propose two security solutions to data communication, which are discussed below.

1) Solution 1: UE-DN's end-to-end security

Users and providers of 5G SCCN usually belong to different trust subjects, especially for some high security level services. Network users do not fully trust the protection mechanism of the network provider, and tend to provide end-to-end data encryption by themselves. Performing user data confidentiality and integrity protection between the UE and the Internet can effectively prevent user data from being eavesdropped and tampered with. At the same time, security functions such as



▲ Figure 5. 5G control plane encryption and integrity protection

encryption, decryption, and integrity verification are performed on the terminal and the ground network. The satellite node does not participate in cryptographic computing but only performs transparent forwarding, which greatly saves the computing power of the satellite node.

However, end-to-end encryption also introduces additional problems, such as illegal interception, and the inability of the IP Multimedia Subsystem (IMS) network to recognize voice over Long-Term Evolution (VoLTE) or guarantee the voice quality. The trusted third-party key management server (KMS) or the deployment of encryption and decryption agents in the core network can help solve the problem of legal interception of encrypted communications and the problem of VoLTE voice recognition after multiple encryptions.

2) Solution 2: UE- NR CU security

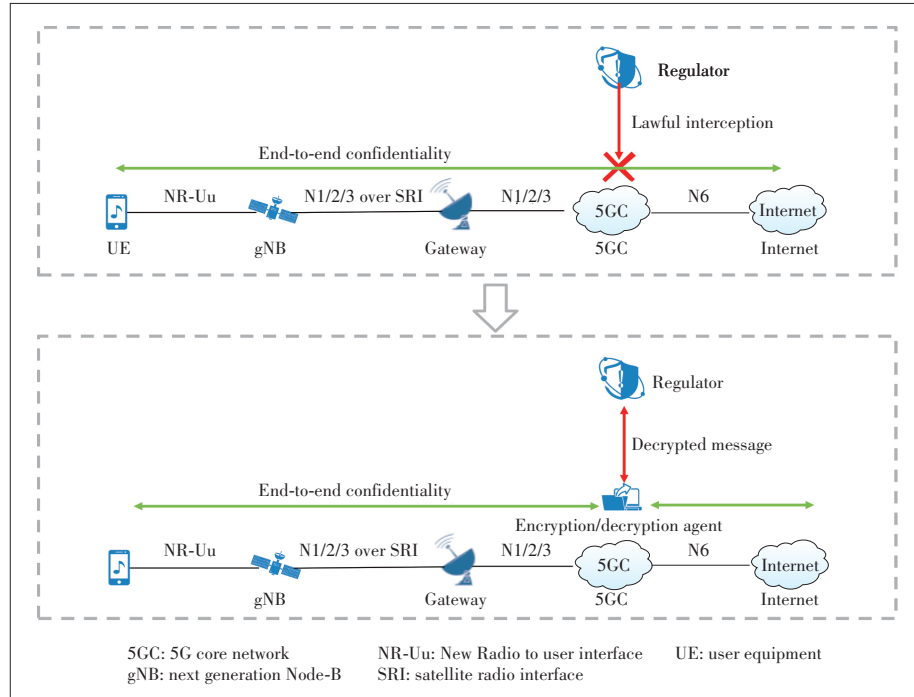
The separate deployment of CU-DU can also effectively solve the contradiction between the protection of user data transmission and the weak processing capability of satellite nodes. As shown in Fig. 7, using the segmented encryption transmission scheme, the air interface enables PDCP-based confidentiality and integrity protection, and the N1/2/3 ports of the backhaul network and the N6 port of the data network can enable IPSec or DTLS protection as needed. The advantage of this solution is that the confidentiality and integrity of user data are processed on the UE and CU on the ground and the DU on the satellite does not participate in the process, so the transmission security of user data can be protected without increasing the cryptographic computing overhead of satellite nodes.

3) Solution comparison

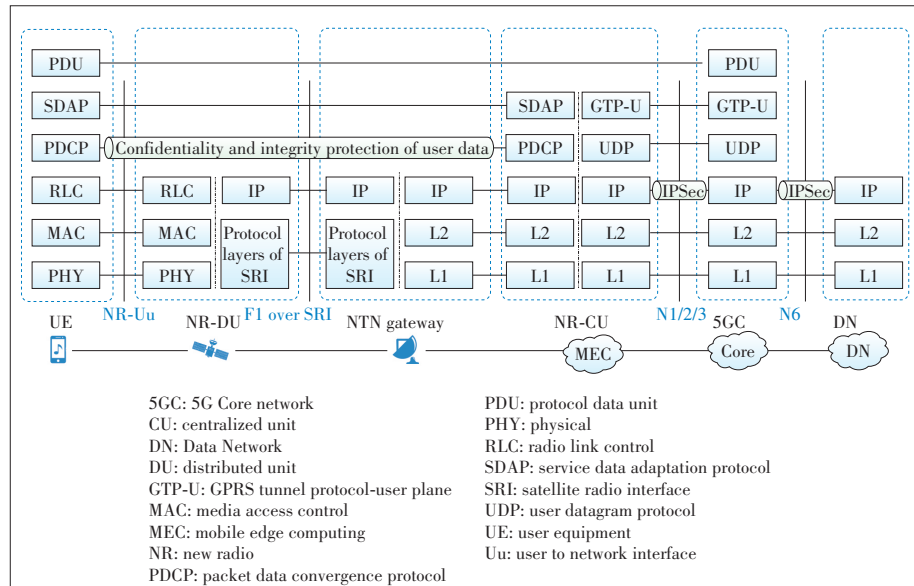
The comparison between Solutions 1 and 2 is shown in Table 2.

The above two solutions can help 5G SCCN solve the contradiction between

user data transmission protection and the weak processing capability of satellite nodes, so that the satellite nodes can avoid heavy cryptographic computing. Solution 1 realizes end-to-end



▲ Figure 6. Schematic diagram of end-to-end encryption in user plane



▲ Figure 7. Schematic diagram of CU-DU separated user plane encrypted communication

▼ Table 2. Comparison of on-board security solutions without user plane encryption

Solution	Advantages	Disadvantages
Solution 1	1) End-to-end encryption 2) The intermediate network cannot obtain data and the level of privacy and security are high	1) The legal interception function is affected 2) VoLTE service quality is affected
Solution 2	1) Segmented encryption easy for security supervision 2) IMS service is not affected	It is possible for the network provider to obtain confidential user data

IMS: IP multimedia subsystem VoLTE: voice over Long-Term Evolution

encryption and integrity protection from terminal to service. Although it achieves higher security and privacy, it has a certain impact on legal monitoring and VoLTE services. While Solution 2 uses segmented encryption and decryption and integrity protection, posing requirements for network deployment.

The two solutions are not contradictory and can be used in combination. For general services, Solution 2 is recommended; for high confidential services, Solutions 1 and 2 can be activated simultaneously, so that critical data can be double protected.

4.3 Network Availability

Availability attacks on existing networks will still exist in 5G SCCN, such as electromagnetic interference, DDoS attacks and signaling storms. Considering the openness of the satellite network environment, lower processing capacity, and high value of the services carried, the security situation is severer.

In addition to the impact of unconscious group behavior on key resources, it is also necessary to combine the characteristics of 5G SCCN global coverage, focusing on the possibility of satellites suffering from availability attacks over the sea or in the air and enhancing the protection of signaling resources at each protocol layer. At present, the industry's protection measures for the availability of satellite networks are not yet systematic or effective, and research needs to be strengthened.

4.3.1 Anti-DDoS Attack

Although identity authentication can make the network reject a large number of unauthorized communications, there are still a certain number of protocol interactions without authentication, or authentication itself can also cause DDoS attacks.

DDoS attacks on the user plane can be effectively prevented by strengthening access authentication and single-session traffic rate limit. DDoS attacks on the control plane can use conventional security defense mechanisms, such as prohibiting Internet control message protocol (ICMP) packets and broadcast packets, adding access control list (ACL) filtering, and adding black and white lists. At the same time, there are still DDoS first-packet attacks on the control plane. A single-packet authorization mechanism can be considered. Meanwhile, special modules and cryptographic chips can be used on the user plane to reduce the consumption of CPU resources.

Considering the limitations of satellite network resources and complex defense strategies, new challenges have been posed to satellite resources. We can consider lightweight DDoS attack detection methods, such as self-organizing map (SOM)^[26] and support vector machine (SVM)-SOM^[27] technologies, building an unsupervised artificial neural network trained by traffic characteristics to detect DDoS attacks, or combining LSTM deep learning models and SVM technologies^[28-29] to perform DDoS detection in spatial networks.

4.3.2 Anti-UE Random Access Attacks

Random competitive access of 5G NR may cause a signaling storm. For example, a base station malfunctions due to mass activities or large-area calls caused by disasters, which can usually be avoided by speed limiting. However, misuse of random access resources or malicious competition access, such as using a UE simulator to make a directional analog call to a specific satellite, will also make the satellite fail to access real and effective calls. Due to the asymmetry of space-ground computing resources, the possibility of such problems erupting in insecure areas also exists.

In order to avoid the aforementioned UE random access attack, the response message can be scrambled. For example, the satellite base station response message can be scrambled, so that the attacker cannot decode it correctly and no longer sends the radio resource request message to avoid occupying resources.

4.3.3 Anti-Wireless Communication Interference

In the sea and sky environment, frequency band suppression attacks may occur. Since the ground transmission power can be several times that of the satellite, the attacker can track and aim the communication satellite and launch strong interference signals to the satellite, including blocking interference and noise interference, which greatly deteriorates the signal-to-noise ratio of the wireless channel. Meanwhile, the 5G frequency band is public, and it is easier for attackers to implement targeted frequency band suppression instead of full frequency band suppression, which will further increase the effectiveness of the attack. Using spot beam and line beam antenna technology to dynamically allocate wireless channels, increasing frequency band guard bands, improving filtering accuracy, and adopting frequency shift can cope with wireless communication interference to a certain extent.

4.4 Network Resource Sharing

Satellite resources are costly and space is limited, so limited satellite resources must be shared among multiple services. 5G SCCN resource sharing can refer to 5G network slicing technology. On the one hand, through the exploration of 5G in the field of Industrial Internet, 5G end-to-end network slicing technology has gradually matured, which provides a good foundation for the feasibility of 5G SCCN resource sharing. On the other hand, the existing 5G network slicing technology shares too large a granularity of resources and is not suitable for direct application on the 5G SCCN. 5G SCCN slicing requires more refined network resource management technology. Meanwhile, the network slicing technology of 5G SCCN also needs to be adaptively designed for other features such as the time-varying topology of satellite communication and network heterogeneity.

The end-to-end security isolation mechanism for network slicing includes RAN slice security isolation, bearer network

slice security isolation, and core network slice security isolation. Fig. 8 compares the 5G SCCN with the slicing technology. It can be seen that even in the manner of separate deployment of CU-DU, the two can be properly mapped. This shows that the 5G network slicing technology has reference significance for the 5G SCCN slicing technology.

4.4.1 RAN Slice Isolation

The core technology of RAN slice isolation is the isolation of wireless spectrum resources, which divides the wireless spectrum into different resource blocks from the time domain, frequency domain, and space domain dimensions for air interface communication. This technology can still be applied in 5G SCCN. According to the requirements of different application scenarios, the use of resource pool reservation and allocation can realize the isolation of the wireless channel for the terminal to access the satellite. Limited by satellite resources, 5G SCCN slicing is likely to be service-oriented rather than industry-oriented.

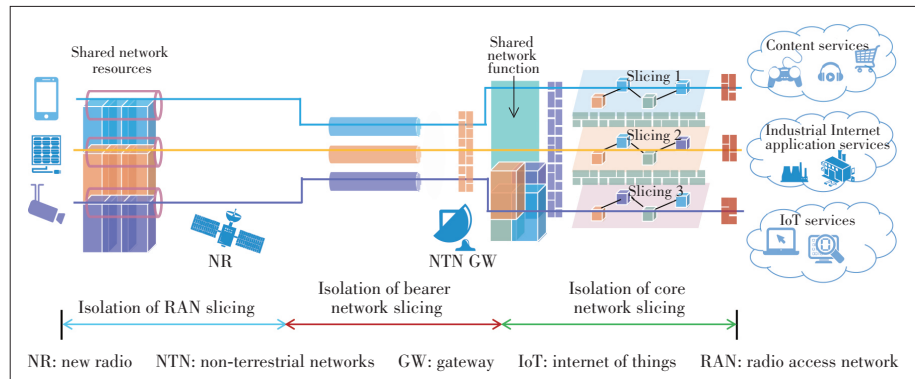
4.4.2 Bearer Network Slice Isolation

The bearer network in the 5G SCCN includes the backhaul network and the middle-haul network, which will cover both the on-board link and the ground link. Currently, bearer network isolation mechanisms include logical virtual local area network (VLAN) isolation, and the Ethernet fragmentation technology (such as FlexE) to achieve physical isolation at the time slot level. Although FlexE has better security, the granularity of 5 Gbit/s is obviously not suitable for applications on satellite links.

Considering that the movement of satellites causes the network topology to dynamically change, the corresponding network slicing also needs to be dynamically adjusted. Although the satellite behavior can be predicted based on the ephemeris information, and the resources of the 5G SCCN bearer network can be planned and deployed in advance, this is still a complicated technical problem. In addition, issues such as the granularity of scheduling between on-board bearer network slicing and terrestrial bearer network slicing, protocol compatibility, and resource docking are also to be studied.

4.4.3 Core Network Slice Isolation

The core network in 5G SCCN is located on the ground, so it can inherit the existing 5G core network slicing technology. This part is relatively mature. Physical isolation can be used to allocate relatively independent physical resources to the slices with higher security requirements. A logical isolation solution can also be used to manage and orchestrate networks and network functions with the help of the virtualization technology.



▲ Figure 8. Schematic diagram of 5G satellite converged communication network (SCCN) slice isolation

5 Conclusions

The 5G network is expected to achieve wide-area coverage and integrated space-ground communications by converging with satellite networks. However, due to the characteristics of openness, dynamics, and low power consumption of satellite networks, 5G SCCN is faced with new security challenges. Based on the network mode proposed by 3GPP TR38.811, this paper conducts a comprehensive analysis from the three dimensions of network structure, network layering and security attributes, and constructs the security architecture of the 5G SCCN, so that readers can get an overall and specific understanding of the 5G SCCN from a security perspective.

By analyzing the security attributes of each layer and segment of the 5G SCCN, four key security technologies can be summarized, namely, strict identity authentication, lightweight data communication security, enhanced network availability, and fine-grained resource sharing and isolation. Based on strict identity authentication, building a relatively independent and trustworthy communication system in an open environment is a key feature that distinguishes the 5G SCCN from the traditional 5G network. Based on the principle of “forwarding on-satellite, processing off-satellite” and NR’s CU-DU separate deployment, a lightweight communication security assurance for 5G SCCN can be provided. This is also a reference for the network design that this paper put forward from a security perspective. Based on new radio and signaling protection technologies, the availability of 5G SCCN services over the sea and in the air is enhanced; Based on the refined network resource management technology, fine-grained resource sharing and isolation for 5G SCCN applications is provided.

All in all, the 5G SCCN can inherit the existing security mechanisms and technologies of 5G networks and IP bearer networks to a large extent. There is no need to start anew for its security or design a completely different set of security architecture and protocol, but it should also be seen that the new network after integrating is faced with huge challenges. It is necessary to systematically design the network security architecture. Meanwhile, it is also necessary to fully consider the network security requirements at the architecture design phase.

References

- [1] 3GPP. Study on new services and markets technology enablers: 3GPP TR 22.891 V14.2.0 [S]. 2016
- [2] XU B Y, HAN M. Study on international standards of satellite communications [J]. Information and communications technology and policy, 2019, (17): 41 – 44
- [3] WANG C T, LI N, ZHAI L J, et al. Preliminary study on the integration of satellite communications and terrestrial 5G network [J]. Satellite & network, 2018, (9): 14 – 21
- [4] KONSTANTINOS L, ALEXANDER G, RAY S, et al. Use cases and scenarios of 5G integrated satellite-terrestrial networks for enhanced mobile broadband: the SaT5G approach [J]. International journal of satellite communications and networking, 2019, 37(2): 91 – 112
- [5] LIU S J, HU Y M, WANG D P. Overview of studies on the satellite-5G integration [J]. Information and communications technology and policy, 2019, (5)
- [6] SHEN Y Y. The Development trend of satellite communications in 5G era. Space international [J]. 2020, (1): 48 – 52
- [7] CRUICKSHANK H, IYENGAR S, SUN Z L. Securing IP multicast over GEO satellites [C]//IEEE Seminar on Broadband Satellite: The Critical Success Factors Technology, Services and Markets. London, UK: IEEE, 2000. DOI: 10.1049/ic: 20000534
- [8] NOUBIR G, ALLMEN LVON. Security issues in Internet protocols over satellite links [C]//50th Vehicular Technology Conference. Amsterdam, Netherlands: IEEE, 1999: 2726 – 2730. DOI: 10.1109/VETECF.1999.800282
- [9] ROY-CHOWDHURY A, BARAS J S, HADJITHEODOSIOU M, et al. Security issues in hybrid networks with a satellite component [J]. IEEE wireless communications, 2005, 12(6): 50 – 61. DOI: 10.1109/MWC.2005.1561945
- [10] JIANG Y W, ZHANG G X, ZHAO L D, et al. Summary of satellite communication and 5G convergence system development [C]//The 15th Annual Conference of satellite Communication. Beijing, China: CIC, 2019: 56 – 65
- [11] CHEN S Z, SUN S H, KANG S L. System integration of terrestrial mobile communication and satellite communication—the trends, challenges and key technologies in B5G and 6G [J]. China communications, 2020, 17(12): 16
- [12] KODHELI O, GUIDOTTI A, VANELLICORALLI A. Integration of Satellites in 5G through LEO Constellaions [C]//IEEE Global Communications Conference. Singapore, Singapore: IEEE, 2017: 1 – 6. DOI: 10.1109/GLOCOM.2017.8255103
- [13] CHEN T T, WANG W J, DING R, et al. Location-based timing advance estimation for 5G integrated LEO satellite communications [C]//IEEE global communications conference. Taipei, China: IEEE, 2020: 1 – 6. DOI: 10.1109/GLOBECOM42002.2020.9322428
- [14] TANG Q Q, XIE R C, LIU X, et al. MEC enabled satellite-terrestrial network: architecture, key technique and challenge [J]. Journal on communications, 2020, 41(4): 162 – 182
- [15] ZHANG Z, ZHANG W, TSENG F H. Satellite mobile edge computing: improving QoS of high-speed satellite-terrestrial networks using edge computing techniques [J]. IEEE Network, 2019, 33(1): 70 – 76
- [16] LI F H, YIN L H, WU W. Research status and development trends of security assurance for space - ground integration information network [J]. Journal on communications, 2016, (11): 160 – 172
- [17] JI X S, LIANG H, HU H C. New thoughts on security technologies for space-ground integration information network [J]. Telecommunications science, 2017, (12): 30 – 41
- [18] Huang Zhan. Research on security protocol of broadband satellite network [D]. Harbin Institute of Technology, 2012
- [19] ITU-R M. Key elements for integration of satellite systems into next generation access technologies [EB/OL]. (2019-07-02) [2021-04-06]. <https://www.itu.int/md/r15-wp5d-c-1263/en>
- [20] 3GPP. Study on new radio (NR) to support non-terrestrial networks: TR 38.811 V15.4.0 [S]. 2020
- [21] 3GPP. Study on using satellite access in 5G: 3GPP TR 22.822 V16.0.0 [S]. 2018
- [22] 3GPP. Study on scenarios and requirements for next generation access technologies: 3GPP TR 38.913 V16.0.0 [S]. 2020
- [23] 3GPP. Study on architecture for next generation system: 3GPP TR 23.799 V14.0.0 [S]. 2016
- [24] 3GPP. Service requirements for the 5G system: 3GPP TS 22.261 V17.2.0 [S]. 2020
- [25] ZHANG Z J, ZHOU Q, ZHANG C. New low-earth orbit satellites authentication and group key agreement protocol [J]. Journal on communications, 2018, (6): 150 – 158. DOI: 10.11959/j.issn.1000 – 436x.2018102
- [26] BRAGA R, MOTA E, PASSITO A. Lightweight DDoS flooding attack detection using NOX/OpenFlow [C]//IEEE Local Computer Network Conference. Denver, USA: IEEE, 2010: 408 – 415. DOI: 10.1109/LCN.2010.5735752
- [27] DEEPA V, SUDAR K M, DEEPALAKSHMI P. Detection of DDoS attack on SDN control plane using hybrid machine learning techniques [C]//International Conference on Smart Systems and Inventive Technology. Tirunelveli, India: IEEE, 2018: 299 – 303. DOI: 10.1109/ICSSIT.2018.8748836
- [28] JIA M, SHU Y J, GUO Q, et al. DDoS attack detection method for space – based network based on SDN architecture [J]. ZTE communications, 2020, 18 (4): 18 – 25. DOI: 10.12142/ZTECOM.202004004
- [29] YANG L F, ZHAO H. DDoS attack identification and defense using SDN based on machine learning method [C]//The 15th International Symposium on Pervasive Systems, Algorithms and Networks (I - SPAN). Yichang, China: IEEE, 2018: 174 – 178. DOI: 10.1109/I – SPAN.2018.00036

Biographies

YAN Xincheng (yan.xincheng@zte.com.cn) received his M.S. degree from South-east University, China in 2004. He is currently the chief system architect and director of the Security Technology Committee of ZTE Corporation, responsible for network security technology planning. He was the leader of the network security sub-project of the “New Generation Broadband Wireless Network Communication Network” and National Science and Technology Major Project “5G Security Overall Architecture Research and Standardization”. He has been awarded a number of scientific and technological awards of Jiangsu Province and Shenzhen.

TENG Huiyun received her M.S. degree from Hohai University, China in 2011. She is currently the senior technical research engineer in ZTE Corporation and has 10 years of professional experience in communication and security.

PING Li received her M.S. degree from Southeast University, China in 2005. She is currently the senior cybersecurity analyst in ZTE Corporation with 8 years of professional experience in cybersecurity policy and technology research and analysis. She got the CISSP in 2018.

JIANG Zhihong received his M.S. degree from Nanjing University of Posts and Telecommunications, China in 2003. He is currently the senior technical research expert in ZTE Corporation.

ZHOU Na received her Ph.D. degree from Nanjing University of Aeronautics and Astronautics University, China in 2004. She is currently the senior technical research expert in ZTE Corporation. She has won Shenzhen Scientific and Technological Award.



Payload Encoding Representation from Transformer for Encrypted Traffic Classification

Abstract: Traffic identification becomes more important, yet more challenging as related encryption techniques are rapidly developing nowadays. Unlike recent deep learning methods that apply image processing to solve such encrypted traffic problems, in this paper, we propose a method named Payload Encoding Representation from Transformer (PERT) to perform automatic traffic feature extraction using a state-of-the-art dynamic word embedding technique. By implementing traffic classification experiments on a public encrypted traffic data set and our captured Android HTTPS traffic, we prove the proposed method can achieve an obvious better effectiveness than other compared baselines. To the best of our knowledge, this is the first time the encrypted traffic classification with the dynamic word embedding has been addressed.

Keywords: traffic identification; encrypted traffic classification; natural language processing; deep learning; dynamic word embedding

HE Hongye, YANG Zhiguo,
CHEN Xiangning

(ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202104010

<https://kns.cnki.net/kcms/detail/34.1294.TN.20211104.1636.002.html>, published online November 5, 2021

Manuscript received: 2021-02-23

Citation (IEEE Format): H. Y. He, Z. G. Yang, and X. N. Chen, "Payload encoding representation from transformer for encrypted traffic classification," *ZTE Communications*, vol. 19, no. 4, pp. 90–97, Dec. 2021. doi: 10.12142/ZTECOM.202104010.

1 Introduction

Traffic classification, a task to identify certain categories of network traffic, is crucial for Internet services providers (ISP) to track the source of network traffic and to further ensure their quality of service (QoS). Also, traffic classification is widely applied in some specific missions, like malware traffic identification and network attack detection. However, this is a challenge since network traffic nowadays is more likely to be hidden with several encryption techniques, making detection hard with a traditional approach.

Typically, there are the following widely applied traffic classification methods: 1) The port-based method, which simply identifies traffic data using specific port numbers, is susceptible to the port number changing and port disguise. 2) Deep packet inspection (DPI), a method which aims to locate patterns and keywords from traffic packets, is not suitable for identifying encrypted traffic because it heavily relies on unencrypted information. 3) The machine learning (ML)-based method focuses on using manually designed traffic statistical features to fit a machine learning model for categorization^[1]. 4) The deep learning (DL)-based method is an extension of the

ML-based approach where neural networks are applied for automatic traffic feature extraction.

Although encrypted traffic packets are hard to identify, an encrypted traffic flow (a flow is a consecutive sequence of packets with the same source IP, source port, destination IP, destination port and protocol) is still analyzable because the first few packets of a flow may contain visible information like handshake details^[2]. In this way, the ML-based and DL-based methods are considered ideal for encrypted traffic classification since they both extract common features from the traffic data. In fact, the ML-based and DL-based methods share the same concept that traffic flows could be vectorized for supervised training according to their feature extraction strategy.

Rather than extracting hand-designed features from the traffic as the ML-based method does, the DL-based method uses a neural network to perform representation learning (RL) for the traffic bytes which allow it to avoid complex feature engineering. It provides an end-to-end solution for encrypted traffic classification where the direct relationship between raw traffic data and its categories is learned. The classification effect of a DL-based method is highly related to its capacity of representation learning.

In this paper, we propose a new DL-based solution named Payload Encoding Representations from Transformers (PERT) in which a dynamic word embedding technique called the Bidirectional Encoder Representations from Transformers (BERT)^[3] is applied during the traffic representation learning phase. Our work is inspired from the great improvements in the natural language processing (NLP) domain that dynamic word embedding brings. We believe that computer communication protocols and natural language have some common characteristics. According to this point, we shall prove that such a strong embedding technique can also be applied to encode traffic payload bytes and provide substantial enhancement while addressing the encrypted traffic classification task.

2 Related Work

We shall introduce some related traffic identification works that involve the DL domain, and further categorize them into two major groups.

1) For feature-engineering: Basically, these methods still use hand-designed features but utilize the DL as a measure of feature processing. For example, JAVAID et al. proposed an approach using the deep belief network (DBN) to make a feature selection before the ML classification^[4]. HOCHST et al. introduced the auto-encoder network to perform dimension reduction for the manually extracted flow features^[5]. REZAEI et al. applied a similar pre-training strategy as we do^[6]. What is different is that this work introduced neural networks to reconstruct time series features. Its pre-training plays a role of re-processing the hand-designed features. Ours instead, is to perform a representation learning for the raw traffic.

2) For representation learning: These works apply DL to learn the encoding representation from raw traffic bytes without manual feature-engineering. These works are also considered as end-to-end implements of traffic classification. WANG et al. proposed this encrypted traffic classification framework for the first time^[7]. They transformed payload data to grayscale images and applied convolutional neural networks (CNN) to perform image processing. Afterward, the emergence of a series of CNN-based works, such as Ref. [8], proved the validity of such an end-to-end classification. LOPEZ-MARTIN et al. further discussed a possible combination for traffic identification where the CNN was still used for representation learning, but a long short-term memory (LSTM) network was introduced to learn the flow behaviors^[9]. It inspired the hierarchical spatial-temporal features-based (HAST) models which obtained a state-of-the-art result in the intrusion detection domain^[10].

Nevertheless, for end-to-end encrypted traffic classification nowadays, CNN is still the mainstream whereas the NLP-related network only works as an supplement to do jobs such as capturing flow information. We can hardly find a full-NLP scheme similar to ours, let alone one which applies current dy-

namic word embedding techniques.

3 Model Architecture

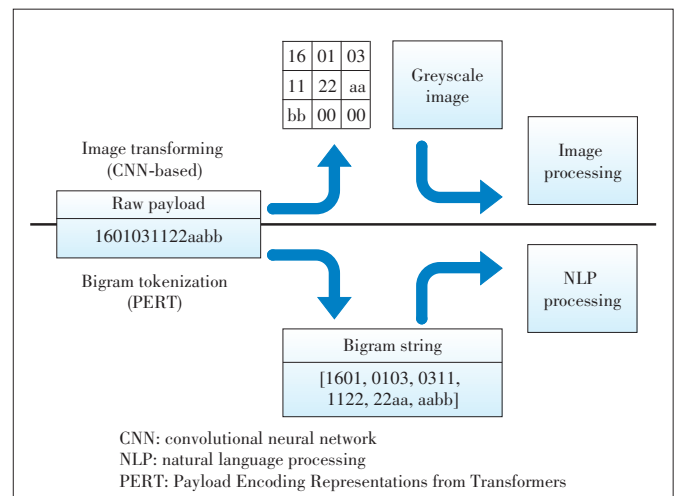
3.1 Payload Tokenization

According to Ref. [2], the payload bytes of a packet are likely to expose some visible information, especially for the first few packets of a traffic flow. Thus, most DL-based methods use this byte data to construct traffic images as the inputs of a CNN model. This is because the byte data is ideal for generating pixel images as its value ranges from 0 to 255, which is just fit for a grayscale image. Rather than applying such an image processing strategy, we treat the payload bytes of a packet as a language-like string for introducing NLP processing.

However, the range of byte value is rather small considering the size of a common NLP vocabulary. To extend the vocab size of traffic bytes, we introduce a tokenization which takes pairs of bytes (with a value range of 0 to 65 535) as basic character units to generate bigram strings (Fig. 1). Afterward, the NLP related encoding methods can be directly applied to the tokenized traffic bytes. Thus, the encrypted traffic identification is transformed to an NLP classification task.

3.2 Representation Learning

While performing representation learning in an NLP task, the word embedding is widely utilized. Recently, a breakthrough was made in this research area as the dynamic word embedding technique overcame the drawback that the traditional word embedding methods such as the Word2Vec^[11] are only capable of mapping words to unchangeable vectors. By contrast, vectors trained by dynamic word embedding can be adjusted according to its context inputs, making it more powerful to learn detailed contextual information. This is just what we need for extracting complex contextual features from the



▲ Figure 1. Comparison of raw payload processing between CNN-based methods and PERT

encrypted traffic data.

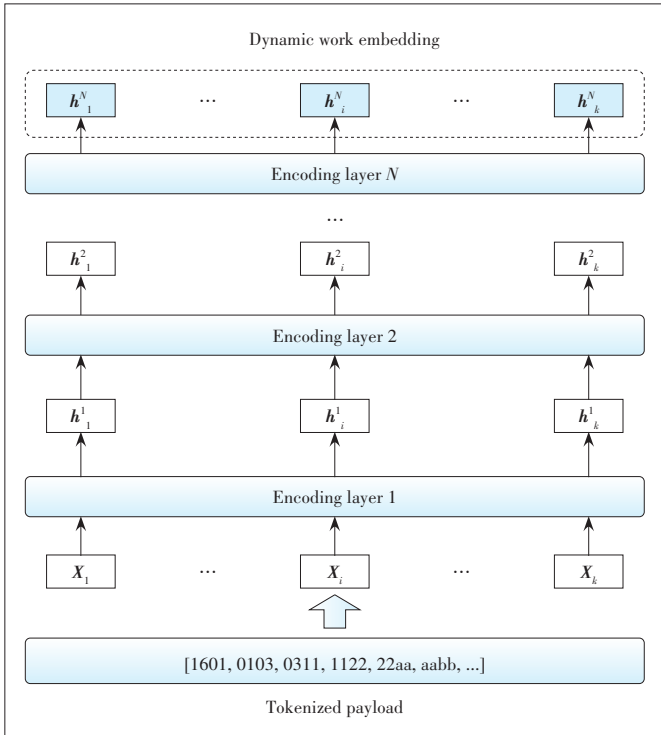
Current popular dynamic word embedding methods such as BERT could be considered as a stack of a certain type of encoding layers. Each encoder takes the outputs of its former layer as inputs and further learns a more abstract representation. In another word, word embedding will be dynamically adjusted while passing through its next encoding layer.

In our work, we take the tokenized payload string $[w_1, w_2, \dots, w_k]$ as our original inputs. The first group of word embedding vectors $[x_1, x_2, \dots, x_k]$ at the bottom of the network is randomly initialized. After N times of dynamic encoding, we obtain the final word embedding outputs $[h_1^N, h_2^N, \dots, h_k^N]$ that imply extremely abstract contextual information of the original payload.

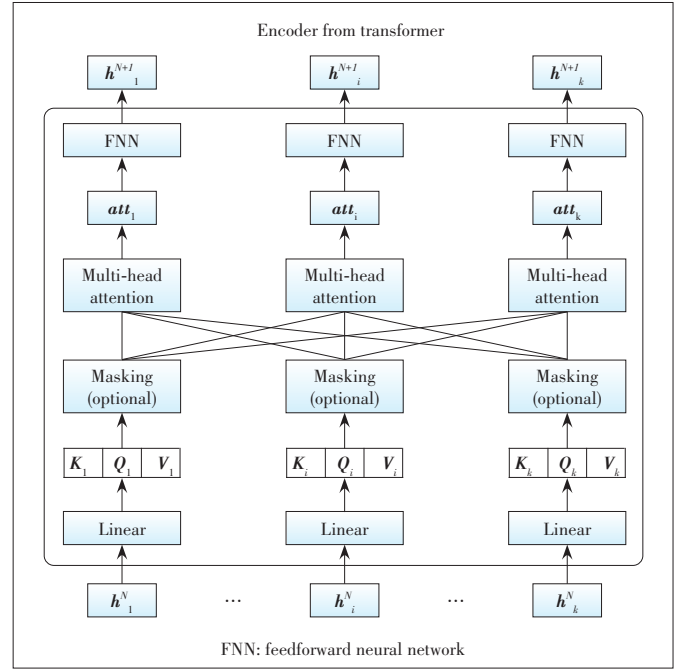
The illustration of our representation learning is shown in Fig. 2.

Earlier dynamic word embedding called the Embeddings from Language Models (Elmo)^[12] uses the bidirectional LSTM as its encoder unit, which is not suitable for large-scale training since the LSTM has a bad support for parallel calculations. To solve this problem, the LSTM was replaced with a self-attention encoder that was firstly applied in the transformer model^[13], and this embedding model was named BERT. This is what we also use for encoding the encrypted payload as shown in Fig. 3. Taking our first embedding vectors $[x_1, x_2, \dots, x_k]$ as examples, there are several steps of the transformer encoding as follows.

1) Linear projections: Each embedding vector x_i will be projected to three vectors using linear transformations:



▲ Figure 2. Representation learning network for payload data using the dynamic word embedding architecture



▲ Figure 3. Detail of the encoding layer

$$\begin{aligned} K_i &= W^K x_i \\ Q_i &= W^Q x_i \\ V_i &= W^V x_i, \end{aligned} \quad (1)$$

where W^K , W^Q and W^V are the three groups of linear parameters.

2) Self-attention and optional masking: The purpose of linear projections is to generate the inputs for the self-attention mechanism. Generally speaking, self-attention is to figure the compatibility between each input x_i and all the other inputs $x_1 - x_k$ via a similarity function, and further to calculate a weighted sum for x_i which implies its overall contextual information. In detail, our self-attention is calculated as follows:

$$att_i = \sum_{j=1}^k \frac{Q_i K_j^T}{Z} \times V_j. \quad (2)$$

The similarity between x_i and x_j is figured by a scaled dot-product operator, where d_k is the dimension of K_j , and Z is the normalization factor. It should be noticed that not every input vector is needed for self-attention calculation. An optional masking strategy that randomly ignores a few inputs while generating attention vectors is allowed to avoid the over-fitting.

3) Multi-head attention: In order to grant encoders the ability of reaching more contextual information, the transformer encoding applies a multi-head attention mechanism. Specifically, linear projections will be done for M times for each x_i to generate multiple attention vectors $[att_{i,1}, att_{i,2}, \dots, att_{i,M}]$. Afterward, a concatenation operator is utilized to obtain the final attention vector:

$$\mathbf{att}_i = \mathbf{att}_{i,1} \oplus \mathbf{att}_{i,2} \oplus \dots \oplus \mathbf{att}_{i,M}. \quad (3)$$

4) Feed-forward network (FFN): A fully-connected network is used to provide the output of current encoder. For \mathbf{x}_i , it is as follows:

$$\mathbf{h}_i = \max(0, \mathbf{W}_1 \mathbf{att}_i + \mathbf{b}_1) + \mathbf{b}_2, \quad (4)$$

where \mathbf{W}_1 , \mathbf{b}_1 , \mathbf{W}_2 and \mathbf{b}_2 are the full-connection parameters and $\max(0, x)$ is a Rectified Linear Unit (ReLU) activation function.

Finally, we get the dynamic embedding \mathbf{h}_i which is encoded from \mathbf{x}_i . It can be further encoded by the next encoding layer or be directly used in downward tasks. Similar to the naming of BERT, we name our encoding network as the PERT considering the application of a transformer encoder.

3.3 Packet-Level Pre-Training

A key factor that makes BERT and its extensive models continuously achieve state-of-the-art results among a wide range of NLP tasks is their “pre-training + fine-tuning” strategy. To the best of our knowledge, our work is the first to introduce such a strategy to an end-to-end encrypted traffic classification architecture. To perform pre-training is to initialize the encoding network and to give it the ability of contextual information extraction before it is applied to downward tasks. The unsupervised language model (LM) is widely used for word embedding pre-training^[14]. BERT, specifically, proposes a masked LM which hides several words from the original string with a unique symbol ‘unk’, and uses the rest of the words to predict those hidden ones.

To demonstrate the procedure of the masked LM, we give a masked traffic bigram string as $\mathbf{w} = [w_1, \dots, \text{'unk'}, \dots, \text{'unk'}, \dots, w_k]$ and a list $\mathbf{msk} = [i_1, i_2, \dots, i_m]$ which indicates the position of bigram units that are masked. After the encoding, for each embedding vector \mathbf{h}_i that is encoded from the i -th position of the original input, a full connection is followed:

$$\mathbf{o}_i = \mathbf{W}' \tanh(\mathbf{W} \mathbf{h}_i + \mathbf{b}) + \mathbf{b}', \quad (5)$$

where \tanh is an activation function like the Relu. The size of the output vector $\mathbf{o}_i = [o_{i,1}, \dots, o_{i,|V|}]$ is the vocab size $|V|$. It stores all the likelihoods about what the corresponding traffic bigram is at the i -th position.

In the end, the masked LM uses partial outputs $\{\mathbf{o}_i | i \in \mathbf{msk}\}$ to perform a large softmax classification with the class number of vocab size. The objective is to maximize the predicted probabilities of all the masked bigrams, which can be simply written as:

$$\arg\max_{\theta} \sum_{i \in \mathbf{msk}} P(w_i | \mathbf{o}_i, \theta), \quad (6)$$

where θ represents the parameters of the entire network.

The LM is considered as a powerful initialization approach for the encoding network using large-scale unlabeled data, yet

it is very time-consuming. Even if we want to perform a flow-level classification for encrypted traffic, we argue that the pre-training should be packet-level considering the possible calculation costs. Particularly, we collect raw traffic packets from the Internet despite their sources and extract their payload bytes to generate an unsupervised data set. Then, the extracted payload bytes are tokenized as bigram strings and are utilized to perform a PERT pre-training. After the training converges, we save the adjusted encoding network.

3.4 Flow-Level Classification

While implementing a certain task like classification, the pre-trained encoding network will be totally reused and be further adjusted to learn the real relationship between the inputs and a specific task objective. This is the concept of “fine-tuning”, where a network is trained based on a proper initialization to achieve a boosted effect in downward tasks.

Fig. 4 shows our encrypted traffic classification framework. Below are the detailed descriptions:

1) Packets extraction: While classifying an encrypted traffic flow, only the first M packets (3 for example in Fig. 4) need to be used. The bigram tokenization is performed for payload bytes in each packet to generate a list of tokenized payload strings $[\mathbf{str}_1, \mathbf{str}_2, \dots, \mathbf{str}_M]$.

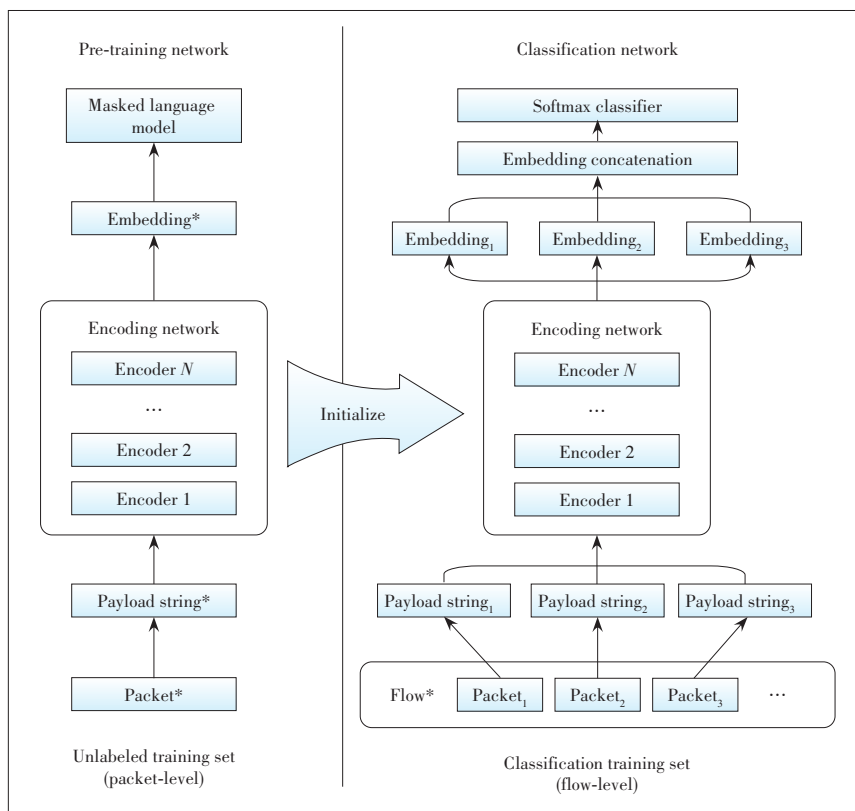
2) Encoding for packets: Before classification, the encoding network of the classifier with the pre-trained counterpart is initialized. As the encoding network is packet-level, each tokenized string will be individually transported to the encoders. According to Ref. [3], while carrying out a classification with BERT, a unique token ‘cls’ should be added at the beginning of the input as the classification mark. For the i -th packet, its tokenized string will be modified as $\mathbf{str}_i' = [\text{cls}, w_{i,1}, w_{i,2}, \dots, w_{i,k}]$. After encoding, a series of embedding vectors $[\mathbf{h}_{i,CLS}^N, \mathbf{h}_{i,1}^N, \dots, \mathbf{h}_{i,k}^N]$ is outputted, yet only the $\mathbf{h}_{i,CLS}^N$ will be picked as the further classification input. We simply represent $\mathbf{h}_{i,CLS}^N$ as \mathbf{emb}_i . In order to make use of all of the information extracted from the first M packets, we apply a concatenation to merge the encoded packets:

$$\mathbf{emb} = \mathbf{emb}_1 \oplus \mathbf{emb}_2 \oplus \dots \oplus \mathbf{emb}_M. \quad (7)$$

3) Final classification: In the end, a softmax classification layer is used to learn the probability distribution of the input flows among possible traffic classes. The objective of the flow-level classification can be written as:

$$\arg\max_{\theta} \sum_{f \in R_{flow}} P(y_f | \mathbf{emb}(f), \theta), \quad (8)$$

where R_{flow} represents the flow-level training set. Given a flow sample f , y_f represents its true label (class) and $\mathbf{emb}(f)$ indicates its concatenated embedding. P is the conditional probability that the softmax layer provides. In a manner of speaking, the objective is to maximize the probability that each encoded flow



▲ Figure 4. Flow-level encrypted traffic classification initialized with a packet-level pre-training

sample is predicted as its corresponding category. The flow-level information is involved in the final softmax classifier, and thus will be used to fine-tune the packet-level encoding network during the back propagation. The main point of such a fine-tuning strategy is to separate the learning of the packets relationship from the time-consuming pre-training procedures.

4 Experiments

4.1 Experiment Settings

4.1.1 Data Sets

1) Unlabeled traffic data set: The data set is utilized for the pre-training of our PERT encoding network. To generate this data set, we capture a large amount of raw traffic data from different sources using different devices through a network sniffer. Typically, there is no special requirement for the unlabeled traffic data except you should make sure your collected samples can cover the mainstream protocols, as many as possible.

2) Information Security Centre of Excellence (ISCX) data set: We chose ISCX2016 VPN-nonVPN¹, a popular encrypted

traffic data set, to make our classification evaluations more persuasive. However, this data set only marks where its encrypted traffic data is captured from and whether the capturing is through a VPN session or not, which means a further labeling should be performed. The ISCX data set is utilized in several works yet the results are rather different even when the same model is applied^[7-8]. This is mainly due to how the raw data is processed and labeled. Because WANG et al.^[7] have provided their pre-processing and labeling procedures in their github², we follow this open source project to process the raw ISCX data set and label it with 12 classes.

3) Android application data set: We find the ISCX data set is not entirely encrypted as it also contains data of some unencrypted protocols like Domain Name System (DNS). To make a better evaluation, in this work, we manually capture traffic samples from 100 Android applications via the Android devices and network sniffer tool-kit. All the captured data belong to the top activated applications of the Chinese Android app markets. Afterward, we exclusively pick the HTTPS

flows to ensure only the encrypted data remain.

4.1.2 Parameters

1) Pre-training: First of all, to perform the packet-level PERT pre-training for our unlabeled traffic data, we introduce public Python library transformers³ which provide implements of the original BERT model and several recently published modified models. In practice, we chose the optimized BERT implement named A Lite BERT (ALBERT)^[15], which is more efficient and less resource-consuming. However, even to be properly optimized, current BERT pre-training is very costly when we use 4 Nvidia Tesla P100 GPU cards.

Table 1 shows the settings of our pre-training and the corresponding description of each parameter. Such settings refer to what common NLP works with BERT encoding use. After sufficient training, the encoding network is saved as a Pytorch⁴ format which can be reused in our classification networks. Also,

▼ Table 1. Pre-training parameter settings

Parameter	Value	Description
hidden_size	768	Vector size of the encoding outputs (embedding vectors)
num_hidden_layers	12	Number of encoders used in the encoding network
num_attention_heads	12	Number of attention heads used in the multi-head attention mechanism
intermediate_size	3 072	Size of the hidden vectors in FNN networks
input_length	128	Amount of tokenized bigrams used in a single packet

1 <https://www.unb.ca/cic/datasets/vpn.htm>

2 <https://github.com/echowei/DeepTraffic>

3 <https://huggingface.co/transformers/>

4 <https://pytorch.org>

all of our other networks are implemented using the Pytorch.

2) Classification: The encoding network used at the classification stage strictly shares the same structure with the pre-trained one. Other settings of the classification layers are shown in Table 2. As fine-tuning the encoding network in a classification task is relatively inexpensive^[3], a single GPU card will be just enough.

4.1.3 Baselines

Below are the baseline classification methods we use for comparison:

1) ML-based: We refer to Ref. [16] to implement our ML-based method using the decision tree classifier (named ML-1). However, it only contains basic flow-statistical features that we consider as not the most optimized ML-based method. Thus, based on ML-1, we further add some time series features as the source ports, destination ports, directions, packet lengths and arrival time intervals of the first 10 packets in a flow to generate the ML-2 model.

2) CNN: The two types of CNN models are the 1D-CNN and the 2D-CNN, provided by Ref. [7]. They both use the first 784 bytes of a traffic flow to perform the classification.

3) HAST: The two HAST models proposed by Ref. [10] are the state-of-art end-to-end methods for intrusion detection. HAST-I uses the first 784 bytes of a flow for direct representation learning. HAST-II, however, only performs packet-level encoding. It further introduces an LSTM to merge the encoded packets.

During the evaluation, we randomly chose 90% of samples from the data set as the training set, and the remaining 10% for validation. Then, three widely used classification metrics are applied:

$$\begin{aligned} \text{Precision}(P) &= \frac{TP}{TP + FP}, \\ \text{Recall}(R) &= \frac{TP}{TP + FN}, \\ F1 - \text{score}(F1) &= \frac{2 \times P \times R}{P + R}. \end{aligned} \quad (9)$$

Take a class y_i as an example, the TP_i is the number of samples correctly classified as y_i , FP_i is the number of samples mistakenly classified as y_i , FN_i is the number of samples mistakenly classified as $\text{nor-}y_i$. As for the overall evaluation for all classes, we use the average values of those metrics.

4.2 Overall Analysis

1) Results on ISCX Data Set

▼Table 2. Classification parameter settings

Parameter	Value	Description
packet_num	alternative (5 by default)	The number of the first packets in a selected flow
softmax_hidden	768	Size of the hidden vectors in the softmax layer
dropout	0.5	The dropout rate of the softmax layer

This group of experiments are used to discuss the classification based on the consistent data settings of Ref. [7]. As we can see in Table 3, our flow-level PERT classification achieves the best classification results where the precision reaches 93.27% and the recall reaches 93.22%. It proves PERT is a power representation learning method for encrypted traffic classification.

As for other models, using the same manner of data preprocessing, the CNN classification results are pretty close to what is provided by Ref. [7]. The CNN methods obviously obtain higher precision and recall than the ML-1 that is implemented based on Ref. [15]. However, the ML-1 can still be improved. When the time series features are added, the precision of the ML-2 classification can exceed 89% which is much better than what the basic CNN methods get. In other words, the basic CNN methods actually have no absolute advantage while classifying the ISCX data set.

HAST-I achieves better results than typical CNN models yet HAST-II with an LSTM works relatively worse. In fact, we think using the first few bytes of a flow to perform a direct deep learning (like HAST-I and CNN-1D) is considered better than merging the packet-level encoded vectors, since the representation learning can directly capture flow-level information. However, the encoding costs on a long string are not affordable for complex dynamic word embedding. At the current stage, the “packet-level encoding + flow-level merging” is the best option for our PERT classification.

2) Results on Android Data Set

These experiments are based on full HTTPS traffic to evaluate the actual encrypted traffic classification ability of each method. As all the data here are HTTPS flows, in comparison with the ISCX data set whose data cover several traffic protocols, it is harder to distinctly locate different flow behaviors among the chosen applications. Consequently, the ML-based methods that strongly rely on flow statistics features work extremely weakly. Even when enhanced by time series features, the ML-2 still obtains a worse result than basic DL methods. As for the original ML-1, we find it is entirely not capable of addressing this 100-class HTTPS classification that we ignore

▼Table 3. Classification results (ISCX data set)

Model	Precision	Recall	F1
ML-1 ^[16]	0.819 4	0.813 6	0.816 4
ML-2	0.890 1	0.889 6	0.889 8
CNN-1D ^[7]	0.861 6	0.860 5	0.861 0
CNN-2D ^[7]	0.842 5	0.842 0	0.842 2
HAST-I ^[10]	0.875 7	0.872 9	0.874 2
HAST-II ^[10]	0.850 2	0.842 7	0.840 9
PERT	0.932 7	0.932 2	0.932 3

CNN: convolutional neural network
HAST: hierarchical spatial-temporal features-based model
ISCX: Information Security Centre of Excellence
ML: machine learning
PERT: Payload Encoding Representation from Transformer

its result in Table 4.

The results on the Android data set demonstrate that the DL-based methods are more suitable for processing full encrypted traffic data. More importantly, PERT again shows its superiority as it introduces a more powerful representation learning strategy. Its F1-score on the 100-class encrypted traffic classification exceeds 90% whereas the HAST can only achieve a result of 81.67%.

4.3 Discussion: Selection of the Packet Number

In a flow-level classification model, the increase in the use of packets will cause significant costs. This is particularly true when the representation learning is applied to traffic packets.

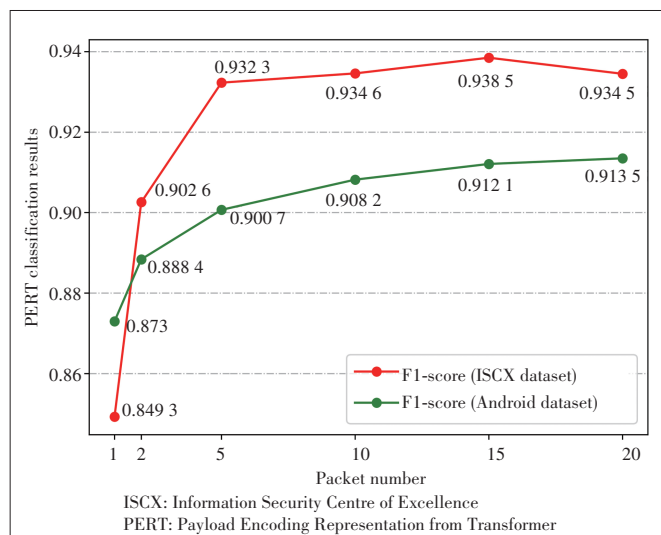
We perform PERT classification multiple times on the two data sets with different settings of the “packet_num” and the results are shown in Fig. 5. As we can see, at the beginning, the classification result on each data set is greatly improved with more packets used. However such increase is slight after the continuous adding of packets. For example, the F1-score is shown to reach 91.35% while classifying the Android data set with 20 packets. However, this result is merely boosted by 1.28% in comparison with using five packets, so it is not recommended considering the costs of PERT encoding for so many packets and such minor further improvements.

We point out that using 5 – 10 packets for our PERT classification will be sufficient. Similar conclusions can be also found in other flow-level classification research works such as Refs. [9] and [10].

4.4 Discussion: Merging of Encoded Packets

A major difference between our PERT classification and the most flow-level DL-based methods such as HAST-II is how the encoded packets are merged. HAST-II constructs a 2-layer LSTM after encoding the packet data whereas we simply apply a concatenation. To make a comparison between these two approaches, we modify our PERT model and the HAST-II model.

Firstly, we refer to HAST-II and construct the PERT_lstm model by using a 2-layer LSTM to follow our PERT encoded packets. Then, we remove the LSTM layer from HAST-II and further generate the HAST_con by concatenating the HAST



▲ Figure 5. Selection of the packet number for PERT classification

encoded packets to fit an ordinary softmax classifier, just as our original PERT model does. For all the compared methods, we consistently select 5 packets for classification based on our former discussion.

We perform validation every training epoch for each classification experiment and record corresponding F1-scores for evaluation. As illustrated in Fig. 6, we cannot actually tell which merging approach is better for classification accuracy. Whether using the concatenation or the LSTM approach for merging, it does not have a major influence on the final classification results.

However, using different merging approaches will have an obvious impact on the converging speed of the classification training. In Fig. 6, it always takes less training rounds before the model converges while introducing the concatenation merging. We believe the LSTM is not a satisfactory option for merging the encoded packets as applying a simple concatenation can reach a very close classification result, yet it is much faster.

5 Conclusions

After a thorough analysis of the possibility of applying the full-NLP scheme for encrypted traffic classification, we point out that the byte data of raw traffic packets can be transformed to character strings by proper tokenization. Based on this, we propose a new method named PERT to encode the encrypted traffic data and to serve as an automatic traffic feature extractor. In addition, we discuss the pre-training strategy of dynamic word embedding in a condition of the flow-level encrypted traffic classification. In accordance with a series of experiments on the public ISCX data set and Android HTTPS traffic, our proposed classification framework can provide significantly better results than current DL-based methods and traditional ML-based methods.

▼ Table 4. Classification results (Android data set)

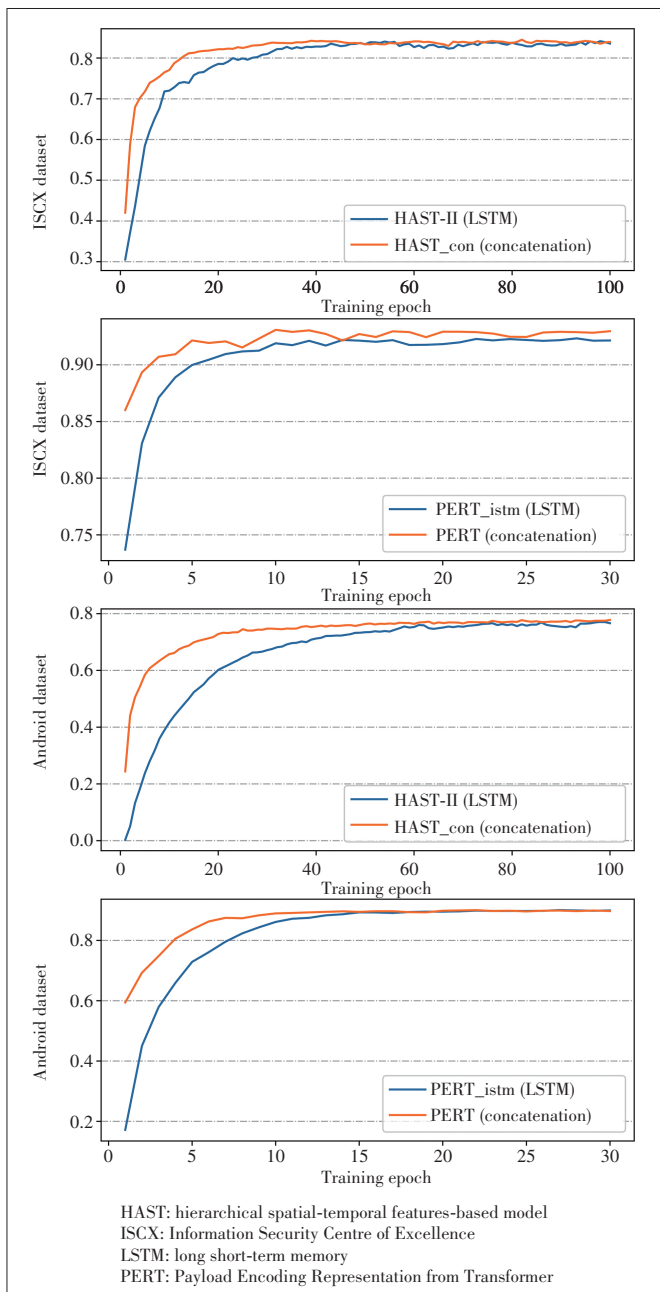
Model	Precision	Recall	F1
ML-1 ^[16]	/	/	/
ML-2	0.735 1	0.733 5	0.732 1
CNN-1D ^[7]	0.770 9	0.768 3	0.766 8
CNN-2D ^[7]	0.768 4	0.765 9	0.764 3
HAST-I ^[10]	0.820 1	0.818 5	0.816 7
HAST-II ^[10]	0.792 4	0.781 3	0.782 6
PERT	0.904 2	0.900 3	0.900 7

CNN: convolutional neural network

HAST: hierarchical spatial-temporal features-based model

ML: machine learning

PERT: Payload Encoding Representation from Transformer



▲ **Figure 6. F1-score converging speed comparison**

References

- [1] VELAN P, CERMAK M, CELEDA P, et al. A survey of methods for encrypted traffic classification and analysis [J]. *International journal of network management*, 2015, 25(5): 355 – 374. DOI: 10.1002/nem.1901
- [2] REZAEI S, LIU X. Deep learning for encrypted traffic classification: an overview [J]. *IEEE communications magazine*, 2019, 57(5): 76 – 81. DOI: 10.1109/MCOM.2019.1800819
- [3] DEVLIN J, CHANG M-W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]//*Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, USA: Association for

- Computational Linguistics, 2019: 4171 – 4186. DOI: 10.18653/v1/N19-1423
- [4] JAVAID A, NIYAZ Q, SUN W Q, et al. A deep learning approach for network intrusion detection system [C]//*Proceedings of the 9th EAI International Conference on Bio-Inspired Information and Communications Technologies*. Brussels, Belgium: ICST, 2016: 21 – 26. DOI: 10.4108/eai.3-12-2015.2262516
- [5] HOCHST J, BAUMGARTNER L, HOLLIICK M, et al. Unsupervised traffic flow classification using a neural autoencoder [C]//*42nd Conference on Local Computer Networks (LCN)*. Singapore, Singapore: IEEE, 2017: 523 – 526. DOI: 10.1109/LCN.2017.57
- [6] REZAEI S, LIU X. How to achieve high classification accuracy with just a few labels: a semi-supervised approach using sampled packets [EB/OL]. (2020-05-16)[2020-06-01]. <https://arxiv.org/abs/1812.09761v2>
- [7] WANG W, ZHU M, WANG J J, et al. End-to-end encrypted traffic classification with one-dimensional convolution neural networks [C]//*IEEE International Conference on Intelligence and Security Informatics (ISI)*. Beijing, China: IEEE, 2017: 43 – 48. DOI: 10.1109/ISI.2017.8004872
- [8] LOTFOLLAHI M, SIAVOSHANI M J, ZADE R S H, et al. Deep packet: a novel approach for encrypted traffic classification using deep learning [J]. *Soft computing*, 2020, 24: 1999 – 2012. DOI: 10.1007/s00500-019-04030-2
- [9] LOPEZ-MARTIN M, CARRO B, SANCHEZ-ESCUEVILLAS A, et al. Network traffic classifier with convolutional and recurrent neural networks for internet of things [J]. *IEEE access*, 2017, 5: 18042 – 18050. DOI: 10.1109/ACCESS.2017.2747560
- [10] WANG W, SHENG Y Q, WANG J L, et al. HAST-IDS: learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection [J]. *IEEE access*, 2017, 6: 1792 – 1806. DOI: 10.1109/ACCESS.2017.2780250
- [11] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [C]//*International Conference on Learning Representation*. Scottsdale, USA: ICLR, 2013
- [12] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations [EB/OL]. (2018-03-22)[2020-06-01]. <https://arxiv.org/abs/1802.05365v1#>
- [13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. (2018-03-22)[2020-06-01]. <https://arxiv.org/abs/1706.03762>
- [14] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. *The journal of machine learning research*, 2000, 3: 1137 – 1155
- [15] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations [EB/OL]. (2020-02-09)[2020-06-01]. <https://arxiv.org/abs/1909.11942v3>
- [16] DRAPER-GIL G, LASHKARI A H, MAMUN M S I, et al. Characterization of encrypted and VPN traffic using time-related features [C]//*2nd International Conference on Information Systems Security and Privacy (ICISSP)*. Rome, Italy: INSTICC, 2016

Biographies

HE Hongye (he.hongye@zte.com.cn) received his M.S. degree from Central South University, China in 2018. He is currently an algorithm engineer working with ZTE Corporation. His research interests include artificial intelligence and network traffic identification.

YANG Zhiguo received his M.S. degree from Hunan University, China in 2015. He is a senior software engineer at ZTE Corporation. His current research interests include Internet traffic identification and network security.

CHEN Xiangning received his bachelor's degree in communication engineering from Hunan University, China in 2004. He is a software engineer at ZTE Corporation. His research interests include big data technology and AI applications.



AI-Based Optimization of Handover Strategy in Non-Terrestrial Networks

Abstract: Complicated radio resource management, e.g., handover condition, will trouble the user in non-terrestrial networks due to the impact of high mobility and hierarchical layouts which co-exist with terrestrial networks or various platforms at different altitudes. It is necessary to optimize the handover strategy to reduce the signaling overhead and improve the service continuity. In this paper, a new handover strategy is proposed based on the convolutional neural network. Firstly, the handover process is modeled as a directed graph. Suppose a user knows its future signal strength, then he/she can search for the best handover strategy based on the graph. Secondly, a convolutional neural network is used to extract the underlying regularity of the best handover strategies of different users, based on which any user can make near-optimal handover decisions according to its historical signal strength. Numerical simulation shows that the proposed handover strategy can efficiently reduce the handover number while ensuring the signal strength.

Keywords: convolutional neural network; directed graph; handover; low earth orbit; non-terrestrial network

ZHANG Chenchen, ZHANG Nan,
CAO Wei, TIAN Kaibo, YANG Zhen

(State Key Laboratory of Mobile Network and
Mobile Multimedia Technology, ZTE Corporation,
Shenzhen 518055, China)

DOI: 10.12142/ZTECOM.202104011

<http://kns.cnki.net/kcms/detail/34.1294.TN.20211028.1117.002.html>, published online
October 28, 2021

Manuscript received: 2021-04-16

Citation (IEEE Format): C. C. Zhang, N. Zhang, W. Cao, et al., "AI-based optimization of handover strategy in non-terrestrial networks," *ZTE Communications*, vol. 19, no. 4, pp. 98 – 104, Dec. 2021. doi: 10.12142/ZTECOM.202104011.

1 Introduction

The non-terrestrial network (NTN) has been regarded as a supplement to the 5G terrestrial mobile network since it provides global coverage and service continuity^[1]. Compared with terrestrial networks, the handover in NTN is more frequent and complex. In this paper, a handover optimization method is proposed and applied to a typical NTN scenario, i. e., a low earth orbit (LEO) satellite network. A LEO is an orbit around the earth with an altitude between 500 km and 2 000 km^[1]. Compared with geostationary earth orbit satellites, the LEO satellites have much lower path-loss and propagation delay. Therefore, the third generation partnership project (3GPP) NTN study item has regarded the LEO satellites as the key to providing global broadband Internet access. Suppose the orbit is circular, the satellite will move around the earth at a constant velocity which is inversely proportional to the square root of the orbit altitude. Be-

cause of the low altitude, the LEO satellites have a high speed with respect to the earth, and terrestrial user equipment (UE) needs to frequently switch to new beams to keep connectivity. In order to ensure the quality of the Internet service, the optimization for NTN handover strategy needs to be carefully investigated.

Previous studies generally make handover decisions based on one or more predefined criteria. The most commonly used criteria include the elevation angle^[2], remaining service time^[3] and the number of free channels^[4], which correspond to the signal strength, handover number and satellite burden, respectively. But these methods cannot get an overall optimization. In Ref. [5], an overall optimization method is proposed by modeling the handover process by a directed graph. Each satellite is denoted by a node, then the best handover strategy is obtained by searching the shortest path. However, in Ref. [5] each satellite node is invariable during the handover process.

UE needs to perform handover as soon as entering the coverage of another beam and cannot choose an appropriate time. Besides, the UE needs to predict its coverage condition in a future time to construct the graph, which may bring unexpected error and is beyond the capability of standard 5G UE.

In recent years, some artificial intelligence (AI) techniques have been applied to search for overall optimization on handover. The most often used technique is Q-learning^[6-8], which is typical model-free reinforcement learning (RL). In Q-learning, some properties of UE are defined as its state, and the handover operation is defined as its action. Numerical simulation is used to iteratively train the Q-table (the reward of each action for each state) until its convergence. Then the UE can decide whether to perform handover according to its state. Furthermore, the Q-table can be replaced by a neural network for an infinite number of states. In Ref. [8], the handover in a LEO scenario is optimized by Q-learning. The state of UE is composed by its position, accessible satellites and whether handover is processed in this time slot. In each time slot, the UE is required to know its state and will choose a satellite for handover, which is a really strong requirement for ordinary UE. Besides RL, a recursive neural network (RNN) can also be used for handover optimization. Refs. [9] and [10] apply RNN for handover optimization in terrestrial millimeter wave mobile systems and vehicular networks, respectively. However, in a LEO scenario, the beam switch is fast, and the signal series of one beam may be too short for the RNN to make decisions.

In practical terms, a handover strategy with a low requirement for UE capability is desired to reduce the handover number while ensuring the reference signal received power (RSRP). In this paper, a convolutional neural network (CNN) based handover strategy optimization is proposed. Firstly, an amount of UE is randomly generated within the coverage of a satellite. The RSRP series of UE is generated based on the channel model in Ref. [1] and the simulation assumption in Ref. [11]. Secondly, the graph-based method in Ref. [5] is improved by setting each satellite in different time slots as different nodes. The improved method is used to find the best handover strategies for each piece of UE. Thirdly, the internal relation between the historical RSRP series and the best handover decision is extracted by a customized CNN. Since standard 5G UE needs to periodically measure the RSRP of the serving cell and adjacent cells, the UE can perform a sub-optimal handover strategy based on the historical measurements. The main contributions of this paper are summarized as follows.

- This paper proposes a novel directed graph model for the handover process. In this model, each beam in different time slots is viewed as different nodes, and the weight of an edge is determined by the RSRP and the beam identities of the two corresponding nodes. Suppose the beam coverage and the RSRP of UE are predictable, the best handover strategy for the UE can be found based on the model.

- A CNN is constructed based on the classical LeNet-5^[12] for handover optimization. The results of the directed graph model are used to train the parameters of the CNN. Using the trained CNN, any UE in the LEO network can perform sub-optimal handover based on its historical RSRP.

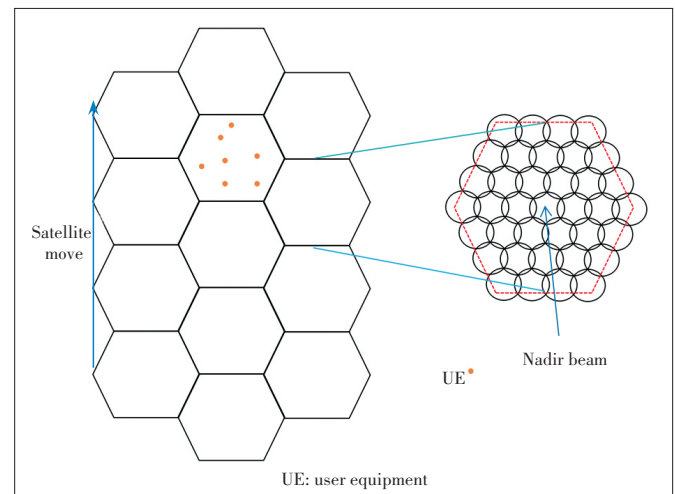
The rest of this paper is organized as follows. Section 2 describes the LEO network model and the motivation of handover optimization. In Section 3, a novel directed graph-based model is proposed for the handover process. A CNN structure is constructed and the results of the directed graph model are used to train the CNN. The effectiveness of the CNN is numerically evaluated in Section 4. Finally, Section 5 concludes this paper.

2 Background

2.1 System Model

A typical LEO satellite network consists of several circular orbits, and each orbit contains several evenly spaced satellites. This paper considers the scenario in Fig. 1 where each hexagon denotes the coverage of a satellite. Referred to the assumptions^[11-13] used in 3GPP NTN study item, each satellite is assumed to have 37 beams that form the hexagon coverage. The UE is assumed to locate within a hexagon in the initial time, and the satellites in the three adjacent orbits are considered to evaluate the RSRPs on the UE. During the flight of the satellites, a piece of UE needs to periodically measure the RSRPs of different beams and make handover decisions.

The beam layout in Fig. 1 decides the center of the 37 beams^[13]. Suppose the satellite is above a plane, then the diameter of the nadir beam on the plane can be computed based on the 3 dB angle. It is easy to compute the other 36 beam centers on the plane according to Fig. 1. Then the bore-sight directions of the 37 beams can be determined. The angles between co-orbital satellites and adjacent orbits are also calcu-



▲ Figure 1. Illustration of system model

lated to fulfill the coverage shown in Fig. 1.

2.2 Motivation

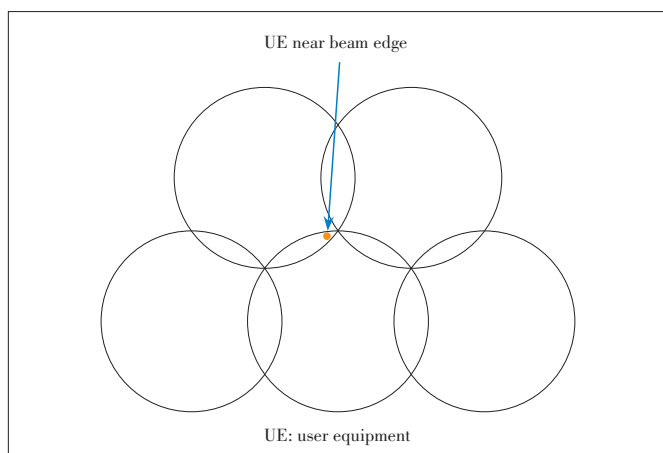
In 3GPP simulation assumption Set-1^[11], a satellite with an altitude of 600 km has a beam diameter of 50 km and a velocity of 7.56 km/s. Therefore, a piece of UE can only connect to one beam in 6.6 s at most. Because of the noise and the overlapping of different beams, the handover will happen more often. In addition, because of the long propagation time, each handover procedure needs a longer time and will consume more time-frequency resources. Therefore in a LEO network, the handover has a time lag and causes a large signaling overhead. To reduce the overhead and improve service continuity, the handover strategy needs to be optimized for the following targets.

- Predict the handover decision to compensate the time lag.
- Reduce the handover caused by noises, including shadow fading, multipath fading, and white Gaussian noise.
- Identify and suppress the handover in this situation. As shown in Fig. 2, a piece of UE near the beam edge may have a short serving time for some beams.

In this paper, an overall handover optimization is obtained in the directed graph model for each piece of UE. The common features of the optimized strategies for different UE are extracted using CNN to fulfill the targets without strong requirements for UE capability.

3 Handover Strategy Optimization Based on CNN

In a LEO satellite network, the satellites fly along predetermined circular orbits, so the change of the RSRP has strong regularity. The regularity can be used to improve the handover decision. Specifically, in each time slot, the previous N RSRP values of UE form a series, and some kinds of RSRP series imply that the UE should start handover. In this section, those kinds of RSRP series are found in two steps. First, the RSRP



▲ Figure 2. An example of UE near beam edge

of each UE during a long period is measured and recorded. A directed graph model is proposed to search for the best handover decision in every time slot. Then the handover decision is regarded as the label for the previous N RSRP values of that time slot to be trained by the CNN. The proposed CNN can efficiently extract the common regularity of handover decisions for different UE.

3.1 A Novel Directed Graph Based Model

The directed graph based model in Ref. [5] is designed to search the optimal handover strategy. However, the UE needs to start handover as soon as entering the coverage of the next satellite, which means it cannot choose a more appropriate handover time. This section proposes an improved directed graph based model to solve the problem.

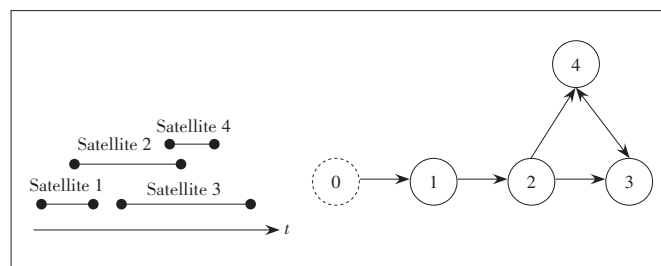
3.1.1 Referenced Model in Ref. [5]

In Ref. [5], every satellite is modeled as a node. If the beginning or end of the coverage of one satellite is between another satellite's coverage period, then there exists a directed edge between the two satellites, which means that a piece of UE can perform a handover between the two satellites. The weight of the edge is determined by the chosen criteria in the two satellites. For example, suppose only the criterion "handover number" is considered, then the weight of every edge should be set to 1. If other criteria such as "number of free channel" and "elevation angle" are considered, the weight can be set according to the two criteria of the target satellite. The Dijkstra's shortest path algorithm can be used to search the path with the smallest or largest weight. By choosing appropriate criteria, the resulting path becomes the overall optimal handover strategy.

An example of satellite coverage time and the corresponding directed graph in the referenced model are shown in Fig. 3. Node 0 denotes the initial time and other nodes denote four satellites. In this model, the node and the edge weight are invariable during the handover process. The weights of the edges cannot reflect the change process of the elevation angles or other criteria of the satellites. The UE can only assume that the handover happens as soon as it enters the coverage of another satellite.

3.1.2 Proposed Model

By considering the variation of satellites during the handover process, a novel model can be constructed to generate



▲ Figure 3. Directed graph in referenced model

more reasonable optimization for handover. The basic idea is to regard a beam in different time slots as different nodes. As shown in Fig. 4, we assume that in each time slot the serving beam of UE is one of the K strongest beams. The beam_{TK} denotes the K -th strongest beam in the T -th time slot. Every two nodes in adjacent time slots are connected by an edge, which means that the handover between them is possible.

Similar to Ref. [5], the weights of the edges can be set according to different criteria for overall optimization. For the sake of simplicity, this paper only considers the RSRP strength and handover number. Then the weight of the edge from beam_{T1K1} to beam_{T2K2} can be defined as

$$w_1 \times \text{RSRP}_{T1K1} - w_2 \times \text{handoverFlag}, \quad (1)$$

where RSRP_{T1K1} denote the RSRP value of beam_{T1K1} , and $\text{handoverFlag} = 1$ if beam_{T1K1} and beam_{T2K2} are two different beams. When the signal-to-noise ratio is small, the channel capacity is proportional to the signal strength.

Therefore $w_1 \times \text{RSRP}_{T1K1}$ in Eq. (1) denotes the benefit of connecting beam_{T1K1} in T_1 -th time slot, where w_1 is a predetermined parameter. Similarly, the parameter w_2 is chosen according to the degree of the negative impact of one handover. By using the Dijkstra's shortest path algorithm, we can find the longest path from the first time slot to the last time slot, which is actually the optimal handover strategy for this UE.

3.2 CNN Based Optimization for Handover

RSRP is defined as the linear average over the power of the resource elements that carry some predefined reference signals. Assume UE can predict the RSRP of different beams for a long period, then the method in Section 3.1.2 can be used to find the optimal handover strategy. However, in most cases, UE only knows its historical RSRP. Standard 5G UE needs to measure the RSRPs of detectable cells and handover to the strongest cell if its RSRP minus a predetermined threshold is larger than the serving RSRP. In this way, the information hid-

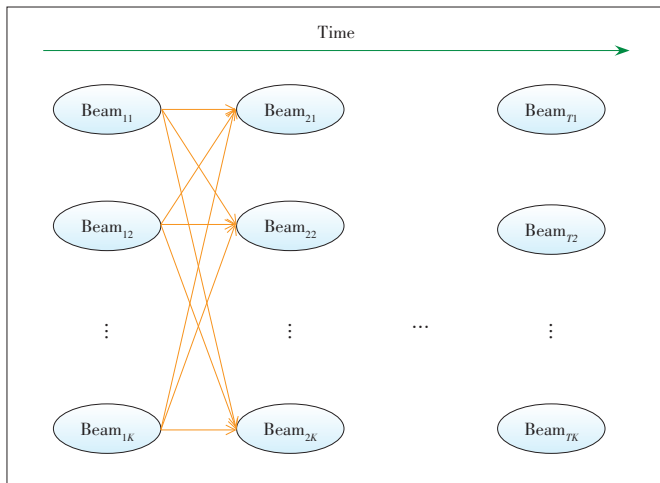
den in the historical RSRP is ignored. Actually, at least in the LEO scenario, the historical RSRP can help UE to make sub-optimal handover decisions. The series of historical RSRPs of the strongest K beams in each time slot forms a two-dimensional matrix. A customized CNN is used to optimize the handover decision based on the matrix in this section.

CNN is an effective tool to elicit information from two-dimensional data. It has been widely used to extract features from images. A classical CNN consists of one or more convolutional layers, pooling layers, and fully-connected layers. The features of the input data are extracted layer by layer, and are summarized in the last fully-connected layer to generate the final output. Compared with the fully-connected layer, the convolutional layer takes advantage of the strong local spatial correlation in natural images and only has a few parameters to be trained. It is worth mentioning that the matrix of RSRP also has the "local spatial correlation", i.e., the cooperation of the RSRP values in adjacent time slots and the RSRP values of the nearest 3 or 4 beams are more likely to contain information for handover decisions. Therefore it is suitable to apply CNN to the problem of handover.

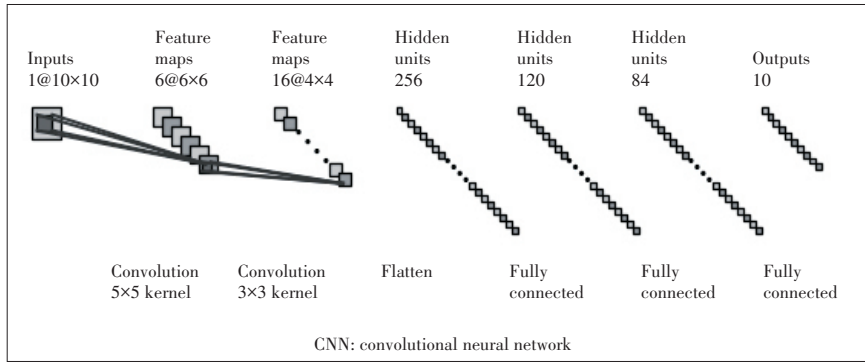
Intuitively, the RSRP series in the LEO network has strong regularity, so a relatively simple neural network structure should be chosen to reduce the training time and prevent overfitting. The LeNet-5^[12] is firstly designed for character recognition and is a relatively simple modern CNN structure. The default input of LeNet-5 is a matrix of the size of 32×32 . However, in the LEO network model presented in Section 2.1, the number of detectable beams for one piece of UE is generally smaller than 32. Therefore the size of the input data needs to be reduced. Actually, in the simulation, the number of considered beams in each time slot is set to be 10. The length of a time slot is set to be 0.5 s and the RSRP values in the previous 10-time slots are used to form the input. Then the input of the CNN is a matrix of the size of 10×10 . In LeNet-5, two convolutional layers are used. The two convolution kernels both have the size of 5×5 . Besides, two pooling layers are used to reduce the number of trained parameters. Because of the reduced input size, some layers in LeNet-5 need to be customized. First, one convolution kernel is reduced to the size of 3×3 . Then the pooling layers are deleted since the number of parameters is not large. The structure of the resulting CNN is presented in Fig. 5. The output of size 10 corresponds to the 10 kinds of handover decisions, i.e., one of the 10 strongest beams is which the UE will connect in the next time slot.

The data preprocessing and the training procedure consist of four steps as follows.

- 1) For every piece of UE, generate the RSRP values of different satellites in every time slot. If one satellite is invisible or its signal is too weak to detect, the RSRP values are regarded as 0.
- 2) Compute the best handover strategy for every UE based on the proposed directed graph-based method in Section 3.1.2.
- 3) For every UE in every time slot, the previous 10 RSRP



▲ Figure 4. Directed graph in the proposed model



▲ Figure 5. CNN structure for handover optimization

values of the 10 strongest beams are used to form 10×10 input data. The best handover decision generated in the previous step is regarded as the corresponding label.

4) The input data and the corresponding labels from different UE are used to train the CNN in Fig. 5. After some epochs, the testing accuracy will converge.

The trained CNN can be used to make suboptimal handover decisions for new UE. In each time slot, the UE extracts the historical RSRP values of the 10 strongest beams as the input of the trained CNN. The output contains 10 values and the index of the largest value is regarded as the serving beam in the next time slot. It is worth noting that the handover decision is actually a prediction for the next time slot, so the time lag in the handover procedure can be compensated.

4 Simulation

The proposed methods are numerically evaluated in this section. The simulation parameters are mainly referred to as the parameter Set-1 in Ref. [11]. Some important parameters are shown in Table 1.

As described in Section 2.1, the LEO network in simulation consists of three orbits. Each satellite has 37 beams which form a hexagon. Some points are randomly generated within one hexagon in the UV plane. The projection of the points on the earth is calculated as the positions of the UE.

4.1 Optimal Handover Strategy Based on Directed Graph Model

With the constructed LEO network, the RSRP values for each UE are calculated in each time slot. The length of a one-

▼ Table 1. Simulation parameters

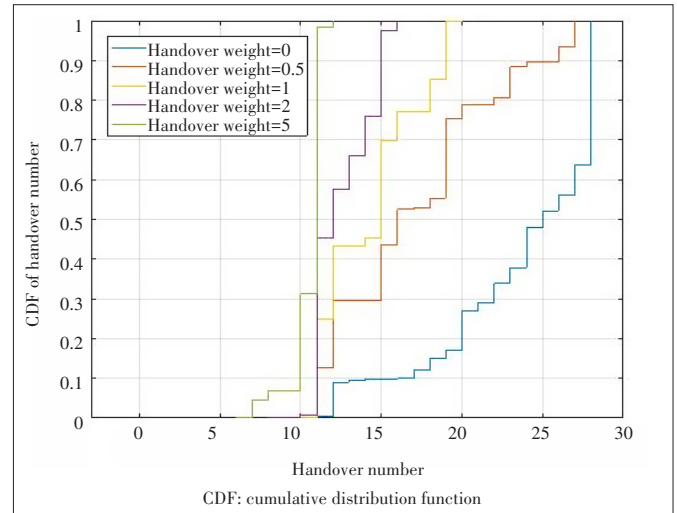
Parameter	Value
Orbit altitude	600 km
Simulation scenario	Rural
Carrier frequency	2 GHz
Antenna type	Bessel antenna
Antenna aperture	2 m
EIRP	34 dBW/MHz

EIRP: effective isotropic radiated power

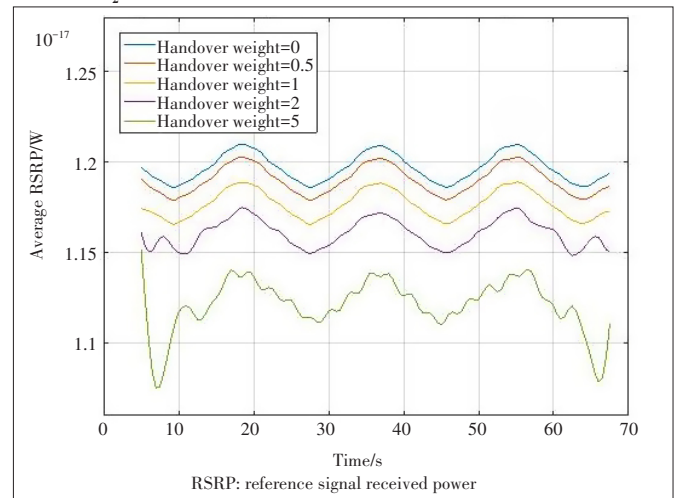
time slot is set to be 0.5 s, and about 140-time slots are considered in the whole simulation. The optimal handover strategy for each UE is generated by using the directed graph-based model in Section 3.1.2.

In the graph-based model, the two parameters w_1 and w_2 form a trade-off between RSRP strength and handover number and need to be predetermined. In this section, w_1 is fixed and different w_2 is evaluated to show the change of handover number and average RSRP strength. Because of the large path loss, the received power of the

strongest beam in one resource element is near 10^{-17} W. Therefore, the w_1 in Eq. (1) is set to be 10^{17} , which means that the benefit of accessing the strongest beam in one-time slot is around 1. Meanwhile, the value of w_2 is set to be 0, 0.5, 1, 2, and 5. When $w_2=0$, the UE will always connect to the strongest beam. As shown in Figs. 6 and 7, with the increase of w_2 ,



▲ Figure 6. Cumulative distribution function of handover number for different w_2



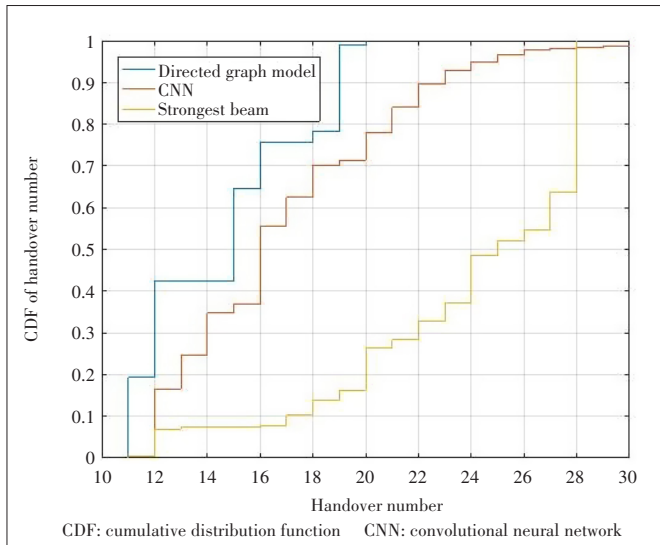
▲ Figure 7. Average RSRP during simulation for different w_2

the handover number and average RSRP will both decrease.

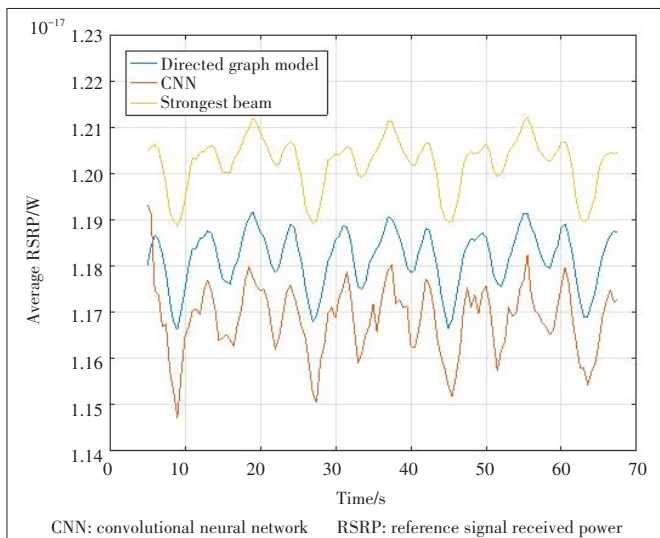
4.2 Performance of CNN in Handover Optimization

Three methods for handover optimization are compared in this section. The first method assumes the UE can predict its RSRP and make handover decisions based on the directed graph model. The second method means that the UE uses the trained CNN to make handover decisions. The CNN is trained by the results of the directed graph model with $w_2 = 1$. In the third method, the UE is always served by the strongest beam.

Compared with the “strongest beam” method, the CNN can largely reduce the number of handovers without a requirement for UE capability. Figs. 8 and 9 show that the handover number of more than 70% of the UE is reduced by more than 1/4, while the average RSRP is only reduced by 3%.



▲ Figure 8. Cumulative distribution function of handover number for different handover optimization methods



▲ Figure 9. Average RSRP during simulation for different handover optimization methods

5 Conclusions

This paper proposes a CNN-based handover optimization method for the LEO satellite network. The CNN structure is customized based on LeNet-5 and is used to extract the hidden information in the historical RSRP. In order to produce the training data for CNN, a novel directed graph-based model is proposed to find the optimal handover strategy when the UE knows its future RSRP. After the training, the CNN can be used to find a suboptimal handover decision based on its historical RSRP. In the simulation, the CNN is verified to be effective in handover optimization. The number of handovers is significantly reduced while the average RSRP is only reduced by 3%.

The optimization of handover in satellite communication is relatively simple because of the strong regulation of the movement of satellites. But the deep learning-based method can also be used in more complex scenarios. In order to extract the hidden regulation, a more advanced neural network structure may be needed, such as the attention-based neural network. The deep Q-learning is also worth investigating for a dynamically changing environment.

References

- [1] 3GPP. Study on new radio (NR) to support non terrestrial networks: 3GPP TR 38.811 [S]. 2018
- [2] GKIZELI M, TAFAZOLLI R, EVANS B. Modeling handover in mobile satellite diversity based systems [C]//The 54th Vehicular Technology Conference. Atlantic City, USA: IEEE, 2001: 131 – 135. DOI: 10.1109/VTC.2001.956570
- [3] DEL RE E, FANTACCI R, GIAMBENE G. Handover queuing strategies with dynamic and fixed channel allocation techniques in low Earth orbit mobile satellite systems [J]. IEEE transactions on communications, 1999, 47(1): 89 – 102. DOI: 10.1109/26.747816
- [4] DEL RE E, FANTACCI R, GIAMBENE G. Efficient dynamic channel allocation techniques with handover queuing for mobile satellite networks [J]. IEEE journal on selected areas in communications, 1995, 13(2): 397 – 405. DOI: 10.1109/49.345884
- [5] WU Z F, JIN F L, LUO J X, et al. A graph-based satellite handover framework for LEO satellite communication networks [J]. IEEE communications letters, 2016, 20(8): 1547 – 1550. DOI: 10.1109/LCOMM.2016.2569099
- [6] YAJNANARAYANA V, RYDÉN H, HÉVIZI L. 5G handover using reinforcement learning [C]//IEEE 3rd 5G World Forum (5GWF). Bangalore, India: IEEE, 2020: 349 – 354. DOI: 10.1109/5GWF49715.2020.9221072
- [7] CHEN Y, LIN X Q, KHAN T, et al. Efficient drone mobility support using reinforcement learning [C]//IEEE Wireless Communications and Networking Conference (WCNC). Seoul, South Korea: IEEE, 2020: 1 – 6. DOI: 10.1109/WCNC45663.2020.9120595
- [8] CHEN M T, ZHANG Y, TENG Y L, et al. Reinforcement learning based signal quality aware handover scheme for LEO satellite communication networks [M]//Human Centered Computing. Cham: Springer International Publishing, 2019: 44 – 55. DOI: 10.1007/978-3-030-37429-7_5
- [9] ALKHATEEB A, BELTAGY I, ALEX S. Machine learning for reliable mmwave systems: Blockage prediction and proactive handoff [C]//IEEE Global Conference on Signal and Information Processing (GlobalSIP). Anaheim, USA: IEEE, 2018: 1055 – 1059. DOI: 10.1109/GlobalSIP.2018.8646438
- [10] ALJERI N, BOUKERCHE A. An efficient handover trigger scheme for vehicular networks using recurrent neural networks [C]//The 15th ACM International

Symposium on QoS and Security for Wireless and Mobile Networks. New York, USA: ACM. 2019: 85 – 91. DOI: 10.1145/3345837.3355963

- [11] 3GPP. Solutions for NR to support non-terrestrial networks (NTN): 3GPP TR 38.821 [S]. 2019
- [12] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278 – 2324. DOI: 10.1109/5.726791
- [13] 3GPP. On beam layout definition for NTN system level simulations [EB/OL]. [2021-04-16]. <https://www.3gpp.org/DynaReport/TDocExMtg--R1-97--32823.htm>

Biographies

ZHANG Chenchen (zhang.cc@zte.com.cn) received his B.S. degree in mathematics from Nankai University, China in 2013, and the Ph.D. degree in computer science and technology from Shanghai Jiao Tong University, China in 2018. He has been with ZTE Corporation since 2018. He is now a senior pre-research engineer in non-terrestrial network. His main research interests include satellite communications, random access, mobility management, neural network and NOMA.

ZHANG Nan received his bachelor's degree in communication engineering

and master's degree in integrated circuit engineering from Tongji University, China in July 2012 and March 2015, respectively. He is now a senior engineer at the Department of Algorithms, ZTE Corporation and works on the standardization of LTE and NR system. His current research interests are in the field of 5G channel modeling, MIMO, NOMA techniques, satellite/ATG communication and network architecture.

CAO Wei is a senior pre-research expert in ZTE Corporation. She received her Ph.D. degree in wireless communication from National University of Singapore in 2008. Her current research interests include non-terrestrial communication network and reconfigurable intelligent surface.

TIAN Kaibo received his master's degree from Xi'an Jiaotong University, China in 2008. Now he is the senior pre-research expert of ZTE Corporation and responsible for the pre-research of the Air-Space-Ground integrated network technology.

YANG Zhen received his B.S. degree in communication and information system from University of Electronic Science and Technology of China in 2012. Since 2012 he has been with ZTE Corporation. He is now a senior pre-research engineer in wireless communications. His main research interests include satellite communications, random access, mobility management, neural network and DPD.



Truly Grant-Free Technologies and Protocols for 6G

Abstract: The further integration of telecommunications and industry has been considerable and is expected to bring significant benefits to society and economics in 6G. It also forms some evolution trends for next-generation communication systems, including further rises in machine-type communications (MTC), uplink-dominated systems, and decentralized structures. However, the existing access protocols are not friendly to these trends. This paper analyzes the problems of existing access protocols and provides novel access technologies to solve them. These technologies include contention-based non-orthogonal multiple access (NOMA), data features, enhanced pilot design and successive interference cancellation (SIC) of diversity. With these key enablers, truly grant-free access can be realized, and some potential modifications of protocols are then analyzed. Finally, this paper uses massive and critical scenarios in digital transformations to show the great necessity of introducing novel access technologies into future communication protocols.

Keywords: decentralization; digital transformations; future access protocols; MTC; uplink

MA Yihua^{1,2}, YUAN Zhifeng^{1,2},
LI Weimin^{1,2}, LI Zhigang^{1,2}

(1. State Key Laboratory of Mobile Network and
Mobile Multimedia Technology, Shenzhen 518055,
China;

2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202104012

<https://kns.cnki.net/kcms/detail/34.1294>.

TN.20211022.1500.004.html, published online

October 22, 2021

Manuscript received: 2021-04-16

Citation (IEEE Format): Y. H. Ma, Z. F. Yuan, W. M. Li, et al., "Truly grant-free technologies and protocols for 6G," *ZTE Communications*, vol. 19, no. 4, pp. 105 – 110, Dec. 2021. doi: 10.12142/ZTECOM.202104012.

1 Introduction

The targets of communications vary from connecting humans to connecting everything, and digital transformations are expected to penetrate many scenarios in 6G^[1-3]. Digital transformations are expected to bring significant benefits to society. Meanwhile, digital transformations raise many new requirements, as well as set the course for future communication protocols.

To support seamless and handy digital services, there are some trends in future communications, including further evolution from human-oriented communications to machine-type communications (MTC), from the downlink-dominated to uplink-dominated, and from centralization to decentralization. This paper focuses on the truly grant-free, or contention-based grant-free, technologies and protocols, which are crucial to fulfilling ultra-low latency and ubiquitous connectivity for future communications. In current 5G standards, the grant-free definition is not truly yet as it requires extra pre-configuration of grants.

Novel access technologies are required to replace the existing ones for realizing these trends, and this paper introduces

four key enablers: the contention-based non-orthogonal multiple access (NOMA), the prior knowledge of data, the enhanced pilot design, and the joint use of diversity and successive interference cancellation (SIC). NOMA was first proposed in Ref. [4], and it usually requires accurate power control and resource allocation. As a comparison, the contention-based NOMA has no such requirements and achieves a higher flexibility. The prior knowledge of data can be used to improve transmission performance. Data-only transmission was proposed in Ref. [5], which is able to remove pilot overheads. Apart from the data-only scheme, a pilot is still crucial to the full use of a large receiving antenna array where the spatial combining search space is huge. Therefore, the pilot should be well designed to gain a good detection and estimation performance^[6-7]. Moreover, the joint use of diversity and SIC^[8] can deal with the fluctuation of loading as it is random in contention-based transmissions. This strategy is able to average the loading in different slots via iterative detection and cancellation.

With these key enablers, users are able to transmit packets without building radio resource control (RRC) connections. That is to say, a connection-free transmission can be achieved,

and the transmission no longer relies on RRC connections. This great enhancement simplifies the transmission procedures a lot, as well as reduces the overheads and latency for building RRC connections. It also leads to some possible modifications of protocols in the future, which are also analyzed in this paper.

The motivation of this paper is to discuss the importance of truly grant-free technology for future communications, summarize key enablers for it, and suggest some protocol evolution considerations. The rest of this paper is organized as follows. Section 2 introduces the main trends of future radio access networks and analyzes the problems of existing access protocols. In Section 3, some novel access technologies are provided to solve the problems. Section 4 briefly introduces the impact of these technologies on future protocols. In Section 5, two important scenarios are discussed to show the advantages of the proposed schemes. Section 6 concludes this paper and provides some future research directions.

2 Three Main Trends

In this section, three main trends of future networks are introduced. The problems of existing access protocols are analyzed under these trends.

2.1 Further Rises in MTC

The first trend is that the main participants of communications vary from humans to machines. Although massive machine-type communication (mMTC) is included in 5G, more massive and critical MTC will develop and be in demand towards 2030 and beyond^[1]. Unlike human communications, the potential users can be massive and the reliability requirement can be very high. However, the classical random access protocol requires a handshaking procedure, which is not suitable for the high-efficiency transmissions of massive potential users or low-latency transmissions of high-mobility networks.

2.2 Uplink-Dominated System

The second trend is that the overall performance is becoming dominated by uplink instead of downlink transmissions. For many MTC applications, uplink is the main bottleneck^[9]. Moreover, in massive multiple input multiple output (MIMO) systems, time division duplex (TDD) is much easier to realize; it uses uplink-downlink reciprocity to obtain downlink channel information. Direct downlink channel estimation is very inefficient as the overheads of downlink pilots increase with the antenna number of the base station (BS)^[10]. Therefore, the pilot in the uplink becomes very important for obtaining the channel information and effectively using the capability of massive antennas of the BS. In the current protocol, the uplink pilot or demodulation reference signal (DMRS) is orthogonal among users. It limits the number of pilots and is not suitable for contention-based transmission.

2.3 Decentralized Structure

The last trend is from centralization to decentralization. This trend has many aspects, including the distributed antenna or cell-free design for ubiquitous connectivity, device-to-device transmission not relying on the central controller, and decentralized information management and control, e.g. the blockchain^[11]. Although network centralization has brought us many good aspects like easy management and global control, the costs should not be neglected, including coverage, latency and privacy risk. These costs greatly limit the performance and credibility of networks.

3 Novel Access Technologies

This section analyzes four novel access technologies as shown in Fig. 1. The basic idea and advantages of these technologies are discussed.

3.1 Contention-Based NOMA

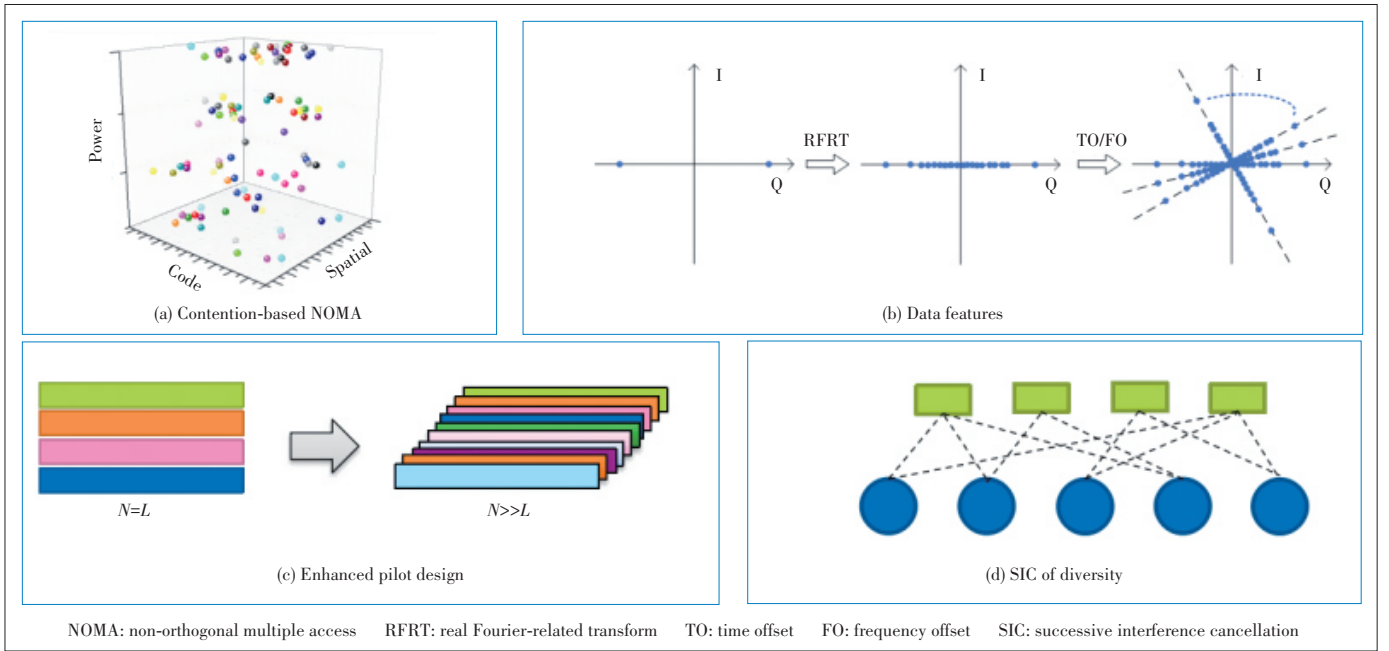
NOMA is a very effective technology to increase spectrum efficiency when there is a near-far effect. Due to the complexity limitation, a non-orthogonal power domain has not been fully utilized especially for the uplink in current protocols. NOMA still acts as an important role in the future, as the complexity limitation is expected to be solved by advanced algorithms and powerful hardware in the future.

It is complex and inefficient to implement accurate power control and resource allocation when there are massive users or the latency requirement is high. Therefore, a scheme allowing users to transmit freely is in high demand. To acquire this convenience for end devices, the transmission itself is inevitable to be non-orthogonal. In this case, extra sensing and local power control can be required to utilize the power domain^[12], but it is not suitable for low-cost and low-power devices. To support a more flexible transmission not relying on any sensing and power control, a joint use of power domain, code domain and spatial domain should be considered as in Fig. 1(a).

3.2 Data Features

To get rid of pilots or reduce the pilot overheads, data features should be used. There are two mainstream ways to realize this. One is to utilize the prior knowledge and statistical information of data^[5,13], e.g. the constellation shape, correlation matrix, constant modulus, etc. This method is compatible with existing protocols, and the modification to existing standards is relatively small. The other is a data-driven method that uses deep learning (DL). The end-to-end auto-encoder is one important application of DL at the physical (PHY) layer, and Ref. [14] shows that pilot-free transmission can also be realized by an auto-encoder.

During the exploitation of data features, novel waveform potentially arises^[15-16]. Discrete Fourier transform spreading orthogonal frequency division multiplexing (DFT-s-OFDM) plays an important role in 5G for its low peak to average power



▲ Figure 1. Key enablers of future access protocols

ratio (PAPR). However, it makes the data feature hard to use. One solution is real Fourier related transform spreading OFDM (RFRT-s-OFDM)^[15]. As shown in Fig. 1(b), this novel waveform can maintain the data feature. Channel equalization and time/frequency offset correction can be conducted using the data features in RFRT-s-OFDM as shown in Fig. 1(b), while the low PAPR advantage is kept.

3.3 Enhanced Pilot Design

To support mMTC and massive MIMO, novel pilot designs are proposed to support more users and reduce allocation overheads as shown in Fig. 1(c). One mainstream research area is non-orthogonal pilots based on compressed sensing (CS)^[6]. The sparsity of user activities, multiple paths and angles of arrival can be used to increase the performance of user detection and channel estimation. However, a CS-based non-orthogonal pilot scheme leads to large computational complexity when the pilot pool is large or the receiving antenna array is large. Also, the inter-cell interference and time/frequency offset problems are hard to solve.

The other research direction is the special non-orthogonal pilot design with partial orthogonality. To be specific, multi-pilot^[7] is proposed which consists of multiple pilots. Multiple orthogonal pilots are usually employed, although multiple non-orthogonal pilots also work. The detection of every orthogonal pilot is very simple, and the detection complexity is reduced a lot compared with general non-orthogonal pilot design. Moreover, an orthogonal pilot has been used and verified during a long period, and many existing engineering methods can be used to ensure the performance of multi-pilot use.

3.4 SIC of Diversity

The joint use of diversity and SIC was first proposed in Ref. [17], and then some novel schemes have improved the performance by optimizing a bipartite graph of Fig. 1(d). The joint use allows users to transmit multiple replica packets at any time slot, and SIC of packets is used after demodulating any packet in each round. This strategy was designed for satellite communications where the round-trip time is very long. The achievable loading of this strategy is approaching 1 with the cost of replicas. This method greatly reduces the transmission delay and increases transmission efficiency. It is also named modern random access, which is seen as one potential next-generation random access protocol^[10].

Unlike satellite communications, the channel gain of different users varies a lot in widely-used terrestrial communications. Therefore, the near-far effect becomes a practical factor that requires consideration. NOMA is able to utilize the near-far effect to separate different users. The combination of NOMA and multiuser diversity with SIC requires a joint design of the medium access control (MAC) and PHY layers. An example was shown in Ref. [18] which jointly uses the code domain NOMA and diversity with SIC.

4 Impacts on Protocols

With the novel access technologies, the transmission procedure can be simplified a lot, which deeply affects the future protocols.

4.1 RRC Idle/Inactive

With the key enablers in Section 3, connection-free trans-

mission becomes possible which enables high-efficiency and instant MTC without any connection establishment. That is to say, the transmission can be realized in an idle or inactive RRC state. These enablers can also evolve around the random access protocols with a much higher successful rate and lower latency especially in dense environment, i.e., they are beneficial for RRC connection establishment. A related standardization work is the two-step random access channel (RACH)^[19]. However, it only acts as a substitution of the four-step RACH, and the collision problem has not been studied as orthogonal multiple access and orthogonal pilots are still in use. In regard to collisions, the protocols relating to multiple access and pilot sequences are expected to evolve.

4.2 Uplink and Sidelink NOMA

As mentioned before, the non-orthogonal domain has not been well utilized in 5G. Actually, the discussion of uplink NOMA has not reached a consensus in 5G standards^[20]. One reason is that the performance and complexity comparison is not enough to decide a winner among different uplink NOMA schemes. The demands of mMTC should be further analyzed, and some crucial key performance indicators should be especially emphasized. The contention-based grant-free feature^[21] is a potential scheme in the future as it is exceedingly friendly to low-cost and low-power MTC devices. Moreover, contention-based NOMA is also a potential standardization direction for sidelink (SL) without central control, e.g., LTE-V SL mode 4^[22].

4.3 Comprehensive Synchronization

The connection-free transmission brings some synchronization (SYNC) challenges. In the current protocols, synchronization is realized by the SYNC signal and measured time advance (TA) from BS^[23]. TA is hard to be obtained in connection-free transmissions, so a comprehensive SYNC is required. Some possible modifications can be based on the information of the UE status (e.g., position and speed), BS position, SYNC signals of multiple BSes, etc. Moreover, an overall framework is required to jointly use all the prior knowledge relating to SYNC which can be obtained by the end device.

5 Case Studies

Two representative cases of both massive and critical MTC are shown in this section. The advantages of novel access technologies show that they may play important roles in future standards.

5.1 Case Study of mMTC

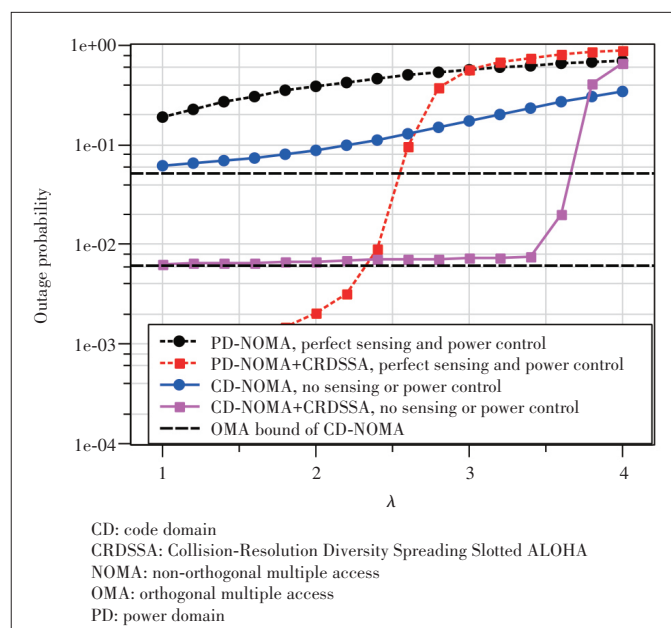
This case study is for mMTC, and a Poisson arrival model is used with an average arrival rate of λ . NOMA provides a large gain when there is a near-far effect, e.g., a near and strong user can reuse the channel used by a far and weak user, which greatly improves the spectrum efficiency. There-

fore, the near-far effect is included in comparison. As shown in Fig. 2, the outage performances of two different contention-based grant-free NOMA schemes^[12,18] are compared. As mentioned before, power domain NOMA with a contention-based feature requires extra sensing of channel gain and power control according to estimated channel gain. The aim is to make the receive power belong to some predetermined power levels. It results in an extremely high transmission power, and Ref. [12] provides sub-channel selection and distance-based methods to solve it. As a comparison, code domain NOMA^[18] is just required to randomly select the spread code and can work without power control.

The simulation results show that code domain NOMA performs better than power domain. For a target outage rate of 0.1, the achievable arrival rate of power domain NOMA is less than 1 while that of code domain NOMA is greater than 2. Also, when combined with the SIC of the diversity strategy, the achievable arrival rates at the outage of 10^{-2} are 2.4 and 3.4 for the power domain and code domain NOMA. In this comparison, power domain NOMA plus SIC of diversity achieves a lower error floor because perfect sensing and power control are assumed. The outage performance of code domain NOMA is bounded by a corresponding orthogonal multiple access (OMA) transmission of perfect scheduling. As the near-far effect is utilized, code domain NOMA plus SIC of diversity achieves an arrival rate greater than 3 with the outage performance very close to the OMA bound.

5.2 Case Study of V2V

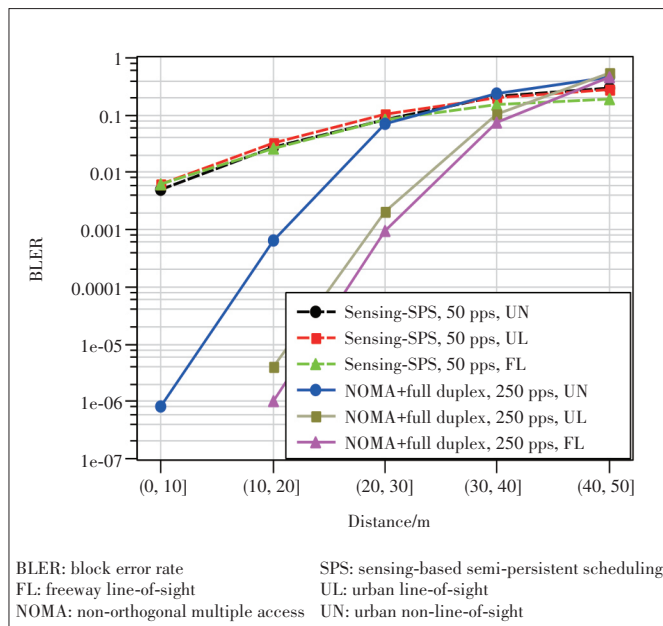
This case study is for vehicle-to-vehicle (V2V) communications without central control. V2V communications should be



▲ Figure 2. Outage comparison of contention-based NOMA and their combination with SIC of diversity

extremely reliable for safety, and V2V without central control is especially important due to the robustness in a non-cellular domain and ultra-low latency requirement. In LTE-V, sensing-based semi-persistent scheduling (SPS)^[22] is employed, and every user sensing the spectrum resources randomly selects idle resources in a given period. A sensing-based method is also studied and a potential candidate in 5G V2V is chosen^[24]. The problem of a sensing-based method is that it is not friendly to multiple antennas, and the reliability is relatively low in ultra-dense environments. To solve these problems, Ref. [25] proposes a novel distributed antenna deployment and full-duplex contention-based NOMA transceiver scheme which jointly use power, code and spatial domains to achieve V2V communications with ultra-low latency and high reliability in ultra-dense scenarios. The block error rate (BLER) comparison is shown in Fig. 3. The channel model is based on LTE-V standards^[22], and 1 500 and 700 vehicles are dropped in the urban and freeway scenarios defined in Ref. [10], respectively.

In Fig. 3, three protocol-defined scenarios are all considered, including urban non-line-of-sight (UN), urban line-of-sight (UL) and freeway line-of-sight (FL). In this comparison, the full-duplex NOMA scheme achieves a much higher reliability than the method in the current protocol. Also, the overheads of the full-duplex NOMA scheme are reduced only 1/5, which means a much more frequent transmission of 250 packets per second (pps) can be supposed, which helps to reduce the end-to-end latency once the information is generated. In this case, NOMA turns the near-far effect into advantages, and achieves a high spectrum efficiency and ultra-high reliability of near vehicles.



▲ Figure 3. BLER comparison of sensing-based SPS in current protocols and full-duplex contention-based NOMA using novel access technologies

6 Conclusions

To fulfill digital transformations, it is necessary to make some great modifications in current protocols. One crucial aspect is to enhance the access protocols as it is a bottleneck of seamless and instant digital services. This paper provides several novel access technologies, and potential impacts on protocols are also briefly analyzed. Moreover, some typical use cases are shown to verify the necessity of these modifications of future protocols. The standardization process of novel access technologies to boost digital transformations requires a much wider application to make it more urgent, more careful considerations given to privacy and security, and a joint work of industry and academics to refine technologies.

References

- [1] MAHMOOD N H, BÖCKER S, MUNARI A, et al. White paper on critical and massive machine type communication towards 6G [EB/OL]. (2020-05-04) [2020-11-11]. <https://arxiv.org/abs/2004.14146>
- [2] ITU-T. FG-NET2030 - Focus group on technologies for network 2030. [EB/OL]. [2020-11-11]. <https://www.itu.int/en/ITU-T/focusgroups/net2030/Pages/default.aspx>
- [3] DAVID K, BERNDT H. 6G vision and requirements: is there any need for beyond 5G? [J]. IEEE vehicular technology magazine, 2018, 13(3): 72 - 80. DOI: 10.1109/MVT.2018.2848498
- [4] SAITO Y, KISHIYAMA Y, BENJEBBOUR A, et al. Non-orthogonal multiple access (NOMA) for cellular future radio access [C]//IEEE VTC Spring, Dresden, Germany: IEEE, 2013. DOI: 10.1109/VTCSpring.2013.6692652
- [5] YUAN Z F, LI W M, HU Y Z, et al. Blind multi-user detection based on receive beamforming for autonomous grant-free high-overloading multiple access [C]//IEEE 2nd 5G World Forum (5GWF). Dresden, Germany: IEEE, 2019: 520 - 523. DOI: 10.1109/5GWF.2019.8911643
- [6] KE M, GAO Z, WU Y, et al. Compressive sensing-based adaptive active user detection and channel estimation: massive access meets massive MIMO [J]. IEEE transactions on signal processing, 2020, 68: 764 - 779. DOI: 10.1109/TSP.2020.2967175
- [7] YUAN Z F, LI W M, LI Z G, et al. Contention-based grant-free transmission with independent multi-pilot scheme [EB/OL]. (2020-04-07) [2020-11-11]. <https://arxiv.org/abs/2004.03225>
- [8] CLAZZER F, MUNARI A, LIVA G, et al. From 5G to 6G: has the time for modern random access come? [C]//6G Wireless Summit. Levi, Finland: 6G Flagship, 2019
- [9] DAWY Z, SAAD W, GHOSH A, et al. Toward massive machine type cellular communications [J]. IEEE wireless communications, 2017, 24(1): 120 - 128. DOI: 10.1109/MWC.2016.1500284WC
- [10] ZHANG W C, XIANG J Y, LI Y-N R Y, et al. Field trial and future enhancements for TDD massive MIMO networks [C]//26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). Hong Kong, China: IEEE, 2015: 2339 - 2343. DOI: 10.1109/PIMRC.2015.7343689
- [11] MACDONALD M, LIU THORROLD L, JULIEN R. The blockchain: a comparison of platforms and their uses beyond bitcoin [EB/OL]. (2020-11-11). https://www.researchgate.net/publication/313249614_The_Blockchain_A_Comparison_of_Platforms_and_Their_Uses_Beyond_Bitcoin
- [12] CHOI J. NOMA-based random access with multichannel ALOHA [J]. IEEE journal on selected areas in communications, 2017, 35(12): 2736 - 2743. DOI: 10.1109/JSAC.2017.2766778
- [13] MA Y H, YUAN Z F, HU Y Z, et al. A data-assisted algorithm for truly grant-free transmissions of future mMTC [C]//IEEE Global Communications Conference. Taipei, China: IEEE, 2020. DOI: 10.1109/GLOBE-COM42002.2020.9348198

- [14] YE H, LI G Y, JUANG B-H F. Bilinear convolutional auto-encoder based pilot-free end-to-end communication systems [C]//IEEE International Conference on Communications (ICC). Dublin, Ireland: IEEE, 2020. DOI: 10.1109/ICC40277.2020.9149030
- [15] MA Y H, YUAN Z F, HU Y Z, et al. A real fourier-related transform spreading OFDM multi-user shared access system [C]//90th Vehicular Technology Conference (VTC2019-Fall). Honolulu, USA: IEEE, 2019: 1 - 5
- [16] YUAN Z, HU Y, MA Y, et al. Autonomous grant-free high overloading multiple access based on conjugated data symbols [C]//WS-20, IEEE International Conference on Communications. Dublin, Ireland: IEEE, 2020
- [17] CASINI E, GAUDENZI R D, HERRERO O D R. Contention resolution diversity slotted ALOHA (CRDSA): an enhanced random access scheme for satellite access packet networks [J]. IEEE transactions on wireless communications, 2007, 6(4): 1408 - 1419. DOI: 10.1109/TWC.2007.348337
- [18] MA Y H, YUAN Z F, LI W M, et al. NOMA based modern random access not relying on sensing or power control [J]. IEEE iInternet of Things journal, 2021, 8(20): 15382 - 15395. DOI: 10.1109/JIOT.2021.3073367
- [19] 3GPP. Consideration on 2-step RACH procedures: 3GPP R1-1903879 [S]. 2019
- [20] 3GPP. Study on non-orthogonal multiple access (NOMA) for NR: 3GPP TR 38.812 Release 16 [S]. 2018
- [21] YUAN Y F, YUAN Z F, TIAN Li. 5G non-orthogonal multiple access (NOMA) study in 3GPP [J]. IEEE communications magazine, 2020, 58(7): 90 - 96. DOI: 10.1109/MCOM.001.1900450
- [22] 3GPP. Vehicle to vehicle (V2V) services based on LTE sidelink; user equipment (UE) radio transmission and reception: 3GPP TS 36.785 Release 14 [S]. 2016
- [23] 3GPP. NR; Physical layer procedures for control: 3GPP TS 38.213 Release 16 [S]. 2020
- [24] 3GPP. Study on NR Vehicle-to-Everything (V2X): 3GPP TR 38.885 Release 16 [S]. 2020
- [25] YUAN Z F, MA Y H, HU Y Z, et al. High-efficiency full-duplex V2V communication [C]//2nd 6G Wireless Summit. Levi, Finland: IEEE, 2020. DOI: 10.1109/6GSUMMIT49458.2020.9083762

Biographies

MA Yihua (yihua.ma@zte.com.cn) received the B.Eng. degree from Southeast University, China in 2015 and the M.Sc. degree from Peking University, China in 2018. Since 2018 he has been with ZTE Corporation, China. He is now a senior research engineer in the Department of Wireless Algorithm, ZTE. His main research interests include mMTC, grant-free transmissions, NOMA and cell-free massive MIMO.

YUAN Zhifeng received the M.S. degree in signal and information processing from Nanjing University of Post and Telecommunications, China in 2005. From 2004 to 2006, he was mainly engaged in FPGA/SOC ASIC design. He has been a member of the wireless technology advance research department at ZTE Corporation since 2006 and has been responsible for the research of the new multiple access group since 2012. His research interests include wireless communications, MIMO system, information theory, multiple access, error control coding, adaptive algorithms and high-speed VLSI design.

LI Weimin received the M.S. degree in communication and information system from Nanjing University of Posts and Telecommunications, China in 2010, and is currently working as a technology pre-research senior engineer at ZTE Corporation. His research interests include power control, interference control, multiple access and grant-free transmission.

LI Zhigang received the B.S. degree from Jiangsu University, China in 2016 and the M.S. degree from Harbin Institute of Technology, China in 2019, and is currently working as a technology pre-research engineer at ZTE Corporation. His current research interests include wireless communication, multiple access and grant-free transmission.

ZTE Communications

Table of Contents, Volume 19, 2021

Volume-Number-Page

Special Topic

Energy Consumption Challenges and Prospects on B5G Communication Systems

Editorial	GE Xiaohu, YANG Yang	19-01-01
Saving Energy for Wireless Transmission: An Important Revelation from Shannon Formula		
.....	ZHU Jinkang, ZHAO Ming	19-01-02
Efficient Network Slicing with Dynamic Resource Allocation		
.....	JI Hong, ZHANG Tianxiang, ZHANG Kai, WANG Wanyuan, WU Weiwei	19-01-11
Enabling Energy Efficiency in 5G Network	LIU Zhuang, GAO Yin, LI Dapeng, CHEN Jiajun, HAN Jiren	19-01-20
Cluster Head Selection Algorithm for UAV Assisted Clustered IoT Network Utilizing Blockchain		
.....	LIN Xinhua, ZHANG Jing, LI Qiang	19-01-30
Green Air-Ground Integrated Heterogeneous Networks in 6G Era	WU Huici, LI Hanjie, TAO Xiaofeng	19-01-39
Kinetic Energy Harvesting Toward Battery-Free IoT: Fundamentals, Co-Design Necessity and Prospects		
.....	LIANG Junrui, LI Xin, YANG Hailiang	19-01-48

Edge Intelligence for Internet of Things

Editorial	LIU Chi	19-02-01
RecCac: Recommendation-Empowered Cooperative Edge Caching for Internet of Things		
.....	HAN Suning, LI Xiuhua, SUN Chuan, WANG Xiaofei, Victor C. M. LEUNG	19-02-02
Cost-Effective Task Scheduling for Collaborative Cross-Edge Analytics		
.....	ZHAO Kongyange, GAO Bin, ZHOU Zhi	19-02-11
BPPF: Bilateral Privacy-Preserving Framework for Mobile Crowdsensing		
.....	LIU Junyu, YANG Yongjian, WANG En	19-02-20
Maximum-Profit Advertising Strategy Using Crowdsensing Trajectory Data		
.....	LOU Kaihao, YANG Yongjian, YANG Funing, ZHANG Xingliang	19-02-29
Speed Estimation Using Commercial Wi-Fi Device in Smart Home		

.....	TIAN Zengshan, YE Chenglin, ZHANG Gongzhui, HE Wei, JIN Yue	19-02-44
Higher Speed Passive Optical Networks for Low Latency Services	ZHANG Weiliang, YUAN Liquan	19-02-61

Wireless Intelligence for Behavior Sensing and Recognition

Editorial	MA Jianhua, Guo Bin	19-03-01
HiddenTag: Enabling Person Identification Without Privacy Exposure		
.....	QIU Chen, DAI Tao, GUO Bin, YU Zhiwen, LIU Sicong	19-03-03
Device-Free In-Air Gesture Recognition Based on RFID Tag Array	WU Jiaying, WANG Chuyu, XIE Lei	19-03-13
Indoor Environment and Human Sensing Via Millimeter Wave Radio: A Review.....		
.....	LIU Haipeng, ZHANG Xingyue, ZHOU Anfu, LIU Liang, MA Huadong	19-03-22
Using UAV to Detect Truth for Clean Data Collection in Sensor-Cloud Systems		
.....	LI Xiuxian, LI Zhetao, OUYANG Yan, DUAN Haohua, XIANG Liyao	19-03-30
Artificial Intelligence Rehabilitation Evaluation and Training System for Degeneration of Joint Disease		
.....	LIU Weichen, SHEN Mengqi, ZHANG Anda, CHENG Yiting, ZHANG Wenqiang	19-03-46
A Survey of Intelligent Sensing Technologies in Autonomous Driving.....		
.....	SHAO Hong, XIE Daxiong, HUANG Yihua	19-03-56

OTFS Modulation for 6G and Future High Mobility Communications

Editorial	YUAN Jinhong, BAI Baoming, FAN Pingzhi, AI Bo	19-04-01
A Survey on Low Complexity Detectors for OTFS Systems.....		
.....	ZHANG Zhengquan, LIU Heng, WANG Qianli, FAN Pingzhi	19-04-03
Signal Detection and Channel Estimation in OTFS		
.....	Ashwitha NAIKOTI, Ananthanarayanan CHOCKALINGAM	19-04-16
Message Passing Based Detection for Orthogonal Time Frequency Space Modulation		
.....	YUAN Zhengdao, LIU Fei, GUO Qinghua, WANG Zhongyong	19-04-34
Performance of LDPC Coded OTFS Systems over High Mobility Channels		
.....	ZHANG Chong, XING Wang, YUAN Jinhong, ZHOU Yiqing	19-04-45
Coded Orthogonal Time Frequency Space Modulation		

.....	LIU Mengmeng, LI Shuangyang, ZHANG Chunqiong, WANG Boyu, BAI Baoming	19-04-54
OTFS Enabled NOMA for MMTTC Systems over LEO Satellite.....		
.....	MA Yiyan, MA Guoyu, WANG Ning, ZHONG Zhangdui, AI Bo	19-04-63
Orthogonal Time Frequency Space Modulation in Multiple-Antenna Systems		
.....	WANG Dong, WANG Fanggang, LI Xiran, YUAN Pu, JIANG Dajie	19-04-71

Review

Next Generation Semantic and Spatial Joint Perception—Neural Metric-Semantic Understanding	ZHU Fang	19-01-61
Analysis of Industrial Internet of Things and Digital Twins	TAN Jie, SHA Xiubin, DAI Bo, LU Ting	19-02-53
QoE Management for 5G New Radio	ZHANG Man, LI Dapeng, LIU Zhuang, GAO Yin	19-03-64
Study on Security of 5G and Satellite Converged Communication Network		
.....	YAN Xincheng, TENG Huiyun, PING Li, JIANG Zhihong, ZHOU Na	19-04-79

Research Paper

Integrating Coarse Granularity Part-Level Features with Supervised Global-Level Features for Person Re-Identification		
.....	CAO Jiahao, MAO Xiaofei, LI Dongfang, ZHENG Qingfang, JIA Xia	19-01-72
Adaptability Analysis of Fluctuating Traffic for IP Switching and Optical Switching		
.....	LIAN Meng, GU Rentao, JI Yuefeng, WANG Dajiang, LI Hongbiao	19-01-82
Differentially Authorized Deduplication System Based on Blockchain		
.....	ZHAO Tian, LI Hui, YANG Xin, WANG Han, ZENG Ming, GUO Haisheng, WANG Dezheng	19-02-67
A Novel De-Embedding Technique of Packaged GaN Transistors		
.....	WEI Xinghui, CHEN Xiaofan, CHEN Wenhua, ZHOU Junmin	19-02-77
Flexible Multiplexing Mechanism for Coexistence of URLLC and EMBB Services in 5G Networks		
.....	XIAO Kai, LIU Xing, HAN Xianghui, HAO Peng, ZHANG Junfeng, ZHOU Dong, WEI Xingguang	19-02-82
Super Resolution Sensing Technique for Distributed Resource Monitoring on Edge Clouds		
.....	YANG Han, CHEN Xu, ZHOU Zhi	19-03-73
Semiconductor Optical Amplifier and Gain Chip Used in Wavelength Tunable Lasers		
.....	SATO Kenji, ZHANG Xiaobo	19-03-81

Feedback-Aware Anomaly Detection Through Logs for Large-Scale Software Systems	
.....	HAN Jing, JIA Tong, WU Yifan, HOU Chuanjia, LI Ying 19-03-88
Payload Encoding Representation from Transformer for Encrypted Traffic Classification	
.....	HE Hongye, YANG Zhiguo, CHEN Xiangning 19-04-90
AI-Based Optimization of Handover Strategy in Non-Terrestrial Networks	
.....	ZHANG Chenchen, ZHANG Nan, CAO Wei, TIAN Kaibo, YANG Zhen 19-04-98
Truly Grant-free Technologies and Protocols for 6G	MA Yihua, YUAN Zhifeng, LI Weimin, LI Zhigang 19-04-105