Ph.D. DISSERTATION

A Study on Sparsity-aware Communications

희소인지를 이용한 전송기술 연구

HYOUNG-JU JI

FEBRUARY 2019

Department of Electrical and Computer Engineering
College of Engineering
Seoul National University

Ph.D. DISSERTATION

A Study on Sparsity-aware Communications

희소인지를 이용한 전송기술 연구

HYOUNG-JU JI

FEBRUARY 2019

Department of Electrical and Computer Engineering
College of Engineering
Seoul National University

A Study on Sparsity-aware Communications

Advisor: Byonghyo Shim

Presented to the Graduate School of Seoul National Univserity
in Partial Fulfilment of the Requirements for

The Degree of Doctor of Philosophy

Nov 2018

by

Hyoungju Ji

Department of Electrical and Computer Engineering
College of Engineering
Seoul National University

This dissertation is approved for

The Degree of Doctor of Philosophy

Dec 2018

Chairman _____
Vice charirman _____
Member _____
Member _____
Member _____

A Study on Sparsity-aware Communications

희소인지를 이용한 전송기술 연구

지도교수 심 병 효
이 논문을 공학박사 학위논문으로 제출함

2018년 11월

서울대학교 대학원

전기 컴퓨터 공학부

지 형 주

지형주의 공학박사 학위 논문을 인준함

2018년 12월

위 원 장: _____
부위원장: _____
위　　원: _____
위　　원: _____
위　　원: _____

Abstract

The new wave of the technology revolution, named the fifth wireless systems, is changing our daily life dramatically. These days, unprecedented services and applications such as driverless vehicles and drone-based deliveries, smart cities and factories, remote medical diagnosis and surgery, and artificial intelligence-based personalized assistants are emerging. Communication mechanisms associated with these new applications and services are way different from traditional communications in terms of latency, energy efficiency, reliability, flexibility, and connection density. Since the current radio access mechanism cannot support these diverse services and applications, a new approach to deal with these relentless changes should be introduced.

This compressed sensing (CS) paradigm is very attractive alternative to the conventional information processing operations including sampling, sensing, compression, estimation, and detection. To apply the CS techniques to wireless communication systems, there are a number of things to know and also several issues to be considered. In the last decade, CS techniques have spread rapidly in many applications such as medical imaging, machine learning, radar detection, seismology, computer science, statistics, and many others. Also, various wireless communication applications exploiting the sparsity of a target signal have been studied. Notable examples include channel estimation, interference cancellation, angle estimation, spectrum sensing, and symbol detection. The distinct feature of this work, in contrast to the conventional approaches exploiting naturally acquired sparsity, is to exploit intentionally designed sparsity to improve the quality of the communication systems.

In the first part of the dissertation, we study the mapping data information into the sparse signal in downlink systems. We propose an approach, called sparse vector coding (SVC), suited for the short packet transmission. In SVC, since the data information is mapped to the position of sparse vector, whole data packet can be decoded by iden-

tifying nonzero positions of the sparse vector. From our simulations, we show that the packet error rate of SVC outperforms the conventional channel coding schemes at the URLLC regime. Moreover, we discuss the SVC transmission for the massive MTC access by overlapping multiple SVC-based packets into the same resources. Using the spare vector overlapping and multiuser CS decoding scheme, SVC-based transmission provides robustness against the co-channel interference and also provide comparable performance than other non-orthogonal multiple access (NOMA) schemes. By using the fact that SVC only identifies the support of sparse vector, we extend the SVC transmission without pilot transmission, called pilot-less SVC. Instead of using the support, we further exploit the magnitude of sparse vector for delivering additional information. This scheme is referred to as enhanced SVC. The key idea behind the proposed E-SVC transmission scheme is to transform the small information into a sparse vector and map the side-information into a magnitude of the sparse vector. Metaphorically, E-SVC can be thought as a standing a few poles to the empty table. As long as the number of poles is small enough and the measurements contains enough information to find out the marked cell positions, accurate recovery of E-SVC packet can be guaranteed.

In the second part of this dissertation, we turn our attention to make sparsification of the non-sparse signal, especially for the pilot transmission and channel estimation. Unlike the conventional scheme where the pilot signal is transmitted without modification, the pilot signals are sent after the beamforming in the proposed technique. This work is motivated by the observation that the pilot overhead must scale linearly with the number of taps in CIR vector and the number of transmit antennas so that the conventional pilot transmission is not an appropriate option for the IoT devices. Primary goal of the proposed scheme is to minimize the nonzero entries of a time-domain channel vector by the help of multiple antennas at the basestation. To do so, we apply the time-domain sparse precoding, where each precoded channel propagates via fewer tap than the original channel vector. The received channel vector of beamformed pilots

can be jointly estimated by the sparse recovery algorithm.

keywords: sparsity-aware, compressed sensing, 5G wireless communications, sparse vector coding, sparse beamforming

student number: 2015-30204

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Background

The new wave of the technology revolution, named the fourth industrial revolution, is changing the way we live, work, and communicate with each other. We are now witnessing the emergence of unprecedented services and applications such as driverless vehicles and drone-based deliveries, smart cities and factories, remote medical diagnosis and surgery, and artificial intelligence-based personalized assistants (see Fig. 1.1). Communication mechanisms associated with these new applications and services are way different from traditional human-centric communications in terms of latency, energy efficiency, reliability, flexibility, and connection density. Therefore, coexistence of human-centric and machine-type services as well as hybrids of these will render emerging wireless environments more diverse and complex. To address diversified services and applications, International Telecommunication Union (ITU) has classified fifth generation (5G) services into three categories: ultra-reliable and low latency communication (URLLC), massive machine-type communication (mMTC), and enhanced mobile broadband (eMBB) [46]. To cope with these new service categories, various performance requirements such as massive connectivity, lower latency, higher reliability, and better energy efficiency have been newly introduced. Since the cur-

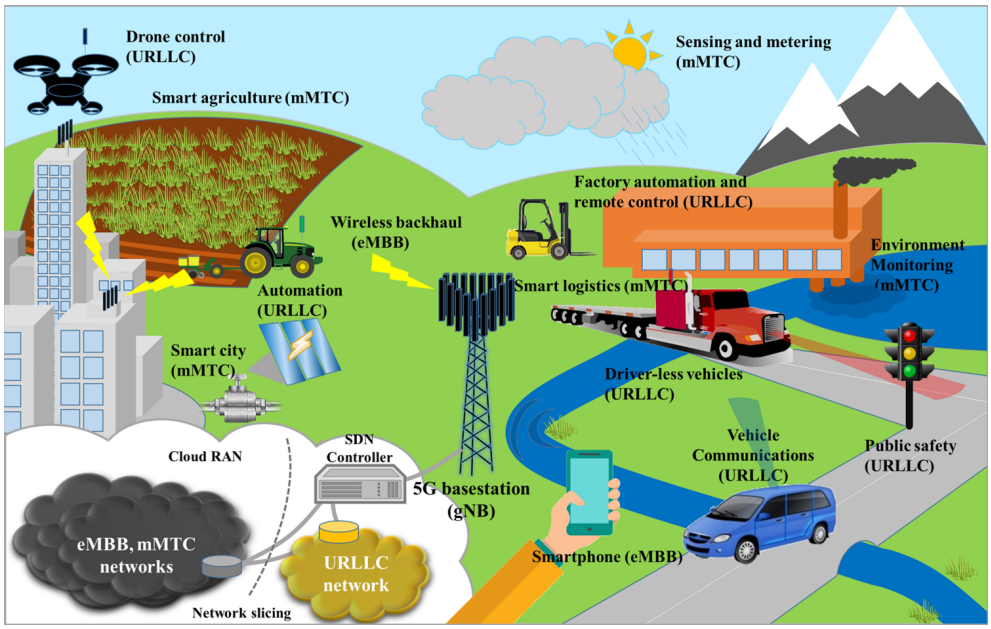


Figure 1.1: Overview of 5G systems: deployment with eMBB, mMTC, and URLLC.

rent radio access mechanism and also conventional approach cannot support these relentless changes, a new approach referred to as *sparse processing* has been studied recently. The primary goal of sparse processing is to bring benefit by projecting the non-sparse signal to the sparse signal that are not fully employed in the traditional 4G systems. Before we proceed, we provide a brief overview of three service categories in 5G, viz., eMBB, mMTC, and URLLC. Next, we discuss the fundamental of sparse processing.

1.1.1 Three Key Services in 5G systems

eMBB is a service category related to high requirements for bandwidth, such as high-resolution video streaming, virtual reality, and augmented reality. The main challenge in 4G systems is to improve the system throughput (e.g., area, average, peak, perceived, and cell-edge throughput). Physical layer technologies introduced to this end include high order modulation transmission, carrier aggregation, cell densification via

heterogeneous network, and multiple-input multiple-output (MIMO) transmission. In essence, the main goal of eMBB is in line with this direction. In order to achieve 100-fold capacity increase over the 4G systems, more aggressive physical layer technologies improving the spectral efficiency and exploiting the unexplored spectrum are needed. Technologies under consideration include full-dimension and massive MIMO [78], millimeter-wave communication [79], and spectrally-localized waveforms [47].

mMTC is a service category to support the access of a large number of machine-type devices. mMTC-based services, such as sensing, tagging, metering, and monitoring, require high connection density and better energy efficiency [49]. Over the years, there have been some trials to support machine-type communications such as NB-IoT in licensed band, SigFox and LoRa in unlicensed band [64]. These approaches are similar in spirit but SigFox and LoRa technologies are suited for the stand alone services while NB-IoT is a good fit for the standard compatible services. These approaches offer some benefits, such as the low power consumption, low operation cost, and improved coverage. However, in the scenario where devices significantly outnumber the resources used for the transmission, an aggressive connection strategy violating the orthogonal transmission principle is required. In recent years, approaches using non-orthogonal spreading sequence or user-specific interleaving have been proposed to accommodate more users than the traditional approach relying on orthogonal multiple access [49].

URLLC is a service category to support the latency sensitive services such as remote control, autonomous driving, and tactile internet [64]. Since the time it takes for the human perception or reaction is in the order of tens of milliseconds (ms), packet transmission time for the mission-critical applications needs to be in the order of tens$\sim$hundreds of microseconds (μs) [31]. While the latency of 4G LTE networks has been significantly improved from 3G networks, the end-to-end latency is still in the range of $30 \sim 100$ ms. This is because the backbone network typically uses the best-effort delivery mechanism and hence is not optimized for the mission-critical service.

To reduce the end-to-end latency, therefore, there should be fundamental changes in both wireless link and backbone network. In the backbone link, software defined network (SDN) and virtual network slicing can be used to construct the private connection to the dedicated URLLC service [64]. Indeed, by using the dedicated network, latency of the backbone link can be reduced significantly. Whereas, in the wireless link, overhead should be reduced and the transmission mechanism needs to be streamlined. In fact, since a large portion of the transmit latency is due to the control signaling (e.g., grant and pilot signal) and it takes almost $0.3 \sim 0.4$ ms per scheduling, it is not so efficient to incorporate a low-latency packet transmission scheme in the current LTE systems. For example, when we design a short packet with the transmission latency being 0.5 ms, more than 60% of resources would be wasted for the control overhead. To support URLLC, therefore, many parts of the physical layer should be re-designed.

1.1.2 Sparse Processing in Wireless Communications

Compressed sensing (CS) is a new paradigm to process or recover the sparse signals. This new approach is very competitive option for information processing operations including sampling, sensing, compression, estimation, and detection. Traditional way to acquire and reconstruct analog signals from sampled signals is based on the Nyquist-Shannon's sampling theorem which states that the sampling rate should be at least twice the bandwidth. While these fundamental principles works well, they might be bottleneck of resource overhead and also complexity in a situation where signals are sparse, meaning that the signals can be represented using a relatively small number of nonzero coefficients. At the heart of the CS lies the fact that a sparse signal vector can be recovered from the underdetermined linear system in a computationally efficient way. In other words, a small number of linear measurements (projections) of the signal contain enough information for its reconstruction. Main wisdom behind the CS is that essential knowledge in the large dimensional signals is just handful, and thus measurements with the size being proportional to the sparsity level of the input sig-

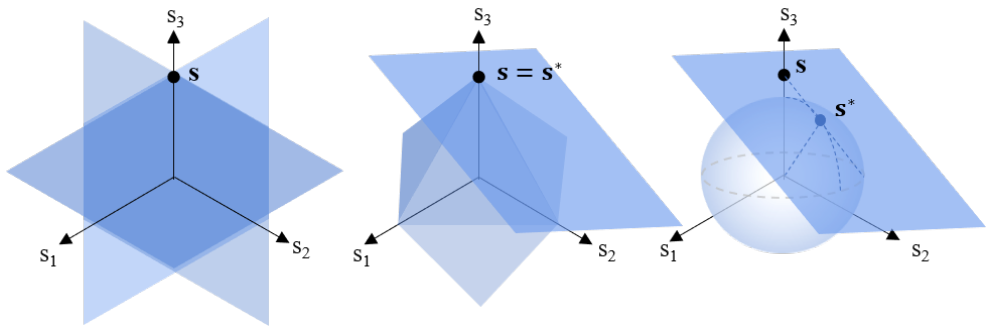


Figure 1.2: Illustration of solver: l_0, l_1, and l_2-norm minimization.

nal are enough to reconstruct the original signal. In the last decade, CS techniques have spread rapidly in many disciplines such as medical imaging, machine learning, radar detection, seismology, computer science, statistics, and many others. Also, various wireless communication applications exploiting the sparsity of a target signal have been proposed in recent years. Notable examples, among many others, include channel estimation, interference cancellation, direction estimation, spectrum sensing, and symbol detection. To understand the principle of CS, we begin with a system given

$$\mathbf{y} \quad = \quad \mathbf{Hs}, \tag{1.1}$$

where $\mathbf{y} \in \mathcal{R}^{m\times 1}$ is the measurement vector, $\mathbf{H} \in \mathcal{R}^{m\times n}$ is the system matrix (a.k.a. the sensing matrix), and $\mathbf{s} \in \mathcal{R}^{n\times 1}$ is the desired signal vector. In the case of overdetermined system ($m \geq n$) and the system matrix is a full rank matrix, one can recover $\mathbf{s}$ using a simple algorithm. However, when the system matrix is underdetermined, finding solution is challenging and not straight-forward. When the desired vector $\mathbf{s}$ is non-sparse signal, one can apply a solution minimizing the l_2-norm of $\mathbf{s}$. That is,

$$\mathbf{s}^* \quad = \quad \arg\min \|\mathbf{s}\|_2 \text{ s.t. } \mathbf{y} = \mathbf{Hs}, \tag{1.2}$$

and one can obtain the estimated desired signal $\mathbf{s}^*$ as

$$\mathbf{s}^* = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{y}. \tag{1.3}$$

When the desired signal is sparse signal, that is l_0-norm of the desired signal $\mathbf{s}$ $\|\mathbf{s}\|_0 = K$ where $K < N$, one can apply to find the l_0-norm from the measurement vector.

$$\mathbf{s}^* = \arg\min \|\mathbf{s}\|_0 \text{ s.t. } \mathbf{y} = \mathbf{H}\mathbf{s}. \tag{1.4}$$

Since the this solution counts the number of nonzero elements in $\mathbf{s}$, one needs to search all possible combinations which is not practical for a large N and K. Alternative approach suggested by [59] is to minimized l_1-norm as follow:

$$\mathbf{s}^* = \arg\min \|\mathbf{s}\|_1 \text{ s.t. } \mathbf{y} = \mathbf{H}\mathbf{s}. \tag{1.5}$$

Using l_1-norm minimization, one can apply convex optimization for finding the solution. While the linear programing to solve l_1-norm minimization problem is effective, it requires substantial computational complexity and not feasible in the practical scenarios. Fig. 1.2. illustrates the difference between them. If the original vector is sparse, while the l_0 and l_1-norm can find the optimal point, the l_2-norm solution has no reason to select the point on the coordinate axes. To overcome this problem, a greedy algorithm has been proposed over the years. By this, one hopes to find the local optimal in each iteration expecting to find the global optimal in the end. Most popular algorithm is the orthogonal matching pursuit (OMP). In each iteration, a column maximally correlated with the observation vector is chosen. Once the solve selects the correct columns, now the system goes to the overdetermined system. There are three key observations from CS recovery problem.

- When the sparse K is much smaller than the size of the desired signal vector, one can find the support of $\mathbf{s}$ more accurately than $\mathbf{s}$.

- When the size of measurement vector m is given, one can recover the desired signal vector more accurately as K decreases.
- When the fact that $K < n$ is given to the solver, one can find more accurate $\mathbf{s}$ than the solver without using the fact.

In this dissertation, we seek to find how we can use these observation for the wireless communication systems. That is, we seek to understand how to transform non-sparse information and signal into the sparse signal. Motivated by this, we aim to design a new sparse processing technique suitable for the data transmission in 5G and beyond. We also discuss the practical benefits of sparse processing such as overhead, throughput, and latency.

1.2 Contributions and Organization

In this dissertation, we introduce a new approach employing the sparse signal in the transmission of data information. To be specific, we propose a transmission technique to transform the information vector into the sparse vector to improve latency and reliability. Next, we extend the proposed scheme in various applications including uplink overlapping and pilot-less transmission. We also present the new beamforming (sparse processing) technique to improve the channel estimation performance by minimizing sparsity in channel (sparser than the original channel). Finally, we discuss joint analog and quantized data transmission scheme to fully enjoy the benefit of estimation performance in sparse recovery.

In Chapter 2, we focus on the downlink data transmission. As a solution for the ultra-short packet transmission scenario, we propose a new type of short packet transmission for URLLC that does not rely on the conventional channel coding principle. Key idea behind the proposed technique, henceforth referred to as *sparse vector coding* (SVC), is to transmit the short-sized information after the sparse vector transformation. To be specific, by mapping the information into the sparse vector and then transmit-

ting it after the random spreading, we obtain an underdetermined sparse system for which the principle of compressed sensing can be applied [72]. It is now well-known from the theory of compressed sensing that an accurate recovery of a sparse vector is guaranteed with a relatively small number of measurements when the system matrix (a.k.a. sensing matrix) is generated at random [59], which is achieved in our case via the random spreading. In fact, since the sparsity of the input vector is guaranteed by the sparse transformation, the SVC decoding is achieved by the sparse signal recovery (more accurately, identification of nonzero positions in the transmit sparse vector). Therefore, the proposed scheme is very simple to implement and can be applied to wide variety of future wireless applications in which the amount of information to be transmitted is sufficiently small.

In Chapter 3, we propose an enhanced SVC to support massive access scenario in uplink transmission. Instead of using the sparsity of active user, we exploit the sparsity of information vector that is concatenated by massive user. We put forth an approach that directly exploits the sparsity of information vector to control the multiuser interference. This, together with the fact that the symbol is spread using randomly generated codebook controls the multiuser interference efficiently. Since the input is a sparse vector and the system matrix is the composite of fading channel and random spreading matrices, we can cast the symbol detection problem into the sparse support recovery problem. In a nutshell, the proposed scheme is simple to implement, robust to interference, and can be readily applied to various mMTC applications where the size of transmit packet is small. From the numerical evaluations in realistic 3GPP LTE-Advanced uplink access scenario, we demonstrate that the proposed technique is very effective in the ultra-short packet transmission and outperforms the LTE physical uplink shared channel (PUSCH) [63] and LDS-based schemes in NOMA. We also demonstrate that the SVC-encoded uplink access is effective in controlling multi-user interference caused by the massive access.

In Chapter 4, we propose an improved SVC that does not require the pilot signal.

A short packet transmission is one of the major operational modes for mission-critical data in ultra reliable and low latency communications (URLLC) and control-type data in massive machine-type communications (mMTC). The central challenge in the short packet transmission is the excessive amount of overhead caused by the pilot signaling. In this dissertation, we propose a novel scheme suitable for the short packet transmission without pilot signals, called pilot-less sparse vector coding (PL-SVC). Key feature of PL-SVC is to map the input as a composite of the sparse vector and the fading channel and to perform decoding by finding out the nonzero positions of the sparse vector. In this setting, the system matrix becomes a pseudo-random spreading matrix and the input vector becomes a channel-scaled sparse vector so that the PL-SVC decoding problem can be cast into the support detection problem in the compressed sensing. We show from the numerical experiments in the 5G uplink scenario that PL-SVC is very effective in the short packet transmission and outperforms conventional schemes.

In Chapter 5, we propose the pilot transmission technique that increases channel estimation accuracy. Key idea behind the proposed technique is to transmit both analog and quantized data using the sparse vector transformation. By mapping the quantized data into positions of sparse vector and then setting the analog data as the magnitude of nonzero elements, two different types of data can be jointly mapped to the sparse vector. After spreading with sequences from multiple Zadoff-chu (ZC) sets, we obtain an underdetermined sparse system and it is now well-known that the theory of compressed sensing guarantees an accurate recovery of a sparse vector and minimize mean square error of recovered vector with a relatively small number of measurement [59]. In fact, the decoding of quantized data is performed by the sparse signal recovery (more accurately, identification of nonzero positions in the transmit sparse vector) and the decoding of analog data is performed by the recovery of support value.

In Chapter 6, we look into sparse processing for pilot transmission and beamforming originated from the MIMO technology: full-dimension MIMO (FD-MIMO). In Chapter 6, we present an overview FD-MIMO systems and how to beamform the

pilot signal for capacity improvement. We then develop the novel beamforming technique named time-domain sparse beamforming to enhance throughput by minimizing overhead in pilot transmission. By designing beamforming weight that minimize non-zero element in time-domain channel impulse response (CIR) in time-domain of pilot signal, the required pilot overhead can be minimized but user full diversity gain in data transmission. Pilot beamforming and CSI acquisition strategy for IoT systems to achieve reduction in the pilot overhead and enhancement in the channel estimation quality are suggested. Key idea of the proposed method, referred to as *time-domain sparse beamforming* (TDSB), is to sparsify the time-domain channel vector using the preprocessing (beamforming) at the basestation. To be specific, using the deliberately designed antenna-domain beamforming, we sparsify the beamformed time-domain channel vector. As a result, only a few samples are required to perform the channel estimation and whole CSI can be acquired with partial pilot symbols in time and frequency. From the numerical evaluations, we show that the proposed scheme outperforms conventional channel estimation schemes [70-73] and achieves N_T-fold reduction in the pilot overhead.

Chapter 7 summaries the contributions of the dissertation and discuss the future research directions based on studies of this dissertation.

1.3 Notation

This dissertation uses the following notation: transpose, hermitian, and inverse of a matrix $\mathbf{A}$ are represented by $\mathbf{A}^T$, $\mathbf{A}^H$, and $\mathbf{A}^{-1}$, respectively. In addition, $E[\cdot]$ indicates the expectation operator. For the matrix representation, $\mathbf{I}_n$ is the $n \times n$ identity matrix, $\mathbf{0}_{m,n}$ indicates an $m \times n$ matrix consisting of all zeros, and $\mathbf{0}_n$ is used for a zero vector of size n. Besides, $\langle \mathbf{a}, \mathbf{b} \rangle$ is the inner product between two vector $\mathbf{a}$ and $\mathbf{b}$. $\mathbb{C}^{m \times n}$, $\mathbb{R}^{m \times n}$, and $\mathcal{CN}(0, \sigma^2)$ denote an $m \times n$ matrix whose components are complex values, an $m \times n$ matrix whose components are real values, and a complex Gaussian

random variable with zero mean and variance σ^2, respectively.

Chapter 2

Sparse Vector Coding for Downlink Ultra-reliable and Low Latency Communications

In this chapter, we propose a new type of short packet transmission for URLLC that does not rely on the conventional channel coding principle. Key idea behind the proposed technique, henceforth referred to as *sparse vector coding* (SVC), is to transmit the short-sized information after the sparse vector transformation. To be specific, by mapping the information into the sparse vector and then transmitting it after the random spreading, we obtain an underdetermined sparse system for which the principle of compressed sensing can be applied. Since the sparsity of the input vector is guaranteed by the sparse transformation, the SVC decoding is achieved by the sparse signal recovery (more accurately, identification of nonzero positions in the transmit sparse vector). Therefore, the proposed scheme is very simple to implement and can be applied to wide variety of future wireless applications in which the amount of information to be transmitted is sufficiently small.

The work of Chapter 2 has been published in part in [53–55].

2.1 Introduction

Ultra reliable and low latency communication (URLLC) is a newly introduced service category in 5G to support delay-sensitive applications such as the tactile internet, autonomous driving, factory automation, cyber-physical system, and remote robot surgery [46]. In order to support this new service category, 3rd Generation Partnership Project (3GPP) sets an aggressive requirement that a packet should be delivered with 10^{-5} packet error rate (PER) within 1 ms period [64]. One notable observation in these applications is that the transmit information is control (command) type information (e.g., move left/right, start/stop, rotate/shift, and speed up/down) so that the amount of information to be delivered is tiny [47]. Since the current wireless transmission strategy designed to maximize the coding gain by transmitting capacity achieving long codeblock is not relevant to these URLLC scenarios, entirely new transmission strategy to support the short packet transmission is required. While there have been some efforts to improve the connection density, the medium access latency, and the reliability of the re-transmission scheme for URLLC [6-9-32-49-50-65], not much work has been done for the short-sized packet transmission except for the consideration of advanced channel coding schemes (e.g., polar code) [48].

In the current 4G systems, reliability of the data transmission is mainly achieved by the channel coding scheme [63]. Encoding at the basestation is done by the convolution coding and the decoding at the mobile terminal is done by the maximum likelihood decoding (MLD) or Turbo decoding. While this approach has shown to be effective in 4G systems, use of this scheme in URLLC scenario would be problematic since there is a stringent limitation on the packet length (and thus the parity size) yet the required reliability (target PER = 10^{-5}) is much higher than the current LTE-Advanced and LTE-Advanced Pro systems (target PER = $10^{-2}\sim10^{-3}$) [48].

The purpose of this chapter is to propose a new type of short packet transmission for URLLC that does not rely on the conventional channel coding principle. Key idea behind the proposed technique, henceforth referred to as *sparse vector coding* (SVC),

is to transmit the short-sized information after the sparse vector transformation. To be specific, by mapping the information into the sparse vector and then transmitting it after the random spreading, we obtain an underdetermined sparse system for which the principle of compressed sensing can be applied [72]. It is now well-known from the theory of compressed sensing that an accurate recovery of a sparse vector is guaranteed with a relatively small number of measurements when the system matrix (a.k.a. sensing matrix) is generated at random [59], which is achieved in our case via the random spreading. In fact, since the sparsity of the input vector is guaranteed by the sparse transformation, the SVC decoding is achieved by the sparse signal recovery (more accurately, identification of nonzero positions in the transmit sparse vector). Therefore, the proposed scheme is very simple to implement and can be applied to wide variety of future wireless applications in which the amount of information to be transmitted is sufficiently small.

Noting that the proposed SVC is conceptually similar to the index modulation (IM) techniques [14-15] in which the indices of the building block of the communication systems, such as transmit antennas at the basestation or subcarrier groups in OFDM systems, are used to convey additional information bits. In the spatial modulation-based IM technique, for example, by using part of transmit antennas in the information transmission, additional information can be delivered. Our work is distinct from these studies in that we fully utilize the physical resources in the data transmission so that the loss, if any, caused by the underutilization of physical resources can be prevented. Also, in contrast to the IM technique where the receiver processing consists of two steps (the index recovery and symbol detection), decoding of the proposed SVC scheme is achieved by the support identification which is in general much simpler than the sparse recovery [72]. The distinction of SVC over the IM technique is further strengthened by the fact that IM has no random sensing mechanism (e.g., random spreading in SVC) so that the compression of the transmit vector is not possible.

From the performance analysis in terms of the decoding success probability and

also numerical evaluations on the realistic setting, we demonstrate that the proposed SVC technique is very effective in short-size packet transmission and outperforms the 4G LTE and LTE-Advanced physical downlink control channel (PDCCH) scheme by a large margin in terms of reliability and transmission latency.

The rest of this chapter is organized as follows. In Section 2.2, we discuss the key requirements for URLLC, and follows physical-layer of URLLC in Section 2.3. In Section 2.4, we briefly explain the short-sized packet transmission in 4G LTE and LTE-Advanced systems. In Section 2.5, we present the proposed SVC scheme and explain the encoding and decoding operations. In Section 2.6, we analyze the success probability of SVC-encoded data transmission. Various implementation issues are discussed in Section 2.7. In Section 2.8, we present simulation results to verify the performance of the proposed scheme. We conclude the dissertation in Section 2.9.

2.2 URLLC Service Requirements

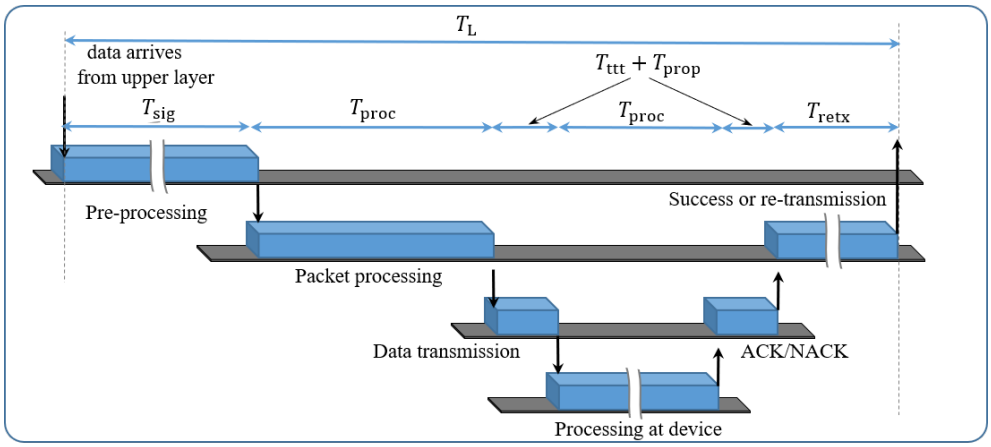
In order to come up with proper solutions to URLLC, it is necessary to understand the key requirements first. In this section, we present the latency and reliability requirements and then discuss the requirement related to the coexistence of URLLC and other services.

2.2.1 Latency

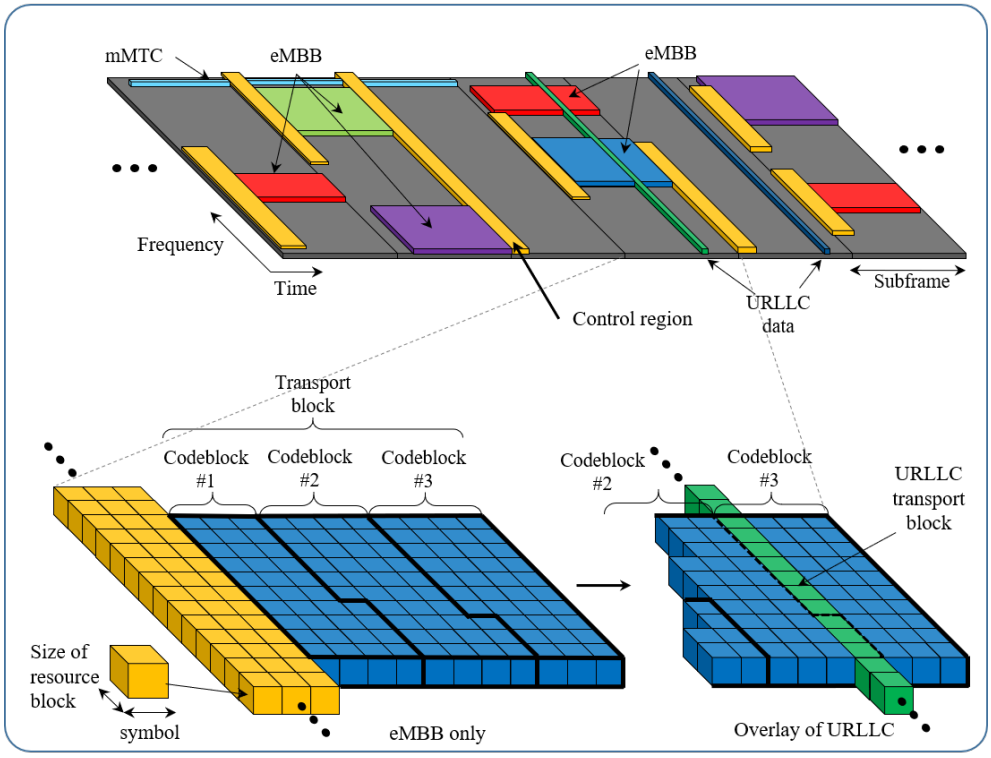
Physical layer latency T_{L} can be divided into the following five components (see Fig. 2(a)):

$$T_{\mathrm{L}} = T_{\mathrm{ttt}} + T_{\mathrm{prop}} + T_{\mathrm{proc}} + T_{\mathrm{retx}} + T_{\mathrm{sig}}, \tag{2.1}$$

- T_{ttt} is the time-to-transmit latency which corresponds to the time to transmit a packet.
- T_{prop} is the signal propagation time from the transmitter to the receiver.



(a)



(b)

Figure 2.1: Physical layer downlink scenarios in URLLC service: a) illustration of latency components; b) transmission of eMBB, mMTC, and URLLC packet in subframe level, and scheduling of URLLC packet into eMBB packet in symbol level.

- T_{proc} is the time to perform the encoding and decoding, and also the channel estimation in the initial transmission.
- T_{retx} is the time caused by the re-transmission.
- T_{sig} is the pre-processing time for the signaling exchange such as connection request, scheduling grant, channel training and feedback, and queuing delay.

In response to the ITU requirements, 3GPP decided that the average latency of URLLC (from L2/L3 ingression to L2/L3 egression) should be less than 0.5 ms [48]. In order to meet this stringent latency constraint, a packet transmission time T_{ttt} should be in the order of hundreds of microseconds. Since T_{ttt} of the current 4G LTE systems is 1 ms period, a new frame structure reducing T_{ttt} should be introduced. Also, since the latency caused by the channel estimation and feedback would be a bottleneck for the URLLC transmission, a transmission scheme that does not rely on the channel information needs to be considered.

2.2.2 Ultra-High Reliability

In 4G LTE systems, typical reliability for a packet transmission is 0.99. Two key ingredients to achieve this goal are the channel coding (convolution and Turbo code) and the partial re-transmission of erroneous transport block called hybrid automatic repeat request (HARQ). URLLC services require much better performance, and in fact, the target reliability within 1 ms period should be at least 0.99999 [48]. Further, in the mission-critical applications such as autonomous driving and remote surgery, the reliability should be as high as $1 - 10^{-7}$ [64]. The first thing to do to meet these stringent requirements is to improve the channel estimation accuracy. This is because the channel coding gain is small for the short packet so that the loss, if any, caused by the channel estimation should be prevented as much as possible. This is done by adding more resources to the pilot and using an advanced channel estimation technique. Even in this case, the required URLLC performance might not be satisfied so that additional

resources in the frequency, antenna, and spatial domains are required to improve the reliability. Further, an advanced channel coding scheme suitable for the short packet transmission should be employed.[1]. In case the slot length is very short, repetitive transmission scheme using time-domain resources can also be a viable option.

2.2.3 Coexistence

When there is a URLLC service request, whether in the scheduling period or in the middle of eMBB or mMTC transmission, the basestation should transmit the URLLC packet immediately [47]. In other words, to support the URLLC packet transmission, ongoing eMBB and mMTC packets should be stopped without notice. As illustrated in the Fig. 2(b), when a transport block consisting of 3 codeblocks is transmitted for the eMBB service, each codeblock is mapped sequentially to the scheduled time-frequency resources. Thus, when the URLLC service is initiated in the middle of the eMBB transport block, part of symbols in the third codeblock are replaced by the symbols of the URLLC packet. Since this interrupt is not reported to the mobile devices in use, reception quality of the eMBB and mMTC services will be degraded severely. This problem, dubbed as a coexistence problem in the 3GPP NR discussion, is a serious concern to the non-URLLC services so that a proper mechanism to protect the ongoing services should be introduced.

2.3 URLLC Physical Layer in 5G NR

In contrast to the 4G LTE systems, latency, reliability, and throughput requirements should be jointly considered in 5G NR so that there should be a fundamental change in the physical layer architecture (packet, slot, and frame). Specifically, a latency-sensitive packet structure for the fast decoding process and a flexible frame structure to support the dynamic change of the resource grid based on the latency requirement

[1] In LTE systems, 12 or 24 resource elements are allocated for demodulation reference signal (DMRS) per resource block (RB)

are needed. Also, when the URLLC service is initiated, the URLLC packet should be transmitted instantly without delay. To do so, a scheduling scheme minimizing the transmit latency of the URLLC packet should be introduced. Further, since the latency requirement might not be satisfied by the HARQ-based re-transmission unless TTI of a packet is very short, a mechanism that significantly reduces the re-transmission latency is required. Besides, an approach to use multiple radio interfaces to reduce the latency can be employed. Basic idea of this approach is to choose the radio access technology (RAT) providing the minimum latency among all possible options including 4G LTE, 5G NR, WiFi, and other IEEE 802.x standards. Using this together with the device-to-device (D2D) communications, the network layer latency can be reduced substantially.

In this section, we put our emphasis on the physical layer solutions for URLLC including packet and frame structure to minimize the latency, multiplexing schemes to overlay the URLLC service into eMBB and mMTC services, and approaches to deal with the coexistence problem. We note that the reliability improvement and the latency reduction are equally important for the success of URLLC. However, there has been a consensus in the 3GPP NR standard meeting that the latency reduction issue should be considered by priority. This is because the reliability improvement can be achieved by the elaboration of 4G techniques such as channel coding, antenna, space, and frequency diversity schemes but such is not the case for the latency reduction effort [32-33].

2.3.1 Packet Structure

The key issue in the URLLC packet design is to minimize the processing latency T_{proc} and the time-to-transmit latency T_{ttt}. Note that T_{proc} consists of the time to receive packets, acquire channel information, extract control (scheduling) information, decode data packets, and check errors. In LTE systems, a square-shaped packet structure is popularly used for the efficient utilization of the spectrum under the channel fading.

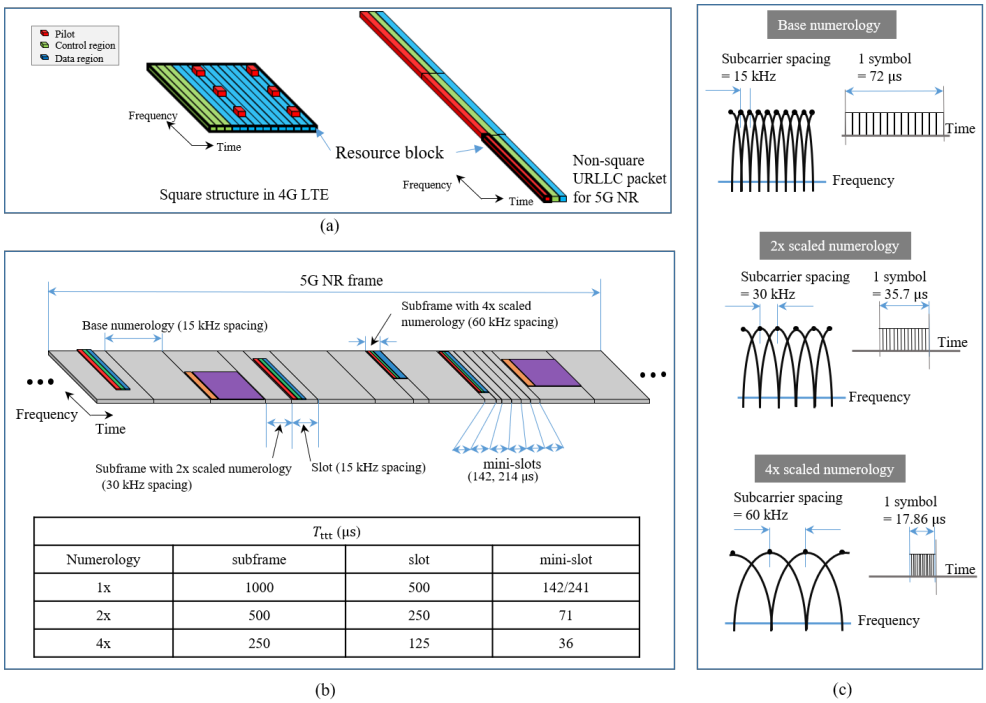
Whereas, in 5G NR systems, a non-square packet stretched in the frequency axis is used as a baseline since this structure minimizes the transmission latency T_{ttt} [34]. Furthermore, in order to reduce the latency T_{proc}, three components of a packet, (pilot, control, and data part) should be grouped together to make a pipelined processing of the channel acquisition, control channel decoding, and data detection (see Fig. 3(a)).

Another important issue to be considered is to use an advanced channel coding scheme. In 4G systems, two types of approaches have been employed to ensure reliability requirement. The first type is the channel coding scheme (Turbo and convolution code) with cyclic redundancy check (CRC) attachment for the large-sized packet. The second type is to use a simple code (the repetition and Reed-Muller code) without CRC attachment for the small-sized packet. In 5G NR, Polar code and low density parity check (LDPC) code have been adopted for the enhancement of control and data channel, respectively. Over the years, many efforts have been made to improve the decoding performance and computational complexity (and hence processing latency) of these codes such as successive cancellation list decoding of Polar code and non-binary LDPC decoding [33].

2.3.2 Frame Structure and Latency-sensitive Scheduling Schemes

One of the main goals in 5G NR is to design a unified frame structure to cover a wide-range of frequency band and various service categories. To this end, flexible frame structure and user scheduling mechanism have been introduced.

One direct option to reduce the time-to-transmit latency T_{ttt} is to reduce the symbol period (see Fig. 3(b)). When the frequency band above 6 GHz (millimeter-wave) is used, due to the path loss, cell radius would be much smaller than that of conventional cellular systems and so will be the channel delay spread. In this case, by controlling the subcarrier spacing, we can reduce the symbol period (see Fig 3(c)). For instance, the symbol length can be reduced by half (from 72 μs to 36 μs) by doubling the subcarrier spacing (from 15 kHz to 30 kHz). In doing so, the time to transmit one subframe



T_{ttt} (μs)			
Numerology	subframe	slot	mini-slot
1x	1000	500	142/241
2x	500	250	71
4x	250	125	36

Figure 2.2: Packet and frame structure for URLLC: (a) packet structure; (b) frame structure; (c) supported numerologies for 5G NR.

can be reduced by half (from 1 ms to 0.5 ms). However, when the frequency band below 6 GHz is used, this option might not be desirable due to the large delay spread. In this case, one can alternatively consider reducing TTI of the packet. For example, using mini-slot level (2∼3 symbols) and slot level (7 symbols) transmission, T_{ttt} can be reduced to 142, 241, and 500 μs, respectively. In short, by controlling the symbol period and also the number of symbols in a packet, T_{ttt} being smaller than 1 ms can be achieved (see Table in Fig. 3(b)). Note that to support this flexible frame structure, an advanced receiver equipped with fast tracking, quick synchronization, and simultaneous decoding functions is needed.

Since the URLLC packet is generated abruptly, how to schedule this into existing services is an important issue in the system design. Two schemes adopted in 3GPP NR standard are the instant scheduling and reservation-based scheduling [34].

- Instant scheduling: Any ongoing data transmission is interrupted to initiate the URLLC packet. This protocol is effective in reducing the URLLC access time but causes a severe performance degradation. Therefore, an approach to mitigate the performance degradation of ongoing services is needed (see Section IV. C).

- Reservation-based scheduling: In this scheme, URLLC resources are reserved prior to the data scheduling. Two types of reservation schemes are *semi-static* and *dynamic* reservations. In the semi-static reservation scheme, the basestation infrequently broadcasts the configuration of the frame structure such as frequency numerology and service period. Whereas, in the dynamic reservation scheme, information on the URLLC resource is updated frequently using the control channel of a scheduled user. For example, if an eMBB packet consists of 14 symbols, then 10 symbols are used for the eMBB transmission and the rest are reserved for the URLLC service. Drawback of this approach is that when there is no URLLC transmission in the scheduled period, resources reserved for the URLLC service will be wasted. When compared to the semi-static reservation, the dynamic reservation requires additional control overhead to indicate the reservation information. Also, an overhead to ensure the reliability of the control signaling itself is unavoidable.

2.3.3 Solutions to the Coexistence Problem

As mentioned, the holy grail of 5G NR is to support diverse service categories and thus how to mitigate the performance degradation of interrupted services is an important issue in the physical layer design. While the flexible frame structure may ease off this problem, due to the implementation complexity and randomness of URLLC packet arrival, a more deliberate solution is required in real deployment scenarios. Two approaches discussed in the 5G NR standard meetings are *reactive* and *proactive* strategies.

The main idea is to give a priority to the URLLC packet while ensuring the re-

liability of the other channel interrupted by URLLC. Two approaches adopted in the 3GPP NR are as follows [47].

- Preemption indicator transmission: In this approach, the basestation indicates which resources are used for the URLLC transmission. Recalling that the URLLC packet is stretched in the frequency axis (see Fig. 3(a)), URLLC transmission will interrupt the whole system bandwidth and thus degrade all data channels in use. To notify this event to the scheduled users, the basestation broadcasts a *preemption indicator* consisting of time and/or frequency information of the interruption. This indicator helps users identify the reason for packet errors and what part of the packet is safe from the interruption.

- Re-transmission of selected codeblocks: When the ongoing service is interrupted by the URLLC transmission, part of the codebook affected by URLLC is re-transmitted. By transmitting *combining indicator* or *flush-out indicator*, the receiver can perform the soft symbol combining of the transmitted and re-transmitted codeblocks. One can further achieve better coding gain by lowering the code rate of the re-transmitted codeblock.

If the URLLC transmission occurs frequently, the efficiency of the reactive approach will be reduced due to the frequent re-transmissions. The main idea of the proactive strategy is to ensure the reliability of ongoing services while supporting the URLLC transmission. Specific schemes to support the proactive strategy include robustness improvement and service sharing.

- Robustness improvement: To reduce the initial packet error of non-URLLC packets, the basestation intentionally lowers the code rate by adding extra parity bits or employing outer error correction code to the non-URLLC packets [34]. Since the URLLC data transmission interferes non-URLLC packets, this approach can help reducing the packet error of non-URLLC packets.

- Resource sharing: This strategy supports the ongoing data channel and the URLLC data channel simultaneously. Multiple antenna or beam-domain techniques are employed for this purpose. Basically, the spatial layer (rank) of the channel are divided into two and then one part of the layers is used for eMBB and the other is used for URLLC. If there is no extra spatial layer, then the power-domain non-orthogonal transmission can be applied [49].

2.4 Short-sized Packet in LTE-Advanced Downlink

In this section, we briefly review the control-type data transmission (PDCCH of 4G LTE systems) to illustrate the short-sized packet transmission in the conventional systems. PDCCH carries essential information for the mobile terminal when it tries to transmit or receive the data. To be specific, PDCCH carries small-sized information needed to decode the data channel (e.g., resource assignment, modulation order, code rate). On top of these, cyclic redundancy check (CRC) is added to test the decoding error [16]. Since the CRC bit stream is scrambled with a user index (called radio network temporary identifier), the only scheduled user can pass the CRC test.

After the channel coding and symbol mapping,[2] the modulated symbol vector $\mathbf{s} \in \mathbb{C}^{N\times 1}$ is transmitted. The corresponding received vector $\mathbf{y} \in \mathbb{C}^{m\times 1}$ is given by

$$\mathbf{y} = \mathbf{HRs} + \mathbf{v}, \tag{2.2}$$

where $\mathbf{H} \in \mathbb{C}^{m\times m}$ is the diagonal matrix whose diagonal entry h_{ii} is the channel component for each resource, $\mathbf{v} \sim \mathcal{CN}(\mathbf{0}, \sigma_v^2\mathbf{I})$ is the additive Gaussian noise, and $\mathbf{R} \in \mathbb{C}^{m\times N}$ is the matrix describing the mapping between the symbol and resource element. For example, when one symbol is mapped to a single resource, $\mathbf{R}$ would be the identity matrix ($\mathbf{R} = \mathbf{I}$). Whereas, if two resources are assigned to one symbol

[2] e.g., convolution coding with rate $\frac{1}{3}$ and quadrature phase shift keying (QPSK) modulation are employed.

for the transmit diversity, then $\mathbf{R}$ would be $2N \times N$ matrix (e.g., if $N = 2$, then $\mathbf{R} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}^T$).

When one tries to improve the reliability with a small modification of current PDCCH, one can think of three options. The first option is to achieve the better coding gain by using lower code rate (i.e., $r = \frac{b}{2N} < r_{pdcch} = \frac{1}{3}$). This option is easy and straightforward but when the coded symbol length N increases, transmission and processing latency will also increase, resulting in the violation of the URLLC requirement. The second option is to use the multiple resources to achieve the diversity gain ($m > N$). By combining multiple versions of the same symbol at the receiver, reliability of the symbol can be improved. The problem of this approach is that a large portion of wireless resources are consumed in achieving the diversity gain so that there would be a severe degradation of the resource utilization efficiency. The third option is to reduce the size of control information b. By removing some of the scheduling parameters, resources used for the control channel can be saved. Even in this case, it is not possible to remove essential information (e.g., CRC and user index) so that one cannot expect a dramatic reduction of control information.

2.5 Sparse Vector Coding

2.5.1 SVC Encoding and Transmission

The key idea of the proposed SVC technique is to map the information into the positions of a sparse vector $\mathbf{s}$. Figuratively speaking, SVC encoding can be thought as marking a few dots to the empty table. As illustrated in Fig. 2.3, if we try to mark dots to two cells out of 9, then there would be $\binom{9}{2} = 36$ choices in total. In general, when we choose K out of N symbol positions, we can encode $\lfloor \log_2 \binom{N}{K} \rfloor$ bits of information. Example of one-to-one mapping between the information bit stream $\mathbf{w}$ and

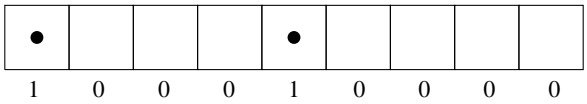


Figure 2.3: Metaphoric illustration of SVC encoding. Information is mapped into the position of a sparse vector.

Table 2.1: Example of mapping between the information $\mathbf{w}$ and the sparse vector $\mathbf{s}$

Input:
Size of sparse vector N, information vector $\mathbf{w}$
Output:
Sparse vector $\mathbf{s}$

$a := 0$
for $i = 2$ **to** N **do**
 for $j = 1$ **to** $i - 1$ **do**
 if $a = (\mathbf{w})_{(10)}$
 $\mathbf{s} := \left(2^i + 2^j\right)_{(2)}$
 end if
 $a := a + 1$
 end for
end for

$(\mathbf{w})_{(10)}$	$\mathbf{s}$
0	000011
1	000101
2	000110
3	001001
4	001010
5	001100
6	010001
7	010010
8	010100
9	011000
10	100001
	:
	:

Note: $(\mathbf{w})_{(10)}$ is decimal expression of binary vector $\mathbf{w}$ and $(w)_{(2)}$ is binary expression of integer w.

transmit sparse vector $\mathbf{s}$ is (see example in Table I)

$$\begin{array}{ccc} 0\,0\,0\,0\,0 & \longleftrightarrow & 0\,0\,0\,0\,0\,0\,0\,1\,1 \\ 0\,0\,0\,0\,1 & \longleftrightarrow & 0\,0\,0\,0\,0\,0\,1\,0\,1 \\ 0\,0\,0\,1\,1 & \longleftrightarrow & 0\,0\,0\,0\,0\,1\,0\,0\,1 \\ \vdots & \vdots & \vdots \\ \underbrace{1\,1\,1\,1\,1}_{b\text{-bit information } \mathbf{w}(b=5)} & \longleftrightarrow & \underbrace{1\,0\,0\,0\,0\,0\,0\,0\,1}_{K-\text{sparse vector } \mathbf{s}\ (K=2)} \end{array} .$$

After the sparse mapping, each nonzero element in $\mathbf{s}$ is spread into m resources using the codeword (spreading sequence) in the spreading codebook $\mathbf{C}$. While it is possible to allocate resources either in time, frequency axis or hybrid of these, in this work, we assume that they are allocated in the frequency axis (see Fig. 2.4(a)). This choice will not affect the system model but minimizes the transmission latency. As a

result of this spreading process called the *multi-code spreading*, the resource mapping matrix $\mathbf{R}$ in (1) is replaced with the codebook matrix $\mathbf{C} = [\mathbf{c}_1 \ \mathbf{c}_2 \ \cdots \ \mathbf{c}_N]$ where $\mathbf{c}_i = [c_{1i} \ c_{2i} \ \cdots \ c_{mi}]^T$ is the spreading sequence. For example, if the first and the third element of $\mathbf{s}$ are nonzero, then the transmit vector after spreading is

$$\begin{aligned} \mathbf{x} &= \mathbf{C}\mathbf{s} \\ &= s_1\mathbf{c}_1 + s_3\mathbf{c}_3. \end{aligned} \tag{2.3}$$

Since the positions of nonzero elements are chosen at random, the codebook matrix $\mathbf{C}$ should be designed such that the transmit vector $\mathbf{x}$ contains enough information to recover the sparse vector $\mathbf{s}$ irrespective of the selection of the nonzero positions. It has been shown that if entries of the codebook matrix $\mathbf{C}$ are generated at random, e.g., sampled from Gaussian or Bernoulli distribution, then an accurate recovery of the sparse vector is possible as long as $m = \mathcal{O}(K \log N)$ [59]. Example of $\mathbf{C}$ for $m = 5$ and $N = 10$, when elements of $\mathbf{c}_i$ are chosen from the Bernoulli distribution, is given by

$$\mathbf{C} = \frac{1}{\alpha}\begin{bmatrix} 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 \\ -1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix}, \tag{2.4}$$

where α is the normalization factor depending on the modulated symbols (see Chapter 2.5.2). The corresponding received signal $\mathbf{y}$ is

$$\begin{aligned} \mathbf{y} &= \mathbf{H}\mathbf{x} + \mathbf{v} \\ &= \begin{bmatrix} \mathbf{H}\mathbf{c}_1 & \mathbf{H}\mathbf{c}_3 \end{bmatrix}\begin{bmatrix} s_1 \\ s_3 \end{bmatrix} + \mathbf{v}. \end{aligned} \tag{2.5}$$

In general, the received vector $\mathbf{y}$ is given by

$$\begin{aligned}\mathbf{y} &= \mathbf{HCs} + \mathbf{v} \\ &= \begin{bmatrix} h_{11} & & \\ & \ddots & \\ & & h_{mm} \end{bmatrix} \begin{bmatrix} | & & | \\ \mathbf{c}_1 & \dots & \mathbf{c}_N \\ | & & | \end{bmatrix} \begin{bmatrix} s_1 \\ \vdots \\ s_N \end{bmatrix} + \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}.\end{aligned} \tag{2.6}$$

It is worth mentioning that an accurate recovery of the sparse vector $\mathbf{s}$ is unnecessary in SVC since the decoding of the information vector is achieved by the identification of nonzero positions, not the actual values of this vector. The fact that the decoding is done by the support[3] identification greatly simplifies the decoding process and also reduces the chance of decoding failure. The overall structure of the proposed SVC is depicted in Fig. 2.4(b).

The benefits of SVC can be summarized as follows; First, the transmission power of the data channel is concentrated on the nonzero elements of an information vector. Thus, when compared to the conventional system in which the transmission power is uniformly distributed across all symbols, effective transmit power per symbol is higher. Second, the SVC decoding process achieved by the sparse recovery algorithm lends itself to the test of decoding success/failure so that the CRC operation is unnecessary. This directly implies that the code rate of SVC can be made smaller than the rate of PDCCH. Specifically, when the number of resources used for the data channel is m and QPSK modulation is used, the code rate of SVC is $r_{svc} = \frac{b_i}{2m}$ (b_i is the number of information bits) and the code rate of PDCCH is $r_{pdcch} = \frac{(b_i+b_c)}{2m}\left(= \frac{1}{3}\right)$. If the number of CRC bits is $b_c = \beta b_i$ ($\beta > 0$), then $m = \frac{3}{2}(b_i + b_c) = \frac{3}{2}(b_i + \beta b_i)$. Thus, the code rate of SVC can be expressed in term of β as

$$r_{svc} = \frac{b_i}{2m} = \frac{1}{3(1+\beta)} < \frac{1}{3} = r_{pdcch}. \tag{2.7}$$

[3]Support is the set of nonzero elements. For example, if $\mathbf{s} = [0\ 0\ 1\ 0\ 0\ 1]$, the $\Omega_{\mathbf{s}} = \{3, 6\}$.

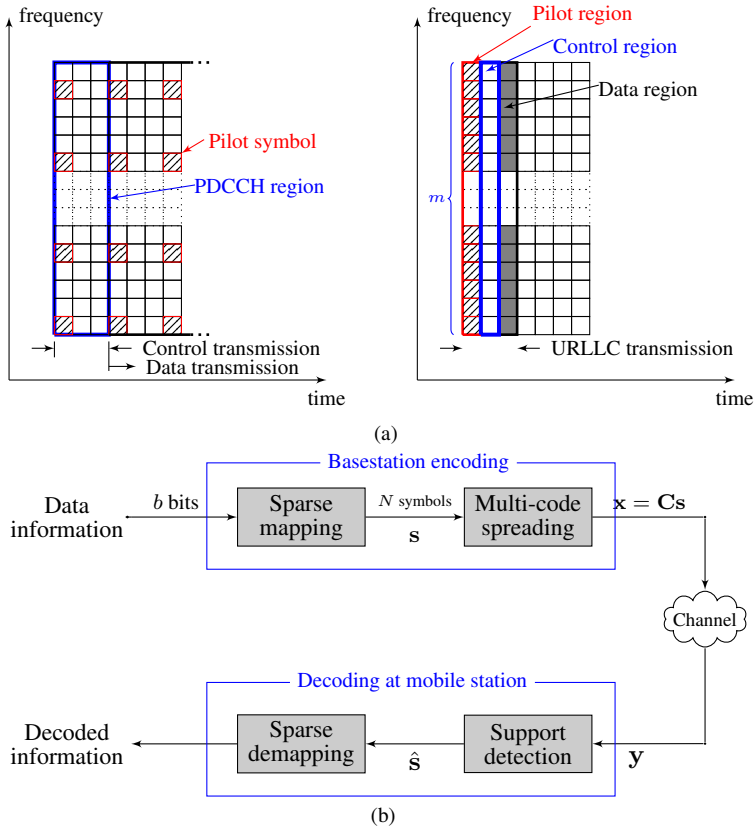


Figure 2.4: SVC-based packet transmission: (a) packet structure of 4G (left) and the URLLC packet (right) and (b) the block diagram for the proposed SVC technique.

Table 2.2: PDCCH versus SVC technique

	PDCCH	SVC technique
Coding (encoding/decoding)	Convolution code ($\frac{1}{3}$ rate) / Viterbi decoding	Sparse encoding / CS recovery algorithm
Transmission	Time/frequency mapping	Spreading in frequency direction
User identification	CRC scrambled with user index	User codebook $\mathbf{C}$
Resource overhead (L repetitions, QPSK)	$L\frac{3b}{2}$	Lm where m is the size of spreading length

Third, when m is sufficiently large, the basestation can easily assign the distinct codebook $\mathbf{C}$ for each user. This is because codebook matrices can be made near orthogonal by using a properly designed codebook generation mechanism.[4] For example, when $m = 42$ and the codebook is generated by the Bernoulli distribution, then there are 2^{42} different spreading sequences $\mathbf{c}_i$. Thus, if $N = 96$, then the basestation can support maximally $2^{35}(\approx \frac{2^{42}}{96})$ devices. Last but not least important benefit of SVC is that the implementation cost is small and the processing latency is low. Encoding is done via a simple injective mapping and spreading, which can be realized by the look-up table and addition/subtraction operations and the decoding is performed by the support detection and demapping. In particular, since the sparsity K is small and also known to the receiver, one can decode the SVC packet using a simple sparse recovery algorithm such as orthogonal matching pursuit (OMP) [72].[5] Comparisons of PDCCH and SVC are summarized in Table. 2.2.

2.5.2 SVC Decoding

As mentioned, the SVC decoding is done by the identification of the support and any sparse recovery algorithm can be employed for this purpose. In this work, we employ the greedy sparse recovery algorithm in the decoding of the SVC-encoded packet. Af-

[4]The correlation between two distinct columns of random matrix decreases exponentially as the dimension of a column increases (see, e.g., [17, Theorem 1]).

[5]Most of CS algorithm finds out the solution without the prior knowledge of the sparsity K. However, when K is known in advance, one can recover the sparse vector more accurately by using the sparse-aware recovery technique [25].

ter pre-multiplying the diagonal matrix constructed by the complex exponential $e^{j\measuredangle h}$, the modified received vector can be expressed as

$$\begin{aligned}\tilde{\mathbf{y}} &= \text{diag}\left[\exp(j\measuredangle h_{11})\ldots\exp(j\measuredangle h_{mm})\right]\mathbf{y}\\ &= \text{diag}\left[\tilde{\mathbf{h}}\right]\mathbf{C}\mathbf{s}+\tilde{\mathbf{v}}\\ &= \tilde{\mathbf{H}}\mathbf{C}\mathbf{s}+\tilde{\mathbf{v}}\\ &= \mathbf{\Phi}\mathbf{C}\mathbf{s}+\tilde{\mathbf{v}},\end{aligned} \tag{2.8}$$

where $\measuredangle h$ is the angle of h, $\tilde{\mathbf{h}} = [h_{11}e^{j\measuredangle h_{11}}\ldots h_{mm}e^{j\measuredangle h_{mm}}]$, $\tilde{\mathbf{H}} = \text{diag}\left[\tilde{\mathbf{h}}\right]$, and $\tilde{\mathbf{v}} = [\tilde{v}_1,\ldots,\tilde{v}_m]$ is the modified noise vector where $\tilde{v}_i = v_i e^{j\measuredangle h_{ii}}$. Since $\mathbf{s}$ has K nonzero elements, the modified received vector $\tilde{\mathbf{y}} = \tilde{\mathbf{H}}\mathbf{C}\mathbf{s}+\tilde{\mathbf{v}}$ can be expressed as a linear combination of K columns of $\mathbf{\Phi} = \tilde{\mathbf{H}}\mathbf{C}$ perturbed by the noise. In view of this, the main task of the SVC decoding is to identify the columns in $\mathbf{\Phi}$ participating in the modified received vector. In each iteration, greedy sparse recovery algorithm identifies one column of $\mathbf{\Phi}$ at a time using a greedy strategy. Specifically, a column of $\mathbf{\Phi}$ that is maximally correlated with the (modified) observation $\mathbf{r}^{j-1}$ is chosen. That is, an index of the nonzero column of $\mathbf{\Phi}$ chosen as j-th iteration is[6]

$$\omega_j = \arg\max_l |<\boldsymbol{\phi}_l, \mathbf{r}^{j-1}>|^2, \tag{2.9}$$

where $\mathbf{r}^{j-1} = \tilde{\mathbf{y}} - \mathbf{\Phi}_{\Omega_{\mathbf{s}}^{j-1}}\hat{\mathbf{s}}^{j-1}$ is the modified observation called the residual and $\hat{\mathbf{s}}^{j-1} = \mathbf{\Phi}^{\dagger}_{\Omega_{\mathbf{s}}^{j-1}}\tilde{\mathbf{y}}$ is the estimate of $\mathbf{s}$ at $(j-1)$-th iteration.[7]

A better way to improve the decoding performance is to use the maximum likelihood (ML) detection. Recalling that the sparsity K is known to both transmitter and receiver, the ML detection problem for the system model in (2.9) is

$$\mathbf{s}^* = \arg\max_{\|\mathbf{s}\|_0=K} P_r(\tilde{\mathbf{y}}|\mathbf{s},\tilde{\mathbf{H}},\mathbf{C}), \tag{2.10}$$

[6] If $\Omega = \{1,3\}$, then $\mathbf{\Phi}_\Omega = [\boldsymbol{\phi}_1\ \boldsymbol{\phi}_3]$.

[7] $\mathbf{\Phi}^{\dagger} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ is the pseudo-inverse of $\mathbf{\Phi}$.

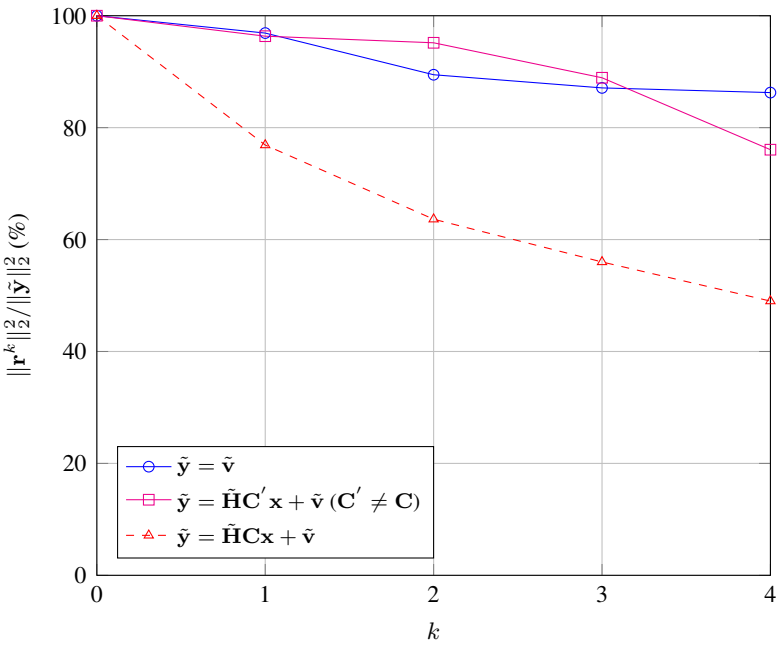


Figure 2.5: Snapshot of the ratio between residual magnitude $\|\mathbf{r}^k\|_2^2$ and $\|\tilde{\mathbf{y}}\|_2^2$ as a function of the number of iterations in the OMP algorithm. Signal-to-noise ratio (SNR) is set to 0 dB and the sparsity K is set to 4.

where $\|\mathbf{s}\|_0$ is the ℓ_0-norm of $\mathbf{s}$ counting the number of nonzero elements in $\mathbf{s}$. Since our goal is to find out the support of $\mathbf{s}$, we alternatively have

$$\Omega_{\mathbf{s}}^* = \arg \max_{|\Omega_{\mathbf{s}}|=K} P_r(\tilde{\mathbf{y}}|\Omega_{\mathbf{s}}, \tilde{\mathbf{H}}, \mathbf{C}), \tag{2.11}$$

where $|\Omega_{\mathbf{s}}|$ is the cardinality of the set $\Omega_{\mathbf{s}}$.

To find out the ML solution, we need to enumerate all possible combinations of candidate supports with cardinality K. Unfortunately, this exhaustive search would not be feasible for most practical scenarios. In this work, we instead use the multipath match pursuit (MMP) algorithm [18], a recently proposed near-ML sparse recovery algorithm, as a baseline for the SVC decoding. In a nutshell, MMP performs an efficient tree search to find out the near-ML solution to the original sparse vector. Unlike the single-path search algorithm, MMP selects multiple promising indices in each it-

eration. Specifically, each candidate chosen in an iteration brings forth multiple new child candidates. After finishing K iterations, candidate $\mathbf{s}^*$ having the smallest cost function among all candidates is chosen as the final output (i.e., $\mathbf{s}^* = \arg\min_{\hat{\mathbf{s}}} J(\hat{\mathbf{s}})$ where $J(\hat{\mathbf{s}}) = \|\tilde{\mathbf{y}} - \mathbf{\Phi}_{\Omega_{\hat{\mathbf{s}}}}\hat{\mathbf{s}}\|_2$). Due to the fact that many candidates are redundant and hence counted only once, an actual number of candidates examined in MMP are quite moderate [18].

One clear advantage of MMP, in the perspective of SVC decoding, is that it deteriorates the quality of incorrect candidate yet does not impose any estimation error to the correct one. This is because the quality of incorrect candidates gets worse due to the error propagation while no such behavior occurs to the correct one. In particular, since nonzero values of an original sparse vector $\mathbf{s}$ are known to the receiver,[8] no estimation error will be introduced in the correct candidate. We note that the computational complexity of the SVC decoding is marginal since the computational complexity of the greedy sparse recovery algorithm is directly proportional to the sparsity K.[9] Accordingly, the processing latency of SVC decoding can also be made sufficiently small. This is in contrast to the Viterbi or Turbo decoding algorithm in which the computational complexity is proportional to the length of a codeblock [19]. In Table 2.3 and Fig. 2.6, we present the comparison of computational complexity [58]. As a metric, we use the number of operation (e.g., add, multiplex) for decoding. Under the assumption in practical regime, we clear see that SVC outperforms the conventional CC below $b = 25$.

2.5.3 Identification of False Alarm

Overall, there are two kinds of false alarm events causing the decoding failure: 1) support detection when the basestation transmits information to the different user and 2)

[8] Since the goal of SVC decoding is to find out the nonzero positions of a sparse vector, we can pre-define values of the nonzero elements in $\mathbf{s}$.

[9] In each iteration, greedy sparse recovery algorithm performs three operations: support identification, nonzero element estimation, and residual update. Since the nonzero values are fixed and known in advance, estimation of the nonzero elements is unnecessary.

Table 2.3: Complexity of PDCCH and SVC technique

	PDCCH (LTE CC $(r, 1, M)$) [58]	SVC technique
The number of operations (OP)	$(r+7.5+r2^{r-1}+2^{M+1})(b+TB)$	$Km(N+1)$
Comparison	When $r = 3, M = 7, TB = 15$ than OP = $1000(b+15)$	When $K = 2, N = 2m$ than OP = $64b^2$

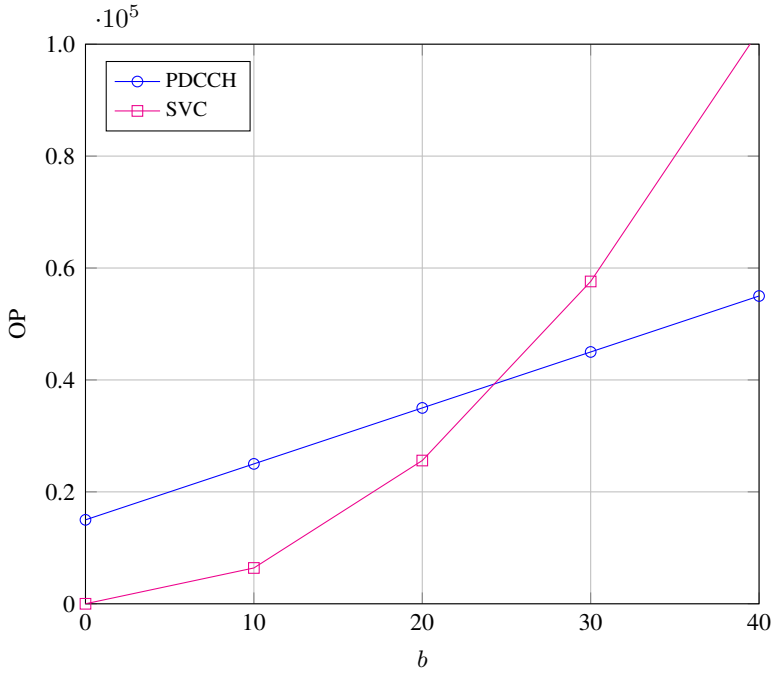


Figure 2.6: The number of operations as the function of information bit b

support detection when there is no transmission at the basestation. In order to prevent these events, we need to examine the residual magnitude in each iteration. Firstly, when a packet for the different user is received, the codebook between two distinct users would be different from each other so that the magnitude of the correlation μ_{ij} between two codewords, each being chosen from two district codebooks would be small. In this case, clearly, one cannot expect a substantial reduction in the residual magnitude. Secondly, when there is no transmission, the received vector will measure the noise only (i.e., $\tilde{\mathbf{y}} = \tilde{\mathbf{v}}$) and thus some column in $\mathbf{\Phi}$, say ϕ_l, will be added to the

Table 2.4: The proposed MMP-based SVC decoding algorithm

Input:
Measurement $\tilde{\mathbf{y}}$, sensing matrix $\mathbf{\Phi} = \tilde{\mathbf{H}}\mathbf{C}$, sparsity K,
number of expansion L, max number of search candidate l_{max},
stop threshold ϵ, detection threshold ε

Output:
Support set $\hat{\Omega}$

Initialization:
$l := 0$ (candidate order), $\rho := \infty$ (minimum magnitude of residual)

While: $l < l_{max}$ and $\epsilon < \rho$ **do**
 $l := l + 1$
 $\mathbf{r}^0 := \tilde{\mathbf{y}}$
 $[p_1, \dots, p_K] :=$ compute_p$_k(l, L)$ (*compute layer order*)
 for $k = 1$ **to** K **do** (*investigate l-th candidate*)
  $\tilde{\omega} :=$compute_$\omega(k, L)$ (*choose L best indices*)
  $\Omega_l^k := \Omega_l^{k-1} \cup \{\tilde{\omega}_{p_k}\}$ (*construct a path in k-th layer*)
  $\mathbf{r}^k := \tilde{\mathbf{y}} - \mathbf{\Phi}_{\Omega_l^k}\mathbf{s}^k$ (*update residual*)
  $\hat{\Omega}^k := \Omega_l^k$ (*update support set*)
 end for
 if $\|\mathbf{r}^K\|_2^2 < \rho$ **then** (*update the smallest residual*)
  $\rho := \|\mathbf{r}^K\|_2^2$
  if $\frac{\|\mathbf{r}^K\|_2^2}{\|\mathbf{y}\|_2^2} > 1 - \varepsilon$ **then** (*false-alarm identification*)
   $\hat{\Omega}^* := \mathbf{0}$
  end if
  $\hat{\Omega}^* := \hat{\Omega}^K$
 end if
end while
return $\hat{\Omega}^*$

function compute_p$_k(l, L)$
 $t := l - 1$
 for $k = 1$ **to** K **do**
  $p_k := \mod(t, L) + 1$
  $t :=$ floor(t/L)
 end for return $[p_1, \dots, p_K]$
end function

function compute_$\omega(k, L)$
 if $k =$ odd **then**
  return $\arg\max_{|\pi|=L} \|(\Re\langle\frac{\boldsymbol{\phi}^T}{\|\boldsymbol{\phi}\|_2}\mathbf{r}^{k-1}\rangle)_\pi\|_2^2$
 else
  return $\arg\max_{|\pi|=L} \|(\Im\langle\frac{\boldsymbol{\phi}^T}{\|\boldsymbol{\phi}\|_2}\mathbf{r}^{k-1}\rangle)_\pi\|_2^2$
 end if
end function

residual in each iteration $\mathbf{r}^i = \mathbf{r}^{i-1} - \boldsymbol{\phi}_l\hat{\mathbf{s}}_l$ (see Fig. 2.5). Based on these observations, we declare the decoding failure when the residual magnitude is outside of the confi-

dence interval of the pure noise contribution. We will say more about the selection of confidence interval Chapter 2.7.5.

The proposed MMP-based SVC decoding algorithm is summarized in Table 2.4.

2.6 SVC Performance Analysis

In this section, we analyze the decoding success probability of the SVC technique. As mentioned, decoding of the SVC-encoded packet is successful when all support elements are chosen by the sparse recovery algorithm so that we analyze the probability that the support is identified accurately. In our analysis, we assume that the greedy sparse recovery algorithm is used in the decoding process and analyze the lower bound of the success probability. For analytic simplicity, we initially consider $K=2$ scenario and then extend to the general case. Without loss of generality, we assume that p and q-th elements of $\mathbf{s}$ are nonzero (i.e., $\Omega_{\mathbf{s}} = \{p, q\}$). Further, by setting the information vector such that $s_p = 1$ and $s_q = j$, we can model the QPSK transmission.

Following lemmas will be useful in our analysis.

Lemma 1. *Consider the vector $\mathbf{a}_i$ ($i = 1, \cdots, N$) whose element is i.i.d. standard Gaussian. Then, $\frac{\mathbf{a}_i^T \mathbf{a}_j}{\|\mathbf{a}_i\|_2}$ is standard Gaussian. That is, $\frac{\mathbf{a}_i^T \mathbf{a}_j}{\|\mathbf{a}_i\|_2} \sim \mathcal{N}(0,1)$.*

Proof. Let $\mathbf{u}_i = \frac{\mathbf{a}_i}{\|\mathbf{a}_i\|_2}$, then it is clear that $\mathbf{u}_i$ is a random vector with zero mean and unit variance. Also, let $X = \frac{\mathbf{a}_i^T \mathbf{a}_j}{\|\mathbf{a}_i\|_2}$, then $X = \mathbf{u}_i^T \mathbf{a}_j$. One can easily show that X conditioned on any realization of $\mathbf{u}_i = u$ is a standard Gaussian. This is because $E[X|\mathbf{u}_i = u] = E[u^T \mathbf{a}_j] = u^T E[\mathbf{a}_j] = 0$ and $Var(X|\mathbf{u}_i = u) = E[u^T \mathbf{a}_j \mathbf{a}_j^T u] = u^T u = 1$. Further,

$$\begin{aligned} f_X(x) &= \int_u f_{X|\mathbf{u}_i}(x|u) f_{\mathbf{u}_i}(u) du \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \int_u f_{\mathbf{u}_i}(u) du \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \end{aligned} \tag{2.12}$$

which is the desired result. □

Lemma 2. *Consider the vector* $\tilde{\mathbf{h}} = [\tilde{h}_{11}\ \tilde{h}_{22}\ \cdots\ \tilde{h}_{mm}]^T$ *where* $\tilde{h}_{ii} = h_{ii}e^{j\angle h_{ii}}$. *The probability density function (PDF) of the* $\|\tilde{\mathbf{h}}\|_2^2$ *is Chi-squared distribution with*

$$f_{\|\tilde{\mathbf{h}}\|_2^2}(x) = \frac{x^{m-1}\exp(-x)}{\Gamma(m)}, \tag{2.13}$$

where $\Gamma(m) = (m-1)!$ *is the Gamma function and* $\mathrm{E}\left[\|\tilde{\mathbf{h}}\|_2^2\right] = m$.

Proof. $\|\tilde{\mathbf{h}}\|_2^2$ can be expressed as $\|\tilde{\mathbf{h}}\|_2^2 = \|\mathbf{h}\|_2^2 = \sum_{i=1}^m |h_{ii}|^2 = \sum_{i=1}^m (\Re(h_{ii})^2 + \Im(h_{ii})^2)$ where $\Re(c)$ and $\Im(c)$ are the real and imaginary part of c, respectively. Since $\Re(h_{ii}), \Im(h_{ii}) \sim \mathcal{N}(0, \frac{\sigma_v^2}{2})$, we can show after some manipulations that $2\|\tilde{\mathbf{h}}\|_2^2$ follows Chi-squared distribution with $2m$ DoF [24]. That is,

$$f_{2\|\tilde{\mathbf{h}}\|_2^2}(x) = \frac{x^{m-1}\exp\left(-\frac{x}{2}\right)}{2^m\Gamma(m)}. \tag{2.14}$$

Since $f_Z(z) = 2f_{2Z}(2z)$, we have

$$f_{\|\tilde{\mathbf{h}}\|_2^2}(x) = \frac{x^{m-1}\exp(-x)}{\Gamma(m)}. \tag{2.15}$$

□

Let $\mathcal{S}^j$ be the success probability that the support element is chosen in the j-th iteration. Since $K = 2$ and thus the required number of iterations to decode the information vector is two, the probability that the SVC packet is successfully decoded can be expressed as

$$\begin{aligned} P_{succ} &= \mathrm{P}(\Omega_{\mathbf{s}}^* = \Omega_{\mathbf{s}}) \\ &= \mathrm{P}\left(\mathcal{S}^1, \mathcal{S}^2\right) \\ &= \mathrm{P}\left(\mathcal{S}^2|\mathcal{S}^1\right)\mathrm{P}\left(\mathcal{S}^1\right). \end{aligned} \tag{2.16}$$

Our main result in this section is as follows.

Theorem 1. *The probability that the SVC-encoded packet is decoded successfully satisfies*

$$P_{succ} \geq \left(1 - \left(1 + \frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m} - \left(1 + \frac{1}{\sigma_v^2}\right)^{-m}\right)^{2N}, \tag{2.17}$$

where m is the number of measurements (resources), N is the size of sparse vectors, σ_v^2 is the noise variance, and $\mu^* = \max_{i \neq j} |\mu_{ij}|$ is the maximum absolute value of correlation between two distinct columns of $\mathbf{\Phi}$.

When m is sufficiently large, we approximately have

$$P_{succ} \gtrsim \left(1 - \left(1 + \frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m}\right)^{2N}. \tag{2.18}$$

Also, since the block error rate is $\text{PER}_{svc} = 1 - P_{succ}$, the upper bound of PER is

$$\text{PER}_{svc} \lesssim 1 - \left(1 - \left(1 + \frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m}\right)^{2N}. \tag{2.19}$$

In Fig. 2.6, we plot the PER performance of SVC as a function of SNR. To judge the effectiveness of Theorem 1, we perform the empirical simulation for $m = 42$, $N = 96$. From the empirical evaluations, we obtain that $\mu^* \approx 0.7$. When we apply this value to the upper bound in (2.19), we could observe that the obtained bound is tight across the board. To better understand the performance of SVC, we plot the PER as a function of μ^*, N, and m in Fig. 2.7, 2.8, and 2.9. First, when the maximum correlation μ^* decreases, we see that the PER gain increase sharply as shown in Fig. 2.7. For example, if μ^* is reduced from 0.4 to 0.2, we can achieve 1.5 dB gain at the target reliability point (PER = 10^{-5}). Next, we test the PER performance for various sparse vector dimensions in Fig. 2.8. Although the PER performance degrades with

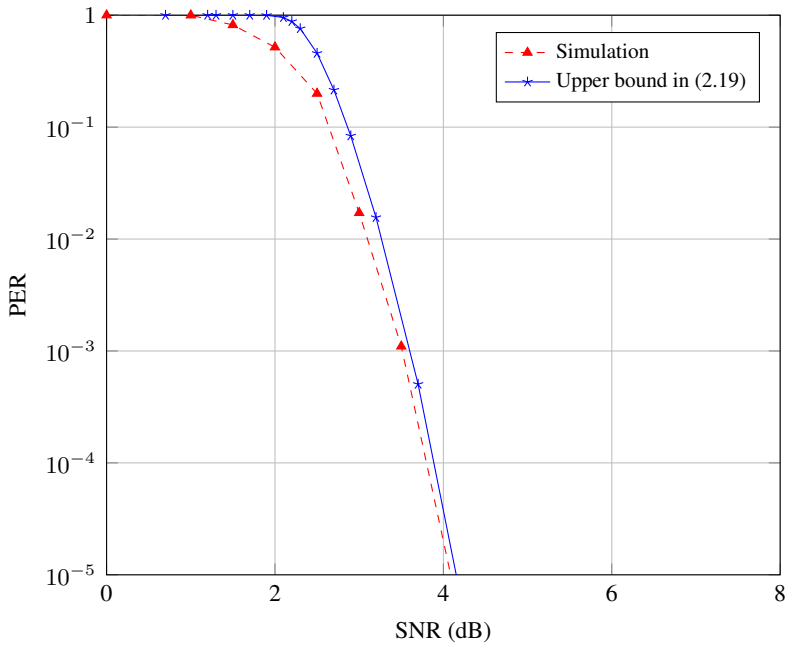


Figure 2.7: The exact PER and bound ($m = 42$, $N = 96$)

N, we see that the degradation is fairly graceful. Whereas, as shown in Fig. 2.9, the PER performance is quite sensitive to the number of measurements.

As a first step to prove Theorem 1, we analyze the success probability $\mathrm{P}\left(\mathcal{S}^1\right)$ for the first iteration.

Lemma 3. *Consider the received signal* $\tilde{\mathbf{y}} = \gamma\mathbf{\Phi}\mathbf{s} + \tilde{\mathbf{v}}$ *where* $\gamma = \frac{\sqrt{\mathsf{SNR}}}{\alpha}$, $\mathbf{\Phi} = [\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2\ \cdots\ \boldsymbol{\phi}_N]$, *and* $\boldsymbol{\phi}_i = [\tilde{h}_{11}c_{1i}\ \tilde{h}_{22}c_{2i}\ \cdots\ \tilde{h}_{mm}c_{mi}]^T$. *The probability that the support element is chosen in the first iteration satisfies*

$$\mathrm{P}(\mathcal{S}^1) \geq \left(1 - \left(1 + \frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m} - \left(1 + \frac{1}{\sigma_v^2}\right)^{-m}\right)^{N-1}. \tag{2.20}$$

Proof. As shown in Table 2.3, N decision statistics $\frac{\boldsymbol{\phi}_l^T}{\|\boldsymbol{\phi}_l\|_2}\mathbf{r}^{k-1}$ $(l = 1, \cdots, N)$ are computed in each iteration. For analytic simplicity, we take the real part of the decision statistic in the first iteration and the imaginary part in the second iteration.[10]

[10]This choice is suboptimal but simplifies the analysis.

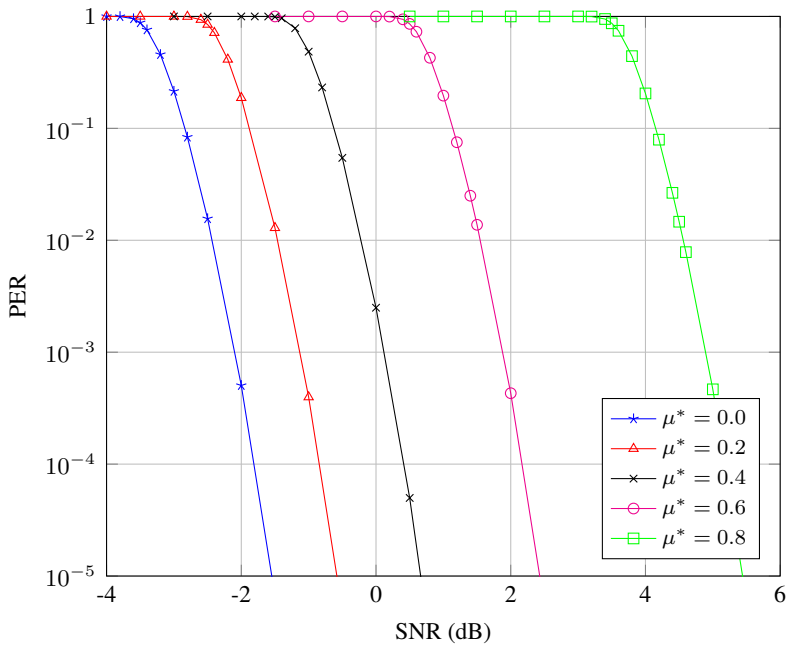


Figure 2.8: PER performancewith μ^* ($m = 42$, $N = 96$)

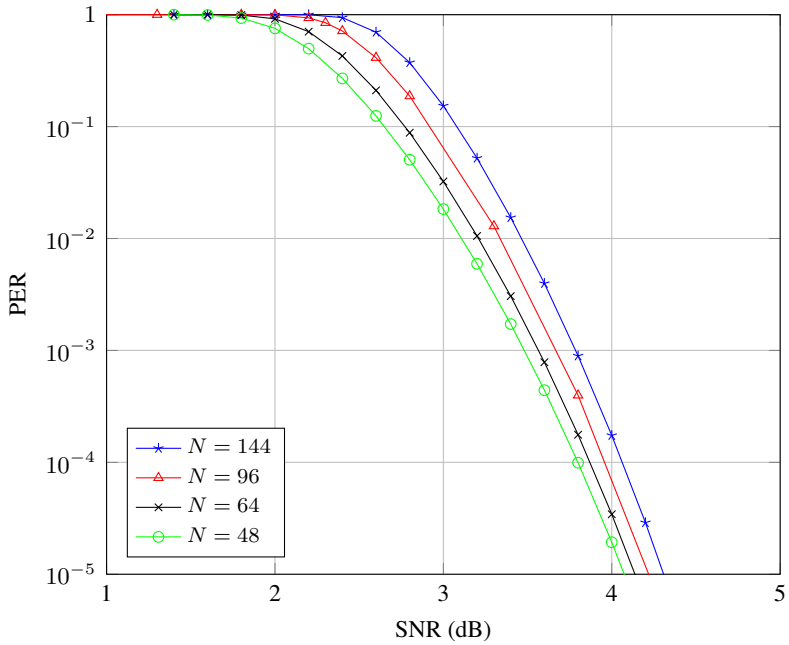


Figure 2.9: PER performance with N ($m = 42$, $\mu^* = 0.7$)

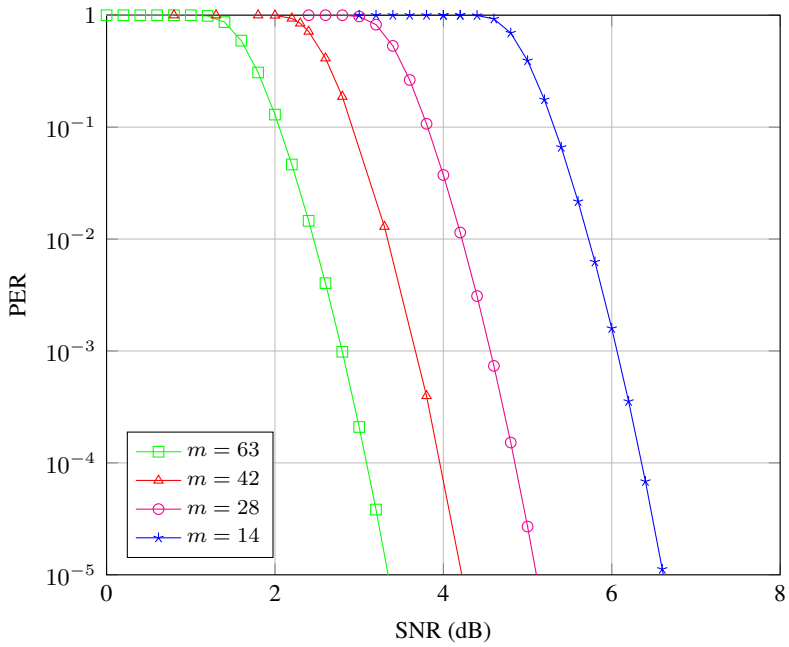


Figure 2.10: PER performance with m ($N = 96$, $\mu^* = 0.7$)

In order to identify the support element in the first iteration, we should have $\left|\Re\langle\frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2}, \mathbf{r}^0\rangle\right| \geq \max_i \left|\Re\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2}, \mathbf{r}^0\rangle\right|$ and thus the success probability for a given channel realization $\mathbf{h}$ is

$$\begin{aligned} \mathrm{P}(\mathcal{S}^1|\mathbf{h}) &= \mathrm{P}\left(\left|\Re\langle\frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2}, \mathbf{r}^0\rangle\right| \geq \max_i \left|\Re\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2}, \mathbf{r}^0\rangle\right|\right) \\ &= \prod_{i=1, i\neq p}^{N} \mathrm{P}\left(\left|\Re\langle\frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2}, \mathbf{r}^0\rangle\right| \geq \left|\Re\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2}, \mathbf{r}^0\rangle\right|\right), \end{aligned} \tag{2.21}$$

where $\langle\mathbf{a}, \mathbf{b}\rangle$ is the inner product between two vector $\mathbf{a}$ and $\mathbf{b}$. First, noting that $s_p = 1$ and $s_q = j$, we have

$$\begin{aligned} \langle\frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2}, \mathbf{r}^0\rangle &= \langle\frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2}, \boldsymbol{\phi}_p s_p + \boldsymbol{\phi}_q s_q + \tilde{\mathbf{v}}\rangle \\ &= \|\tilde{\mathbf{h}}\|_2 + j\|\tilde{\mathbf{h}}\|_2 \mu_{qp} + \frac{\boldsymbol{\phi}_l^T}{\|\boldsymbol{\phi}_l\|_2}\tilde{\mathbf{v}}, \end{aligned} \tag{2.22}$$

where the equality follows from the followings:

Noting that $\boldsymbol{\phi}_i = [\tilde{h}_{11}c_{1i}\ \tilde{h}_{22}c_{2i}\ \cdots\ \tilde{h}_{mm}c_{mi}]^T$ and $\mu_{ij} = \frac{\boldsymbol{\phi}_i^T\boldsymbol{\phi}_j}{\|\tilde{\mathbf{h}}\|_2^2}$, we have

$$\langle \frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2}, \boldsymbol{\phi}_j \rangle \quad = \quad \frac{\boldsymbol{\phi}_i^T\boldsymbol{\phi}_j}{\|\boldsymbol{\phi}_i\|_2}. \tag{2.23}$$

Since $\|\boldsymbol{\phi}_i\|_2 = \sqrt{|\tilde{h}_{11}c_{1i}|^2 + \cdots + |\tilde{h}_{mm}c_{mi}|^2} = \|\tilde{\mathbf{h}}\|_2$, we have

$$\begin{aligned} \langle \frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2}, \boldsymbol{\phi}_j \rangle &= \|\tilde{\mathbf{h}}\|_2 \frac{\boldsymbol{\phi}_i^T\boldsymbol{\phi}_j}{\|\tilde{\mathbf{h}}\|_2^2} \\ &= \|\tilde{\mathbf{h}}\|_2 \mu_{ij}. \end{aligned} \tag{2.24}$$

In particular, $i = j$, $\mu_{ij} = 1$ and thus

$$\langle \frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2}, \boldsymbol{\phi}_i \rangle \quad = \quad \|\tilde{\mathbf{h}}\|_2. \tag{2.25}$$

From (2.24) and (2.25), we have $\langle \frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2}, \boldsymbol{\phi}_j \rangle = \begin{cases} \|\tilde{\mathbf{h}}\|_2 & \text{for } i = j \\ \|\tilde{\mathbf{h}}\|_2 \mu_{ij} & \text{for } i \neq j. \end{cases}$

$$\langle \frac{\boldsymbol{\phi}_k}{\|\boldsymbol{\phi}_k\|_2}, \boldsymbol{\phi}_l \rangle \quad = \quad \begin{cases} \|\tilde{\mathbf{h}}\|_2 & \text{for } k = l \\ \|\tilde{\mathbf{h}}\|_2 \mu_{kl} & \text{for } k \neq l \end{cases}. \tag{2.26}$$

Let $z_p = \Re\left(\frac{\boldsymbol{\phi}_p^T}{\|\boldsymbol{\phi}_p\|_2}\tilde{\mathbf{v}}\right)$, then

$$\Re\langle \frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2}, \mathbf{r}^0 \rangle \quad = \quad \|\tilde{\mathbf{h}}\|_2 + z_p. \tag{2.27}$$

In a similar way, we have

$$\Re\langle \frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2}, \mathbf{r}^0 \rangle \quad = \quad \|\tilde{\mathbf{h}}\|_2 \mu_{ip} + z_i, \tag{2.28}$$

and hence

$$\begin{aligned}
&\mathrm{P}\left(\left|\Re\langle\frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2},\mathbf{r}^0\rangle\right| \geq \left|\Re\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2},\mathbf{r}^0\rangle\right|\right)\\
&= \mathrm{P}\left(\left|\|\tilde{\mathbf{h}}\|_2+z_p\right| \geq \left|\|\tilde{\mathbf{h}}\|_2\mu_{il}+z_i\right|\right)\\
&\overset{(a)}{=} \mathrm{P}\left(\|\tilde{\mathbf{h}}\|_2+z_p > \left|\|\tilde{\mathbf{h}}\|_2\mu_{ip}+z_i\right|\right)\mathrm{P}\left(\|\tilde{\mathbf{h}}\|_2+z_p>0\right)\\
&\quad +\mathrm{P}\left(-\|\tilde{\mathbf{h}}\|_2-z_p > \left|\|\tilde{\mathbf{h}}\|_2\mu_{ip}+z_i\right|\right)\mathrm{P}\left(\|\tilde{\mathbf{h}}\|_2+z_p<0\right)\\
&\geq \mathrm{P}\left(\|\tilde{\mathbf{h}}\|_2+z_p > \|\tilde{\mathbf{h}}\|_2|\mu_{ip}|+|z_i|\right)\mathrm{P}\left(\|\tilde{\mathbf{h}}\|_2+z_p>0\right)\\
&\geq \mathrm{P}\left(\|\tilde{\mathbf{h}}\|_2+z_p > \mu^*\|\tilde{\mathbf{h}}\|_2+|z_i|\right)\mathrm{P}\left(\|\tilde{\mathbf{h}}\|_2+z_p>0\right), \qquad (2.29)
\end{aligned}$$

where (a) follows from

$$\mathrm{P}\left(|A|\geq|B|\right) = \mathrm{P}\left(A>|B|\right)\mathrm{P}\left(A>0\right)+\mathrm{P}\left(-A>|B|\right)\mathrm{P}\left(A<0\right). \qquad (2.30)$$

Since $z_i\sim\mathcal{N}(0,\frac{\sigma_v^2}{2})$, the second term in (2.29) is lower bounded as

$$\begin{aligned}
\mathrm{P}\left(\|\tilde{\mathbf{h}}\|_2+z_p>0\right) &= \mathrm{P}\left(z_p>-\|\tilde{\mathbf{h}}\|_2\right)\\
&= 1-Q\left(-\frac{\|\tilde{\mathbf{h}}\|_2}{\frac{\sigma_v}{\sqrt{2}}}\right)\\
&\geq 1-\exp\left(-\frac{\|\tilde{\mathbf{h}}\|_2^2}{\sigma_v^2}\right), \qquad (2.31)
\end{aligned}$$

where the last inequality follows from $Q(x)\leq\exp\left(-\frac{x^2}{2}\right)$. In a similar way, the first term in (2.29) is lower bounded as

$$\begin{aligned}
\mathrm{P}\left(\|\tilde{\mathbf{h}}\|_2+z_p>\mu^*\|\tilde{\mathbf{h}}\|_2+|z_i|\right) &= 1-\mathrm{P}\left(|z_i|-z_p\geq(1-\mu^*)\|\tilde{\mathbf{h}}\|_2\right)\\
&= 1-\mathrm{P}\left(z_i-z_p\geq(1-\mu^*)\|\tilde{\mathbf{h}}\|_2\right)\mathrm{P}\left(z_i>0\right)\\
&\quad -\mathrm{P}\left(-z_i-z_p\geq(1-\mu^*)\|\tilde{\mathbf{h}}\|_2\right)\mathrm{P}\left(z_i<0\right)\\
&\overset{(a)}{=} 1-2\mathrm{P}\left(z_i-z_p\geq(1-\mu^*)\|\tilde{\mathbf{h}}\|_2\right)\mathrm{P}\left(z_i>0\right)
\end{aligned}$$

$$\begin{aligned} &\overset{(b)}{\geq} \quad 1 - Q\left(-\frac{\|\tilde{\mathbf{h}}\|_2(1-\mu^*)}{\sigma_v}\right) \\ &\geq \quad 1 - \exp\left(-\frac{\|\tilde{\mathbf{h}}\|_2^2(1-\mu^*)^2}{2{\sigma_v}^2}\right), \end{aligned} \tag{2.32}$$

where (a) is because $-z_i \sim \mathcal{N}(0, \frac{\sigma_v^2}{2})$ and (b) is because $z_i - z_p \sim \mathcal{N}(0, \sigma_v^2)$. By plugging (2.31) and (2.32) into (2.29), we have

$$\begin{aligned} &\mathrm{P}\left(\left|\Re\langle\frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2}, \mathbf{r}^0\rangle\right| \geq \left|\Re\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2}, \mathbf{r}^0\rangle\right|\right) \\ &\geq \quad \left(1 - \exp\left(-\frac{\|\tilde{\mathbf{h}}\|_2^2(1-\mu^*)^2}{2{\sigma_v}^2}\right)\right)\left(1 - \exp\left(-\frac{\|\tilde{\mathbf{h}}\|_2^2}{\sigma_v^2}\right)\right) \\ &\geq \quad 1 - \exp\left(-\frac{\|\tilde{\mathbf{h}}\|_2^2(1-\mu^*)^2}{2{\sigma_v}^2}\right) - \exp\left(-\frac{\|\tilde{\mathbf{h}}\|_2^2}{\sigma_v^2}\right). \end{aligned} \tag{2.33}$$

Note that $\mathrm{P}(\mathcal{S}^1|\mathbf{h})$ in (2.21) is the success probability in the first iteration for a given channel realization $\mathbf{h}$. In order to obtain the unconditional probability, we need to take expectation with respect to the channel $\mathbf{h}$. That is,

$$\mathrm{P}(\mathcal{S}^1) = \int \mathrm{P}(\mathcal{S}^1|\mathbf{h}) f_{\mathbf{h}}(x)dx = \mathrm{E}_{\mathbf{h}}\left[\mathrm{P}(\mathcal{S}^1|\mathbf{h})\right]. \tag{2.34}$$

Thus,

$$\begin{aligned} &\mathrm{P}(\mathcal{S}^1) \\ &= \quad \mathrm{E}_{\mathbf{h}}\left[\prod_{i=1, i\neq p}^{N} \mathrm{P}\left(\left|\Re\langle\frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2}, \mathbf{r}^0\rangle\right| \geq \left|\Re\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2}, \mathbf{r}^0\rangle\right|\right) \mid \mathbf{h}\right] \\ &= \quad \prod_{i=1, i\neq p}^{N} \mathrm{E}_{\mathbf{h}}\left[\mathrm{P}\left(\left|\Re\langle\frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2}, \mathbf{r}^0\rangle\right| \geq \left|\Re\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2}, \mathbf{r}^0\rangle\right|\right) \mid \mathbf{h}\right] \\ &\geq \quad \prod_{i=1, i\neq p}^{N} \mathrm{E}_{\mathbf{h}}\left[1 - \exp\left(-\frac{\|\tilde{\mathbf{h}}\|_2^2(1-\mu^*)^2}{2{\sigma_v}^2}\right) - \exp\left(-\frac{\|\tilde{\mathbf{h}}\|_2^2}{\sigma_v^2}\right) \mid \mathbf{h}\right] \end{aligned}$$

$$= \prod_{i=1,i\neq p}^{N} \left(1 - \mathrm{E}_{\mathbf{h}}\left[\exp\left(-\frac{\|\tilde{\mathbf{h}}\|_2^2(1-\mu^*)^2}{2{\sigma_v}^2}\right) \Big| \mathbf{h}\right] - \mathrm{E}_{\mathbf{h}}\left[\exp\left(-\frac{\|\tilde{\mathbf{h}}\|_2^2}{\sigma_v^2}\right) \Big| \mathbf{h}\right]\right). \tag{2.35}$$

Since $\|\tilde{\mathbf{h}}\|_2^2$ follows Chi-squared distribution with $2m$ DoF (see Lemma 2), we have

$$\begin{aligned} \mathrm{E}_{\tilde{\mathbf{h}}}\left[\exp\left(-\frac{\|\tilde{\mathbf{h}}\|_2^2}{{\sigma_v}^2}\right) \Big| \mathbf{h}\right] &= \int_0^\infty \exp\left(-\frac{x}{{\sigma_v}^2}\right)\frac{x^{m-1}\exp(-x)}{(m-1)!}dx, \\ &= \frac{1}{\left(\frac{1}{\sigma_v^2}+1\right)^m}, \end{aligned} \tag{2.36}$$

where the equality follows from $\int_0^\infty x^n \exp(-ax)dx = \frac{n!}{a^{n+1}}$ for $n = 0, 1, 2, ..., a > 0$.

In a similar way, we have

$$\mathrm{E}_{\mathbf{h}}\left[\exp\left(-\frac{\|\tilde{\mathbf{h}}\|_2^2(1-\mu^*)^2}{2{\sigma_v}^2}\right) \Big| \mathbf{h}\right] = \left(1+\frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m}. \tag{2.37}$$

Finally, by plugging (2.36) and (2.37) into (2.35), we obtain the lower bound of $\mathrm{P}(\mathcal{S}^1)$ as

$$\begin{aligned} \mathrm{P}(\mathcal{S}^1) &= \mathrm{E}_{\mathbf{h}}\left[\prod_{i=1,i\neq p}^{N} \mathrm{P}\left(\left|\Re\langle\frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2}, \mathbf{r}^0\rangle\right| \geq \left|\Re\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2}, \mathbf{r}^0\rangle\right|\right) \Big| \mathbf{h}\right] \\ &= \prod_{i=1,i\neq p}^{N}\left(1 - \left(1+\frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m} - \left(1+\frac{1}{\sigma_v^2}\right)^{-m}\right) \\ &\geq \left(1 - \left(1+\frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m} - \left(1+\frac{1}{\sigma_v^2}\right)^{-m}\right)^{N-1}. \end{aligned} \tag{2.38}$$

□

We now move to the success probability for the second iteration when the first iteration is successful.

Lemma 4. *The probability that the support element is chosen at the second iteration under the condition that the first iteration is successful satisfies*

$$\mathrm{P}\left(\mathcal{S}^2|\mathcal{S}^1\right) \geq \left(1-\left(1+\frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m}-\left(1+\frac{1}{\sigma_v^2}\right)^{-m}\right)^{N-2}. \tag{2.39}$$

Proof. When the first iteration is successful, the residual $\mathbf{r}^1$ can be expressed as

$$\begin{aligned}\mathbf{r}^1 &= \mathbf{r}^0-\mathbf{\Phi}_{\Omega_{\mathbf{s}}^1}\hat{\mathbf{s}}^1\\ &\overset{(a)}{=} \mathbf{r}^0-\boldsymbol{\phi}_p s_p\\ &= \boldsymbol{\phi}_q s_q+\tilde{\mathbf{v}},\end{aligned} \tag{2.40}$$

where (a) is because the transmit symbols are known in advance ($\hat{\mathbf{s}}^1 = s_p$). After taking similar steps to Lemma 3, one can show that $\mathrm{P}\left(\mathcal{S}^2|\mathcal{S}^1\right)$ satisfies (we skip the detailed steps for brevity)

$$\begin{aligned}\mathrm{P}\left(\mathcal{S}^2|\mathcal{S}^1\right) &= \mathrm{P}\left(\left|\Im\langle\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|_2},\mathbf{r}^1\rangle\right|\geq\max_i\left|\Im\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2},\mathbf{r}^1\rangle\right|\right)\\ &= \prod_{i=1,i\neq p,q}^{N}\mathrm{P}\left(\left|\Im\langle\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|_2},\mathbf{r}^1\rangle\right|\geq\left|\Im\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2},\mathbf{r}^1\rangle\right|\right)\\ &\geq \left(1-\left(1+\frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m}-\left(1+\frac{1}{\sigma_v^2}\right)^{-m}\right)^{N-2}.\end{aligned} \tag{2.41}$$

□

It is worth mentioning that the lower bounds of $\mathrm{P}\left(\mathcal{S}^1\right)$ and $\mathrm{P}\left(\mathcal{S}^2|\mathcal{S}^1\right)$ have the same form except for the exponent. We are now ready to prove the main theorem.

Proof of Theorem 1. By combining Lemma 3 and 4, we can obtain the lower bound of the success probability P_{succ} as

$$P_{succ} = \mathrm{P}\left(\mathcal{S}^2|\mathcal{S}^1\right)\mathrm{P}\left(\mathcal{S}^1\right)$$

$$
\begin{aligned}
&= \mathrm{P}\left(\left|\Im\langle\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|_2},\mathbf{r}^1\rangle\right| \geq \max_i \left|\Im\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2},\mathbf{r}^1\rangle\right|\right) \\
&\quad \times \mathrm{P}\left(\left|\Re\langle\frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2},\mathbf{r}^0\rangle\right| \geq \max_i \left|\Re\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2},\mathbf{r}^0\rangle\right|\right) \\
&= \prod_{i=1,i\neq p,q}^{N} \mathrm{P}\left(\left|\Im\langle\frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|_2},\mathbf{r}^1\rangle\right| \geq \left|\Im\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2},\mathbf{r}^1\rangle\right|\right) \\
&\quad \times \prod_{i=1,i\neq p}^{N} \mathrm{P}\left(\left|\Re\langle\frac{\boldsymbol{\phi}_p}{\|\boldsymbol{\phi}_p\|_2},\mathbf{r}^0\rangle\right| \geq \left|\Re\langle\frac{\boldsymbol{\phi}_i}{\|\boldsymbol{\phi}_i\|_2},\mathbf{r}^0\rangle\right|\right) \\
&\geq \left(1-\left(1+\frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m}-\left(1+\frac{1}{\sigma_v^2}\right)^{-m}\right)^{(N-2)+(N-1)} \\
&\geq \left(1-\left(1+\frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m}-\left(1+\frac{1}{\sigma_v^2}\right)^{-m}\right)^{2N}, \qquad (2.42)
\end{aligned}
$$

which completes the proof. □

Finally, we present the decoding success probability bound for general sparsity K.

Theorem 2. *The probability that the SVC-encoded packet can be successfully decoded for a given K satisfies*

$$
P_{succ} \gtrsim \left(1-\left(1+\frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m}\right)^{KN}. \qquad (2.43)
$$

Proof. The success probability P_{succ} is expressed as

$$
\begin{aligned}
P_{succ} &= \mathrm{P}\left(\mathcal{S}^1,\mathcal{S}^2,\cdots,\mathcal{S}^K\right) \\
&= \mathrm{P}\left(\mathcal{S}^K|\mathcal{S}^{K-1},\cdots,\mathcal{S}^1\right)\cdots\mathrm{P}\left(\mathcal{S}^2|\mathcal{S}^1\right)\mathrm{P}\left(\mathcal{S}^1\right) \\
&\geq \left(1-\left(1+\frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m}-\left(1+\frac{1}{\sigma_v^2}\right)^{-m}\right)^{(N-K)+\cdots+(N-2)+(N-1)} \\
&\geq \left(1-\left(1+\frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m}-\left(1+\frac{1}{\sigma_v^2}\right)^{-m}\right)^{KN}. \qquad (2.44)
\end{aligned}
$$

Since the proof is similar to the proof of Theorem 1, we skip the detailed steps. □

If $m \gg 1$, we approximately have

$$P_{succ} \gtrsim \left(1 - \left(1 + \frac{(1-\mu^*)^2}{\sigma_v^2}\right)^{-m}\right)^{KN}. \tag{2.45}$$

It is clear from (2.45) that the decoding success probability decreases when the information vector is less sparse (i.e., K is large), which matches with our expectation.

2.7 Implementation Issues

In this section, we discuss the implementation issues including codebook design, high-order modulation, diversity transmission, pilot-less transmission, and threshold selection to prevent the false alarm event.

2.7.1 Codebook Design

From our analysis in the previous section, we clearly see that codebook with small correlation is important to improve the decoding success probability. As mentioned, as m increases, the correlation between two randomly generated codewords decreases, and thus we can basically use any kind of randomly generated sequence. For example, if we use the Bernoulli random matrix, then the maximum correlation satisfies $\mu^* \leq \sqrt{4m^{-1} \ln \frac{N}{\delta}}$ with probability exceeding $1 - \delta^2$ [20].

Instead of relying on the random sequence, we can alternatively use the deterministic sequences. Well-known deterministic sequences include chirps, BCH, DFT, and second-order Reed-Muller (SORM) sequences [21]. For example, SORM is a sequence designed to generate low correlation sequences. SORM of length 2^m is defined as

$$\phi_{\mathbf{P},\mathbf{b}(\mathbf{a})} = \frac{(-1)^{w(\mathbf{b})}}{\sqrt{2^p}} i^{(2\mathbf{b}+\mathbf{P}\mathbf{a})^T \mathbf{a}}, \tag{2.46}$$

where $\mathbf{P}$ is a $d \times d$ binary symmetric matrix, $\mathbf{a} = [a_0\ a_1\ \cdots\ a_{d-1}]^T$ and $\mathbf{b} = [b_0\ b_1\ \cdots\ b_{d-1}]^T$ are binary vectors in $\mathbb{Z}_2^d$, and $w(\mathbf{b})$ is the weight (number of ones)

of $\mathbf{b}$. The corresponding SORM matrix can be expressed as

$$\Phi_{rm} = \begin{bmatrix} \mathbf{U}_{\mathbf{P}_1} & \mathbf{U}_{\mathbf{P}_2} & \cdots & \mathbf{U}_{\mathbf{P}_{2^{d(d-1)/2}}} \end{bmatrix}, \tag{2.47}$$

where $\mathbf{U}_{\mathbf{P}_j}$ is the $2^d \times 2^d$ orthogonal matrix whose columns are the SORM sequences. The maximum correlation ν^* of the SORM sequence is

$$\nu^* = \begin{cases} \frac{1}{\sqrt{2^l}}, & l = \text{rank}(\mathbf{P}_i - \mathbf{P}_j) \\ \frac{1}{\sqrt{m}}, & l = d \end{cases}. \tag{2.48}$$

For example, if $m = 64$ and $l = d$, then $\nu^* = 0.125$. The benefit of using SORM sequence is that the correlation between any two codewords is a constant and thus the performance variation can be minimized.

2.7.2 High-order Modulation

Since the ensuring reliability is the top priority in URLLC, QPSK modulation would be the popular option in practice. In order to use the QPSK modulation in SVC, we set one of the nonzero entries in $\mathbf{s}$ to 1 and the other to j. For example, if the nonzero positions are 5 and 7, then we set $\mathbf{s} = [0\ 0\ 0\ 0\ 1\ 0\ j\ 0\ 0\ 0]^T$ and thus the transmit vector $\mathbf{x}$ can be expressed as

$$\mathbf{x} = 1\mathbf{c}_5 + j\mathbf{c}_7. \tag{2.49}$$

From (2.49), we can easily see that elements of the transmit vector $\mathbf{x}$ are mapped to the QPSK symbol (i.e., $x_i \in \{1{+}j, 1{-}j, {-}1{+}j, {-}1{-}j\}$). It is worth mentioning that one additional bit can be encoded by differentiating two possible choices (i.e., $[1, j]$ and $[j, 1]$). However, this choice will increase the computational overhead of the decoding algorithm and also degrade the performance little bit. When the higher sparsity is used, this mapping can be readily extended to the high order modulation (e.g., $K = 4$ for 16-QAM and $K = 6$ for 64-QAM). Specifically, if $K = 4$, we map the element

in $\mathbf{x}$ to the 16-QAM symbol by setting two of the nonzero entries to 1, 2 and the remaining nonzero entries to j, $2j$. In a similar way, if $K = 6$, then we can transmit 64-QAM symbols by setting three of the nonzero entries to 1, 2, 3 and the remaining ones to j, $2j$, and $3j$. The normalization factor (α in (2.4)) corresponding M-QAM is $\alpha = \sqrt{\frac{2(M-1)}{3}}$.

2.7.3 Diversity Transmission

One can easily integrate the diversity scheme to SVC to further improve the reliability. The first option is to use the frequency diversity in which the SVC-encoded packet is repeated L times in L distinct frequency bands. The benefit of the frequency diversity is that the diversity gain can be achieved without increasing the transmission latency. Specifically, by applying the maximal-ratio combining at the receiver for the same symbol of the repeated packets, effective SNR can be increased and thus the PER performance can be improved [67]. For example, when the SVC-encoded packet is repeated for $L = 8$ times, due to the power gain of the combined symbol, the required SNR to achieve the desired URLLC performance (e.g., 10^{-5} PER) can be reduced from 3 dB to $3 - 10\log_{10}(L) =$ -6dB in AWGN environments. On top of the frequency diversity, other diversity schemes such as time, antenna, and space diversity can also be easily incorporated.

2.7.4 SVC without Pilot

When the channel is constant or channel variation is very small (i.e., $h \approx$ const.), which is true for mobile devices under static or slowly moving environments, decoding of the SVC packet can be performed without pilot transmission, resulting in a substantial reduction of the resources, transmission power, receiver processing time, and also implementation cost. In fact, since the packet length is smaller than channel coherence time, this assumption hold true in many realistic scenarios. Pilot-less transmission is done by slightly modifying the system model such that the system matrix equals the

codebook $\mathbf{C}$ and the sparse vector contains the channel component ($\mathbf{s}^{'} = h\mathbf{s}$). That is,

$$\begin{aligned} \mathbf{y} &= \mathbf{HCs} + \mathbf{v} \\ &= \mathbf{Cs}^{'} + \mathbf{v} \\ &= \begin{bmatrix} | & & | \\ \mathbf{c}_1 & \dots & \mathbf{c}_N \\ | & & | \end{bmatrix} \begin{bmatrix} hs_1 \\ \vdots \\ hs_N \end{bmatrix} + \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}. \end{aligned} \tag{2.50}$$

Recalling that the goal of the SVC decoding is to find out the nonzero positions of $\mathbf{s}^{'}$ vector, we can perform the decoding without the channel knowledge. When the channel variation is flat in the frequency axis, tall packet structure (stretched in frequency axis) is preferred. Whereas, if the channel variation is very small in time-domain, horizontal packet (stretched in time axis) would be a good choice.

2.7.5 Threshold to Prevent False Alarm Event

To distinguish the false alarm event from the normal decoding process, we examine the probability that the residual after the sparse recovery algorithm is not pure noise. In fact, if the SVC decoding is finished successfully, the residual contains the noise contribution only ($\mathbf{r}^K = \mathbf{v}$) so that the residual power $\|\mathbf{r}^K\|_2^2$ can be readily modeled as a Chi-squared random variable with $2m$ degree of freedom. Naturally, one can reject this hypothesis if the residual power is too large and lies outside of the pre-defined confidence interval. In other words, if $\|\mathbf{r}^K\|_2^2 > F^{-1}_{\|\mathbf{v}\|_2^2}(1 - P_{th})$ where P_{th} is the pre-defined probability threshold (e.g., $P_{th} = 0.01$) and $F^{-1}_{\|\mathbf{v}\|_2^2}$ is the inverse cumulative distribution function of Chi-squared random variable, then we declare the hypothesis is not true (i.e., decoding is not successful) and discard the decoded packet. To evaluate the effectiveness of this thresholding approach, we simulate the probability of false alarm as a function of SNR for the conventional 16-bit CRC and the proposed residual-based thresholding. As is clear from Fig. 2.11, the residual-based thresholding

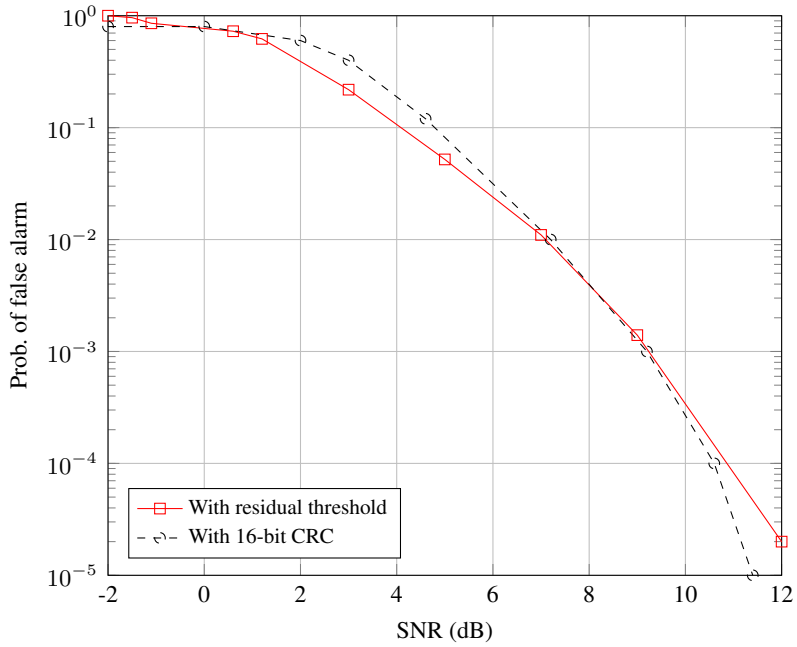


Figure 2.11: Decoding failure as a function of SNR ($P_{th} = 10^{-5}$).

performs similarly to the CRC-based error checking.

2.8 Simulations and Discussions

2.8.1 Simulation Setup

In this section, we examine the performance of the proposed SVC technique. Our simulation setup is based on the downlink OFDM system in the 3GPP LTE-Advanced Rel.13 [63]. As a channel model, we use AWGN and realistic ITU channel models including extended typical urban (ETU) and extended pedestrian-A (EPA) channel model [63]. For comparison, we also investigate the performance of the conventional PDCCH of LTE-Advanced system, polar code-based PDCCH of 5G systems [23], and AWGN lower bound. We test the transmission of b bit information which consists of information bit b_i and CRC bit b_c. In the conventional PDCCH method, the convolution code with rate $\frac{1}{3}$ with the 16-bit CRC is employed. Since the block size of the polar

code is not flexible, we set the rate $\frac{1}{4}$ to test similar conditions ($b = 2^4$ and $m = 32$). In the proposed SVC algorithm, we set the random binary spreading codebook with $N = 96$ and $K = 2$. To ensure the fair comparison, we use the same number of resources ($m = 42$ with $L = 8$ repetitions) in the control packet transmission. As a performance measure, we use PER of the code blocks.

2.8.2 Simulation Results

In Fig. 2.12(a), we investigate the PER performance of the proposed SVC method and competing schemes under AWGN channel condition. We observe that the proposed SVC technique outperforms the conventional PDCCH and polar code-based scheme, achieving more than 4 dB gain over the conventional PDCCH and about 1.1 dB gain over the polar code-based scheme at 10^{-5} PER point. Even in realistic scenarios such as EPA and EVA channels in LTE-Advanced, we observe that the performance gain of the proposed SVC scheme over competing schemes is maintained (see Fig. 2.12(b)).

In Fig. 2.13, we evaluate the PER performance of PDCCH and SVC as a function of SNR for various information bit size ($b_i = 12, 24, 48$, and 96). These results demonstrate that the proposed SVC technique can deliver more information bits than the conventional PDCCH can support. For example, SVC can deliver twice more information than PDCCH in the low SNR region (for example, $b_i = 12$ of PDCCH and $b_i = 24$ of SVC in Fig. 2.13). To further investigate this, we plot the minimum SNR to achieve the target PER as a function of the information bit size in Fig. 2.14. For example, to achieve 10^{-5} PER with $b = 10$, it requires -2.9 dB for PDCCH while -6.2 dB SNR for SVC, resulting in 3.3 dB gain in performance. It is worth mentioning that the coding gain of the conventional PDCCH improves with the codeblock size so that the gap between the SVC and PDCCH diminishes gradually as the number of information bits increases so that PDCCH outperforms SVC scheme (e.g., $b > 160$).

Next, we evaluate the latency performance of the SVC and PDCCH (see Fig. 2.15). In this experiments, we plot the distribution of transmission latency to achieve 10^{-5}

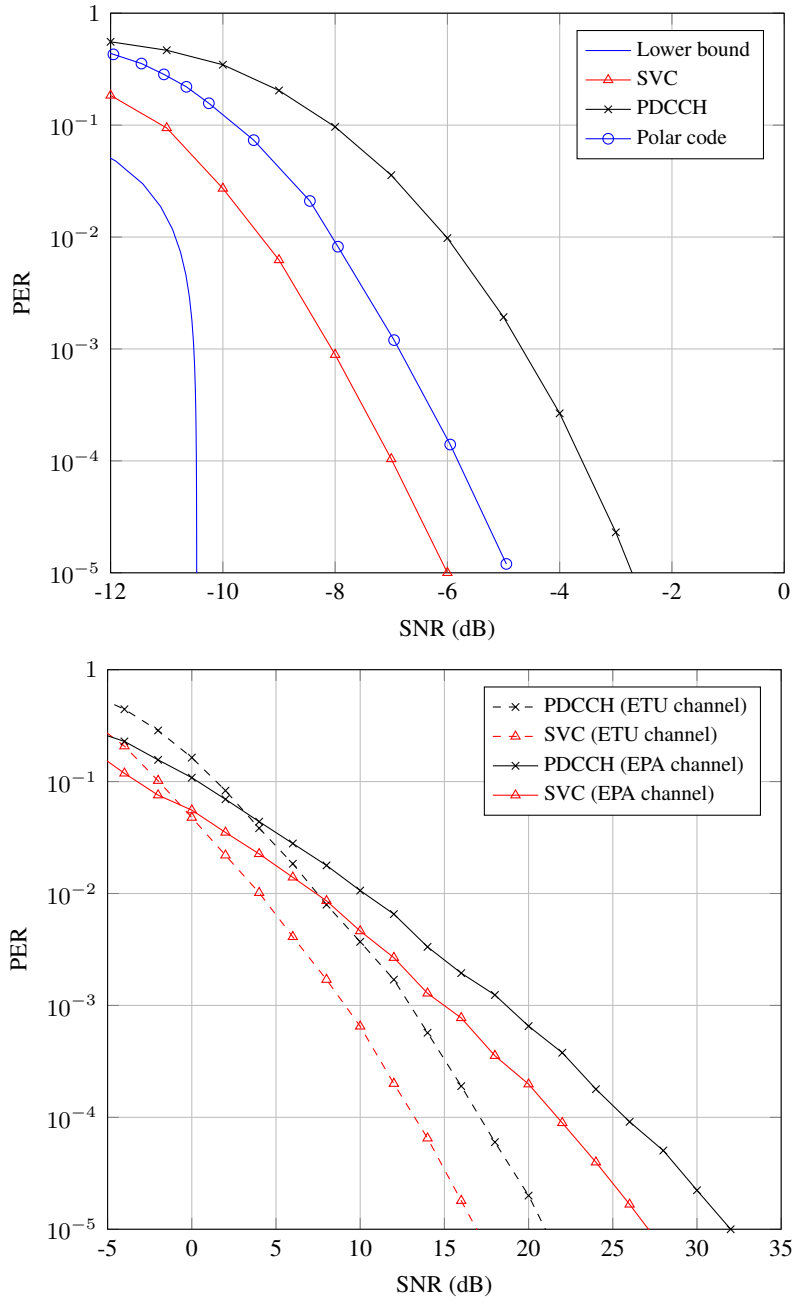


Figure 2.12: PER performance as a function of SNR ($b_i = 12$, $b_c = 16$, $m = 42$, $L = 8$, and $N = 96$) for (a) AWGN channel (upper) and (b) ETU and EPA channel (lower).

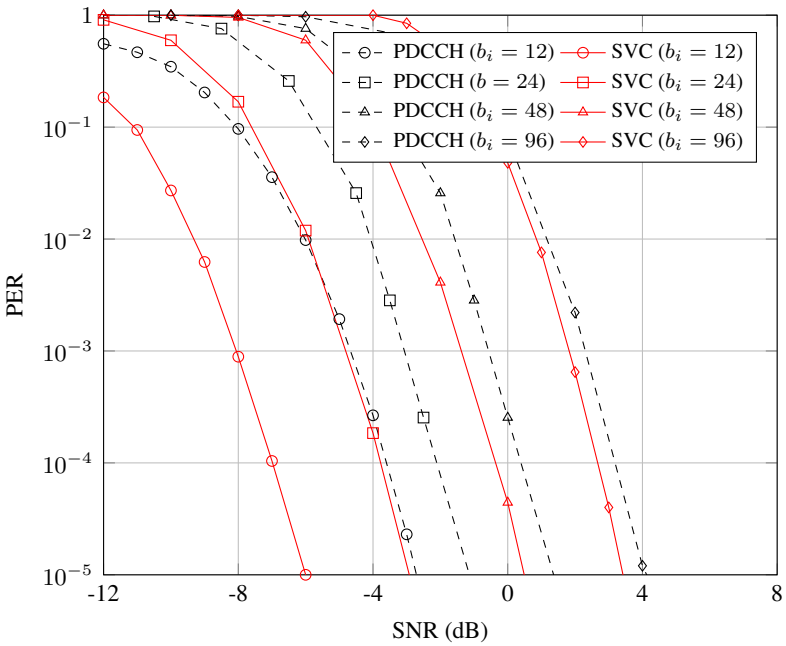


Figure 2.13: PER performance for various size of control information ($L = 8$).

PER when n-repetition scheme is employed. Transmission latency is defined as the time from the initial transmission to the time that the packet is successfully decoded at the mobile terminal.[11] These results demonstrate that most of SVC packets satisfy the URLLC requirement (1 ms latency).

Finally, we investigate the performance of SVC in the small cell scenarios where the received signal contains a considerable amount of interference from adjacent basestations. Note that densely deployed small cell (pico, femto, and micro) environments will play a key role to enhance the cell throughput in 5G and how to manage the interference is the key to the success of small cell networks. In our simulations, we set the power level of interference to half of the desired cell signal. Since the SVC transmission is based on the multi-code spreading and also the effective transmit power per symbol is large, SVC can effectively manage the interference. Whereas, since the conventional PDCCH has no such interference protection mechanism, error correction

[11] In our experiments, we ignored the decoding latency.

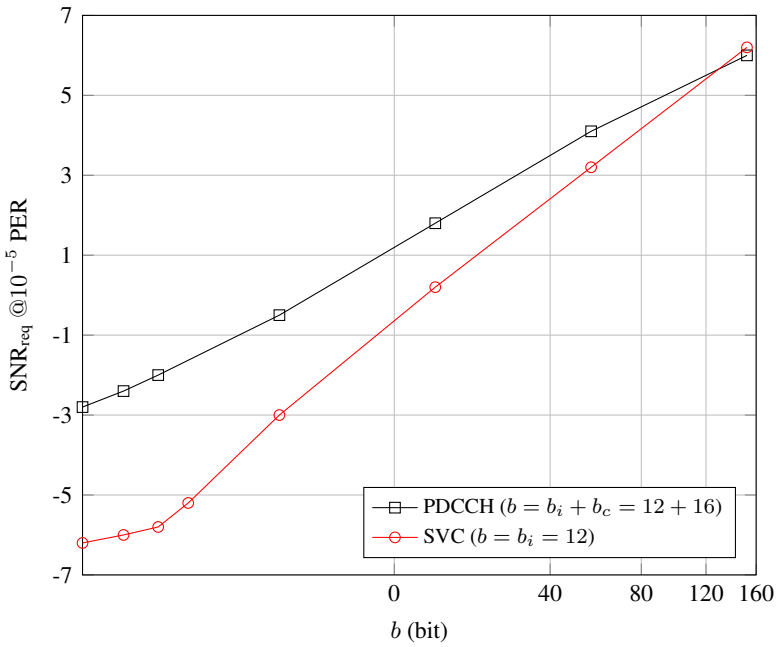


Figure 2.14: Minimum required SNR for achieving 10^{-5} PER ($m = 42$, $L = 8$, and $N = 96$)

capability of PDCCH is degraded significantly and thus the PDCCH performs very poor as shown in Fig. 2.16.

2.9 Summary

In this paper, we have proposed the short packet transmission strategy for URLLC. The key idea behind the proposed SVC technique is to transform an information vector into the sparse vector in the transmitter and to exploit the sparse recovery algorithm in the receiver. Metaphorically, SVC can be thought as a marking dots to the empty table. As long as the number of dots is small enough and the measurements contain enough information to figure out the marked cell positions, accurate decoding of SVC packet can be guaranteed. We showed from the numerical evaluations that the proposed SVC scheme is very effective in URLLC scenarios. In this chapter, we restricted our attention to the URLLC scenario but we believe that there are many other

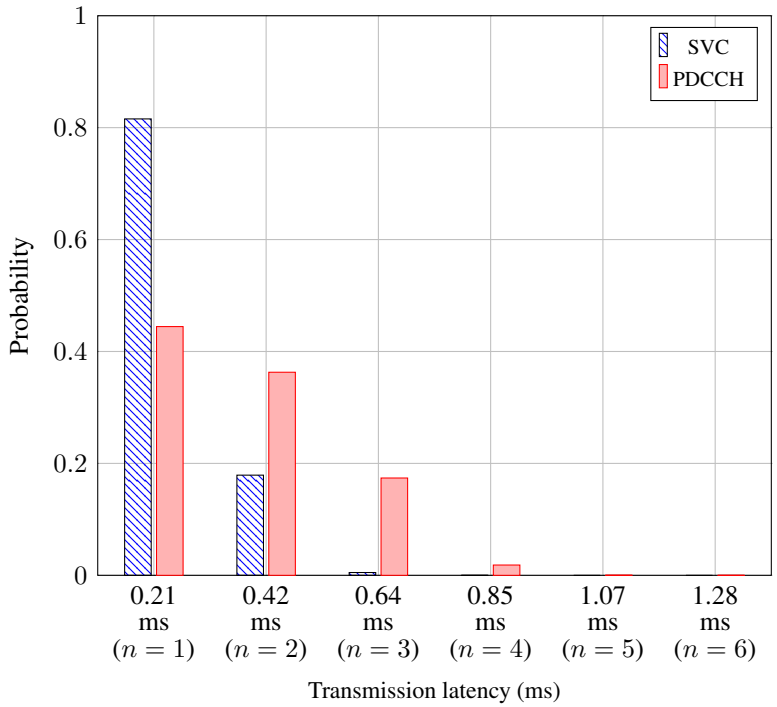


Figure 2.15: Probability of transmission latency for achieving 10^{-5} PER ($b_i = 12$, $m = 42$, $L = 8$, $N = 96$, and SNR=-12 dB).

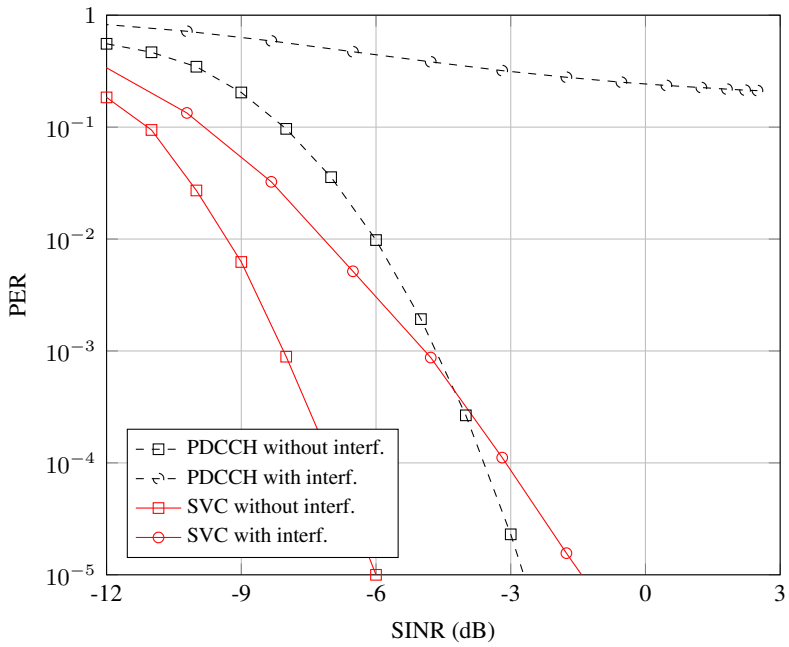


Figure 2.16: PER performance as a function of SINR ($b_i = 12$, $m = 42$, $L = 8$, $N = 96$, and interference power is half of the signal power).

applications that the SVC technique can be applied to. Also, there are many interesting extensions and variations worth investigating, such as the information embedding in nonzero positions, channel aware sparse vector coding, and combination of SVC and error correction codes.

Chapter 3

Sparse Vector Coding for Uplink Massive Machine-type Communications

This chapter proposed the enhanced SVC in application of massive access scenario in uplink transmission. Instead of using sparsity of active user, we exploit the sparsity of information vector that is concatenated by massive user. I put forth an approach that directly exploits the sparsity of information vector to control the multiuser interference. This, together with the fact that the symbol is spread using randomly generated codebook controls the multiuser interference efficiently. Since the input is a sparse vector and the system matrix is the composite of fading channel and random spreading matrices, I can cast the symbol detection problem into the sparse support recovery problem. In a nutshell, the proposed scheme is simple to implement, robust to interference, and can be readily applied to various mMTC applications where the size of transmit packet is small. I demonstrate that the proposed technique is very effective in the ultra-short packet transmission and outperforms the LTE physical uplink shared channel (PUSCH) [63] and LDS-based schemes in NOMA. I also demonstrate that the SVC-encoded uplink access is effective in controlling multi-user interference caused by the massive access.

This work of Chapter 3 has been published in [56].

3.1 Introduction

Massive machine type communications (mMTC) is a new service category in 5G and beyond to support Internet of Things (IoT), Internet of Everything (IoE), and Industry 4.0 [46]. mMTC has many distinct features over the conventional human-to-human (H2H) communications. For example, communication traffic is uplink dominated and machine-type devices transmit the low volume of information sporadically. Over the years, various licensed and unlicensed technologies to support mMTC have been proposed. Well-known examples include narrowband-IoT (NB-IoT) in the licensed band and SigFox and LoRA in the unlicensed band. While primary goals in these approaches are low-power and low-cost operation and enhanced coverage, not much work to support the ultra short packet transmission has been proposed [49]. One notable observation is mMTC is that the transmit information is either control type information (e.g., move, shift, rotate, start, and stop) or sensing information (e.g., temperature, moisture, air pressure, and wind speed) so that the amount of information is in general very tiny. Since the information transmission principle of today, largely based on Shannon's channel coding theorem, requires long codeblock to maximize the coding gain and hence is not adequate for ultra-short packet transmission, new transmission strategy optimized for the mMTC scenario is required [50].

Further, since the massive number of users access to the networks simultaneously in mMTC environments, a simple yet effective transmission scheme to control the multiuser interference is needed [48]. In recent years, an approach to support a large number of mMTC devices using a limited number of resources, collectively called non-orthogonal multiple access (NOMA) technology, has been proposed [40–43]. For example, in [40], an approach to use sparse spreading code, called low-density signature (LDS) has been proposed. In [41], an enhanced version of LDS technique referred to as sparse code multiple access (SCMA) technique has been proposed. Also, variants of SCMA technique have been proposed in [42-43]. Since these NOMA-based techniques exploit the sparsity of codebook, multiuser interference can be controlled by the

number and location of punctured elements in the codebook. In fact, due to the codebook sparsity, NOMA tecniques enjoy reduced signal-to-interference ratio per symbol.

3.2 Uplink NOMA transmission for mMTC

In uplink multiple access scenario where the basestation receives data transmission from active devices, the received signal at the basestation can be expressed as

$$\mathbf{y} = \sum_{i \in \mathbf{u}} h_i \mathbf{s}_i x_i + \mathbf{v}, \tag{3.1}$$

where $\mathbf{u}$ is the set of active user index, $\mathbf{y} = [y_1, \ldots, y_L]^T$ is the received signal, L is the number of resources for data transmission, $x_i \in \mathbb{X}$, $h_i \in \mathbb{C}$ and $\mathbf{s}_i = [s_{i,1}, \ldots, s_{i,L}]^T \in \mathbb{C}^L$ are data symbol, the channel and codeword vector of ith device, and $\mathbf{v} \in \mathbb{C}^L$ is the complex Gaussian noise vector ($\mathbf{v} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$), respectively. With LDS or SCMA based access, the sparse codebook $\mathbf{C}_{(M,L)}$ is designed to have a set of codes with L-length and M codes (i.e., $\mathbf{C}_{(M,L)} = [\mathbf{c}_1, \cdots, \mathbf{c}_M] \in \mathbb{C}^{L \times M}$, $\mathbf{c}_i = [c_{i,1}, \cdots, c_{i,L}]^T \in \mathbb{C}^L$, and $M > L$). Then, each user randomly selects $\mathbf{s_i}$ in the LDS codebook $\mathbf{C}_{(M,L)}$. Then, the received signal is given by

$$\begin{aligned} \mathbf{y}^{(p)} &= \sum_{i=1}^{N_T} h_i \mathbf{d}_i p_i \psi_i + \mathbf{v}^{(p)} && (3.2) \\ &= \widetilde{\mathbf{D}}\mathbf{g} + \mathbf{v}^{(p)}, && (3.3) \end{aligned}$$

where $\widetilde{\mathbf{D}} = [p_1\mathbf{d}_1, \cdots, p_{N_T}\mathbf{d}_{N_T}] \in \mathbb{C}^{KL \times N_T}$ is uplink pilot signals ($\mathbf{d}_i \in \mathbb{C}^{KL \times 1}$), $\mathbf{g} = [h_1\psi_1, \cdots, h_{N_T}\psi_{N_T}]^T \in \mathbb{C}^{N_T}$, ψ_i for $i = 1, 2, \cdots, N_T$ is a logical variable to indicate activity of user i, $\mathbf{y}^{(p)}$ is received signal, and $\mathbf{v}^{(p)} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$ is noise vector, respectively.

Since the number of active users is relatively smaller than the number of total users, the vector $\mathbf{g} \in \mathbb{C}^{N_T}$ becomes sparse. By the process of reconstructing the sparse

vector, we obtain the indices of estimated active users. For the multiuser data detection, NOMA uses the message passing algorithm (MPA). MPA can update message passing factor graph by placing colliding users at the code node.

For all $x \in \mathbb{X}$ the message values (i.e., $r_{R_l \to U_n}$ and $q_{U_n \to R_l}$) between user node and resource node at the i-th iteration are

$$r^i_{R_l \to U_n}(x) = \sum_{\substack{\mathbf{x}^{[l]} \in \mathbb{X}^{|\xi_l|} \\ x_n = x}} \left(G_l(\mathbf{x}^{[l]}) \prod_{n' \in \xi_l \setminus n} q^{i-1}_{U_{n'} \to R_l}(x_{n'}) \right) \tag{3.4}$$

$$q^i_{U_n \to R_l}(x) = \gamma \prod_{l' \in \zeta_n \setminus l} r^i_{R_{l'} \to U_n}(x), \tag{3.5}$$

where $\mathbf{x}^{[l]}$ is the vector of symbols which contribute to y_l, $G_l(\mathbf{x}^{[l]})$ is the probability of y_l given $\mathbf{x}^{[l]}$, ξ_l is set of user index which contribute to resource element l, ζ_n is set of resource index which user n spread its symbol on, and γ is normalizing factor to satisfy $\sum_x q^i_{U_n \to R_l}(x) = 1$. The value $q^i_{U_n \to R_l}(x)$ is the probability that symbol x_n has the value x, given the information obtained via resource elements other than resource element l. i.e., $p(x_n = x | \{y^{(d)}_{l'}\} : l' \in \zeta_n \setminus l)$. The value $r^i_{R_l \to U_n}(x)$ is the probability that the value of resource element l is satisfied if symbol x_n is considered fixed at x. i.e., $p(y^{(d)}_l | x_n = x)$.

After these message values have propagated a number of iterations, symbol decisions are made by

$$\hat{x}_n = \arg\max_x \{ \prod_{l \in \zeta_n} r^i_{R_l \to U_n}(x) \}. \tag{3.6}$$

3.3 Sparse Vector Coding based NOMA for mMTC

3.3.1 System Model

In the uplink transmission of the conventional wireless systems, the transmit vector $\mathbf{x} \in \mathbb{C}^{m\times 1}$ is generated after the channel coding and symbol mapping. The received vector $\mathbf{y} \in \mathbb{C}^{m\times 1}$ for a single user (say user l) is given by

$$\mathbf{y}^l = \mathbf{H}^l\mathbf{x}^l + \mathbf{v}, \tag{3.7}$$

where $\mathbf{H}^l \in \mathbb{C}^{m\times m}$ is the diagonal matrix whose diagonal element h_{ii} is the channel corresponding to i-th resource, $\mathbf{v} \sim \mathcal{CN}(\mathbf{0}, \sigma_v^2\mathbf{I})$ is the additive Gaussian noise, $\mathbf{x}^l = \mathbf{C}^l\mathbf{s}^l$ is the transmit vector where $\mathbf{C}^l$ is the spreading codebook matrix describing the mapping between the symbol and resource elements, and $\mathbf{s}^l \in \mathbb{C}^{N\times 1}$ is the modulated symbol vector. When one symbol is mapped to a single resource, $\mathbf{C}^l$ is the square matrix constructed from the spreading sequences (e.g., Walsh matrix).

The SVC encoding is achieved by the mapping of an information vector $\mathbf{w}^l$ into the *position* of the sparse vector $\mathbf{s}^l$. By way of analogy, SVC can be thought as placing a few balls into the empty boxes. When we try to put K balls in N boxes, we have $\binom{N}{K}$ choices. This implies that $b = \lfloor \log_2 \binom{N}{K} \rfloor$ bits of information can be encoded by this strategy. For example, when $N = 9$ and $K = 2$, we can encode 5 bit information. The mapping between the information vector $\mathbf{w}^l$ and the transmit sparse vector $\mathbf{s}^l$ can

be expressed as (see Table 3.1)

$$
\begin{array}{ccc}
w_5\ w_4\ w_3\ w_2\ w_1 & & s_9\ s_8\ s_7\ s_6\ s_5\ s_4\ s_3\ s_2\ s_1 \\
0\ \ 0\ \ 0\ \ 0\ \ 0 & \longleftrightarrow & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1 \\
0\ \ 0\ \ 0\ \ 0\ \ 1 & \longleftrightarrow & 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 1 \\
0\ \ 0\ \ 0\ \ 1\ \ 0 & \longleftrightarrow & 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 0 \\
\vdots & \vdots & \vdots \\
1\ \ 1\ \ 1\ \ 1\ \ 1 & \longleftrightarrow & 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\
\underbrace{\qquad\qquad}_{\text{5-bit information } \mathbf{w}^l} & & \underbrace{\qquad\qquad\qquad\qquad}_{K-\text{sparse vector } \mathbf{s}^l}
\end{array}.
$$

In order to transmit the compressed version of the sparse vector, we further apply the random spreading to the sparse vector. As a result of sparse mapping and spreading, we obtain an underdetermined sparse system ($\mathbf{y}^l = \mathbf{H}^l\mathbf{C}^l\mathbf{s}^l + \mathbf{v}$) for which the principle of CS can be applied. In the spreading step, each nonzero element in $\mathbf{s}$ is spread into m resources using the spreading codeword $\mathbf{c}_i^l$. Since the position of nonzero elements in $\mathbf{s}^l$ can be chosen at random, the codebook matrix $\mathbf{C}^l$ should be designed such that the transmit vector $\mathbf{x}$ contains enough information to recover the sparse vector $\mathbf{s}^l$. Toward this end, random spreading matrices (e.g., Gaussian, Bernoulli) are popularly used. Also, to transmit the 2^K-QAM modulated symbol, we set half of the nonzero entries in $\mathbf{s}$ to 1 and the rest to j. Since the QPSK modulation ($K = 2$) is a typical choice in the mMTC scenario, by assigning one nonzero entry to 1 and the other to j, each element in the transmit vector $\mathbf{x}^l$ is mapped to the QPSK symbol (i.e., $x_i \in \alpha\{1+j, 1-j, -1+j, -1-j\}$).

Based on the principle of CS, if entries of the codebook matrix $\mathbf{C}^l$ are generated at random, then an accurate recovery of the sparse vector is possible as long as $m = \mathcal{O}(K \log N)$ [59]. From this fact, typical setting of the spreading sequence length (i.e., measurement vector size) m to ensure the robust recovery of the sparse vector can be

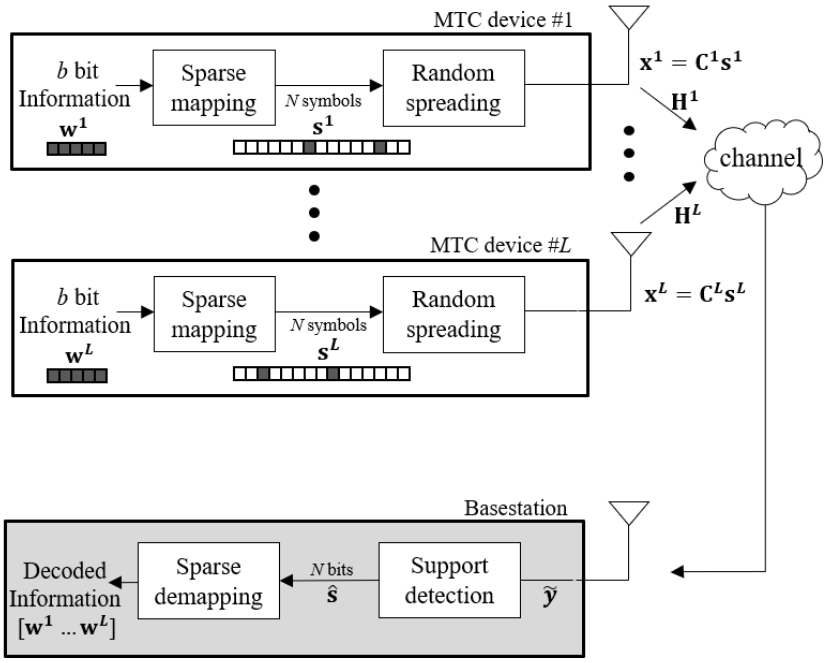


Figure 3.1: The block diagram for the proposed SVC-encoded uplink transmission.

derived. To be specific, since $2^b \approx \binom{N}{K}$ and $K = 2$, we approximately have

$$N \approx 2^{\frac{b+1}{2}}. \tag{3.8}$$

As mentioned, the dimension of the measurement vector should satisfy $m = cK \log N$, which means

$$\begin{aligned} m &= 2c \log 2^{\frac{b+1}{2}} \\ &= c(b+1). \end{aligned} \tag{3.9}$$

As a rule of thumb, $c = 4$ is commonly used as a ballpark number [59]. This means that if the measurement m satisfies $m \geq 4(b+1)$, one can readily apply CS technique at the receiver. In fact, in our simulation, we designed m based on (3) (e.g., $b = 12$ and $m = 50$).

Table 3.1: One-to-one mapping between information and sparse vector $\mathbf{s}$ in increasing order ($K = 2$)

Input:
Size of sparse vector N, information vector $\mathbf{w}^l$
Output:
Sparse vector $\mathbf{s}^l$

$k := 0$
for $i = 2$ **to** N **do**
 for $j = 1$ **to** $i - 1$ **do**
 if $k = \left(\mathbf{w}^l\right)_{(10)}$
 $\mathbf{s}^l := \left(2^i + 2^j\right)_{(2)}$
 end if
 $k := k + 1$
 end for
end for

Note: $(\mathbf{a})_{(10)}$ is decimal expression of binary vector $\mathbf{a}$ and $(a)_{(2)}$ is binary expression of integer a.

3.3.2 Joint Multiuser Decoding

The uplink system model after the spreading can be expressed as

$$\begin{aligned} \tilde{\mathbf{y}} &= \sum_{l \in U} \mathbf{H}^l \mathbf{x}^l + \mathbf{v} \\ &= \sum_{l \in U} \mathbf{H}^l \mathbf{C}^l \mathbf{s}^l + \mathbf{v}, \end{aligned} \tag{3.10}$$

where U is the index set of uplink accessing users. When L users are accessing simultaneously, the received vector can be expressed as the sum of a linear combination of LN columns. That is,

$$\begin{aligned} \tilde{\mathbf{y}} &= \begin{bmatrix} \mathbf{H}^1\mathbf{C}^1 & \cdots & \mathbf{H}^L\mathbf{C}^L \end{bmatrix} \begin{bmatrix} \mathbf{s}^1 \\ \vdots \\ \mathbf{s}^L \end{bmatrix} + \mathbf{v} \\ &= \mathbf{\Phi}\tilde{\mathbf{s}} + \mathbf{v}, \end{aligned} \tag{3.11}$$

where $\mathbf{\Phi}$ is a concatenation of $\mathbf{H}^l\mathbf{C}^l$ and $\tilde{\mathbf{s}}$ is the stacked vector of the sparse vector $\mathbf{s}^l$. As mentioned, since the sparse vector for each user has only K nonzero elements, the

stacked sparse vector $\tilde{\mathbf{s}}$ has LK nonzero elements. In light of this, the main task of the receiver is to identify LK columns in $\mathbf{\Phi}$ participating in the received vector.

In a nutshell, greedy sparse recovery algorithm such as orthogonal matching pursuit (OMP) algorithm finds the support of $\tilde{\mathbf{s}}$ in an iterative fashion.[1] When columns of $\mathbf{\Phi}$ are approximately orthogonal, $\hat{\mathbf{s}} = \mathbf{\Phi}_{\mathcal{S}_{\tilde{\mathbf{s}}}}^H \tilde{\mathbf{y}} = \mathbf{\Phi}_{\mathcal{S}_{\tilde{\mathbf{s}}}}^H (\mathbf{\Phi}\tilde{\mathbf{s}} + \mathbf{v})$ can be a reasonable approximation of $\tilde{\mathbf{s}}$. Based on this observation, OMP finds the support element one at each iteration. To be specific, let $\mathbf{\Phi}_{\mathcal{S}_{\tilde{\mathbf{s}}}^{j-1}}$ be the submatrix of $\mathbf{\Phi}$ that only contains columns indexed by $\mathcal{S}_{\tilde{\mathbf{s}}}^{j-1}$, then an index ω_j chosen at the j-th iteration of the greedy sparse recovery algorithm is given by[2]

$$\omega_j = \arg\max_l |\mathbf{\Phi}_l^H \mathbf{r}^{j-1}|^2, \tag{3.12}$$

where $\mathbf{r}^{j-1} = \tilde{\mathbf{y}} - \mathbf{\Phi}_{\mathcal{S}_{\tilde{\mathbf{s}}}^{j-1}} \hat{\mathbf{s}}^{j-1}$ is the residual vector and $\hat{\mathbf{s}}^{j-1}$ is either 1 or j. Between these two, we choose the value that minimizes the residual magnitude.

As mentioned, one advantage of SVC is that an accurate recovery of the sparse vector $\mathbf{s}^l$ is unnecessary since the decoding is simply done by the identification of nonzero positions of $\mathbf{s}^l$. Thus, by presetting the nonzero values, we can avoid effort to find out actual values. Another advantage is that it requires low computational complexity. Once K nonzero elements are detected in $\mathbf{s}^l$, the system matrix is updated by removing all columns of user l (i.e., $\mathbf{H}^l\mathbf{C}^l$ in $\mathbf{\Phi}$). This is because the sparsity K of $\mathbf{s}^l$ is known to the receiver, no further investigation on user l's column is needed. In doing so, the computation complexity of the decoding process can be reduced substantially.

The proposed CS-based joint receiver algorithm for SVC-encoded short packets is summarized in Table 3.2.

[1] Support is the set of nonzero elements. For example, if $\mathbf{s}^l = [0\ 0\ 1\ 0\ 0\ 1]$, the $\mathcal{S}_{\mathbf{s}^l} = \{3, 6\}$.
[2] If $\mathcal{S} = \{2, 5\}$, then $\mathbf{\Phi}_{\mathcal{S}} = [\Phi_2\ \Phi_5]$.

Table 3.2: The proposed joint user detection algorithm

Input:
Measurement $\tilde{\mathbf{y}}$, sensing matrix $\mathbf{\Phi}$, sparsity LK, stop threshold ϵ
expansion size E, maximum candidate l_{max}, false-alarm threshold ε
Output:
Support set $\mathcal{S}_{\tilde{\mathbf{s}}}$
Initialization:
$l := 0$ (candidate order), $\rho := \infty$ (minimum magnitude of residual)

While: $l < l_{max}$ and $\epsilon < \rho$ **do**
 $l := l + 1$
 $\mathbf{r}^0 := \mathbf{y}$
 $[p_1, \dots, p_{LK}] :=$ compute_p$_k(l, E)$ (*compute layer order*)
 for $k = 1$ **to** LK **do** (*investigate l-th candidate*)
  $\tilde{\omega}_{p_k} := \arg\max_{|\pi|=E} \|(\langle \frac{\mathbf{\Phi}^H}{\|\mathbf{\Phi}\|_2} \mathbf{r}^{k-1} \rangle)_\pi\|_2^2$
  $\mathcal{S}_l^k := \mathcal{S}_l^{k-1} \cup \{\tilde{\omega}_{p_k}\}$ (*construct a path in k-th layer*)
  $\mathbf{r}^k := \tilde{\mathbf{y}} - \mathbf{\Phi}_{\mathcal{S}_l^k} \hat{\mathbf{s}}^k$ (*update residual*)
  $\hat{\mathcal{S}}^k := \mathcal{S}_l^k$ (*update support set*)
 end for
 if $\|\mathbf{r}^{LK}\|_2^2 < \rho$ **then** (*update the smallest residual*)
  $\rho := \|\mathbf{r}^{LK}\|_2^2$
  if $\frac{\|\mathbf{r}^{LK}\|_2^2}{\|\mathbf{y}\|_2^2} > 1 - \varepsilon$ **then** (*false-alarm identification*)
   $\hat{\mathcal{S}}^* := \mathbf{0}$
  end if
  $\mathcal{S}_{\tilde{\mathbf{s}}}^* := \hat{\mathcal{S}}^{LK}$
 end if
end while
return $\mathcal{S}_{\tilde{\mathbf{s}}}^*$

function compute_p$_k(l, E)$
 $t := l - 1$
 for $k = 1$ **to** LK **do**
  $p_k :=$ mod $(t, E) + 1$
  $t :=$ floor(t/E)
 end for
return $[p_1, \dots, p_{LK}]$
end function

3.4 Simulations and Discussions

3.4.1 Simulation Setup

In this section, we examine the performance of the proposed SVC technique in grant-based uplink access scenario. Our simulation setup is based on the OFDM system in the 3GPP LTE-Advanced Rel.13 [63]. For comparison, we investigate the performance of the PUSCH, an uplink data channel in LTE-Advanced system. PUSCH is encoded

with the rate $\frac{1}{3}$ Turbo code along with the 16-bit cyclic redundancy check (CRC). In the proposed SVC algorithm, we set $N = 96$ and $K = 2$. To ensure the fair comparison, we use the QSPK modulation and the same number of resources for all schemes under test ($m = \frac{50}{12}b$). We also assume that the channel coefficients are known at the basestation. As a performance measure, we use the packet error rate (PER) of the uplink transmissions.

3.4.2 Simulation Results

In Fig. 3.2, we evaluate the PER performance PUSCH and SVC as a function of SNR for various information bit size ($b = 12, 24, 48,$ and 96). We observe that the proposed SVC technique outperforms the conventional PUSCH, achieving 3.2 dB gain when the packet size is 12 bit. However, since the channel coding gain of the conventional PUSCH improves with the codeblock size, the gap between the SVC and PUSCH is reduced gradually as the number of information bits increases. For example, a gain of the SVC is reduced to 2.1 dB at $b = 24$.

In order to test the capability to support the massive connection of machine devices, we next consider the performance of overloading scenarios where the number of users is larger than the number of resources. For example, 200% overloading means than L users transmit packets using $\frac{L}{2}$ resources. In this test, we observe the performance of PUSCH, LDS-based NOMA, and SVC. In LDS-based NOMA, we use $\frac{L}{2} \times L$ sized LDS codebook for the packet transmission and the MPA decoding algorithm at the basestation [45]. In the simulations, we observe the performance at the reference PER point (around 10^{-1} PER) of uplink mMTC transmission. We observe from Fig. 3.3 that the SVC technique outperforms the conventional PUSCH and LDS, achieving 1.7 dB gain over LDS. As mentioned, since the sparse vector is transmitted after the random spreading, SVC is effective in controlling the multi-user interference. Whereas, the conventional PUSCH has no such interference protection mechanism so that the performance degradation is severe in high SNR.

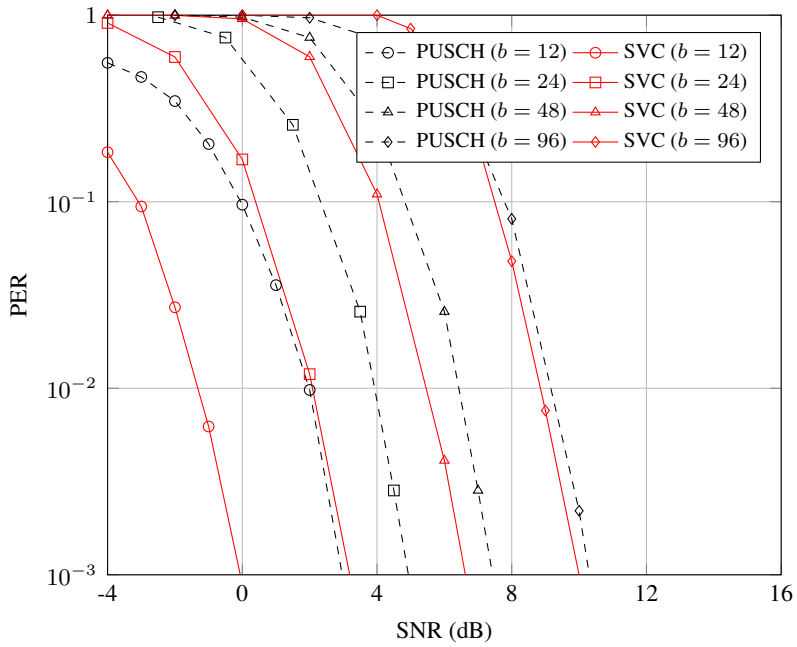


Figure 3.2: PER performance for various packet size for a single user transmission.

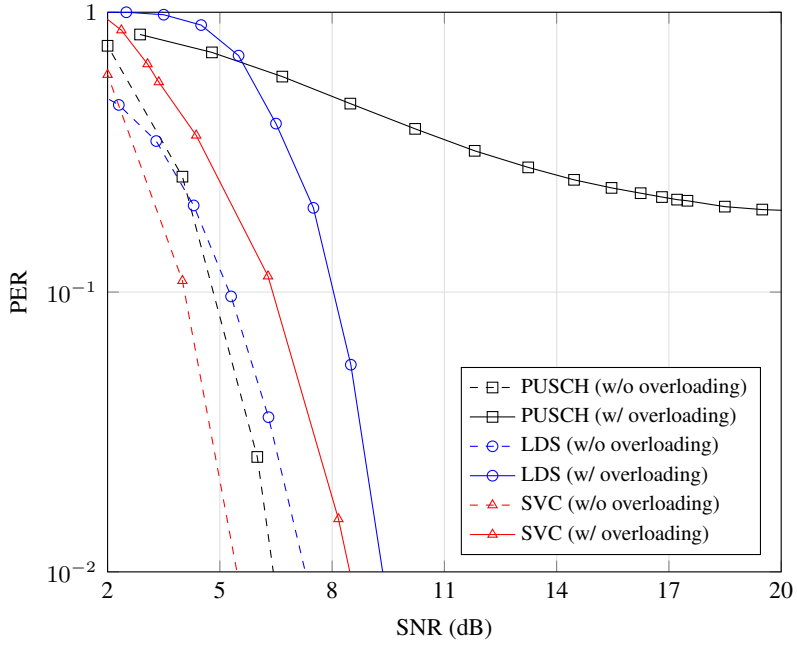


Figure 3.3: PER performance for 200% overloading as a function of SNR ($b = 48$).

3.5 Summary

In this chapter, we have proposed a new short packet transmission scheme called sparse vector coding (SVC) for the mMTC uplink scenario. The key idea behind the proposed SVC transmission scheme is to convert an information vector into the sparse vector and use the support identification algorithm as a decoder in the receiver. The SVC transmission scheme is easy to implement, robust to noise and multiuser interference, and also scalable to the massive access scenario in mMTC. We showed from the simulations in the LTE uplink scenario and massive access scenario in 5G that the proposed SVC technique is very effective in short packet transmissions.

Chapter 4

Pilot-less Sparse Vector Coding for Short Packet Transmission

This chapter proposes SVC for short packet transmission without pilot symbols. A short packet transmission is one of the major operational modes for mission-critical data in ultra reliable and low latency communications (URLLC) and control-type data in massive machine-type communications (mMTC). The central challenge in the short packet transmission is the excessive amount of overhead caused by the pilot signaling. In this chapter, we propose a novel scheme suitable for the short packet transmission without pilot signals, called pilot-less sparse vector coding (PL-SVC). Key feature of PL-SVC is to map the input as a composite of the sparse vector and the fading channel and to perform decoding by finding out the nonzero positions of the sparse vector. In this setting, the system matrix becomes a pseudo-random spreading matrix and the input vector becomes a channel-scaled sparse vector so that the PL-SVC decoding problem can be cast into the support detection problem in the compressed sensing. We show from the numerical experiments in the 5G uplink scenario that PL-SVC is very effective in the short packet transmission and outperforms conventional schemes.

This work of Chapter 4 has been submitted in part in [66]

4.1 Introduction

These days, short packet transmission is becoming increasingly popular [46]. Short packets are commonly used as a means to deliver sensor generated information (e.g., temperature, moisture, pressure, gas density), mission-critical data in ultra reliable and low latency communications (URLLC), and control-type information (e.g., start, stop, rotate, shift, turn left/right) in machine-type communications (MTC) [47]. The central challenge in the short packet transmission is the excessive amount of overhead caused by the pilot signaling. This issue has not been a serious concern of the conventional systems since the amount of payload (data information) is much larger than the amount of pilot signals [49–51]. However, when the packet size is short, pilot overhead is no longer negligible in size so that the pilot overhead has become a major hindrance in the short packet transmission.

In recent years, there have been some efforts to reduce pilot overhead in the short packet transmission [51–53]. For example, a channel estimation technique exploiting the most reliable data symbols as pilot signals has been proposed in [51]. In [52], an algorithm to minimize the pilot overhead under the block length and error probability constraints has been proposed. Recently, an approach to support a short packet transmission based on the principle of compressed sensing, called sparse vector coding (SVC), has been proposed [53]. The basic idea of SVC is that the data information is mapped into the position of a sparse vector and then transmitted after the spreading. The packet decoding is done by finding out the nonzero positions of the sparse vector. It has been shown that the packet error rate (PER) of SVC outperforms the conventional channel coding schemes for both single and multi-user transmission scenarios [53].

An aim of this study is to propose an approach to transmit the short packet without pilot signals. The proposed technique, referred to as pilot-less SVC (PL-SVC), performs the *sparse vector conversion* and *random spreading* at the encoder and the *support identification* at the decoder. Key distinctive feature of the proposed PL-SVC over the conventional SVC transmission is to map the input as a composite of the

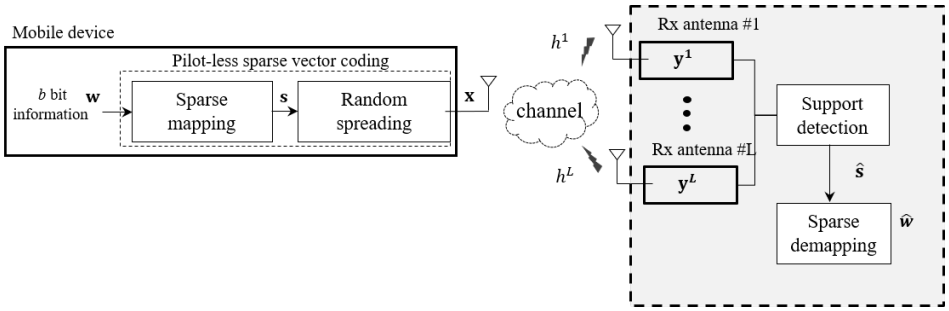


Figure 4.1: The block diagram for the proposed PL-SVC technique.

sparse vector and the fading channel. Under this formulation, the system matrix becomes a pseudo-random spreading matrix and the input vector becomes a channel-scaled sparse information vector. Therefore, the PL-SVC decoding problem can be cast into the sparse signal recovery (more accurately support detection) problem in the compressed sensing (CS) [72]. Somewhat surprisingly, this is done without channel information and with small computational complexity. This is because what we need is the identification of nonzero positions of the sparse input vector, not the nonzero value. This direction means that pilot transmission to acquire the channel estimation is unnecessary. When multiple antennas at the basestation are available, the proposed PL-SVC can improve the reliability by negligible computational overhead, which cannot be easily obtained by the conventional transmission schemes. In a nutshell, the proposed scheme is effective in various URLLC and MTC applications where the packet size is very short (e.g., $10 \sim 100$ bytes).

From the numerical evaluations in URLLC and mMTC uplink transmission scenario, we demonstrate that the proposed PL-SVC is effective in the short packet transmission, achieving 1 dB gain at 10^{-4} PER point. In the scenario where the multiple receiving antennas at the basestation available, PL-SVC achieves diversity gain (approximately 0.5 dB gain per antenna) over single antenna.

4.2 Pilot-less Sparse Vector Coding Processing

4.2.1 SVC Processing with Pilot Symbols

We consider the uplink transmission of URLLC and mMTC scenarios where a user uses m physical resources to transmit b-bit information. In the conventional wireless systems, the transmit vector $\mathbf{x} \in \mathbb{C}^{m\times 1}$ is generated via the channel coding and symbol mapping of data information. The received vector $\mathbf{y} \in \mathbb{C}^{m\times 1}$ is given by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}, \tag{4.1}$$

where $\mathbf{H} = \text{diag}(h_{11}, \cdots, h_{mm})$ is the diagonal matrix where h_{ii} is the channel response for the i-th resource, $\mathbf{v} \sim \mathcal{CN}(\mathbf{0}, \sigma_v^2\mathbf{I})$ is the additive Gaussian noise, $\mathbf{x} = \begin{bmatrix}(\mathbf{Cs})^T & \mathbf{p}^T\end{bmatrix}^T$ is the transmit vector where $\mathbf{C}$ is the matrix describing the mapping between the symbol and resource elements, $\mathbf{s}$ is the modulated symbol vector, and $\mathbf{p}$ is the pilot vector.

In contrast, the transmit vector $\mathbf{x}$ of SVC is generated by two operational steps: *sparse mapping* and *pseudo-random spreading* (see Fig. 4.1). First, an information vector $\mathbf{w}$ is mapped to the *position* of the sparse vector $\mathbf{s}$. For example, if 5-bit information vector is mapped into the m-dimensional sparse vector ($m = 9$) with the sparsity $K = 2$, then

$$\begin{aligned}
\mathbf{w} = [0\,0\,0\,0\,0] &\longrightarrow \mathbf{s} = [0\,0\,0\,0\,0\,0\,0\,1\,1] \\
\mathbf{w} = [0\,0\,0\,0\,1] &\longrightarrow \mathbf{s} = [0\,0\,0\,0\,0\,0\,1\,1\,0] \\
&\vdots \\
\mathbf{w} = [1\,1\,1\,1\,1] &\longrightarrow \mathbf{s} = [1\,1\,0\,0\,0\,0\,0\,0\,0].
\end{aligned}$$

When we try to put K nonzero elements in N positions, we have $\binom{N}{K}$ choices in total and thus $b = \lfloor \log_2 \binom{N}{K} \rfloor$ bits information can be encoded by this sparse mapping (in the example above, $N = 9$ and $K = 2$, and thus $b = 5$). In order to transmit the

compressed version of the sparse vector, we further spread the sparse vectors using the pseudo-random codebook $\mathbf{C}$. Note that since the position of nonzero elements in $\mathbf{s}$ can be chosen at random, the codebook matrix $\mathbf{C} = [\mathbf{c}_1 \ \mathbf{c}_2 \ \cdots \ \mathbf{c}_N]$ should be designed such that the transmit vector $\mathbf{x}$ contains enough information to recover the sparse vector $\mathbf{s}$. In this work, we consider the random Bernoulli sequences for the codebook design for simplicity. Nevertheless, by using more sophisticated codebook based on Reed-Muller or Zadoff-chu sequences, one can further improve the decoding performance.

As a result of the sparse mapping and spreading, the overall system can be modeled as an underdetermined sparse system given by

$$\mathbf{y} = \mathbf{HCs} + \mathbf{v}. \tag{4.2}$$

It has been shown from the theory of CS that if entries of the codebook matrix $\mathbf{C}$ are generated at random, then an accurate recovery of the sparse vector is possible with the measurement size (resource size) m being proportional to the sparsity K, i.e., $m = \mathcal{O}(K \log N)$ [59].

4.2.2 Pilot-less SVC

When the channel is a constant or channel variation is very small (i.e., $\mathbf{h} \approx$ const.), which is true for mobile device under static or slowly varying environments, decoding of the SVC packet can be performed without channel information, resulting in the savings of the pilot resources, transmission power, receiver processing time and cost. This scenario can also happen when the short packet is constructed in a narrowband channel. In fact, when the packet size is short, the packet transmission time nT_s (n is the number of symbols in a packet and T_s is the symbol duration) is in general much smaller than the channel coherence time T_c for the moderate mobility ν. ($T_c \gg nT_s$).

For example, when the carrier frequency $f_c = 3.5$ GHz and the mobile speed is $\nu = 15$ Km/h, then $T_c = \frac{9c}{16\pi\nu f_c} = 1.52$ ms is much larger than $T_s = 0.07$ ms [67][1]. Thus the channel remains unchanged or at least channel variation is very marginal for short packet (small n) and hence we can readily assume that $\mathbf{H}$ as $\mathbf{H} = \text{diag}(h, \cdots, h)$. Pilot-less transmission is done by slightly modifying the system model such that the system matrix equals the codebook $\mathbf{C}$ and the sparse vector contains the channel component as well as the sparse information vectors. To be specific,

$$\begin{aligned} \mathbf{y} &= \mathbf{H}\mathbf{x} + \mathbf{v} \\ &= \begin{bmatrix} h & & \\ & \ddots & \\ & & h \end{bmatrix} \mathbf{x} + \mathbf{v} \\ &= h\mathbf{C}\mathbf{s} + \mathbf{v} \\ &= \sum_i (hs_i)\, c_i + \mathbf{v}. \end{aligned} \tag{4.3}$$

Let $\check{\mathbf{s}} = [hs_1 \ \cdots \ hs_N]^T$, then

$$\mathbf{y} = \mathbf{C}\check{\mathbf{s}} + \mathbf{v}. \tag{4.4}$$

Recalling that the goal of the PL-SVC decoding is to identify the nonzero positions of $\mathbf{s}$ vector (equivalently, $\check{\mathbf{s}}$ vector), we can perform the decoding without the channel knowledge, which directly implies that the pilot transmission is unnecessary.

In the PL-SVC decoding, basically, any sparse signal recovery algorithm can be used. In this work, we employ an orthogonal matching pursuit (OMP), most popular greedy sparse recovery algorithm [72]. In essence, the role of a decoding algorithm is to identify the support Ω, locations of nonzero elements. For example, if $\check{\mathbf{s}} = [0 \ 0 \ hs_3 \ 0 \ hs_5 \ 0]$, then $\Omega = \{3, 5\}$. When columns of $\mathbf{C}$ are approximately orthogonal, $\hat{\mathbf{s}} = \mathbf{C}_\Omega^H \mathbf{y} = \mathbf{C}_\Omega^H(\mathbf{C}\check{\mathbf{s}} + \mathbf{v})$ can be a reasonable approximation of $\tilde{\mathbf{s}}$. Based

[1] We use LTE symbol length, i.e., 14 symbols in 1 subframe (1 ms) [63]

on this observation, OMP finds the support element one at each iteration. To be specific, let $\mathbf{C}_{\Omega^j}$ be the submatrix of $\mathbf{C}$ that only contains columns indexed by the element of Ω^j (support of j-th iteration), then an index ω_{j+1} chosen at the $j+1$-th iteration of the OMP algorithm is

$$\omega_{j+1} = \arg\max_{l} |\mathbf{C}_l^H \mathbf{r}^j|^2, \tag{4.5}$$

where $\mathbf{r}^j = \mathbf{y} - \mathbf{C}_{\Omega^j}\hat{\mathbf{s}}^j$ is the residual vector and $\hat{\mathbf{s}}^j = \mathbf{C}_{\Omega^j}^{\dagger}\mathbf{y}$.

4.2.3 PL-SVC Decoding in Multiple Basestation Antennas

As discussed, an element of $\hat{\mathbf{s}}^j$ is hs_i where h is the channel component and thus the detection performance depends heavily on the power of h. When the channel is under deep fading, therefore, the performance would not be satisfactory. When multiple antennas are available at basestation, which is true for most of wireless systems, we can achieve the diversity gain and therefore improve the reliability of the proposed scheme.[2] When multiple, say L, antennas are receiving the PL-SVC packet, a received vector $\tilde{\mathbf{y}}$ can be expressed as the sum of a linear combination of LN columns. That is,

$$\begin{aligned} \tilde{\mathbf{y}} &= \begin{bmatrix} \mathbf{y}^1 & \mathbf{y}^2 & \cdots & \mathbf{y}^L \end{bmatrix}^T \\ &= \underbrace{\begin{bmatrix} \mathbf{C} & & & \\ & \mathbf{C} & & \\ & & \ddots & \\ & & & \mathbf{C} \end{bmatrix}}_{\tilde{\mathbf{C}}} \underbrace{\begin{bmatrix} h^1\mathbf{s} \\ h^2\mathbf{s} \\ \vdots \\ h^L\mathbf{s} \end{bmatrix}}_{\tilde{\mathbf{s}}} + \underbrace{\begin{bmatrix} \mathbf{v}^1 \\ \mathbf{v}^2 \\ \vdots \\ \mathbf{v}^L \end{bmatrix}}_{\tilde{\mathbf{v}}}, \end{aligned} \tag{4.6}$$

where $\mathbf{y}^l$ is the received signal at the l-th antennas, $\mathbf{h}^l$ is the channel response at the l-th antenna, and $\mathbf{v}^l$ is the corresponding noise vector. In order to exploit the block sparsity of the aggregated sparse vector $\tilde{\mathbf{s}}$, we rearrange the $\tilde{\mathbf{s}}$ vector as $[h^1\mathbf{s} \ \cdots \ h^L\mathbf{s}]^T$

[2] Equivalently, by the cooperation of more than one reception points (e.g., basestation, radio remote head, radio unit) in the decoding of PL-SVC, we can achieve the diversity gain and thus further improve the performance.

$\longrightarrow [h^1 s_1 \cdots h^L s_1 \ \cdots \ h^1 s_L \cdots h^L s_L]^T$. Thus,

$$\tilde{\mathbf{y}} \ = \ \begin{bmatrix} \bar{\mathbf{C}}^{(1)} & & & \\ & \bar{\mathbf{C}}^{(2)} & & \\ & & \ddots & \\ & & & \bar{\mathbf{C}}^{(N)} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{h}} s^1 \\ \bar{\mathbf{h}} s^2 \\ \vdots \\ \bar{\mathbf{h}} s^N \end{bmatrix} + \begin{bmatrix} \bar{\mathbf{v}}^1 \\ \bar{\mathbf{v}}^2 \\ \vdots \\ \bar{\mathbf{v}}^N \end{bmatrix}, \tag{4.7}$$

where $\bar{\mathbf{C}}^{(i)} = \text{diag}(\mathbf{c}_i \ \mathbf{c}_i \ \cdots \ \mathbf{c}_i)$ is a block diagonalization of i-th spreading sequence, $\bar{\mathbf{h}} = [h^1 \ h^2 \ \cdots \ h^L]^T$ is the stacked channel response, and $\bar{\mathbf{v}}^i = [\mathbf{v}^1(i) \ \mathbf{v}^2(i) \ \cdots \ \mathbf{v}^N(i)]^T$ is the rearranged noise vector. One can easily see that the sparse recovery algorithm exploiting the block sparsity (e.g., block OMP algorithm [72]) can be used to find a nonzero block. The nonzero block selection rule of BOMP at $(j+1)$-th iteration is

$$\omega_{j+1} = \arg\max_{l} \|\bar{\mathbf{C}}^{(l)} \bar{\mathbf{r}}^j\|_2^2, \tag{4.8}$$

where $\bar{\mathbf{r}}^j = \tilde{\mathbf{y}} - \bar{\mathbf{C}}^{(\Omega^j)} \hat{\mathbf{s}}^j$ is the residual vector.

When the channels among multiple antennas are independent, diversity gain proportional to the number of antennas can be obtained. Note that the computational burden to process this SIMO channel decoding is fairly small because the decoding operation is simply finished in K-th iteration regardless of the number of antennas. In fact, since the sparsity K is known to the receiver, computational overhead to convert the multiple measurement vector problem in (6) to the single measurement vector problem in (7) is negligible. The proposed PL-SVC decoding algorithm is summarized in Table 4.1.

Table 4.1: The proposed PL-SVC decoding algorithm

Input:
Measurement $\tilde{\mathbf{y}}$, sensing matrix $\tilde{\mathbf{C}}$, sparsity K, stop threshold ϵ
expansion size E, maximum candidate l_{max}
Output:
Support set $\mathcal{S}_{\tilde{\mathbf{s}}}$

Initialization:
$l := 0$ (candidate order)
$\rho := \infty$ (minimum magnitude of residual)
While: $l < l_{max}$ and $\epsilon < \rho$ **do**
$l := l + 1$, $\bar{\mathbf{r}}^0 := \tilde{\mathbf{y}}$
$[p_1, ..., p_K] :=$ compute_p$_k(l, E)$ (*compute layer order*)
for $k = 1$ **to** K **do** (*investigate l-th candidate*)
$\omega_{p_k} := \arg \max_{|\pi|=E} \|(\langle \bar{\mathbf{C}} \bar{\mathbf{r}}^{k-1} \rangle)_\pi\|_2^2$
$\mathcal{S}_l^k := \mathcal{S}_l^{k-1} \cup \{\omega_{p_k}\}$ (*construct a path in k-th layer*)
$\hat{\mathbf{s}}^k := \bar{\mathbf{C}}^{(\mathcal{S}_l^k)\dagger} \tilde{\mathbf{y}}$
$\bar{\mathbf{r}}^k := \tilde{\mathbf{y}} - \bar{\mathbf{C}}^{(\mathcal{S}_l^k)} \hat{\mathbf{s}}^k$ (*update residual*)
$\hat{\mathcal{S}}^k := \mathcal{S}_l^k$ (*update support set*)
end for
if $\|\bar{\mathbf{r}}^K\|_2^2 < \rho$ **then** (*update the smallest residual*)
$\rho := \|\bar{\mathbf{r}}^K\|_2^2$
$\mathcal{S}_{\tilde{\mathbf{s}}}^* := \hat{\mathcal{S}}^K$
end if
end while
return $\mathcal{S}_{\tilde{\mathbf{s}}}^*$

function compute_p$_k(l, E)$
$t := l - 1$
for $k = 1$ **to** K **do**
$p_k :=$ mod $(t, E) + 1$ and $t :=$ floor(t/E)
end for
return $[p_1, ..., p_K]$
end function

4.3 Simulations and Discussions

4.3.1 Simulation Setup

In this section, we examine the performance of the proposed PL-SVC technique in 5G uplink scenario. Our simulation setup is based on the OFDM systems in the New Radio (NR). For comparison, we investigate the performance of the conventional SVC with pilot transmission and physical uplink share channel (PUSCH) transmission in 4G

LTE [63].[3] In the SVC processing, we set $N = 96$ and $K = 2$. For the conventional SVC transmission with pilot, $m = 48$ resources are occupied for spreading resource and the remaining resources are used for pilot symbols so that $r = 84$ resources are used in total. Whereas, in the PL-SVC transmission scheme, we set $m = r = 42$ and $m = r = 84$ for the fair comparison. As a performance measure, we use the packet error rate (PER).

4.3.2 Simulation Results

In Fig. 4.2, we evaluate the PER performance of PL-SVC, conventional SVC, and PUSCH with and without pilot transmission as a function of SNR ($b = 12$). First, we compare the performance of the conventional SVC and PUSCH (dotted lines in Fig. 4.2). We observe that SVC outperforms the conventional PUSCH, achieving more than 3 dB gain at 10^{-4} PER point. Second, we compare PER performance of PL-SVC and PUSCH without pilot transmission. We observe that the PUSCH-based transmission is simply not working while PL-SVC performs slightly worse than the conventional SVC at PER $= 10^{-4}$. Next, we compare performance of the conventional SVC and PL-SVC. When the spreading lengths are equivalent ($m = 42$), the PER performance of PL-SVC is degraded in low SNR regime but the performance gap decreases with SNR. Meanwhile, when the same amount of resources is used ($r = 84$), PL-SVC outperforms SVC, achieving more than 1 dB gain at PER $= 10^{-4}$.

In order to test the performance of SVC in SIMO channel scenarios, we consider the three distinct cases (i.e., number of received antennas is $L = 1, 2$, and 4. From Fig. 4.3, when we observe that the performance of the PL-SVC technique improves with L. For example, $L = 4$, we observe 1.5 dB gain over the case with $L = 1$. We note that the conventional transmission scheme might achieve similar the performance gain, but at the expense of large computational overhead caused by multiple channel estimations.

[3]The information is encoded with the code rate $\frac{1}{3}$ Turbo code.

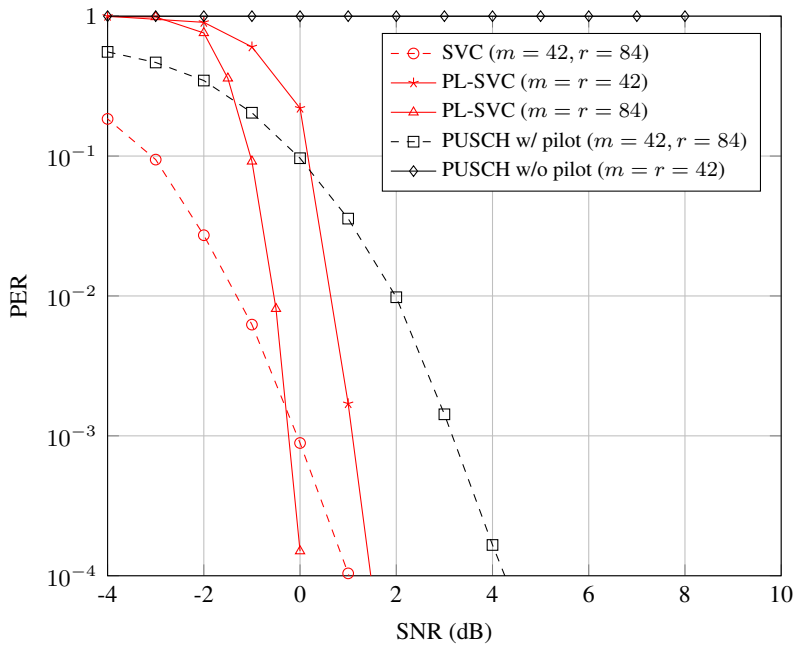


Figure 4.2: PER performance of PL-SVC and the conventional schemes ($b = 12$).

4.4 Summary

In this chapter, we proposed the pilot-less sparse vector coding (PL-SVC) suitable for the short packet transmission in URLLC and mMTC scenarios. The key idea behind the proposed PL-SVC is to transmit the location information in the form of a sparse vector and then to decode the location information via the compressed sensing technique. The PL-SVC transmission scheme does not require pilot transmission, also very easy to implement, and can be easily extended to the SIMO and MIMO configurations. We show from the numerical experiments in the 5G OFDM uplink scenario that the proposed PL-SVC is effective in the short packet transmission and outperforms the conventional SVC and PUSCH transmission schemes.

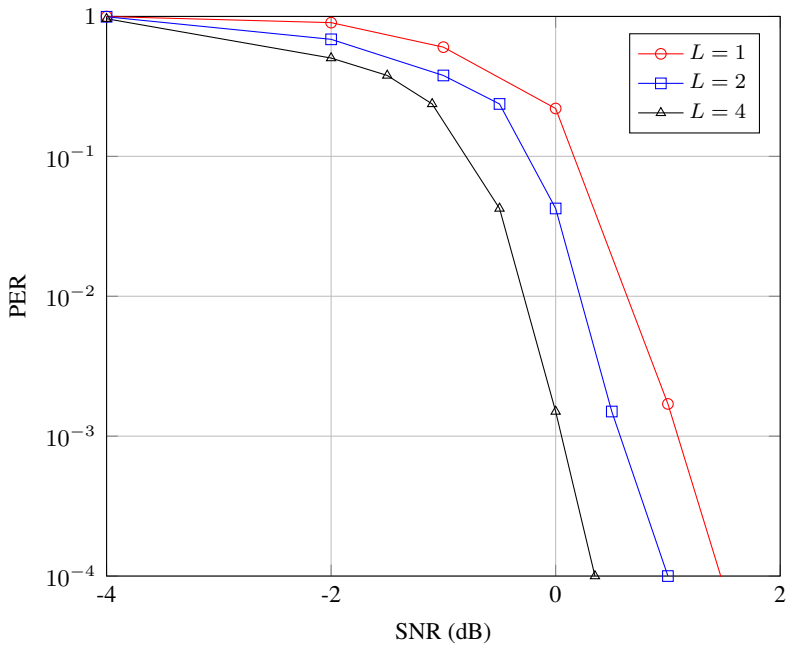


Figure 4.3: PER performance of PL-SVC for the SIMO channels as a function of SNR ($b = 12, m = r = 42$).

Chapter 5

Joint Analog and Quantized Feedback via Sparse Vector Coding

In this chapter, I explore a potential benefit of estimation accuracy in sparse recovery. Key idea behind the proposed technique is to transmit both analog and quantized data using the sparse vector transformation. By mapping the quantized data into positions of sparse vector and then setting the analog data as the magnitude of nonzero elements, two different types of data can be jointly mapped to the sparse vector. After spreading with sequences from multiple Zadoff-chu (ZC) sets, we obtain an underdetermined sparse system and it is now well-known that the theory of compressed sensing guarantees an accurate recovery of a sparse vector and minimize mean square error of recovered vector with a relatively small number of measurement [59]. In fact, the decoding of quantized data is performed by the sparse signal recovery (more accurately, identification of nonzero positions in the transmit sparse vector) and the decoding of analog data is performed by the recovery of support value.

5.1 Introduction

Mission-critical machine type communications is one of service categories that newly introduced in 5G systems to support various applications including factory automa-

tions, unmanned vehicles, and smart cites [46]. One notable observation in these applications is that the transmit information of a packet is control (command) type information (e.g., on/off, open/close, move left/right, start/stop, rotate/shift, and speed up/down) so that the amount of information to be delivered is much smaller than the packet of 4G systems [64]. Since the current wireless transmission strategy designed to maximize the coding gain by transmitting capacity achieving long codeblock is not relevant to these scenarios, entirely new transmission strategy to support the ultra-short packet is required [6-49-50].

Recently, sparse vector coding (SVC) has been proposed for transmitting ultra-short packet which is suitable for compressed sensing (CS)-based decoding. Since the data information is mapped into the position of sparse vector, whole data packet can be decoded by detection of a few support. In [3], SVC-based packet is used for mMTC transmission by overlapping multiple SVC-based packets into the same resource. Using multiuser CS decoding, SVC-based transmission has robustness against the co-channel interference and also provide comparable performance than other NOMA schemes. However, [2][3] are somewhat limited to use the position of sparse vector, not the magnitude of the nonzero elements which has potential to transmit additional information.

The purpose of this paper is to propose a new technique for transmitting ultra-short packet called joint sparse vector coding (JSVC). Key idea behind the proposed technique is to map the transmit data into positions of sparse vector and set the analog side-information as the magnitude of nonzero elements. By doing so, two different types of information can be jointly mapped to the sparse vector. After spreading with multiple sequences from non-square Zadoff-chu (ZC) codebook, we obtain an underdetermined sparse system for which the principle of compressed sensing can be applied [72]. It is now well-known that the theory of compressed sensing guarantees an accurate recovery of a sparse vector with a relatively small number of measurement [59], which can be achieved in our case via the ZC-based multi-sequence spreading. The decoding of

JSVC is performed by the support recovery (more accurately, identification of nonzero positions in the transmit sparse vector) and the estimation of analog side-information is performed by the recovery of support values. Noting that the side-information is used for various purpose such as improving the accuracy of information or indicating the additional information. The proposed scheme is very simple to implement and can be applied to wide variety of wireless environments. From the numerical evaluations and decoder performance analysis, we demonstrate that the JSVC-based packet is very effective in ultra-short packet transmission of 5G communications and outperforms the conventional approaches by a large margin.

The remainder of this chapter is organized as follows. In Section 5.2, we introduce the system model of SVC-based data transmission. In Section 5.3, we present the proposed SVC schemes for joint analog and quantized data and explain the operations of transmission and reception. In Section 5.4, we analyze the performance of SVC-based uplink transmission. In section 5.5, we explain three key applications for the proposed scheme for supporting eMBB, URLLC, and mMTC. In section 5.6, we present simulation results to verify the performance of the proposed scheme, and conclude our work in Section 5.7.

5.2 System Model for Joint Spase Vector Coding

We consider the data transmission in orthogonal frequency division multiplexing (OFDM) system. In the transmitter, data information is transmitted in a form of modulated symbol vector $\mathbf{x} \in \mathbb{C}^{N\times 1}$. The corresponding received vector $\mathbf{y} \in \mathbb{C}^{m\times 1}$ is given by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{5.1}$$

where $\mathbf{H} \in \mathbb{C}^{m\times m}$ is the diagonal matrix whose diagonal element h_{ii} is the channel component for each resource, and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_v^2\mathbf{I})$ is the additive Gaussian noise.

In this work, we use joint encoding for data information with analog side-information,

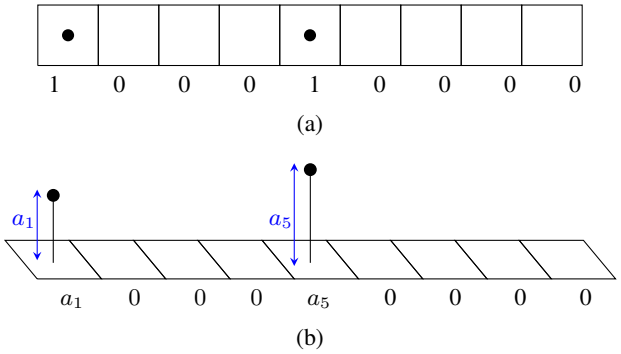


Figure 5.1: Illustration of JSVC encoding: (a) Basic SVC encoding where data is mapped into the position of a sparse vector only; (b) JSVC encoding where transmit data is mapped into the position of a sparse vector and additional side-information is mapped to the magnitudes of support.

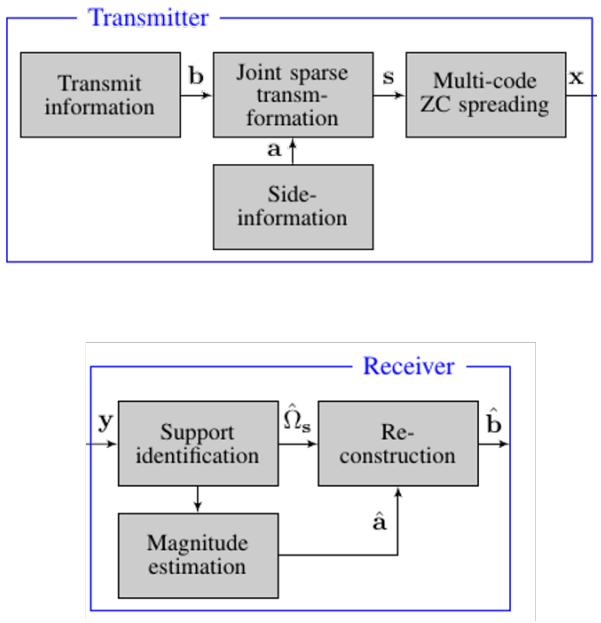


Figure 5.2: System model of proposed JSVC-based ultra-short packet transmission.

referred as *joint SVC encoding* (JSVC). As illustrated in Fig. 5.1, the basic of the SVC technique is to map the information into the positions of a sparse vector $\mathbf{s}$. When we choose K out of N symbol positions, we can encode $b_{svc} = \lfloor \log_2 \binom{N}{K} \rfloor$ bits of information. The proposed JSVC is to extend SVC for using the height of dots in the boxes. By using height information, one can open another dimension to deliver additional data ($b_{jsvc} > b_{svc}$).

The first step of JSVC encoding is mapped the transmit data at K nonzero positions among N symbol vector $\mathbf{s}$.

$$\begin{array}{ccc} 0\,0\,0\,0\,0 & \longleftrightarrow & 0\,0\,0\,0\,0\,0\,0\,0\,1\,1 \\ 0\,0\,0\,0\,1 & \longleftrightarrow & 0\,0\,0\,0\,0\,0\,0\,1\,0\,1 \\ \vdots & \vdots & \vdots \\ \underbrace{1\,1\,1\,1\,1}_{b\text{-bit digital information}(b=5)} & \longleftrightarrow & \underbrace{1\,0\,0\,0\,0\,0\,0\,0\,0\,1}_{K-\text{sparse vector } \mathbf{s}\ (K=2)} \end{array} \tag{5.2}$$

Next step is to map analog side-information as the magnitude of nonzero elements. That is,

$$\begin{array}{ccc} 0\,0\,0\,0\,0\,0\,0\,0\,1\,1 & \longleftrightarrow & 0\,0\,0\,0\,0\,0\,0\,0\,a_9\,a_{10} \\ 0\,0\,0\,0\,0\,0\,0\,1\,0\,1 & \longleftrightarrow & 0\,0\,0\,0\,0\,0\,0\,a_8\,0\,a_{10} \\ \vdots & \vdots & \vdots \\ \underbrace{1\,0\,0\,0\,0\,0\,0\,0\,0\,1}_{K-\text{sparse vector } \mathbf{s}\ (K=2)} & \longleftrightarrow & \underbrace{a_1\,0\,0\,0\,0\,0\,0\,0\,0\,a_{10}}_{\text{Joint}K-\text{sparse vector } \mathbf{s}\ (K=2)} \end{array} \tag{5.3}$$

where $0 < a_i \leq 1$ for $i = 0, ..., N$.

After the sparse mapping, each nonzero element in $\mathbf{s}$ is spread into m resources using the spreading codebook $\mathbf{C}$. As a result of the multi-code spreading, the modulated symbol vector $\mathbf{s}$ in (5.1) is replaced with $\mathbf{x} = \mathbf{C}\mathbf{s}$ the codebook matrix $\mathbf{C} = [\mathbf{c}_1\ \mathbf{c}_2\ \cdots\ \mathbf{c}_N]$ where $\mathbf{c}_i = [c_{i1}\ c_{i2}\ \cdots\ c_{im}]^T$ is the spreading sequence.

For accurate recovery, we consider Zadoff-chu (ZC) sequence based codebook. ZC is a complex-valued mathematical sequence where cyclically shifted versions of

the sequence imposed on a signal result in zero correlation with one another. Also, ZC provides low correlation between sequence from difference root set. Due to these reason, ZC sequence is used for many uplink transmission including random access signal, pilot signal, control channel, and data channel. Let $\mathbf{Z}_q \in \mathbb{C}^{m\times m}$ be the ZC sequence set with root number q, that is,

$$\mathbf{Z}_q = [\mathbf{z}_q^0 \cdots \mathbf{z}_q^{m-1}], \tag{5.4}$$

where $z_q^{\alpha}(n) = \exp\left(-j\frac{\pi q(n+\alpha)(n+1+\alpha)}{m}\right)$. Since $\mathbf{Z}_q$ is a square matrix, the codebook matrix $\mathbf{C}$ uses more than two root sequence sets having low correlation between sets as follows:[1]

$$\mathbf{C} = [\mathbf{Z}_{q_1} \cdots \mathbf{Z}_{q_l}]. \tag{5.5}$$

Using ZC based codebook, the system matrix provide approximately zero correlation between columns and also recover the support accurately. For example, when $K = 2$ and f-th and g-th elements are non-zero, the received vector $\mathbf{y}$ is given by

$$\mathbf{y} \quad = \quad h(\mathbf{c}_f a_f + \mathbf{c}_g a_g) + n. \tag{5.6}$$

In general, the received vector $\mathbf{y}$ is given by

$$\begin{aligned}\mathbf{y} &= \mathbf{HCs} + \mathbf{v} \\ &= \begin{bmatrix} h_{11} & & \\ & \ddots & \\ & & h_{mm} \end{bmatrix} \begin{bmatrix} c_{11} & \cdots & c_{Nm} \\ \vdots & \ddots & \vdots \\ c_{1m} & \cdots & c_{Nm} \end{bmatrix} \begin{bmatrix} s_1 \\ \vdots \\ s_N \end{bmatrix} + \begin{bmatrix} n_1 \\ \vdots \\ n_m \end{bmatrix},\end{aligned} \tag{5.7}$$

where the system matrix $\mathbf{HC}$ is the underdetermined system with $m < N$. Since

[1] Root number of ZC sequence should be a prime number. To use multiple root sequence set, one can apply computational search for selecting root numbers as in 4G LTE systems.

the transmitter is often power-limited due to heavy pathloss, we assume that transmission bandwidth is narrow. Thus, the channel is constant or channel variation in frequency-domain is small, which is true for mobile devices under static or slowly moving environments ($\mathbf{s}^{'} = h\mathbf{s}$) and thus we further have

$$\mathbf{y} = h\mathbf{C}\mathbf{s} + \mathbf{n}. \tag{5.8}$$

Since the positions of nonzero elements are chosen at random, the codebook matrix $\mathbf{C}$ should be designed such that the transmit vector $\mathbf{x}$ contains enough information to recover the sparse vector $\mathbf{s}$ irrespective of the selection of the nonzero positions.

5.3 Sparse Recovery Algorithm and Performance Analysis

As mentioned, JSVC decoding is done by two step: the identification of the support and estimation of supports. To this end, any sparse recovery algorithm can be employed for this purpose. In this work, we employ the greedy sparse recovery algorithm in the decoding of the JSVC-encoded packet. The received signal $\mathbf{y}$ can be modified by multiplying $\frac{h^H}{\|h\|_2^2}$ and thus we have

$$\begin{aligned} \tilde{\mathbf{y}} &= \frac{h^H}{\|h\|_2^2}\mathbf{y} \\ &= \mathbf{C}\mathbf{s} + \tilde{\mathbf{n}}, \end{aligned} \tag{5.9}$$

where $\tilde{\mathbf{n}} = \frac{h^H}{\|h\|_2^2}\mathbf{n}$ is modified noise vector. The modified received signal $\tilde{\mathbf{y}}$ and the system matrix $\mathbf{\Phi}$ are decoupled into their real and imaginary parts.

$$\begin{bmatrix} \Re\{\tilde{\mathbf{y}}\} \\ \Im\{\tilde{\mathbf{y}}\} \end{bmatrix} = \begin{bmatrix} \Re\{\mathbf{\Phi}\} & -\Im\{\mathbf{\Phi}\} \\ \Im\{\mathbf{\Phi}\} & \Re\{\mathbf{\Phi}\} \end{bmatrix} \begin{bmatrix} \Re\{\mathbf{s}\} \\ \Im\{\mathbf{s}\} \end{bmatrix} + \begin{bmatrix} \Re\{\tilde{\mathbf{n}}\} \\ \Im\{\tilde{\mathbf{n}}\} \end{bmatrix}, \tag{5.10}$$

where $\Re\{\cdot\}$ and $\Im\{\cdot\}$ are the real and imaginary operations, respectively. Since $\Im\{\mathbf{s}\} = \mathbf{0}$, we further have

$$\begin{aligned} \begin{bmatrix} \Re\{\tilde{\mathbf{y}}\} \\ \Im\{\tilde{\mathbf{y}}\} \end{bmatrix} &= \begin{bmatrix} \Re\{\mathbf{\Phi}\} \\ \Im\{\mathbf{\Phi}\} \end{bmatrix} \Re\{\mathbf{s}\} + \begin{bmatrix} \Re\{\tilde{\mathbf{n}}\} \\ \Im\{\tilde{\mathbf{n}}\} \end{bmatrix}, \\ \check{\mathbf{y}} &= \check{\mathbf{\Phi}}\check{\mathbf{s}} + \check{\mathbf{n}}, \end{aligned} \tag{5.11}$$

where $\check{\mathbf{s}}$ is the real part of the transmit symbol vector. Since $\check{\mathbf{s}}$ has only K nonzero elements, the modified received vector $\check{\mathbf{y}}$ can be expressed as a linear combination of K columns in $\check{\mathbf{\Phi}}$ perturbed by the modified noise. In view of this, main task of the SVC decoding is to identify the columns in $\check{\mathbf{\Phi}}$ participating in the received vector. In each iteration, greedy sparse recovery algorithm identifies the column of $\check{\mathbf{\Phi}}$ one at a time using a greedy strategy. Specifically, let $\check{\mathbf{\Phi}}_{\Omega_{\mathbf{s}}^{j-1}}$ be the submatrix of $\check{\mathbf{\Phi}}$ that only contains columns indexed by $\check{\Omega}_{\mathbf{s}}^{j-1}$, then the index $\check{\omega}_j$ chosen at the j-th iteration of the greedy algorithm is given by[2]

$$\omega_j = \arg\max_l |\check{\mathbf{\Phi}}_l^T \mathbf{r}^{j-1}|^2, \tag{5.12}$$

where $\mathbf{r}^{j-1} = \mathbf{y} - \check{\mathbf{\Phi}}_{\Omega_{\mathbf{s}}^{j-1}} \hat{\mathbf{s}}^{j-1}$ is the residual vector and $\hat{\mathbf{s}}^{j-1} = \check{\mathbf{\Phi}}^{\dagger}_{\Omega_{\mathbf{s}}^{j-1}} \mathbf{y}$ is the estimate of $\mathbf{s}$ at $(j-1)$-th iteration.[3]

For analysis simplicity, we set $K = 2$ and then extend to the general case. Without loss of generality, we assume that f and g-th elements of $\mathbf{s}$ are nonzero (i.e., $\Omega_{\mathbf{s}} = \{f, g\}$). Further, we set the information vector such that $s_l = 1$ and $s_m = a$ ($0 < s_m < s_l$). Let $\mathcal{S}^1$ and $\mathcal{S}^2$ be the success probability that the support element s_f and s_g are chosen in the first and second iteration, respectively. The probability that the data information is successfully decoded is expressed

$$P_{succ} = \mathrm{P}(\Omega_s^* = \Omega_s)$$

[2] If $\Omega = \{1, 3\}$, then $\mathbf{\Phi}_\Omega = [\boldsymbol{\phi}_1 \ \boldsymbol{\phi}_3]$.

[3] $\mathbf{\Phi}^\dagger = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ is the pseudo-inverse of $\mathbf{\Phi}$.

$$
\begin{aligned}
&= \mathrm{P}(\mathcal{S}^1, \mathcal{S}^2) \\
&= \mathrm{P}(\mathcal{S}^2|\mathcal{S}^1)\mathrm{P}(\mathcal{S}^1). \qquad (5.13)
\end{aligned}
$$

In each iteration, N decision statistics $|\frac{\check{\boldsymbol{\phi}}_l^T}{m}\mathbf{r}^{k-1}|$ for $l = 1, \cdots, N$ are computed. In order to identify the support element s_f in the first iteration, we should have $|\frac{\check{\boldsymbol{\phi}}_f^T}{m}\mathbf{r}^0| \geq |\frac{\check{\boldsymbol{\phi}}_g^T}{m}\mathbf{r}^0|$. Also, in order to identify the support element s_g in the second iteration, we should have $|\frac{\check{\boldsymbol{\phi}}_g^T}{m}\mathbf{r}^1| \geq \max_l |\frac{\check{\boldsymbol{\phi}}_l^T}{m}\mathbf{r}^1|$. Thus the success probability for a given channel realization h are

$$
\begin{aligned}
\mathrm{P}(\mathcal{S}^1|h) &= \mathrm{P}\left(\left|\langle\frac{\check{\boldsymbol{\phi}}_f}{m}, \mathbf{r}^0\rangle\right|^2 \geq \left|\langle\frac{\check{\boldsymbol{\phi}}_g}{m}, \mathbf{r}^0\rangle\right|^2\right), \\
\mathrm{P}(\mathcal{S}^2|\mathcal{S}^1, h) &= \prod_{i=1, i\neq f}^{N} \mathrm{P}\left(\left|\langle\frac{\check{\boldsymbol{\phi}}_g}{m}, \mathbf{r}^1\rangle\right|^2 \geq \left|\langle\frac{\check{\boldsymbol{\phi}}_i}{m}, \mathbf{r}^1\rangle\right|^2\right), \qquad (5.14)
\end{aligned}
$$

where $\langle \mathbf{a}, \mathbf{b}\rangle$ is the inner product between vector $\mathbf{a}$ and $\mathbf{b}$. First, we consider the first iteration. Noting that $s_f = 1$ and $s_g = a$, the left element in (13) can be expressed as

$$
\begin{aligned}
\left|\langle\frac{\tilde{\boldsymbol{\phi}}_f}{m}, \mathbf{r}^0\rangle\right| &= \left|\langle\frac{\tilde{\boldsymbol{\phi}}_f}{m}, \tilde{\boldsymbol{\phi}}_f s_f + \tilde{\boldsymbol{\phi}}_g s_g + \tilde{\mathbf{n}}\rangle\right| \\
&= \left|\frac{\mathbf{c}_f^H \mathbf{c}_f}{m} x_f + \frac{\mathbf{c}_f^H \mathbf{c}_g}{m} x_g + \frac{\tilde{\boldsymbol{\phi}}_f^T \check{\mathbf{n}}}{m}\right|. \qquad (5.15)
\end{aligned}
$$

Let $\frac{\mathbf{c}_f^H \mathbf{c}_g}{m} = \mu_{fg}$ and $\frac{\tilde{\boldsymbol{\phi}}_f^T \tilde{\mathbf{n}}}{m} = z_f$, we have

$$
\begin{aligned}
\left|\langle\frac{\tilde{\boldsymbol{\phi}}_f}{m}, \mathbf{r}^0\rangle\right| &= \left|\frac{\mu_{ff}}{m} x_f + \frac{\mu_{fg}}{m} x_g + z_f\right| \\
&= \left|1 + \mu_{fg} a + z_f\right|, \qquad (5.16)
\end{aligned}
$$

where $\mu_{ii} = m$ for any i. In a similar way, we have

$$
\left|\langle\frac{\tilde{\boldsymbol{\phi}}_g}{m}, \mathbf{r}^0\rangle\right| = |a + \mu_{gf} + z_g|. \qquad (5.17)
$$

Since the cross-correlation between two ZC sequences is theoretically zero, (14) can be approximated as

$$\begin{aligned}
&\mathrm{P}\left(\left|\langle\frac{\tilde{\boldsymbol{\phi}}_f}{m},\mathbf{r}^0\rangle\right|^2 \geq \left|\langle\frac{\tilde{\boldsymbol{\phi}}_g}{m},\mathbf{r}^0\rangle\right|^2\right) \\
&\approxeq \mathrm{P}\left(|1+z_f|^2 \geq |a+z_g|^2\right) \\
&= \mathrm{P}\left(1+z_f > |a+z_g|\right)\mathrm{P}\left(1+z_f>0\right) \\
&\qquad\qquad + \mathrm{P}\left(-1-z_f > |a+z_g|\right)\mathrm{P}\left(1+z_f<0\right) \\
&> \mathrm{P}\left(1+z_f > a+z_g\right)\mathrm{P}\left(1+z_f>0\right)\mathrm{P}\left(a+z_g>0\right), \qquad (5.18)
\end{aligned}$$

Let $z_f \sim \mathcal{N}(0, c\sigma_v^2)$ where $c = \frac{1}{h_r^2+h_i^2}$, the first term in (18) is lower bounded as

$$\begin{aligned}
\mathrm{P}\left(1+z_f > a+z_g\right) &= \mathrm{P}\left(z_f - z_g > -(1-a)\right) \\
&= 1 - Q\left(-\frac{1-a}{\sqrt{2c}\sigma_v}\right) \\
&\geq 1 - \exp\left(-\frac{(1-a)^2}{4c\sigma_v^2}\right). \qquad (5.19)
\end{aligned}$$

In a similar way, the second and third term are lower bounded as

$$\begin{aligned}
\mathrm{P}\left(1+z_f>0\right) &= \mathrm{P}\left(z_f > -1)\right) \\
&= 1 - Q\left(\frac{-1}{\sqrt{c}\sigma_v}\right) \\
&\geq 1 - \exp\left(-\frac{1}{2c\sigma_v^2}\right). \qquad (5.20)
\end{aligned}$$

$$\mathrm{P}\left(a+z_g>0\right) \geq 1 - \exp\left(-\frac{a^2}{2c\sigma_v^2}\right). \qquad (5.21)$$

By plugging (19), (20), and (21) into (14), we have

$$\mathrm{P}\left(\left|\langle\frac{\boldsymbol{\phi}_f}{m},\mathbf{r}^0\rangle\right| \geq \left|\langle\frac{\boldsymbol{\phi}_g}{m},\mathbf{r}^0\rangle\right|, h\right) \qquad (5.22)$$

$$\geq \left(1-e^{\frac{-(1-a)^2}{4c\sigma_v^2}}\right)\left(1-e^{-\frac{1}{2c\sigma_v^2}}\right)\left(1-e^{-\frac{a^2}{2c\sigma_v^2}}\right).$$

After taking exception with respect to the channel h, we have

$$\mathrm{P}(\mathcal{S}^1) \geq 1-\left(1+\frac{(1-a)^2}{2\sigma_v^2}\right)^{-2}, \tag{5.23}$$

where a is large and

$$\mathrm{P}(\mathcal{S}^1) \geq 1-\left(1+\frac{a^2}{\sigma_v^2}\right)^{-2}, \tag{5.24}$$

where a is small.

Next, we move to the success probability for the second iteration when the first iteration is successful. Since s_f is known to the receiver, the estimation of s_f is not required. Then, the residual $\mathbf{r}^1$ can be expressed as

$$\begin{aligned}\mathbf{r}^1 &= \mathbf{r}^0 - \mathbf{\Phi}_{\Omega_{\mathbf{s}}^1}\hat{\mathbf{s}}^1 \\ &= \mathbf{r}^0 - \boldsymbol{\phi}_f s_f \\ &= \boldsymbol{\phi}_g s_g + \tilde{\mathbf{n}}.\end{aligned} \tag{5.25}$$

After taking similar steps, we have that $\mathrm{P}(\mathcal{S}^2|\mathcal{S}^1)$ satisfies (we skip the detailed steps for brevity)

$$\begin{aligned}\mathrm{P}(\mathcal{S}^2|\mathcal{S}^1) &= \prod_{i=1,i\neq f}^{N}\mathrm{P}\left(\left|\langle\frac{\check{\boldsymbol{\phi}}_g}{m},\mathbf{r}^1\rangle\right|^2 \geq \left|\langle\frac{\check{\boldsymbol{\phi}}_i}{m},\mathbf{r}^1\rangle\right|^2\right) \\ &\geq \left(1-\left(1+\frac{1}{2\sigma_v^2}\right)^{-2}\right)^{N-1},\end{aligned} \tag{5.26}$$

when a is large and

$$\mathrm{P}(\mathcal{S}^2|\mathcal{S}^1) \;\geq\; \left(1-\left(1+\frac{a^2}{\sigma_v^2}\right)^{-2}\right)^{N-1}, \tag{5.27}$$

when a is small. Combining (23) and (26) and also (24) and (27), we have the success probability of JSVC-based packet. That is,

$$P_{succ} \geq \left(1-\left(1+\frac{(1-a)^2}{2\sigma_v^2}\right)^{-2}\right)\left(1-\left(1+\frac{1}{2\sigma_v^2}\right)^{-2}\right)^{N-1}, \tag{5.28}$$

for large a and we have

$$P_{succ} \geq \left(1-\left(1+\frac{a^2}{2\sigma_v^2}\right)^{-2}\right)^{N}, \tag{5.29}$$

for small a. This reveal that when a is small, the decoding performance is bounded with the detection performance of nonzero element having smallest magnitude (s_g). After detecting all nonzero elements, one can estimate analog side-information vector $\hat{\mathbf{a}}$ as

$$\hat{\mathbf{a}} \;=\; \check{\mathbf{\Phi}}^{\dagger}_{\Omega_{\hat{\mathbf{s}}}}\mathbf{y}. \tag{5.30}$$

5.4 Applications

In this section, we present three examples for using JSVC transmission. In each example, analog side-information is used exclusively. In the first example, side-information is used for minimize quantization error. The second one, side-information is used to indicate coefficient of operating function. The last one, side-information is to indicate how to interpret the data information.

5.4.1 Linear Interpolation of Sensing Information

In this example, the analog side-information is used for recovering the data information. This can be used for IoT applications reporting the sensing information (e.g., temperature, pressure, wind speed, and traffic load) and also device's status reports (e.g., pathloss, buffer status, and power margin). For example, when the device reports 2.45 degree of Celsius temperature scale with quantization level of 2 degree (0, 2, 4 degrees), the information vector is mapped to indicate 4 degree with 2 nonzero support $\mathbf{\Omega_b}$. In addition, by setting analog side-information $a_1 = 1$ and $a_2 = 0.775$, the receiver can interpolate the sensing information using the side-information (e.g., $2.45 = 2 + 2 \times (a_1 - a_2)$).

5.4.2 Linear Combined Feedback

The analog side-information can be used to indicate how to combine multiple information vectors. When there are two dominant paths between the transmitter and receiver, the channel vector h can be feedback with indicating two directions θ_1 and θ_2. To feed these angles, information vectors $\mathbf{b}_i = f(\theta_i)$ are mapped to the support where $f(\cdot)$ is mapping function between angle to codebook. Using analog side-information $a_1 = 1$ and $a_2 = a$, the channel vector $h = \frac{1}{1+a_2} f^{-1}(b_1) + \frac{a_2}{1+a_2} f^{-1}(b_2)$ where the function f^{-1} is remapping function can be reconstructed. Instead of transmitting one dominant precoder index $\mathbf{b}_1$ in 4G systems, multiple indexes $(\mathbf{b}_1, \mathbf{b}_2)$ can be reported as in LTE-Advanced Pro systems and provide combining weights of reported indexes through the analog side-information.

5.4.3 One-shot Packet Transmission

By using relative magnitude of analog side-information, one can control the detection order of non-zero element. Let f-th and g-th elements be the nonzero elements having $a_f > a_g$, the receiver would detect f-th element first. There are two options for using the detecting order. The first option is to use one-shot packet where pilot and data

symbols are superposed and transmitted together. When the pilot symbols are mapped to the f-th element with larger magnitude than g-th element, (8) can be expressed as

$$\begin{aligned} \mathbf{y} &= \mathbf{C}h\mathbf{s} + \mathbf{n} \\ &= \mathbf{C}\mathbf{u} + \mathbf{n} \end{aligned} \tag{5.31}$$

where $\mathbf{u} = [0 \cdots 0\ h\ 0 \cdots 0\ ha_g\ 0 \cdots 0]$ where $a_f = 1, a_g < 1$. In this case, at the first iteration, the j-th indexed nonzero element can be detected firstly with high probability and thus $\hat{h}$ can be estimated. Then, the residual $\mathbf{r}^1$ can be expressed as

$$\begin{aligned} \mathbf{r}^1 &= \mathbf{r}^0 - \hat{h}\boldsymbol{\phi}_f \\ &= \hat{h}\boldsymbol{\phi}_g a_g + \mathbf{n}. \end{aligned} \tag{5.32}$$

Since the channel is estimated in the first iteration, detection performance in each iteration would be enhanced than the case without knowing the channel coefficient.

5.5 Simulations

5.5.1 Assumptions

In this section, I examine the performance of the proposed JSVC technique. Our simulation setup is based on the downlink OFDM system in the 3GPP LTE-Advanced Rel.13 [63] under AWGN channel. For comparison, we also investigate the performance of the repetition code. I test the transmission of b bit information which consists of quantizing the original information using b bit level. In the conventional repetition coding, the information block is repeated before resource mapping. Since the block size of the repetition code is not flexible, we set the rate $\frac{1}{2}$ and $\frac{1}{3}$ to test similar conditions. In the proposed JSVC algorithm, I set the ZC spreading codebook with $N = 12$, $m = 12$ and $K = 2$. To ensure the fair comparison, I use the same number of resources ($m = 12$) in the ultra short packet transmission. As a performance measure, I

use PER of the code blocks and mean square error of the original information. Also, the effective throughput is compared in the function of SNR.

5.5.2 Results and Discussions

In order to observe the performance of JSVC, we test PER performance as depicted in Fig. 5.3. As the conventional short packet transmission, $\frac{1}{2}$ and $\frac{1}{3}$ rate of repetition code are used. That is, 12 and 8 bits of information is transmitted over 12 resources, respectively. In case of JSVC, 9 bit of information is encoded into $K = 2$ sparse vector and transmitted over 12 resources and side information is mapped to the magnitude of sparse vector. Interestingly, we observed that PER performance is comparable to the conventional short packet transmission.

To confirm the effectiveness of JSVC, we test MSE performance of original information in Fig. 5.4. As shown in the results, the MSE performance of JSVC is improved as SNR while the conventional approaches are same. The benefit of JSVC can be better understood by observing the effective throughput. In Fig. 5.5. we plot the effective throughput as a function of SNR. In this case, we plot throughput as function MSE that can be achieved by each transmission scheme. When SNR is higher than 3 dB, the effective throughput of JSVC is higher than the conventional approaches for transmitting short packet. This is because the MSE of the original information is improved so that dense quantization level can be used for the proposed scheme.

5.6 Summary

In this chapter, we have proposed a new data transmission strategy for the ultra-shot packet by transmitting side (analog) information. The key idea behind the proposed JSVC transmission scheme is to transform the small information into a sparse vector and map the side-information into a magnitude of the sparse vector. Metaphorically, JSVC can be thought as a standing a few poles to the empty table and measure the

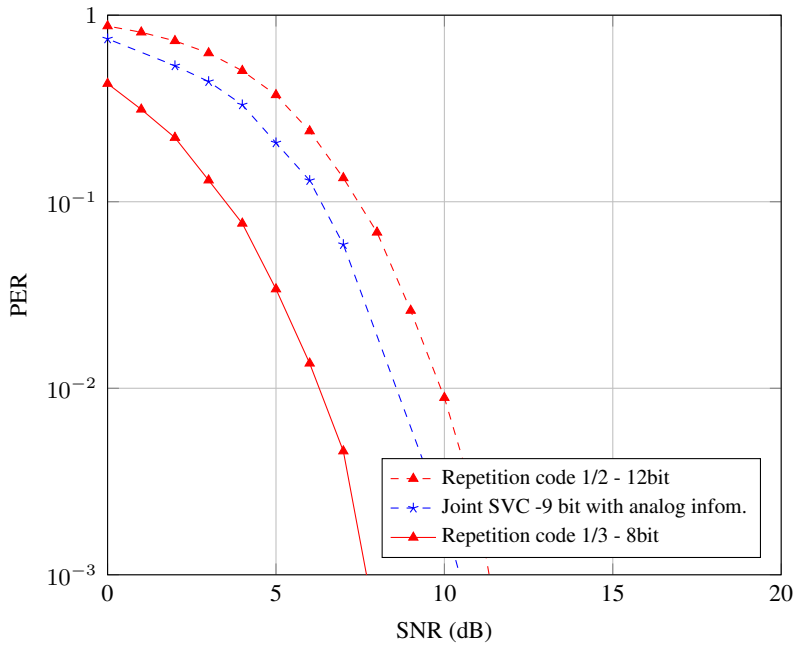


Figure 5.3: PER performance of joint SVC ($K = 2, N = 24, m = 12, b_{SVC} = 8$).

height of pole at the receiver. As long as the number of poles is small enough and the measurements contains enough information to find out the marked cell positions, accurate recovery of JSVC packet can be guaranteed. I showed from the numerical evaluations and analysis of decoding performance that the proposed JSVC scheme is very effective in ultra-short packet transmission such as analog feedback, MIMO feedback, and control-type channel.

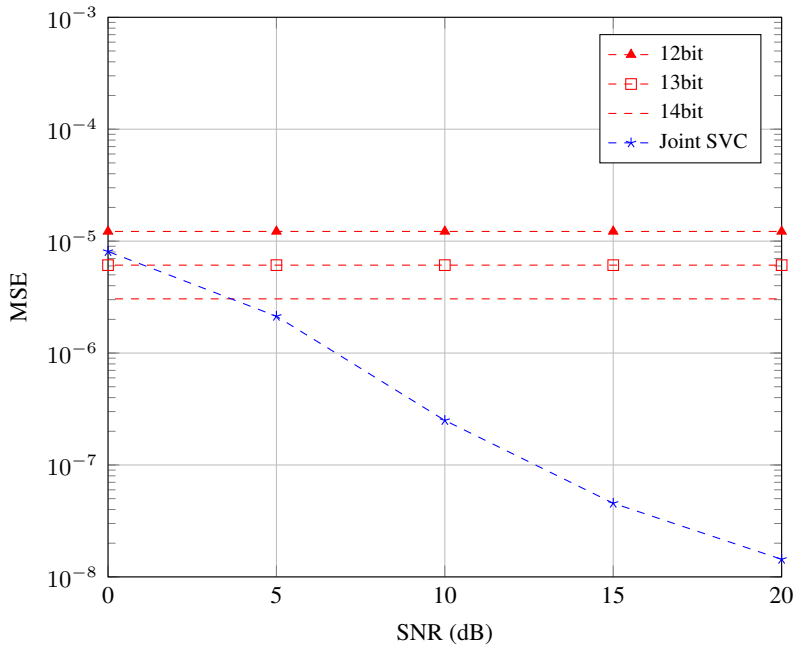


Figure 5.4: MSE performance of joint SVC ($K = 2, N = 24, m = 12, b_{SVC} = 8$).

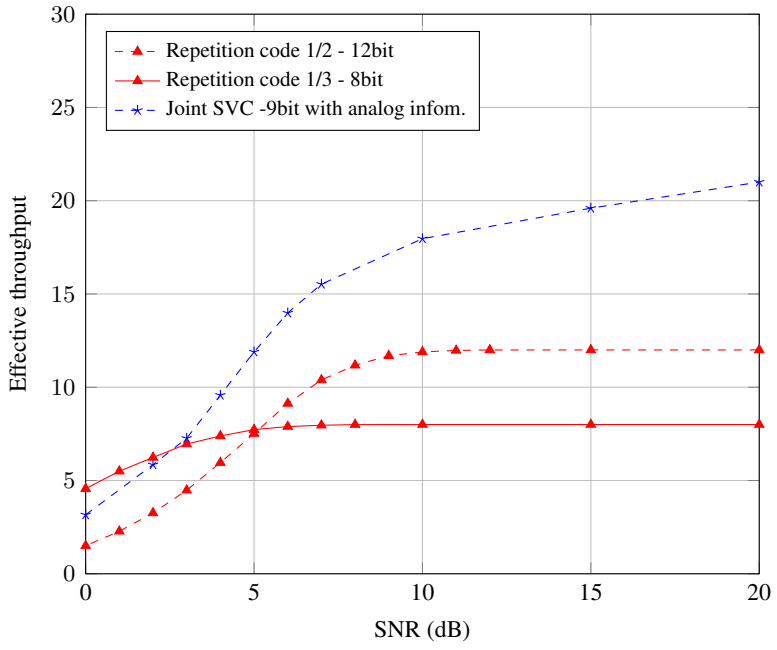


Figure 5.5: Effective throughput performance of joint SVC ($K = 2, N = 24, m = 12, b_{SVC} = 8$).

Chapter 6

Sparse Beamforming for Enhanced Mobile Broadband Communications

In this Chapter, I propose the novel beamforming technique named time-domain sparse beamforming to enhance throughput by minimizing overhead in pilot transmission. By designing beamforming weight that minimize non-zero element in time-domain channel impulse response (CIR) in time-domain of pilot signal, the required pilot overhead can be minimized but user full diversity gain in data transmission. Pilot beamforming and CSI acquisition strategy for IoT systems to achieve reduction in the pilot overhead and enhancement in the channel estimation quality are suggested. Key idea of the proposed method, referred to as *time-domain sparse beamforming* (TDSB), is to sparsify the time-domain channel vector using the preprocessing (beamforming) at the basestation. To be specific, using the deliberately designed antenna-domain beamforming, we sparsify the beamformed time-domain channel vector. As a result, only a few samples are required to perform the channel estimation and whole CSI can be acquired with partial pilot symbols in time and frequency. From the numerical evaluations, we show that the proposed scheme outperforms conventional channel estimation schemes [70-73] and achieves N_T-fold reduction in the pilot overhead. .

The work of Chapter 6 has been published in part in [81]

6.1 Introduction

In this section, we discuss key features of FD-MIMO systems. These include a large number of basestation antennas, 2D active antenna array, 3D channel propagation, and new pilot transmission with CSI feedback. In what follows, we will use LTE terminology exclusively: enhanced node-B (eNB) for basestation, user equipment (UE) for the mobile terminal, and reference signal (RS) for pilot signal.

6.1.1 Increase the number of transmit antennas

One of the main features of FD-MIMO systems distinct from the MIMO systems of the current LTE and LTE-Advanced standards is to use a large number of antennas at eNB. In theory, as the number of eNB antennas N_T increases, cross-correlation of two random channel realizations goes to zero [82] so that the inter-user interference in the downlink can be controlled via a simple linear precoder. Such benefit, however, can be realized only when the perfect CSI is available at the eNB. While the CSI acquisition in time division duplex (TDD) systems is relatively simple due to the channel reciprocity, such is not the case for frequency division duplex (FDD) systems. Note that in the FDD systems, time variation and frequency response of the channel are measured via the downlink RSs and then sent back to the eNB after the quantization. Even in TDD mode, one cannot solely rely on the channel reciprocity because the measurement at the transmitter does not capture the downlink interference from neighboring cells or co-scheduled UEs. As such, downlink RSs are still required to capture the channel quality indicator (CQI) for the TDD mode, and thus the downlink RS and the uplink CSI feedback are essential for both duplex modes. Identifying the potential issues of CSI acquisition and developing the proper solutions are, therefore, of great importance for the successful commercialization of FD-MIMO systems. Before we go into detail, we briefly summarize two major problems related to the CSI acquisition.

- **Degradation of CSI accuracy**: One well-known problem for the MIMO sys-

tems, in particular for FDD-based systems, is that the quality of CSI is affected by the limitation of feedback resources. As the CSI distortion increases, quality of the multiuser MIMO (MU-MIMO) precoder to control the inter-user interference is degraded and so will be the performance of the FD-MIMO systems. In general, the amount of CSI feedback, determining the quality of CSI, needs to be scaled with N_T to control the quantization error so that the overhead of CSI feedback increases in FD-MIMO systems.

- **Increase of pilot overhead**: An important problem related to the CSI acquisition at eNB, yet to be discussed separately, is the pilot overhead problem. UE performs the channel estimation using the RS transmitted from the eNB. Since RSs need to be assigned in an orthogonal fashion, RS overhead typically grows linearly with N_T. For example, if $N_T = 64$, RS will occupy approximately 48% of resources, eating out substantial amount of downlink resources for the data transmission.

6.1.2 2D active antenna system (AAS)

Another interesting feature of the FD-MIMO system is an introduction of the active antenna with 2D planar array. In the active antenna-based systems, gain and phase are controlled by the active components, such as power amplifier (PA) and low noise amplifier (LNA), attached to each antenna element. In the 2D structured antenna array, one can control the radio wave on both vertical (elevation) and horizontal (azimuth) direction so that the control of the transmit beam in 3D space is possible. This type of wave control mechanism is also referred to as the 3D beamforming. Another important benefit of 2D AAS is that it can accommodate a large number of antennas without increasing the deployment space. For example, when 64 linear antenna arrays are deployed in a horizontal direction, under the common assumption that the antenna spacing is half wavelength ($\frac{\lambda}{2}$) and the system is using LTE carrier frequency (2 GHz), it requires a horizontal room of 3m. Due to the limited space on a rooftop or mast, this

space would be burdensome for most of the cell sites. In contrast, when antennas are arranged in a square array, relatively small space is required for 2D antenna array (e.g., 1.0×0.5m with dual-polarized 8×8 antenna array).

6.1.3 3D channel environment

When basic features of the FD-MIMO systems are determined, the next step is to design a system maximizing performance in terms of throughput, spectral efficiency, and peak data rate in the realistic channel environment. There are various issues to consider in the design of practical systems, such as investigation and characterization of the realistic channel model for the performance evaluation. While the conventional MIMO systems consider the propagation in the horizontal direction only, FD-MIMO systems employing 2D planar array should consider the propagation in both vertical and horizontal direction. To do so, geometric structure of the transmitter antenna array and propagation effect of the 3D positions between the eNB and UE should be reflected in the channel model. Main features of 3D channel propagation obtained from real measurement are as follows [86]:

- Height and distance-dependent line-of-sight (LOS) channel condition: LOS probability between eNB and UE increases with the UE's height and also increases when the distance between eNB and UE decreases.
- Height-dependent pathloss: UE experiences less pathloss on a higher floor (e.g., 0.6dB/m gain for macro cell and 0.3dB/m gain for micro cell).
- Height and distance-dependent elevation spread of departure angles (ESD): When the location of eNB is higher than the UE, ESD decreases with the height of the UE. It is also observed that the ESD decreases sharply as the UE moves away from the eNB.

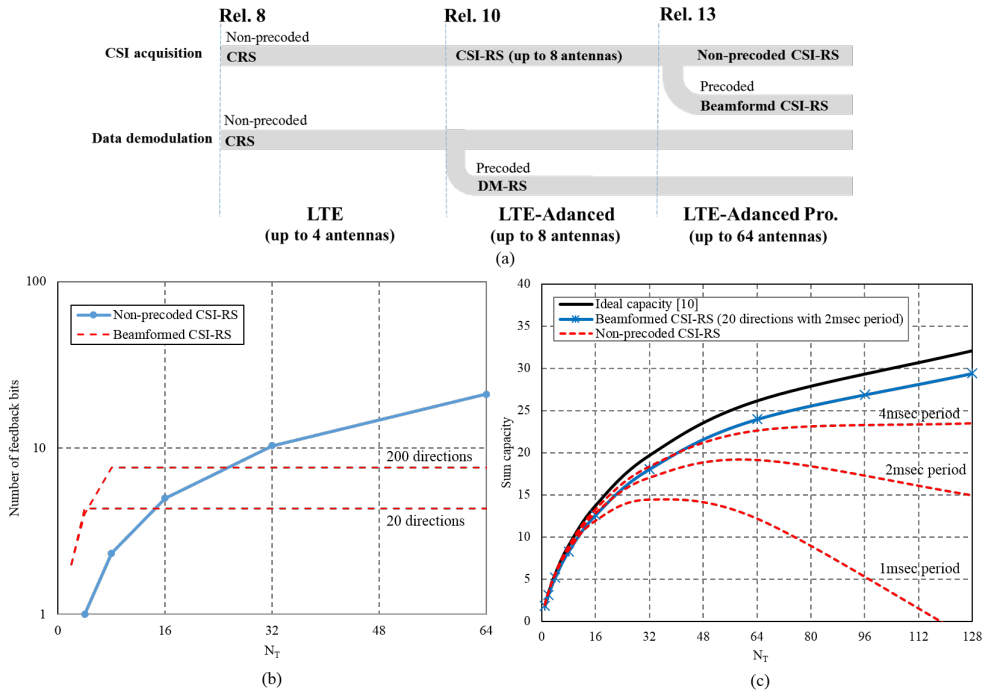


Figure 6.1: MIMO evaluation: (a) RS evolution in LTE systems, (b) uplink feedback overhead (SNR=10dB [88]), (c) MU-MIMO capacity with considering CSI-RS overhead (ideal CSI and ZFBF MU-precoding with 10 UEs and SNR=10dB).

6.1.4 RS transmission for CSI acquisition

From the LTE to LTE-Advanced, there has been substantial improvement in the RS scheme for MIMO systems (see Fig. 6.1(a)). From the common RS (CRS) to the channel state information RS (CSI-RS), various RSs to perform the CSI acquisition have been introduced. While these are common to all users in a cell and thus un-precoded, the demodulation RS (DM-RS) is UE-specific (i.e., dedicated to each UE) so that it is precoded by the same weight applied for the data transmission. Since the DM-RS is present only on time/frequency resources where the UE is scheduled, this cannot be used for CSI measurements [87].

One of the new features of the FD-MIMO systems is to use a beamformed RS, called beamformed CSI-RS, for the CSI acquisition. Beamformed RS transmission is a channel training technique that uses multiple precoding weights in spatial domain. In this scheme, UE picks the best weight among transmitted and then feeds back its index. This scheme provides many benefits over non-precoded CSI-RS, in particular when N_T is large. Some of the benefits are summarized as follows:

- **Less uplink feedback overhead:** In order to maintain a rate comparable to the case with perfect CSI, feedback bits used for the channel vector quantization should be proportional to N_T [88]. Whereas, the amount of feedback for the beamformed CSI-RS scales logarithmic with the number of RSs N_B since this scheme only feeds back an index of the best beamformed CSI-RS. Thus, as depicted in Fig. 6.1(b), the benefit of beamformed CSI-RS is pronounced when N_T is large.

- **Less downlink pilot overhead:** When the non-precoded CSI-RS is used, pilot overhead increases with N_T, resulting in a substantial loss of the sum capacity in the FD-MIMO regime (see Fig. 6.1(c)). Whereas, pilot overhead of the beamformed CSI-RS is proportional to N_B and independent of N_T so that the rate loss of the beamformed CSI-RS is marginal even when N_T increases.

- **Higher quality in RS:** If the transmit power is P watt, P/N_T watt is needed for each non-precoded CSI-RS transmission, while P/N_B watt is used for the beamformed CSI-RS. For example, when $N_T = 32$ and $N_B = 12$, beamformed CSI-RS provides 4.3dB gain in signal power over the non-precoded CSI-RS.[1]

In order to support the beamformed CSI-RS scheme, new transmitter architecture called transceiver unit (TXRU) architecture has been introduced. By TXRU architecture, we mean a hardware connection between the baseband signal path and antenna array elements. Since this architecture facilitates the control of phase and gain in both digital and analog domain, more accurate control of the beamforming direction is possible. One thing to note is that the conventional codebook cannot measure the CSI of the beamformed transmission so that a new channel feedback mechanism supporting the beamformed transmission is required.

6.2 System Design and Standardization of FD-MIMO Systems

The main purpose of the Rel. 13 study item is to identify key issues to support up to 64 transmit antennas placed in the form of a 2D antenna array. Standardization of the systems supporting up to 16 antennas is an initial target of Rel. 13 and issues to support more than 16 antennas will be discussed in subsequent releases. In the study item phase, there has been extensive discussion to support 2D array antennas, elaborated TXRUs, enhanced channel measurement and feedback schemes, and also an increased number of co-scheduled users (up to eight users). Among these, an item tightly coupled to the standardization is the CSI measurement and feedback mechanism. In this subsection, we discuss the deployment scenarios, antenna configurations, TXRU structure, new RS strategy, and feedback mechanisms.

[1] In 3D channel model, the typical number of multi-paths (clusters) is 12 [86].

6.2.1 Deployment scenarios

For the design and evaluation of FD-MIMO systems, a realistic scenario in which antenna array and UEs are located in different height is considered. To this end, two typical deployment scenarios, viz., 3D urban macro scenario (3D-UMa) and 3D urban micro (3D-UMi), are introduced (see Fig. 6.2). In the former case, transmit antennas are placed over the rooftop, and in the latter case, they are located below the rooftop. In case of 3D-UMa, diffraction over the rooftop is a dominant factor for the propagation so that down-tilted transmission in the vertical direction is desirable (see Fig. 6.2(b)). In fact, by transmitting beams with different steering angles, eNB can separate channels corresponding to multiple UEs. In the 3D-UMi scenario, on the other hand, the location of users is higher than the height of the antenna so that direct signal path is dominant (see Fig. 6.2(c)). In this scenario, both up and down-tilting can be used to schedule UEs in different floors. Since the cell radius of the 3D-UMi scenario is typically smaller than that of 3D-UMa, LOS channel condition is predominant, and thus more UEs can be co-scheduled without increasing the inter-user interference [86]. Although not as strong as the 3D-UMi scenario, LOS probability in the 3D-UMa scenario also increases when the distance between eNB and UE decreases.

6.2.2 Antenna configurations

Unlike the conventional MIMO systems relying on the passive antenna, systems based on the active antenna can dynamically control the gain of an antenna element by applying the weight of low-power amplifiers attached to each antenna element. Since the radiation pattern depends on the antenna arrangement, such as the number of the antenna elements and antenna spacing, the antenna system should be modeled in an element-level. As shown in Fig. 6.3(a), there are three key parameters characterizing the antenna array structure (M, N, P): the number of elements M in vertical direction, the number of elements N in horizontal direction, and the polarization degree P ($P = 1$ is for co-polarization and $P = 2$ is for dual-polarization). As a benchmark set-

ting, 2D planar array using dual polarized antenna ($P = 2$) configuration with $M = 8$ (0.8λ spacing in vertical direction) and $N = 4$ (0.5λ spacing in horizontal direction) is suggested.[2] In this setting, null direction, an angle to make the magnitude of beam pattern to zero, for the elevation beam pattern is 11° and that for the horizontal beam pattern is 30° (see Fig. 6.3(c)). Since the null direction in the vertical domain is much smaller than that of the horizontal domain, scheduling UEs in the vertical domain is more effective in controlling the inter-user interference. Also, a tall or fat array structure ($M \gg N$ or $M \ll N$) is favorable since it will generate a sharp beam but it might be less flexible in the situation where the surrounding environment is changed. Further, large antenna spacing is not always a desirable option since it can increase the inter-cell interference due to the narrow beamforming for cell edge UEs (this phenomenon is called flash-light effect). For this reason, in a real deployment scenario, the design parameters should be carefully chosen by considering various factors, such as user location, cell radius, building height, and antenna height.

6.2.3 TXRU architectures

As mentioned, one interesting feature of the active antenna systems is that each TXRU contains PA and LNA so that eNB can control the gain and phase of an individual antenna element. In order to support this, a power feeding network between TXRUs and antenna elements called TXRU architecture is introduced [90]. TXRU architecture consists of three components: TXRU array, antenna array, and radio distribution networks (RDN). A role of the RDN is to deliver the transmit signal from PA to antenna array elements and the received signal from antenna array to LNA. Depending on the CSI-RS transmission and feedback strategy, two representative options, array partitioning and array connected architecture, are suggested. The former is for the con-

[2]Note that the total number of antenna elements in this setup is the same as that of 8Tx antennas in conventional systems and thus FD-MIMO eNB can provide backward compatibility [89]. The vertical configuration is to ensure the same cell coverage and the horizontal configuration is for the conventional MIMO operation for LTE.

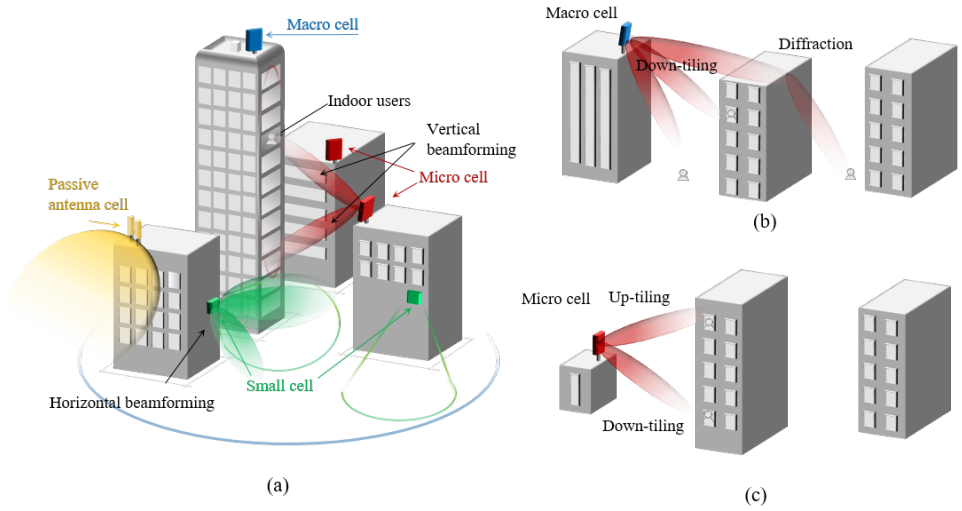


Figure 6.2: FD-MIMO deployment scenarios: (a) 3D macro cell site (placed over the rooftop) and 3D micro cell site (placed below the rooftop) with small cell, (b) beamforming for 3D macro cell, and (c) beamforming in 3D micro cell.

ventional codebook scheme and the latter is for the beamforming scheme.

In the array partitioning architecture, antenna elements are divided into multiple groups and each TXRU is connected to one of them (see Fig. 6.3(d)). Whereas, in the array connected structure, RDN is designed such that RF signals of multiple TXRUs are delivered to the single antenna element. To mix RF signals from multiple TXRUs, additional RF combining circuitry is needed as shown in Fig. 6.3(e). The difference between the two can be better understood when we discuss the transmission of the CSI-RS. In the array partitioning architecture, N_T antenna elements are partitioned into L groups of TXRU and orthogonal CSI-RS is assigned for each group. Each TXRU transmits its own CSI-RS so that the UE measures the channel h from the CSI-RS observation $y = hx + n$. In the array connected architecture, each antenna element is connected to $L^{'}$ (out of L) TXRUs and orthogonal CSI-RS is assigned for each TXRU. Denoting $\mathbf{h} \in \mathbb{C}^{1\times N_c}$ as the channel vector and $\mathbf{v} \in \mathbb{C}^{N_c\times 1}$ as the precoding weight ($N_T\frac{L^{'}}{L} = N_c$) for each beamformed CSI-RS, the beamformed CSI-RS observation is $y = \mathbf{hv}x + n$ and the UE measures the precoded channel $\mathbf{hv}$ from this. Due to

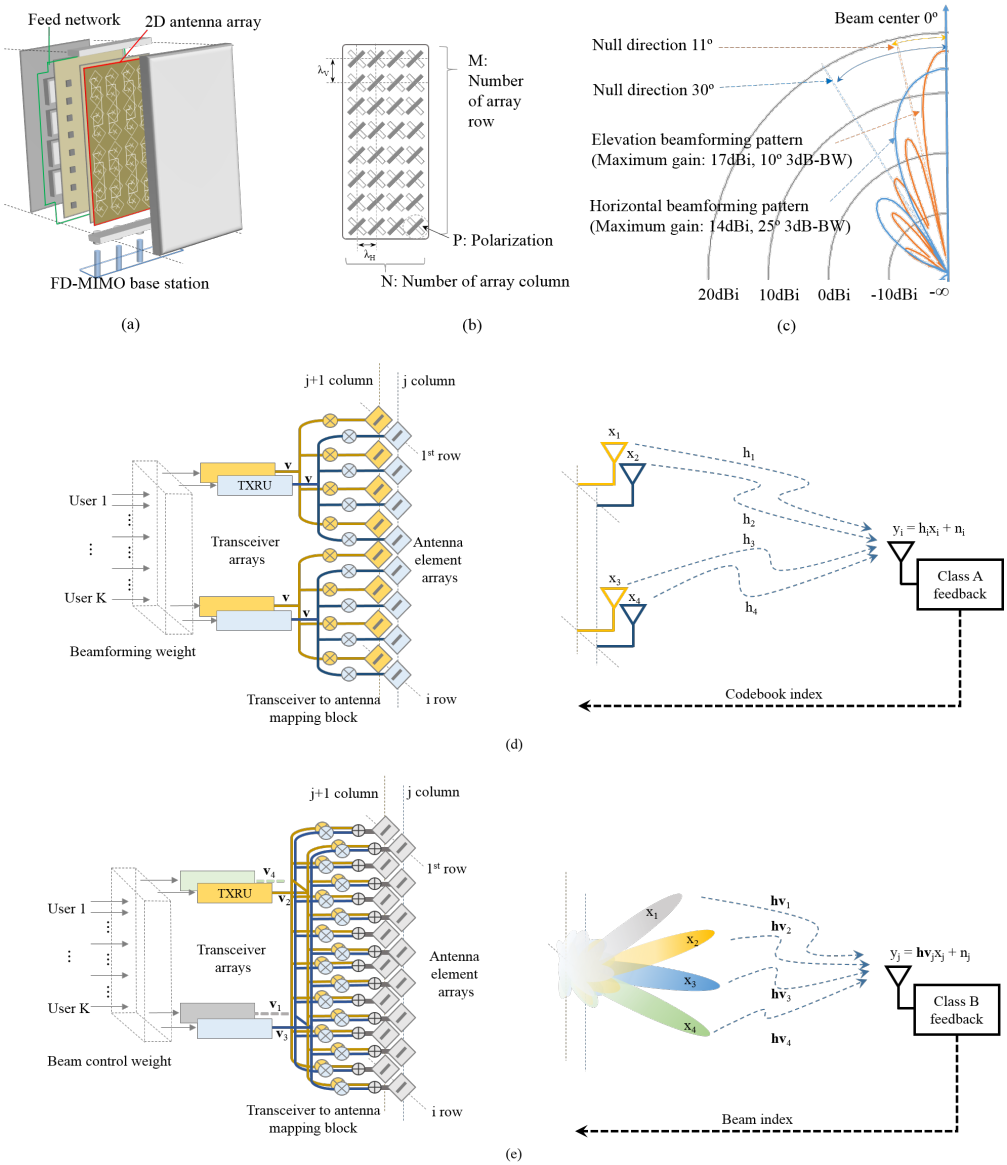


Figure 6.3: FD-MIMO systems: (a) concept of FD-MIMO systems, (b) 2D array antenna configuration, (c) vertical and horizontal beamforming patterns, (d) array partitioning architecture with the conventional CSI-RS transmission, and (e) array connected architecture with beamformed CSI-RS transmission.

the narrow and directional CSI-RS beam transmission with a linear array, SNR of the precoded channel is maximized at the target direction.[3]

6.2.4 New CSI-RS transmission strategy

In the standardization process, two CSI-RS transmission strategies, i.e., extension of the conventional non-precoded CSI-RS and the beamformed CSI-RS, are suggested. In the first strategy, UE observes the non-precoded CSI-RS transmitted from each of partitioned antenna arrays (see Fig.6.3(d)). By sending the precoder maximizing the properly designed performance criterion to the eNB, UE can adapt to the channel variation. In the second strategy, eNB transmits multiple beamformed CSI-RS (we call it beam for simplicity) using connected arrays architecture. Among these, UE selects the preferred beam and then feeds back its index. When the eNB receives the beam index, the weight corresponding to the selected beam is used for the data transmission.

Overall downlink precoder for data transmission $\mathbf{W}_{\text{data}}$ and CSI-RS transmission $\mathbf{W}_{\text{rs}}$ can be expressed as

$$\mathbf{W}_{\text{data}} = \mathbf{W}_T \mathbf{W}_{\text{rs}} \text{ and } \mathbf{W}_{\text{rs}} = \mathbf{W}_P \mathbf{W}_U, \tag{6.1}$$

where $\mathbf{W}_T \in \mathbb{C}^{N_T \times L}$ is the precoder between TXRU and the antenna element, $\mathbf{W}_P \in \mathbb{C}^{L \times N_P}$ is the precoder between the CSI-RS port and the TXRU (N_P is the number of antenna ports), and $\mathbf{W}_U \in \mathbb{C}^{N_P \times r}$ is the precoder between data channel to CSI-RS port.

In the following, we summarize details of two strategies.

- Conventional CSI-RS transmission: One option to maximize the capacity is to do one-to-one mapping of the TXRU and the CSI-RS resource (i.e., $\mathbf{W}_P = \mathbf{I}_{N_{TXRU}}$). To achieve the same coverage for each CSI-RS resource, an identical

[3] SNR $= \frac{|\mathbf{hv}(\phi)|^2}{\sigma^2}$, where ϕ is the beam direction and σ^2 is the noise power.

weight $\mathbf{v}$ is applied to L groups.[4] Each UE measures the CSI-RS resources and then chooses the preferred codebook index i^* maximizing the channel gain for each subband:

$$i^* = \arg\max_i \|\bar{\mathbf{h}}^H \mathbf{W}_U^i\|_2^2, \tag{6.2}$$

where $\|\mathbf{a}\|_2 = \sqrt{\sum_i |a_i|^2}$ and $\bar{\mathbf{h}} = \mathbf{h}/\|\mathbf{h}\|_2$ is the estimated channel direction vector, and $\mathbf{W}_U^i$ is the ith precoder between the data channel and CSI-RS ports. This scheme is called class-A CSI feedback.

- Beamformed CSI-RS transmission: In order to acquire the spatial angle between the eNB and UE, eNB transmits multiple beamformed CSI-RSs. Let N_B be the number of CSI-RSs, then we have $\mathbf{W}_T = [\mathbf{v}_1 \mathbf{v}_2 \dots \mathbf{v}_{N_B}]$ where $\mathbf{v}_i \in \mathbb{C}^{N_T \times 1}$ is the 3D beamforming weight for the ith beam. For example, when the rank-1 beamforming is applied, we have $\mathbf{W}_P = \mathbf{1}_{N_B}$ and $\mathbf{W}_U = 1$. Among all possible beams $\mathbf{v}_1, ..., \mathbf{v}_{N_B}$, UE selects and feeds back the best beam index j^* maximizing the received power:

$$j^* = \arg\max_j |\bar{\mathbf{h}}^H \mathbf{v}_j|^2. \tag{6.3}$$

This scheme is called class-B CSI feedback. Under the rich scattering environment, dominant paths between eNB and UE depend on the direction and width of the transmit signal. In the multiple-input single-output (MISO) channel, for example, the channel vector in an angular domain is expressed as $\mathbf{h} = \sum_i e_r \mathbf{e}_t(\phi_i)^*$, where $e_r = 1$ and $\mathbf{e}_t(\phi_i) = [1 \;\; e^{-j2\pi\gamma\phi_i} \;\; ... \;\; e^{-j2\pi(N_T-1)\gamma\phi_i}]^T$ is the spatial signature of the transmitter (ϕ_i is direction of ith path and γ is normalized antenna spacing) [91]. When the RS is transmitted in a direction ϕ_j, the beamforming weight would be $\mathbf{v} = \mathbf{e}_t(\phi_j)$ so that the resulting beam-

[4] In this section, we assume that discrete Fourier transform (DFT) weights are used as $\mathbf{W}_T$ for mapping between TXRU and antenna elements for simplicity. For example, $\mathbf{W}_T$ can be expressed as $\mathbf{W}_T = [\mathbf{v}\ \mathbf{v}; \mathbf{v}\ \mathbf{v}]$ in Fig. 6.3(d).

Table 6.1: Comparison between CSI-RS transmission and CSI feedback classes

Category	Class-A CSI feedback (Conventional CSI-RS)	Class-B CSI feedback (Beamformed CSI-RS)
Feedback design	Need to design codebook for 2D antenna layout and feedback mechanism for adapting channel variation	Need to devise a method to feed back beam index for adapting both weight changes and channel variation
UL Feedback overhead	Depend on resolution of codebook and the number of antennas	Depend on the number of operating beam N_B
CSI-RS overhead	Require N_T CSI-RS resources	Scale linearly with the number of beam N_B
Backward compatibility	Supportable with virtualization between TXRUs and antenna ports	Supportable with vertical 1D beamforming weight
Forward compatibility	Scalable to larger TXRU system if CSI-RS resources are allowed	Scalable to larger TXRU system if long-term channel statistics are acquired

formed channel is readily expressed as one or at most a few dominant taps ($\mathbf{e}_t(\phi_i)^T\mathbf{e}_t(\phi_i) \approx 0$ when $i \neq j$). In fact, by controlling the weight applied to CSI-RS, the effective dimension of the channel vector can be reduced so that the feedback overhead can be reduced substantially.

In Table 6.1, we summarize two CSI-RS transmission schemes discussed in the FD-MIMO.

6.2.5 CSI feedback mechanisms for FD-MIMO systems

In the study item phase, various RS transmission and feedback schemes have been proposed. As shown in Fig. 6.1, capacity and overhead of class-A and class-B feedback schemes are more or less similar in the initial target range ($N_t = 16$) so that Rel. 13 has decided to support both classes. In this subsection, we briefly describe the CSI feedback schemes associated with TXRU architectures. Among various schemes, composite codebook and beam index feedback have received much attention as main ingredients for class-A and class-B CSI feedback. The rest will be considered in a future release.

Composite codebook: In this scheme, overall codebook is divided into two (vertical and horizontal codebooks) and thus the channel information is separately delivered to the eNB. By combining two codebooks (e.g., Kronecker product of two codebooks $\mathbf{W}_U = \mathbf{W}_{U,V} \otimes \mathbf{W}_{U,H}$), eNB reconstructs whole channel information. Considering that the angular spread of the vertical direction is smaller than that of the horizontal direction, one can reduce the feedback overhead by setting a relatively long reporting period to the vertical codebook.

Beam index feedback: To obtain the UE's channel direction information (CDI) from beamformed CSI-RSs, eNB needs to transmit multiple beamformed CSI-RSs. When the channel rank is one, feedback of a beam index and corresponding CQI is enough. Whereas, when the channel rank is two with dual-polarized antennas, co-phase information is additionally required for adapting channel orthogonalization between layers. For example, once eNB obtains the CDI, this can be used for the beamforming vector of two-port CSI-RS and each CSI-RS port is mapped to the different polarized antennas. UE then estimates and feeds back short-term co-phase information between two ports.

Other CSI feedback schemes: In the partial CSI-RS transmission, CSI-RS overhead can be reduced by partitioning the 2D antenna array into horizontal and vertical ports, say N_H ports in the row and N_V ports in the column. In doing so, the total number of CSI-RS can be reduced from $N_H \times N_V$ to $N_H + N_V$. Overall channel information can be reconstructed by exploiting spatial and temporal correlation among antenna elements [92]. In the adaptive CSI feedback scheme, benefits of the beamformed and non-precoded CSI-RS transmission can be combined. First, in order to acquire long-term channel information, eNB transmits N_T non-precoded CSI-RSs. After receiving sufficient long-term channel statistics from UE, eNB determines spatial direction roughly and then transmits the beamformed CSI-RSs used for short-term and subband feedbacks. The flexible codebook scheme can support various 2D antenna layouts without increasing the number of codebooks. In this approach, one master

codebook is designed for a large number of TXRUs, say 16 TXRUs, and the specific codebook (e.g., (2×8), (4×4), or (1×16)) is derived based on this. To support this, the eNB needs to send the layout information via separate signaling.

6.3 System Model

6.3.1 Basic System Model

We consider the downlink IoT systems with N_T antennas at the FD-MIMO basestation and a single antenna at the IoT device. Typically, small portion of the system bandwidth, say B_u [Hz] out of B_s [Hz] ($B_u \ll B_s$), is assigned for each IoT device. For examples, 1.4 MHz out of 20 MHz bandwidth is assigned to each device in eMTC standard and 180 kHz bandwidth in NB-IoT standard [74]. In the OFDM-based system, pilot signals are inserted in the time-frequency resource. Pilot signals are used for the channel estimation, data demodulation, and channel state information (CSI) feedback.

Let $\mathbf{y}_n^i \in \mathbb{C}^{N_P \times 1}$ be the received pilot vector in the frequency domain extracted from the nth time-symbol and ith antenna (see Fig. 1(b)). Then $\mathbf{y}_n^i$ is expressed as

$$\mathbf{y}_n^i = \text{diag}\left(\mathbf{p}_n^i\right) \mathbf{\Phi}_n^i \mathbf{g}_n^i + \mathbf{z}_n^i, \tag{6.4}$$

where $\mathbf{p}_n^i \in \mathbb{C}^{N_P \times 1}$ is the vector constructed from the pilot symbols, $\mathbf{\Phi}_n^i \in \mathbb{R}^{N_P \times N_F}$ is the selection matrix containing only one element being one in each row and rest being zero[5], $\mathbf{g}_n^i \in \mathbb{C}^{N_F \times 1}$ is the frequency-domain channel vector, and $\mathbf{z}_n^i \in \mathbb{C}^{N_P \times 1}$ is the additive white Gaussian noise ($\mathbf{z}_n^i \sim \mathcal{CN}(0, \sigma_w^2 \mathbf{I}_{N_P})$). The relationship between the frequency-domain channel vector $\mathbf{g}_n^i$ and the time-domain channel vector $\mathbf{h}_n^i = [h_{n,1}^i, ..., h_{n,N_F}^i] \in \mathbb{C}^{N_F \times 1}$ is

$$\mathbf{g}_n^i = \mathbf{F}\mathbf{h}_n^i, \tag{6.5}$$

[5]For example, if the 1st and 3rd subcarriers are used for pilot, then $\mathbf{\Phi}_n^i = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \end{bmatrix}$ and N_F is the number of subcarriers.

where $\mathbf{F} \in \mathbb{C}^{N_F \times N_F}$ is the DFT matrix. From (6.4) and (6.5), we have

$$\mathbf{y}_n^i = \text{diag}\left(\mathbf{p}_n^i\right) \mathbf{\Phi}_n^i \mathbf{F} \mathbf{h}_n^i + \mathbf{z}_n^i. \tag{6.6}$$

In characterizing the channel vector $\mathbf{h}$, we assume that there are K clusters and N_{sp} sub-paths for each cluster. When the basestation antennas have a linear array structure (e.g., 1D linear array or 2D uniform planner array) and the receiver has a single antenna, the channel for each antenna can be expressed in terms of the angle of departure (AoD). A time-varying channel tap h_m^i of the ith antenna and delay bin m is expressed as

$$h_m^i = \frac{1}{N_T} \sum_{l=1}^{N_{sp}} \sqrt{\frac{h_m^1}{N_{sp}}} e^{j\phi_m} e^{j\kappa(i-1)\sin\theta_{m,l}}, \tag{6.7}$$

where $\kappa = \frac{2\pi}{\lambda}$ is the wavenumber, ϕ_m is the random phase of mth element, and $\theta_{m,l} = \theta_m + \Delta\theta_l$ is AoD of sub-paths in mth cluster where $\Delta\theta_l \sim \mathcal{N}(0, \sigma_m^2)$, and σ_m is standard deviation of angular spread for $m = 1, ..., K$ [75-76].

In this work, we assume that the support of the channel vector $\mathbf{h}^i$ is common for all antennas. In fact, $\text{supp}(\mathbf{h}^i) = \text{supp}(\mathbf{h}^j)$ for i and $j = 1, ..., N_T$. Note that the scale of the antennas at the basestation is much smaller than the signal transmission distance in typical multi-antenna geometry so that channels associated with different transmit-receive antenna pairs usually share the common support [77].

6.3.2 Beamformed Pilot Transmission

When the number of transmit antennas is large, orthogonal pilot transmission cannot be a desirable option due to the pilot overhead and computation complexity of signal processing operations. One viable approach to reduce the pilot overhead is to transmit multiple beamformed pilots (we henceforth call it *beam* [78]). Main idea of this approach is to transmit the pilot signals after applying the predefined beam patterns. In doing so, multiple beams having different beam directions can be transmitted simultaneously (e.g., AoD of line-of-sight path in [79] and angular separation in [80]). Since

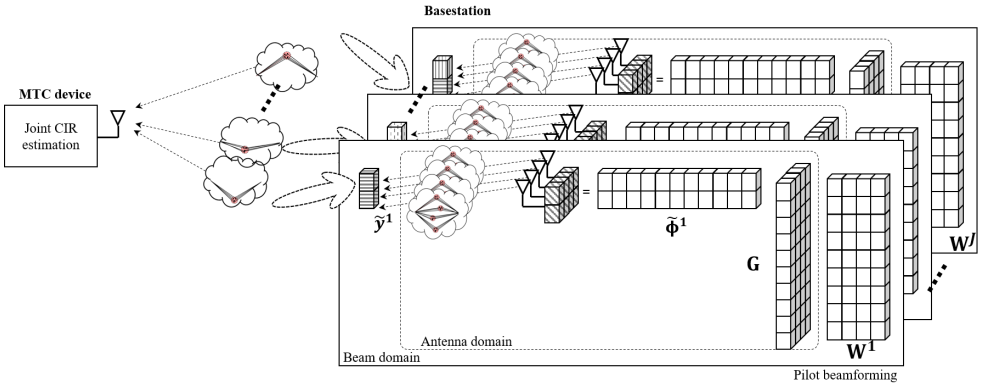


Figure 6.4: Beamforming of the proposed TDSB scheme.

the effective dimension of the channel vector (cardinality of $\mathbf{h}$) can be reduced, pilot and feedback overhead are also reduced substantially.

When the pilot signal is transmitted without any modification, the channel vector in an angular domain is expressed as $[h^1 \ ... \ h^{N_T}]^T = \sum_i w_r \mathbf{h}_t(\phi_i)^*$, where $w_r = 1$ and $\mathbf{h}_t(\phi_i) = [1 \ \ e^{-j2\pi\kappa\phi_i} \ ... \ \ e^{-j2\pi(N_T-1)\kappa\phi_i}]^T$ is the spatial signature of the transmitter (ϕ_i is direction of ith path). When the pilot signal is transmitted in a direction ϕ_j, we set the beamforming weight to $\mathbf{w} = \mathbf{h}_t(\phi_j)$ so that the resulting beamformed channel can be readily expressed as one or at most a few dominant taps ($\mathbf{w}_t(\phi_j)^T \mathbf{h}_t(\phi_i) \approx 0$ for $i \neq j$). In fact, key idea of the proposed scheme is to design the beamforming weight to sparsify the beamformed channel vector $\tilde{\mathbf{h}}$.

6.4 Sparsification of Pilot Beamforming

Unlike the conventional scheme where the pilot signal is transmitted without modification (e.g., common and channel state information reference signal in LTE-Advanced), the pilot signals are sent after the beamforming in the proposed TDSB technique (see Fig. 6.4). Primary goal of the proposed TDSB scheme is to minimize the nonzero entries of a time-domain channel vector by the help of multiple antennas at the basestation. In our work, we assume that the nonzero tap information is available at the

basestation via the channel feedback or channel reciprocity.

6.4.1 Time-domain System Model without Pilot Beamforming

Let $\mathbf{h}_n^i = [h_{n,1}^i, ..., h_{n,N_{cir}}^i]^T$ be the time-domain CIR vector associated with ith antenna. The pilot observation $\mathbf{y}_n^i$ for the conventional wireless systems is

$$\begin{aligned}\mathbf{y}_n^i &= \text{diag}\left(\mathbf{p}_n^i\right)\mathbf{\Phi}_n^i\mathbf{g}_n^i + \mathbf{z}_n^i \\ &= \text{diag}\left(\mathbf{p}_n^i\right)\mathbf{\Phi}_n^i\mathbf{F}\begin{bmatrix}\mathbf{h}_n^i \\ \mathbf{0}_{N_F-N_{cir}}\end{bmatrix} + \mathbf{z}_n^i \\ &= \text{diag}\left(\mathbf{p}_n^i\right)\mathbf{\Phi}_n^i\mathbf{F}\mathbf{\Pi}\mathbf{h}_n^i + \mathbf{z}_n^i,\end{aligned} \tag{6.8}$$

where $\mathbf{\Pi} \in \mathbb{R}^{N_F \times N_C}$ is the selection matrix containing only one element being one in each column and rest being zero.[6] Let $\mathbf{U}_n^i = \text{diag}\left(\mathbf{p}_n^i\right)\mathbf{\Phi}_n^i\mathbf{F}\mathbf{\Pi}$, then we have

$$\mathbf{y}_n^i = \mathbf{U}_n^i\mathbf{h}_n^i + \mathbf{z}_n^i. \tag{6.9}$$

In many wireless environments, the channel can be readily expressed using a small number of taps (say K). That is, K is much smaller than the length of CIR vector N_{cir} (i.e., $K \ll N_{cir}$). Since $\mathbf{h}_n^i$ is a sparse vector, one can recover it using only small number of measurements via the CS technique [72]. The drawback of CS-based approaches, in the perspective of multiple antenna systems, is that the channel estimation is performed per antenna so that the pilot overhead and computational complexity increase linearly with the number of transmit antenna N_T. When N_T is large, therefore, this scheme might not be a good fit for IoT systems.

[6] If the first two taps are nonzero elements, then $\mathbf{\Pi} = \begin{bmatrix}1 & 0 & 0 & ... \\ 0 & 1 & 0 & ... \\ 0 & 0 & 0 & ... \\ ... & ... & ... & ...\end{bmatrix}$.

6.4.2 Pilot Beamforming

When the beamforming weight $\mathbf{w}_n(k) = [w_n^1(k)...w_n^{N_T}(k)]^T$ is applied, the beamformed (scalar) channel for k-th subcarrier is

$$\tilde{g}_n(k) = \mathbf{w}_n^T(k) \begin{bmatrix} g_n^1(k) \\ \vdots \\ g_n^{N_T}(k) \end{bmatrix}. \tag{6.10}$$

Then the received vector $\tilde{\mathbf{y}}_n \in \mathbb{C}^{\tilde{N}_P \times 1}$ after aggregating all subcarriers can be expressed as

$$\begin{aligned} \tilde{\mathbf{y}}_n &= \mathrm{diag}\left(\tilde{\mathbf{p}}_n\right) \tilde{\mathbf{\Phi}}_n \begin{bmatrix} \tilde{g}_n(1) \\ \vdots \\ \tilde{g}_n(N_F) \end{bmatrix} + \tilde{\mathbf{z}}_n \\ &= \mathrm{diag}\left(\tilde{\mathbf{p}}_n\right) \tilde{\mathbf{\Phi}}_n \mathrm{diag}\left(\mathbf{W}_n \mathbf{G}_n\right) + \tilde{\mathbf{z}}_n \\ &= \mathrm{diag}\left(\tilde{\mathbf{p}}_n\right) \tilde{\mathbf{\Phi}}_n \mathbf{F} \begin{bmatrix} \tilde{\mathbf{h}}_n^i \\ \mathbf{0}_{N_F - N_{cir}} \end{bmatrix} + \tilde{\mathbf{z}}_n, \end{aligned} \tag{6.11}$$

where $\tilde{\mathbf{h}}_n^i = [\tilde{h}_{n,1}^i, ..., \tilde{h}_{n,N_{cir}}^i]^T$ is the time-domain CIR vector after the beamforming, $\tilde{\mathbf{p}}_n \in \mathbb{C}^{\tilde{N}_P \times 1}$ is the pilot symbol vector, $\tilde{\mathbf{\Phi}}_n \in \mathbb{R}^{\tilde{N}_P \times N_F}$ is the selection matrix, and $\tilde{\mathbf{z}}^j \in \mathbb{C}^{\tilde{N}_P \times 1}$ is the additive white Gaussian noise ($\tilde{\mathbf{z}}_n \sim \mathcal{CN}(0, \sigma_w^2 \mathbf{I}_{\tilde{N}_P})$). Also, $\mathbf{G}_n = \left[\mathbf{g}_n^1 \ ... \ \mathbf{g}_n^{N_T}\right]^T \in \mathbb{C}^{N_T \times N_F}$ is the matrix consisting of frequency-domain channel vectors of N_T antennas, and $\mathbf{W}_n = [\mathbf{w}_n(1) \ ... \ \mathbf{w}_n(N_F)]^T \in \mathbb{C}^{N_F \times N_T}$ is the matrix constructed by stacking beamforming vectors of all subcarriers.

The beamforming matrix $\mathbf{W}_n$ is designed to minimize the cardinality of beam-

formed channel vector $\tilde{\mathbf{h}}_n$. That is,

$$\begin{aligned}\mathbf{W}_n &= \arg\min_{\tilde{\mathbf{W}}_n} \left\|\tilde{\mathbf{h}}_n\right\|_0 \\ &= \arg\min_{\tilde{\mathbf{W}}_n} \left\|\frac{1}{N_F}\mathbf{F}^* \text{diag}\left(\tilde{\mathbf{W}}_n \mathbf{G}_n\right)\right\|_0 .\end{aligned} \tag{6.12}$$

Using the deliberately designed beamforming, we can sparsify the (precoded) time-domain channel vector. For example, suppose the support (index of nonzero elements) of $\mathbf{h}_n^i$ is $\Gamma = \{n_1, n_2, n_3\}$ and the number of antennas N_T is 4. Then the frequency-domain channel for the i-th antenna and k-th subcarrier is

$$\begin{aligned}g_n^i(k) &= \sum_{n=1}^{N_F} h_{n,n}^i e^{j\frac{2\pi k(n-1)}{N_F}} \\ &= \begin{bmatrix} h_{n,n_1}^i & h_{n,n_2}^i & h_{n,n_3}^i \end{bmatrix} \begin{bmatrix} e^{j\frac{2\pi k n_1}{N_F}} \\ e^{j\frac{2\pi k n_2}{N_F}} \\ e^{j\frac{2\pi k n_3}{N_F}} \end{bmatrix} .\end{aligned} \tag{6.13}$$

As mentioned in Section II.A, the channel gain of the i-th antenna can be expressed in terms of h^1 (i.e., $h^i = h^1 e^{j(i-1)\kappa \sin\theta_n}$) and thus

$$\begin{bmatrix} g_n^1(k) \\ g_n^2(k) \\ g_n^3(k) \\ g_n^4(k) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ e^{j\kappa\sin\theta_{n_1}} & e^{j\kappa\sin\theta_{n_2}} & e^{j\kappa\sin\theta_{n_3}} \\ e^{j2\kappa\sin\theta_{n_1}} & e^{j2\kappa\sin\theta_{n_2}} & e^{j2\kappa\sin\theta_{n_3}} \\ e^{j3\kappa\sin\theta_{n_1}} & e^{j3\kappa\sin\theta_{n_2}} & e^{j3\kappa\sin\theta_{n_3}} \end{bmatrix} \begin{bmatrix} h_{n,n_1}^1 e^{j\frac{2\pi k n_1}{N_F}} \\ h_{n,n_2}^1 e^{j\frac{2\pi k n_2}{N_F}} \\ h_{n,n_3}^1 e^{j\frac{2\pi k n_3}{N_F}} \end{bmatrix} . \tag{6.14}$$

Let $\breve{\mathbf{g}}_n(k) = \begin{bmatrix} g_n^1(k) & g_n^2(k) & g_n^3(k) & g_n^4(k) \end{bmatrix}^T$, $\mathbf{\Omega}(k) = \begin{bmatrix} 1 & 1 & 1 \\ e^{j\kappa\sin\theta_{n_1}} & e^{j\kappa\sin\theta_{n_2}} & e^{j\kappa\sin\theta_{n_3}} \\ e^{j2\kappa\sin\theta_{n_1}} & e^{j2\kappa\sin\theta_{n_2}} & e^{j2\kappa\sin\theta_{n_3}} \\ e^{j3\kappa\sin\theta_{n_1}} & e^{j3\kappa\sin\theta_{n_2}} & e^{j3\kappa\sin\theta_{n_3}} \end{bmatrix}$

and $\bar{\mathbf{h}}_n = \begin{bmatrix} h^1_{n,n_1} e^{j\frac{2\pi k n_1}{N_F}} \\ h^1_{n,n_2} e^{j\frac{2\pi k n_2}{N_F}} \\ h^1_{n,n_3} e^{j\frac{2\pi k n_3}{N_F}} \end{bmatrix}$, then we have

$$\breve{\mathbf{g}}_n(k) = \mathbf{\Omega}(k)\bar{\mathbf{h}}_n. \tag{6.15}$$

After the beamforming using $\mathbf{w}^i_n(k)$, the beamformed channel $\tilde{g}(k)$ is given by

$$\begin{aligned} \tilde{g}_n(k) &= \mathbf{w}^T_n(k)\breve{\mathbf{g}}_n(k) \\ &= \mathbf{w}^T_n(k)\mathbf{\Omega}(k)\bar{\mathbf{h}}_n. \end{aligned} \tag{6.16}$$

In this case, by setting the beamforming vector $\mathbf{w}^T_n(k)$ as

$$\mathbf{w}^T_n(k) = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \mathbf{\Omega}^\dagger(k),, \tag{6.17}$$

where $\mathbf{\Omega}^\dagger = (\mathbf{\Omega}^T\mathbf{\Omega})^{-1}\mathbf{\Omega}^T$ is the pseudo-inverse of $\mathbf{\Omega}$[7], we can annihilate all taps of $\mathbf{h}^i_n$ except for n_1-th position. After the beamforming, we obtain the beamformed frequency-domain (scalar) channel $\tilde{g}_n(k)$ as

$$\begin{aligned} \tilde{g}_n(k) &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \mathbf{\Omega}^\dagger(k)\mathbf{\Omega}(k)\bar{\mathbf{h}}_n \\ &= h^1_{n,n_1} e^{j\frac{2\pi k n_1}{N_F}}. \end{aligned} \tag{6.18}$$

From (6.18), one can observe that the channel $\tilde{g}_n(k)$ for the subcarrier k is a function

[7]Since the dimension of matrix $\mathbf{\Omega}(k)$ is $N_T \times K$ and $\mathbf{\Omega}(k)$is a full rank matrix, one can obtain the pseudo-inverse as long as the number of antenna N_T is larger than or equal to K. Other than IoT scenarios, this condition is also well suited for millimeter wave communication scenarios.

of only h^1_{n,n_1}. Thus, one can show that

$$\begin{bmatrix} \tilde{g}_n(1) \\ \vdots \\ \tilde{g}_n(N_F) \end{bmatrix} = \mathbf{F} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ h^1_{n,n_1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{6.19}$$

Without loss of generality, to select k taps out of K taps, we can use binary vector $\mathbf{e}$ for selecting taps for channel sparsification.

$$\mathbf{w}^T_n(k) = \mathbf{e}^T \mathbf{\Omega}^\dagger(k). \tag{6.20}$$

It is worth mentioning that one can further improve the performance by transmitting multiple beams simultaneously and then choosing the beam with the best quality. When the number of beam is J, the aggregated time-domain channel vectors of beams can be expressed as

$$\begin{aligned} \left[\tilde{\mathbf{h}}^1_n \ ... \ \tilde{\mathbf{h}}^J_n\right] &= \frac{1}{N_F}\mathbf{F}^* \left[\text{diag}\left(\mathbf{W}^1_n \mathbf{G}_n\right) \quad ... \quad \text{diag}\left(\mathbf{W}^J_n \mathbf{G}_n\right)\right] \\ &= \frac{1}{N_F}\mathbf{F}^* \begin{bmatrix} [\mathbf{w}^1_n(1) \ ... \ \mathbf{w}^J_n(1)]\mathbf{\Omega}(1)\bar{\mathbf{h}}_n \\ \vdots \\ [\mathbf{w}^1_n(N_F) \ ... \ \mathbf{w}^J_n(N_F)]\mathbf{\Omega}(N_F)\bar{\mathbf{h}}_n \end{bmatrix}. \end{aligned} \tag{6.21}$$

In each subcarrier k, the precoding weights are expressed as

$$[\mathbf{w}^1_n(k) \ ... \ \mathbf{w}^J_n(k)] = \mathbf{E}\mathbf{\Omega}(k)^\dagger \text{ for } K \leq N_T. \tag{6.22}$$

where $\mathbf{E} = [\mathbf{e}^1 \ ... \ \mathbf{e}^J]^T$. For example, by setting

$$\text{supp}\left\{\mathbf{e}^i\right\} \neq \text{supp}\left\{\mathbf{e}^j\right\} \text{ for } i, j = 1, ..., J., \tag{6.23}$$

J beamforming pilots can exploit orthogonal taps from the original time-domain channel vector $\mathbf{h}$. This setting has benefit when IoT device performs channel estimation for multiple beamformed pilot jointly. Since the receiver has prior knowledge of non-overlapped tab location between beams, the estimation complexity can be reduce significantly.

6.5 Channel Estimation of Beamformed Pilots

6.5.1 Recovery using Multiple Measurement Vector

The received vector of the beamformed pilot signals $\tilde{\mathbf{y}}_n$ is

$$\begin{aligned} \tilde{\mathbf{y}}_n &= \text{diag}\left(\tilde{\mathbf{p}}_n\right) \tilde{\mathbf{\Phi}}_n \mathbf{F} \begin{bmatrix} \tilde{\mathbf{h}}_n \\ \mathbf{0}_{N_F - N_{cir}} \end{bmatrix} + \tilde{\mathbf{z}}_n \\ &= \tilde{\mathbf{U}}_n \tilde{\mathbf{h}}_n + \tilde{\mathbf{z}}_n, \end{aligned} \tag{6.24}$$

where $\tilde{\mathbf{U}}_n = \text{diag}\left(\tilde{\mathbf{p}}_n\right) \tilde{\mathbf{\Phi}}_n \mathbf{F} \mathbf{\Pi}$. In estimating the channel $\tilde{\mathbf{h}}_n$, we use the sparse recovery algorithm (such as an orthogonal matching pursuit). In many IoT scenarios, the position of dominant components in the CIR vector remain unchanged during the coherence time so that we can use multiple observations to estimate the beamformed sparse channel. Let L be the number of pilot observations in a local block, then the

stacked vector $\bar{\mathbf{y}}_n = [\tilde{\mathbf{y}}_{n-L} ... \ \tilde{\mathbf{y}}_n]^T$ of beamformed observations is expressed

$$\underbrace{\begin{bmatrix} \tilde{\mathbf{y}}_{n-L} \\ \vdots \\ \tilde{\mathbf{y}}_n \end{bmatrix}}_{\bar{\mathbf{y}}_n} = \underbrace{\begin{bmatrix} \tilde{\mathbf{U}}_{n-L} & & \\ & \ddots & \\ & & \tilde{\mathbf{U}}_n \end{bmatrix}}_{\bar{\mathbf{U}}_n} \underbrace{\begin{bmatrix} \tilde{\mathbf{h}}_{n-L} \\ \vdots \\ \tilde{\mathbf{h}}_n \end{bmatrix}}_{\bar{\mathbf{h}}_n} + \underbrace{\begin{bmatrix} \tilde{\mathbf{z}}_{n-L} \\ \vdots \\ \tilde{\mathbf{z}}_n \end{bmatrix}}_{\bar{\mathbf{z}}_n}. \tag{6.25}$$

After rearranging columns of $\bar{\mathbf{h}}_n$, we have

$$\bar{\mathbf{y}}_n = \bar{\mathbf{U}}_n^{'} \bar{\mathbf{h}}_n^{'} + \bar{\mathbf{z}}_n^{'}, \tag{6.26}$$

where $\bar{\mathbf{h}}_n^{'} = [\tilde{\mathbf{h}}_{n-L}(1) \ ... \ \tilde{\mathbf{h}}_n(1) \tilde{\mathbf{h}}_{n-L}(2) \ ... \ \tilde{\mathbf{h}}_n(2) \ ... \ \tilde{\mathbf{h}}_n(N_{cir})]^T$ and $\bar{\mathbf{U}}_n^{'} = [\tilde{\mathbf{U}}_{n-L}(: , 1) \ ... \ \tilde{\mathbf{U}}_n(:, 1) \tilde{\mathbf{U}}_{n-L}(:, 2) \ ... \ \tilde{\mathbf{U}}_n(:, 2) \ ... \ \tilde{\mathbf{U}}_n(:, N_{cir})]$.

Let $\mathbf{d}_{n,k} = [\tilde{\mathbf{h}}_{n-L}(k) \ ... \ \tilde{\mathbf{h}}_n(k)]^T$ and $\mathbf{\Sigma}_{n,k} = [\tilde{\mathbf{U}}_{n-L}(:, k) \ ... \ \tilde{\mathbf{U}}_n(:, k)]$, we can rewrite $\bar{\mathbf{y}}_n$ using $\mathbf{d}_{n,k}$ and $\mathbf{\Sigma}_{n,k}$ as

$$\bar{\mathbf{y}}_n = \begin{bmatrix} \mathbf{\Sigma}_{n,1} & ... & \mathbf{\Sigma}_{n,N_{cir}} \end{bmatrix} \begin{bmatrix} \mathbf{d}_{n,1} \\ ... \\ \mathbf{d}_{n,N_{cir}} \end{bmatrix} + \begin{bmatrix} \bar{\mathbf{z}}_{n,1} \\ ... \\ \bar{\mathbf{z}}_{n,N_{cir}} \end{bmatrix}. \tag{6.27}$$

The output (nonzero tap index and channel estimate) of the block OMP algorithm is given by [72]

$$\begin{aligned} \hat{k} &= \arg\max_k \|\mathbf{\Sigma}_{n,k} \bar{\mathbf{y}}_n\|_2^2 \\ \hat{\mathbf{d}}_{n,\hat{k}} &= E[\mathbf{d}_{n,\hat{k}} \bar{\mathbf{y}}_n^H] E^{-1}[\bar{\mathbf{y}}_n \bar{\mathbf{y}}_n^H] \bar{\mathbf{y}}_n \\ &= [\mathbf{R}_n \mathbf{\Sigma}_{n,\hat{k}}^H][\mathbf{\Sigma}_{n,\hat{k}} \mathbf{R}_n \mathbf{\Sigma}_{n,\hat{k}}^H + \sigma_{\tilde{z}}^2 \mathbf{I}]^{-1} \bar{\mathbf{y}}_n, \end{aligned} \tag{6.28}$$

where $\mathbf{R}_n$ is the covariance matrix of channel tap [70]. Rearranging back to the original

Algorithm 1 Joint time-domain channel vector recovery of multiple beamformed pilots

Input: CS matrix $\mathbf{\Sigma}$, measurement vector $\bar{\mathbf{y}}$
Output: Estimated CIR of beamformed pilots $\hat{\mathbf{h}}^j$
Initialize: Estimated channel vector $\hat{\mathbf{d}}^{(0)} = \mathbf{0}$, residual vector $\Delta = \bar{\mathbf{y}}$, joint support vector $\mathbf{S} = \mathbf{0}$, beam support vector $\mathbf{b} = \mathbf{0}$, tap support vector $\mathbf{c} = \mathbf{0}$, $i = 0$, tap searching set $\Psi = \{1, ..., N_{cir}\}$, beam searching set $\Upsilon = \{1, ..., J\}$
while stopping criterion false **do**
$i \leftarrow i + 1$
$l, k \leftarrow \underset{l \in \Upsilon, k \in \Psi}{\arg\max} \|\mathbf{\Sigma}_{n,l,k}\Delta\|_2^2$, //select beam and tap index
$s_{b_l c_k} \leftarrow 1$, $s_{b_l c_m} \leftarrow 0$ when $m \neq k$, $s_{b_t c_k} \leftarrow 0$ when $l \neq t$ //update support vector
$\Psi \leftarrow \Psi - \{k\}$, $\Upsilon \leftarrow \Upsilon - \{l\}$ //remove detected supports from the searching sets
$E[\mathbf{d}_{n,q,p}\bar{\mathbf{y}}_n^H] \leftarrow [\mathbf{s}_{b_q,c_p}\mathbf{R}_n\mathbf{\Sigma}_{n,q,p}^H]$
$E^{-1}[\bar{\mathbf{y}}_n\bar{\mathbf{y}}_n^H] \leftarrow [\sum_p \sum_q \mathbf{s}_{b_q,c_p}\mathbf{R}_n\mathbf{\Sigma}_{n,q,p}\mathbf{\Sigma}_{n,q,p}^H + \sigma_{\tilde{z}}^2\mathbf{I}]^{-1}$
$\hat{\mathbf{d}}_{n,q,p}^{(i)} \leftarrow E[\mathbf{d}_{n,q,p}\bar{\mathbf{y}}_n^H]E^{-1}[\bar{\mathbf{y}}_n\bar{\mathbf{y}}_n^H]\bar{\mathbf{y}}$ for $p = 1, ..., N_{cir}$ and $q = 1, ..., J$ //
$\Delta \leftarrow \bar{\mathbf{y}} - \sum_{p=1}^{p=N_{cir}} \sum_{q=1}^{q=J} \mathbf{\Sigma}_{n,q,p}\mathbf{s}_{b_q c_p}\hat{\mathbf{d}}_{n,q,p}$ //update residual
$\hat{\mathbf{h}}^l \leftarrow \left[(\hat{\mathbf{d}}_{n,l,1}^{(i)})_L ... (\hat{\mathbf{d}}_{n,l,N_{cir}}^{(i)})_L\right]^T$ //Estimated channel tap
end while

index, we obtain the beamformed time-domain channel vector as

$$\hat{\mathbf{h}}_n = \left[(\hat{\mathbf{d}}_{n,1})_L ... (\hat{\mathbf{d}}_{n,N_{cir}})_L\right]^T. \tag{6.29}$$

Since the number of iterachannel estimation of TDSB requires a which is feasible for IoT solutions. Since the complexity of channel estimation with TDSB is extremely low, we can estimate multiple beams jointly. When we exploit multiple beams (see Fig. 6.5), we can estimate channels of multiple beams jointly. Let $\tilde{\mathbf{y}}_n^j$ be the received beamformed pilot observations of j-th beamformed pilot and $\hat{\mathbf{y}}_n = \left[\tilde{\mathbf{y}}_n^1 ... \ \tilde{\mathbf{y}}_n^J\right]^T$ be the stacked beamformed pilots. Let L be the number of observation block, then we have

$$\bar{\mathbf{y}}_n = \begin{bmatrix} \hat{\mathbf{y}}_{n-L} & \dots & \hat{\mathbf{y}}_n \end{bmatrix}^T. \tag{6.30}$$

Again, after proper rearranging, we have

$$\bar{\mathbf{y}}_n = \bar{\mathbf{U}}_n''\bar{\mathbf{h}}_n'' + \bar{\mathbf{z}}_n'', \tag{6.31}$$

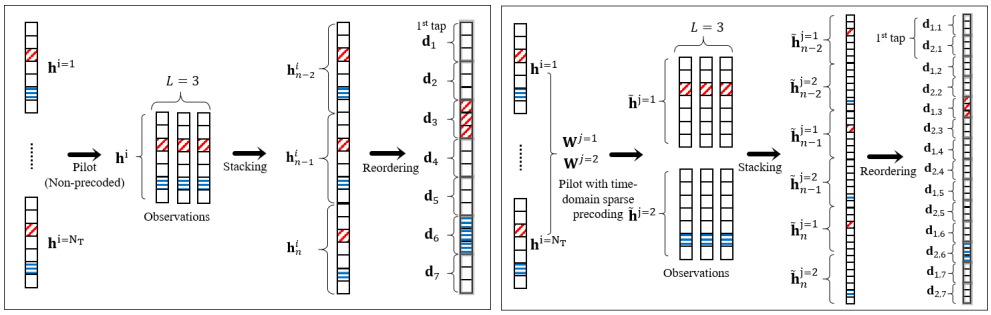


Figure 6.5: Comparison between channel vector recovery when $L = 3$ and $K = 2$: (a) the conventional time-domain CS recovery (find 2 support out of 7 candidates), and (b) The proposed CS recovery with time-domain sparse precoding (find 2 support out of 14 candidates).

where $\bar{\mathbf{h}}_n'' = [\tilde{\mathbf{h}}_{n-L}^1(1) \ ... \ \tilde{\mathbf{h}}_n^1(1)\tilde{\mathbf{h}}_{n-L}^1(2) \ ... \ \tilde{\mathbf{h}}_n^1(2) \ ... \ \tilde{\mathbf{h}}_n^J(N_{cir})]^T$ and $\bar{\mathbf{U}}_n'' = [\tilde{\mathbf{U}}_{n-L}^1(:,1) \ ... \ \tilde{\mathbf{U}}_n^1(:,1)\tilde{\mathbf{U}}_{n-L}^1(:,2) \ ... \ \tilde{\mathbf{U}}_n^1(:,2) \ ... \ \tilde{\mathbf{U}}_n^J(:,N_{cir})]$.

Let $\mathbf{d}_{n,j,k} = [\tilde{\mathbf{h}}_{n-L}^j(k) \ ... \ \tilde{\mathbf{h}}_n^j(k)]^T$ and $\mathbf{\Sigma}_{n,j,k} = [\tilde{\mathbf{U}}_{n-L}^j(:,k) \ ... \ \tilde{\mathbf{U}}_n^j(:,k)]$, MMV can be rewritten as

$$\bar{\mathbf{y}}_n = \begin{bmatrix} \mathbf{\Sigma}_{n,1,1} & ... & \mathbf{\Sigma}_{n,J,N_{cir}} \end{bmatrix} \begin{bmatrix} s_{1,1}\mathbf{d}_{n,1,1} \\ ... \\ s_{J,N_{cir}}\mathbf{d}_{n,J,N_{cir}} \end{bmatrix} + \begin{bmatrix} \bar{\mathbf{z}}_{n,1,1} \\ ... \\ \bar{\mathbf{z}}_{n,J,N_{cir}} \end{bmatrix}, \tag{6.32}$$

where $\mathbf{\Sigma}_{n,j,k}$ is the re-ordered sensing matrix of signal vectors $\mathbf{d}_{n,j,k}$. In order to identify whether signal vector $\mathbf{d}_n$ contains energy, we define a binary vector $\mathbf{s} = [s_{1,1} \ ... \ s_{1,N_{cir}} \ ... \ s_{i,j} \ ... \ s_{J,N_{cir}}]$ where $s_{i,j} = 1$ indicates that j-th tap in i-th beam is nonzero and $s_{i,j} = 0$ otherwise. In each iteration of the recovery algorithm, which is essentially the block OMP algorithm, the index of the component with the largest energy in the channel vector is chosen. We note that all taps in the selected beam and all beam in the selected tap are removed from the search set in the next iteration. We repeat the iterations until K dominant components are identified. The detailed steps of algorithm are depicted in Algorithm 1.

6.5.2 MSE Analysis

In this section, we investigate the MSE performance of beamformed pilot using compressive sensing. For analysis simplicity, the followings are assumed: 1) Orthogonal matching pursuit (OMP) algorithm is used, 2) The probability of time-domain channel tap locations are equally likely between N_{cir} taps, and 3) $\tilde{K}$ sparsity is known to the receiver. Since we use OMP algorithm, the estimator would like to detect correct tap when the measurements are enough. In fact, if the algorithm works correctly, the set of detected tap would be equal or subset of actual taps with high probability. Although there are many combinatorial cases for detected set and active sets, we can simplify as follows:

- Case A: the number of detected tap $t_{det}(= t_{corr} + t_{inc})$ is less than the sparsity of beamformed pilot $\tilde{K}$. For example, $\tilde{K} - t_{det} = n$, there are t_{det} taps are corrected detected and n taps are mis-detected.

- Case B: the number of detected tap t_{det} is larger than the sparsity of beamformed pilot $\tilde{K}$ and actual taps are subset of detected set. For example, $t_{det} - \tilde{K} = m$, there are $\tilde{K}$ taps are corrected detected and m taps are mis-detected.

Let $\mathsf{MSE}_j = E(|\Delta \mathbf{g}_j|^2)$ be the MSE of jth CIR tap, where $\Delta \mathbf{g}_j = \hat{\mathbf{g}}_j - \mathbf{g}_j$. When jth tap is active and detected correctly, MSE of jth CIR tap is distorted by noise only ($\mathsf{MSE}_j = \sigma_z^2$). Otherwise, MSE_j is occurred from either of two cases; detected but not active and non-detected but active. When the jth tap is detected but not active tap, Then MSE_j can be summarized when the number of detected tap k_{det} is less than $\tilde{K}$ and vice versa. That is

$$\mathsf{MSE}_j = \begin{cases} \sigma_z^2 + \sigma_{s,j}^2, & \text{if } k_{det} < \tilde{K} \\ 2\sigma_z^2 + \sigma_{s,j}^2, & \text{otherwise} \end{cases}, \tag{6.33}$$

where $\sigma_{s,j}^2$ is the signal power of jth active tap. The overall MSE is given by

$$\begin{aligned}\mathsf{MSE} &= P_s \left(\sum_{correct} \sigma_z^2 \right) \\ &\quad + (1 - P_s) \sum_{l=1}^{N_{cir}} P_{tap,l} \left(\sum_{correct} \sigma_z^2 + \sum_{incorrect} \mathsf{MSE}_j \right) \\ &\approx P_s \left(\sum_{correct} \sigma_z^2 \right) + (1 - P_s) \\ &\quad \times \sum_{n=1}^{\tilde{K}} P_{m,n}^{\tilde{K}} \left(\sum_{correct} \sigma_z^2 + \sum_{incorrect} \mathsf{MSE}_j \right),\end{aligned} \tag{6.34}$$

where P_s is probability of perfect support detection with $\tilde{K}$ sparse, $P_{tap,l}$ is probability of l tap detection, and $P_{m,n}^{\tilde{K}}$ is probability when n taps are miss-detected among $\tilde{K}$ taps. With these relaxed assumptions, $P_{m,n}^{\tilde{K}}$ is $\binom{\tilde{K}}{n}\binom{N_{cir}-\tilde{K}}{\tilde{K}-n}/\binom{N_{cir}}{\tilde{K}}$ and we can rewrite (6.33) using the Corollary 4.2 in [94] as follows

$$\begin{aligned}\mathsf{MSE} = \sigma_E^2 &= P_s \left(\sum_{correct}^{\tilde{K}} \sigma_z^2 \right) + (1 - P_s) \\ &\quad \times \sum_{n=1}^{\tilde{K}} P_{m,n}^{\tilde{K}} \left((\tilde{K} - n)\sigma_z^2 + n(\sigma_z^2 + \sigma_{s,j}^2) \right) \\ &\leq \tilde{K}\sigma_z^2 + C\delta^{1/2} \frac{\tilde{K}^2}{N_{cir}} \sigma_s^2, \quad \text{if } \frac{\sigma_s^2}{\sigma_z^2} \geq \delta_{2\tilde{K}}^{-3/2}.\end{aligned} \tag{6.35}$$

6.6 Simulations and Discussion

6.6.1 Simulation Setup

In this section, we compare the MSE and spectral efficiency of the conventional schemes and the proposed TDSB scheme. We consider OFDM-based IoT systems where $B_s = 20$ MHz, subcarrier spacing of 15 KHz, and DFT size of $N_F = 2048$. The maxi-

mum delay spread of the multipath channel is assumed to be 0.467 μs, which yields $N_{cir} = 144$. Duration of OFDM symbol is 7.2 μs and the time interval between adjacent pilot symbols is 5 ms. We assume that 5 taps in the channel vector are nonzero with uniform energy. One out of 12 subcarriers is randomly chosen for the pilot purpose. In the proposed TDSB scheme 5 beams are used for the pilot transmission. As a channel model, we use Jakes' model for the temporal correlation of the complex gains of the tap is $J_0(2\pi f_D \tau)$ where $J_0(\cdot)$ is the 0-th order Bessel function of the fist kind, $f_D = v f_c/c$ denotes the maximum Doppler frequency, and τ is the pilot interval. We also assume that nonzero elements of time-domain channel are uncorrelated.

6.6.2 Simulation Results

We first compare the MSE performances of the proposed TDSB scheme with the conventional channel estimation algorithms (LMMSE [73]) in Fig. 6.6. In order to observe the performance of narrowband observation, we assign 12 consecutive subcarriers ($B_u = 150$ KHz) to each IoT device. Since the receiver has only partial observation in frequency-domain at each pilot observation, the conventional LMMSE does not work properly. Further, pilot subcarriers are locally placed within narrow bandwidth of B_s, the conventional CS-based channel estimation scheme is also not working well. We observe that the proposed scheme outperforms the conventional schemes and achieves more than 20 dB gain in high SNR regime. This result reveals that the proposed TDSB scheme is feasible for the MTC or IoT devices equipped with stringent hardwares.

In order to observe the effectiveness of channel sparsification, we next measure the MSE performances with and without using TDSB as depicted in Fig. 6.7. As the conventional pilot transmission, beamforming is not used so that all nonzero taps are estimated from the pilot measurements. That is, pilots are transmitted per antenna without beamforming and time-domain channel vectors $\mathbf{h}_n^i$ are estimated for each antenna i. In case of TDSB, each beamformed pilot employ one of nonzero taps in $\mathbf{h}_n$ and beamformed time-domain channel vectors $\tilde{\mathbf{h}}_n^j$ are estimated jointly at the device. Since the

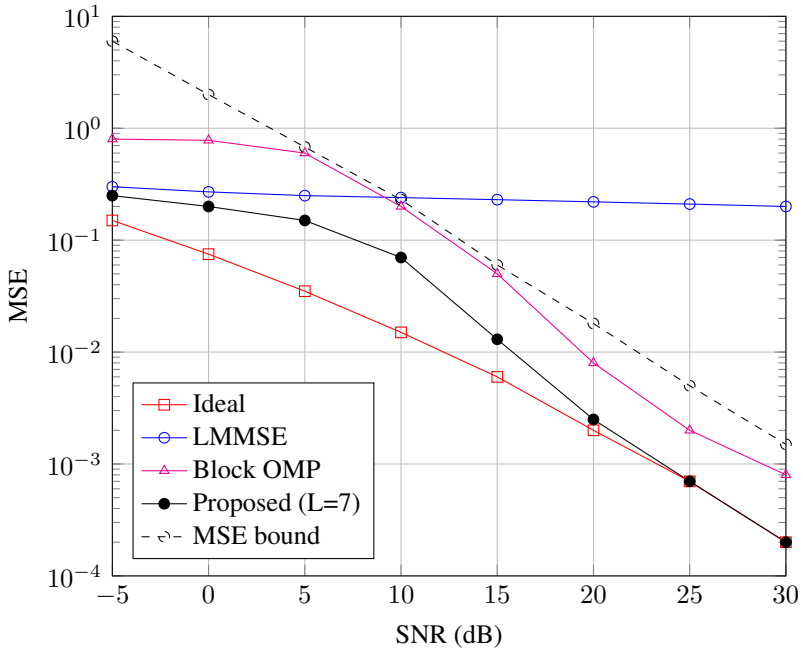


Figure 6.6: MSE performance of TDSB scheme as a function of SNR ($L = 7$, $J = 5$, $K = 5$ and $N_T = 16$).

number of nonzero taps is reduced by TDSB, the proposed scheme outperforms the cases without using TDSB scheme at the basestation. Interestingly, we observed that MSE performance is improved as the nonzero taps are decreased in the time-domain channel vector and this reveals same results mentioned in Section V. The benefit of sparsification can be better understood by observing the normalized throughput. In Fig. 6.9, we plot the normalized throughput as a function of SNR. As a performance metric, we define the normalized throughput η as

$$\eta\,(bps/Hz) = \frac{N_{bit}}{(N_t - N_p)T_d B_s}, \tag{6.36}$$

where N_{bit} is the number of successive bit during T_d duration, N_t is the number of REs used for transmission, and N_p is the number of REs used for pilot transmission. In this case, we set $N_T = 32$ and compare the performance of the TSDB scheme with

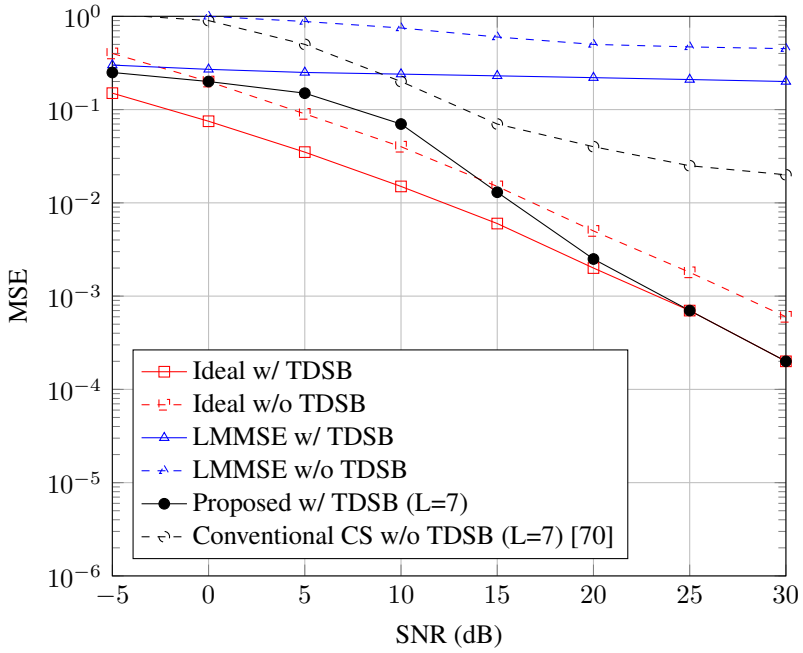


Figure 6.7: MSE performance with and without TDSB scheme at the basestation ($L = 7$, $J = 5$, $K = 5$ and $N_T = 16$).

the following two scenarios; 1) a system use the proposed beamforming and channel estimation algorithm, and 2) a system use the conventional (without beamforming) pilot transmission and CS-based channel estimation. For pilot overhead, 2 and 170 subcarriers are used for the proposed scheme and the conventional pilot transmission, respectively. For the data transmission, we use the best beam feed back from the IoT device [78]. Since the proposed scheme use fewer pilot symbols, the TDSB scheme achieves 2-fold of gain in terms of normalized throughput.

In order to observe the effect of pilot overhead, we plot the normalized throughput as a function of pilot overhead in Fig. 6.10. The pilot overhead is the ratio of the pilot resources to the LTE-Advanced systems. For example, 100% pilot overhead indicates that 8.3% of whole OFDM resources are used for pilot symbols. As pilot overhead decreases, the conventional LMMSE scheme performs poorly so that normalized throughput decreases. While MSE performance is improved as pilot overhead

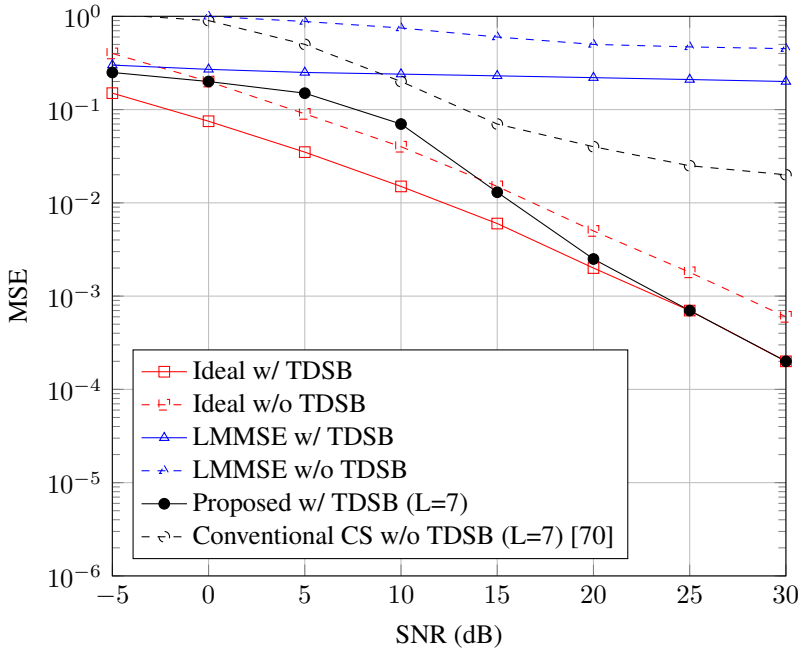


Figure 6.8: Test.

increase in the TDSB scheme, normalized throughput is decreased due to pilot overhead. Overall, results shows that TSDB scheme outperforms the conventional scheme without TDSB even when the pilot overhead is accounted for and achieves 2 folds of gain in 10 dB SNR regime with N_T-fold reduction in pilot overhead. Also, the results reveal that the proposed scheme achieves substantial reduction in pilot overhead.

6.7 Summary

In this chapter, we proposed a pilot beamforming scheme for the IoT systems by Sparsification of pilot beamforming. Our work is motivated by the observation that the pilot overhead must scale linearly with the number of taps in CIR vector and the number of transmit antennas so that the conventional pilot transmission is not an appropriate option for the IoT devices. The key feature of the proposed TDSB scheme is to minimize the nonzero elements of a channel vector using the antenna-domain beamforming. To

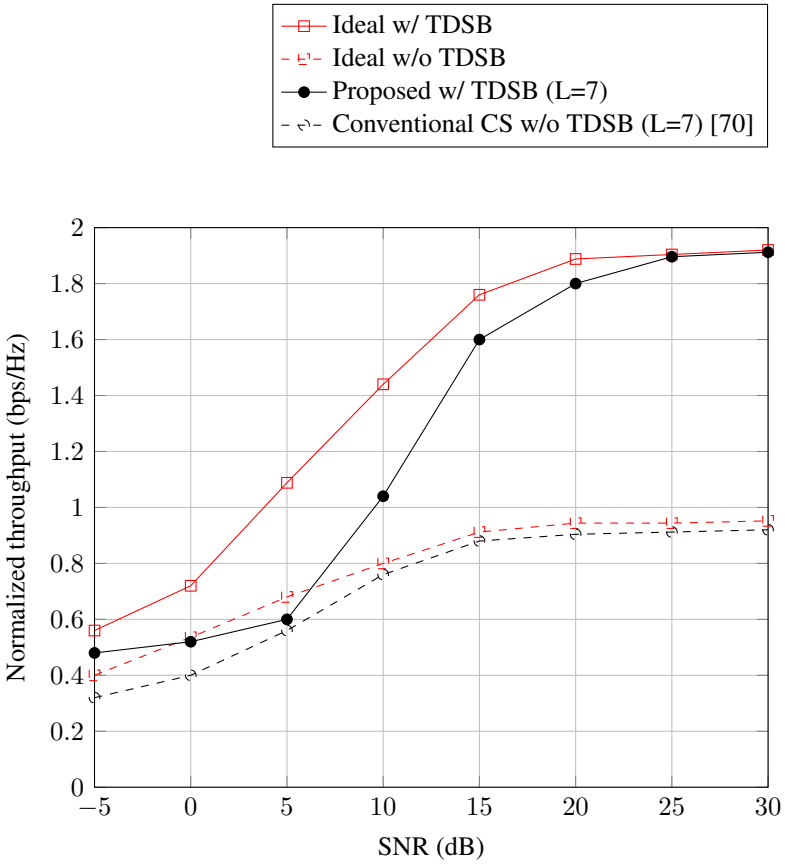


Figure 6.9: Normalized throughput (QPSK with $N_T = 32$, K is 5).

do so, we apply the time-domain sparse precoding, where each precoded channel propagates via fewer tap than the original channel vector. The received channel vector of beamformed pilots can be jointly estimated by the sparse recovery algorithm. From numerical simulations, we observed that the proposed TDSB scheme achieves significant reduction in pilot overhead. Having a few nonzero taps with orthogonal supports between the beamformed pilots, CS estimator worked effectively with less pilot overhead. It has been shown by MSE analysis and evaluation results that the proposed scheme is effective in achieving an extreme overhead reduction in the pilot overhead in the IoT scenarios.

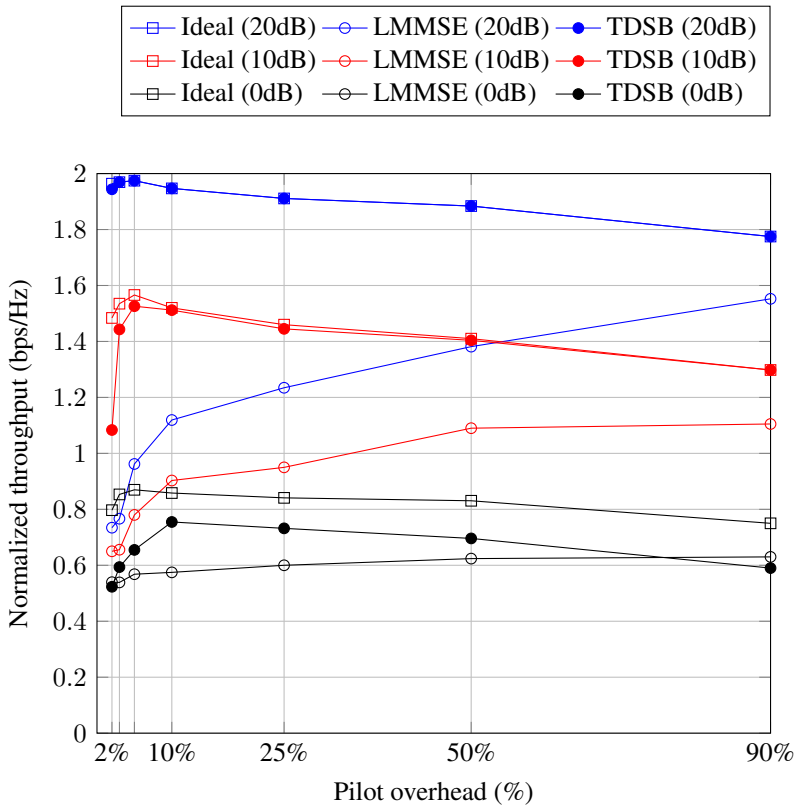


Figure 6.10: Throughput as a function of pilot overhead (SNR=10dB, $N_T = 32$).

Chapter 7

Conclusion

7.1 Summary

In this dissertation, new strategies to employ sparse-ware recovery have been extensively studied. Although the sparse recovery has been used in the recovery of channel and sensing the sporadic signal phenomenon, I focused on the various applications including packet transmission, multiuser access, beamforming, and feedback. These new extensions enables to transmit data efficiently, manage interuser interference, recover beamformed channel with far less pilot overhead, and transmit highly accurate information, which provide guidelines for designing 5G systems. Specifically, I make the following contributions:

- In Chapter 2, we have proposed the short packet transmission strategy for URLLC. The key idea behind the proposed SVC technique is to transform an information vector into the sparse vector in the transmitter and to exploit the sparse recovery algorithm in the receiver. Metaphorically, SVC can be thought as a marking dots to the empty table. As long as the number of dots is small enough and the measurements contain enough information to figure out the marked cell positions, accurate decoding of SVC packet can be guaranteed. We showed from the numerical evaluations that the proposed SVC scheme is very effective in URLLC

scenarios. In this dissertation, we restricted our attention to the URLLC scenario but we believe that there are many other applications that the SVC technique can be applied to. Also, there are many interesting extensions and variations worth investigating, such as the information embedding in nonzero positions, channel aware sparse vector coding, and combination of SVC and error correction codes.

- In Chapter 3, we have proposed a new short packet transmission scheme called sparse vector coding (SVC) for the mMTC uplink scenario. The key idea behind the proposed SVC transmission scheme is to convert an information vector into the sparse vector and use the support identification algorithm as a decoder in the receiver. The SVC transmission scheme is easy to implement, robust to noise and multiuser interference, and also scalable to the massive access scenario in mMTC. We showed from the simulations in the LTE uplink scenario and massive access scenario in 5G that the proposed SVC technique is very effective in short packet transmissions.

- In Chapter 4, we have proposed the pilot-less sparse vector coding (PL-SVC) suitable for the short packet transmission in URLLC and mMTC scenarios. The key idea behind the proposed PL-SVC is to transmit the location information in the form of a sparse vector and then to decode the location information via the compressed sensing technique. The PL-SVC transmission scheme does not require pilot transmission, also very easy to implement, and can be easily extended to the SIMO and MIMO configurations. We show from the numerical experiments in the 5G OFDM uplink scenario that the proposed PL-SVC is effective in the short packet transmission and outperforms the conventional SVC and PUSCH transmission schemes.

- In Chapter 5, we have proposed a new data transmission strategy for the ultra-shot packet by transmitting side (analog) information. The key idea behind the proposed JSVC transmission scheme is to transform the small information into

a sparse vector and map the side-information into a magnitude of the sparse vector. Metaphorically, JSVC can be thought as a standing a few poles to the empty table and measure the height of pole at the receiver. As long as the number of poles is small enough and the measurements contains enough information to find out the marked cell positions, accurate recovery of JSVC packet can be guaranteed. We showed from the numerical evaluations and analysis of decoding performance that the proposed JSVC scheme is very effective in ultra-short packet transmission such as analog feedback, MIMO feedback, and control-type channel.

- In Chapter 6, we have proposed a pilot beamforming scheme for the IoT systems by Sparsification of pilot beamforming. Our work is motivated by the observation that the pilot overhead must scale linearly with the number of taps in CIR vector and the number of transmit antennas so that the conventional pilot transmission is not an appropriate option for the IoT devices. The key feature of the proposed TDSB scheme is to minimize the nonzero elements of a channel vector using the antenna-domain beamforming. To do so, we apply the time-domain sparse precoding, where each precoded channel propagates via fewer tap than the original channel vector. The received channel vector of beamformed pilots can be jointly estimated by the sparse recovery algorithm. From numerical simulations, we observed that the proposed TDSB scheme achieves significant reduction in pilot overhead. Having a few nonzero taps with orthogonal supports between the beamformed pilots, CS estimator worked effectively with less pilot overhead. It has been shown by MSE analysis and evaluation results that the proposed scheme is effective in achieving an extreme overhead reduction in the pilot overhead in the IoT scenarios.

7.2 Future Research Directions

In this section, we discuss future research directions related to the topics of this dissertation.

- Combined with channel coding and sparse vector coding: one of drawback of sparse encoding is the accept of error correction capability. Since the number of nonzero element is a few, when the decoder miss the correct support, the whole packet is corrupted, so that bit error detection or correction is not possible. However, when we combine the channel coding mechanism into the sparse mapping, the decoder can correct the false detected support so that the performance would be improved.

- Grant-free access with blind active user detection: when sparse mapping is used in uplink transmission, the packet structure as well as the user sparsity can jointly used for decoder. When a few users are active among the user set, the actual number of transmitted user is only a few and the support of each user would be small as well. When we consider double sparsity, the active user detection performance as well as packet decoding performance would be improved.

- Multi-dimensional sparsity mapping for short packet transmission: in this dissertation, we only considered the sparsity in code dimension. However, when we extend sparse mapping in other dimension, sparse mapping can be extended into such as antenna ports, frequency, time, and spatial layers. Using multi-dimensional sparse mapping, one can improve the reliability of ultra short packet transmission.

- Simultaneous wireless information and power transmission: when wireless power transmission is used together with information, the information is transmitted without changing its original form so that the efficiency of power transmission is very small. Meanwhile, to maximize the efficiency of power transmission,

the rate of information is extremely low so that it is not possible to communicate. When the sparse mapping is used for information, the transmitter will have freedom to design the signal for power efficiency and both information rate and power efficiency will improve.

Bibliography

[1] Rec. ITU-R M.2083-0, "IMT Vision - Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," Sep, 2015.

[2] P. Schulz, M. Matthé, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Müller, T. Elste, and M. Windisch, "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70-78, 2017.

[3] 3GPP Technical Report 38.802, "Study on New Radio Access Technology Physical Layer Aspects (Release 14)," v14.1.0, 2017.

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

[5] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra Reliable and Low Latency Communications in 5G: Physical Layer Aspects," *to appear in IEEE Wireless Commun.*

[6] B. Lee, S. Park, D. Love, H. Ji, and B. Shim, "Packet Structure and Receiver Design for Low Latency Wireless Communications with Ultra Short Packets," *IEEE Trans. on Commun.*, vol. 66, no. 2, pp. 796-807, Feb. 2018.

[7] O. Yilmaz, Y. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, and Joachim Sachs, “Analysis Ultra-Reliable and Low-Latency 5G Communication for a Factory Automation Use Case," *In Proc. IEEE Int. Conf. on Comm. (ICC) Workshop*, pp. 1190 - 1195, 2015.

[8] N. A. Johansson, Y.Wang, E. Eriksson, and M. Hessler, “Radio Access for Ultra-Reliable and Low-Latency 5G Communications," *In Proc. IEEE Int. Conf. on Comm. (ICC) Workshop*, pp. 1185 - 1189, 2015.

[9] X. Chen, T. Chen, and D. Guo, “Capacity of Gaussian Many-Access Channels,” *IEEE Trans. Inform. Theory*, vol. 63. no. 6, pp. 3516-3539, 2017.

[10] 3GPP Technical Report 38.913, “Study on Scenarios and Requirements for Next Generation Access Technologies (Release 14),” v14.2.0, 2017.

[11] S. Sesia, M. Baker, and I. Toufik, *LTE-the UMTS Long Term Evolution: from Theory to Practice,* John Wiley & Sons., 2012.

[12] J. W. Choi, B. Shim, Y. Ding, B. Rao, and D. I. Kim, “Compressed Sensing for Wireless Communications: Useful Tips and Tricks.,” *IEEE Commun. Survey and Tutorials,*, vol. 19, pp. 1527-1550, 2017.

[13] D. L. Donoho, “Compressed Sensing,” *IEEE Trans. on Inform. Theory*, vol.52, no. 4, pp. 1289-1306, 2006.

[14] E. Basar, M. Wen, R. Mesleh, M. D. Renzo, Y. Xiao, and H. Haas, “Index Modulation Techniques for Next-Generation Wireless Networks,” *IEEE Access*, vol. 5, pp. 16693-16746, Aug. 2017.

[15] G. Kaddoum, Y. Nijsure, and H. Tran, “Generalized code index modulation technique for high-data-rate communication systems”, *IEEE Trans. Veh. Technol.,* vol. 65, no. 9, pp. 7000-7009, Sept. 2016.

[16] G. Castagnoli, S. Bräuer, and M. Herrmann, "Optimization of Cyclic Redundancy-Check Codes with 24 and 32 Parity Bits," *IEEE Trans. on Commun.*, vol. 41, no. 6, pp. 883-892, 1993.

[17] T. T. Cai and J. Tiefeng, "Phase Transition in Limiting Distributions of Coherence of High-dimensional Random Matrices," *Journal of Multivariate Analysis,* vol. 107, pp. 24-39, 2012.

[18] S. Kwon, J. Wang and B. Shim, "Multipath Matching Pursuit," *IEEE Trans. Inform. Theory*, vol. 60, no. 5, pp. 2986-3001, May 2014.

[19] A. J. Viterbi, and K. O. Omura. *Principles of digital communication and coding,* Courier Corporation, 2013.

[20] J. A. Tropp and A. C. Gilbert, "Signal Recovery from Random Measurements via Orthogonal Matching Pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655-4666, 2007.

[21] S. D. Howard, A. R. Calderbank, and S. J. Searle,"A Fast Reconstruction Algorithm for Deterministic Compressive Sensing using Second Order Reed-Muller Codes," *Proc. of IEEE Conf. on Information Sciences and Systems*, pp. 11-15, 2008.

[22] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, 2005.

[23] H. Vangala, Y. Hong, and E. Viterbo, "Efficient Algorithms for Systematic Polar Encoding," *IEEE Commun. Lett.*, vol. 20, no. 1, pp. 17-20, Jan. 2016.

[24] K. Venkatarama, *Probability and Random Processes*, John Wiley & Sons., 2006.

[25] R. Xie, H. Yin, X. Chen, and Z. Wang, "Many Access for Small Packets Based on Precoding and Sparsity-Aware Recovery," *IEEE Trans. Commun.*, vol. 64. no. 11, pp. 4680-4694, 2016.

[26] H. Ji, Y. Kim, J. Lee, O. Eko, Y. Nam, Z. Zhang, B. Lee, and B. Shim, “Overview of Full-Dimension MIMO in LTE-Advanced Pro," *IEEE Commun. Mag.*, vol.55, no.2, Feb. 2017, pp. 176-184.

[27] S. Han, C. I, Z. Xu, and C. Rowell, “Large-scale Antenna systems with Hybrid Analog and Digital Beamforming for Millimeter Wave 5G," *IEEE Commun. Mag.*, vol.53, no.1, Jan. 2015, pp. 186-193.

[28] 3GPP Technical Report 38.802, “Study on New Radio Access Technology Physical Layer Aspects (Release 14),” v14.1.0, 2017.

[29] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, and A. Dekorsy, “Massive Machine-type Communications in 5G: Physical and MAC-layer Solutions,” *IEEE Commun. Mag.*, vol.54, no.9, Sept. 2016, pp. 59-65.

[30] P. Schulz, M. Matthé, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Müller, T. Elste, and M. Windisch, “Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture,” *IEEE Commun. Mag.*, vol.55, no.2, Feb. 2017, pp. 70-78.

[31] ITU-T Watch Report, “Tactile Internet,” Aug. 2014.

[32] N. A. Johansson, Y. Wang, E. Eriksson, and M. Hessler, “Radio Access for Ultra-Reliable and Low-Latency 5G Communications," *Proc. IEEE Int. Conf. on Comm. (ICC) Workshop*, June 2015, pp. 1185-1189.

[33] M. Sybis, K. Wesolowski, K. Jayasinghe, V. Venkatasubramanian, and V. Vukadinovi, “Channel Coding for Ultra-Reliable Low-Latency Communication in 5G Systems,” *Proc. IEEE Vehicular Tech. Conf. (VTC)*, Sept. 2016, pp.1-5.

[34] 3GPP Technical Specifications 38.211, “Technical Specification Group Radio Access Network, NR (Release 15)," v15.0.0, 2017.

[35] H. Ji, S. Park, and B. Shim, “Sparse Vector Coding of Control Channel for Ultra-Reliable and Low Latency Communications," *submitted to IEEE Trans. Wireless Commun.*, 2017.

[36] J. W. Choi, B. Shim, Y. Ding, B. Rao, and D. I. Kim, “Compressed Sensing for Wireless Communications: Useful Tips and Tricks," *IEEE Commun. Survey and Tutorials,*, vol.19, Feb. 2017, pp. 1527-1550.

[37] T. J. Richardson and T. L. Urbanke, “Efficient Encoding of Low-Density Parity-Check Codes,” *IEEE Trans. Inform. Theory,* vol. 47, no.2, Feb. 2001.

[38] S. Park, H. Seo, H. Ji, and B. Shim, “Joint Active User Detection and Channel Estimation for Massive Machine-Type Communications,” *Proc. of SPAWC*, Feb. 2017.

[39] Y. H. Polyanskiy, P. Vincent, and V. Sergio, “Channel Coding Rate in the Finite Blocklength Regime,” *IEEE Trans. Inform. Theory,* vol. 56, no. 5, pp. 2307-2359, 2010.

[40] R. Hoshyar, F. P. Wathan and R. Tafazolli, “Novel Low-Density Signature for Synchronous CDMA Systems Over AWGN Channel,” *IEEE Trans. Sig. Proc.,* vol. 56, no. 4, pp. 1616-1626, 2008.

[41] S. Zhang, X. Xu, L. Lu, Y. Wu, G. He, and Y. Chen, “Sparse code multiple access: An energy efficient uplink approach for 5G wireless systems,” *Proc. of IEEE Globecom*, 2014.

[42] H. Nikopour and H. Baligh, “Sparse code multiple access,” *Proc. of IEEE Personal Indoor and Mobile Radio Communications (PIMRC)* pp. 332-336.

[43] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, “Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends,” *IEEE Communications Magazine,* vol. 53, no.9, pp. 74-81.

[44] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, K. Niu, "Pattern Division Multiple Access (PDMA) - A Novel Non-orthogonal Multiple Access for 5G Radio Networks," *IEEE Trans. on Veh. Tech.,* vol.PP, no.99, pp.1-1, Jul. 2016.

[45] S. Park, H. Seo, H. Ji, and B. Shim, "Joint Active User Detection and Channel Estimation for Massive Machine-Type Communications," *Proc. of IEEE SPAWC* 2017.

[46] Rec. ITU-R M.2083-0, "IMT Vision - Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," Sep, 2015.

[47] 3GPP Technical Report 38.802, "Study on New Radio Access Technology Physical Layer Aspects (Release 14)," v14.1.0, 2017.

[48] 3GPP Technical Report (TR) 38.913, "Study on Scenarios and Requirements for Next Generation Access Technologies (Release 14)," v14.2.0, 2017.

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

[50] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra Reliable and Low Latency Communications in 5G: Physical Layer Aspects," *IEEE Wireless Commun.*, vol.25, no.3, pp.124-130, Jun 2018.

[51] B. Lee, S. Park, D. Love, H. Ji, and B. Shim, "Packet structure and receiver design for low latency wireless communications with ultra short packets," *IEEE Trans. on Commun.,* vol. 66, no. 2, pp.796-807, Feb. 2018.

[52] M. Mousaei and B. Smida, "Optimizing Pilot Overhead fro Ultra-reliable Short-packet Transmission," *Proc. on IEEE ICC*, May 2017.

[53] H. Ji, S. Park, and B. Shim, "Sparse Vector Coding for Ultra Reliable and Low Latency Communications," *to appear in IEEE Trans. on Wireless Commun.*

[54] H. Ji and B. Shim, "Sparse Vector Coding for 5G Ultra Reliable and Low Latency Communications," in *Proc. of ICC*, May 2018.

[55] H. Ji, S. Kim, and B. Shim, "Sparse Vector Coding for Ultra Short Packet Transmission,", in *Proc. of ITA Workshop*, Feb. 2018.

[56] H. Ji, W. Kim, and B. Shim, "Sparse Vector Coding for Short Packet Transmission in Massive Machine Type Communications ," *Proc. of APCC*, Oct 2018.

[57] J. W. Choi, B. Shim, Y. Ding, B. Rao, and D. I. Kim, "Compressed Sensing for Wireless Communications: Useful Tips and Tricks.," *IEEE Commun. Survey and Tutorials,*, vol.19, pp.1527-1550, 2017.

[58] Z. We, C. Gong, and D. Liu, "Computational Complexity Analysis of FEC Decoding on SDR Platforms," *Journal of Sign. Process Syst.*, vol.89, 2017.

[59] D. L. Donoho, "Compressed Sensing," *IEEE Trans. on Inform. Theory*, vol.52, no.4, pp.1289-1306, 2006.

[60] J. Wang, S. Kwon, and B. Shim, "Generalized orthogonal matching pursuit," *IEEE Trans. on Sig. Proc.,* vol. 60, no. 12, pp. 6202-6216, Dec. 2012.

[61] S. Kwon, J. Wang and B. Shim, "Multipath matching pursuit," *IEEE Trans.s on Inform. Theory,* vol. 60, no. 5, pp. 2986-3001, May 2014.

[62] S. Chang, B. Shim, and J. Choi, "Downlink pilot reduction for massive MIMO systems via compressed sensing," *IEEE Commun. Let.*, vol. 19, no. 11, pp. 1889-1892, Nov. 2015.

[63] S. Sesia, M. Baker, and I. Toufik, *LTE-the UMTS Long Term Evolution: from Theory to Practice,* John Wiley & Sons., 2012.

[64] P. Schulz, M. Matthé, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Müller, T.

Elste, and M. Windisch, “Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture,” *IEEE Commun. Mag.*, vol.55, no.2, pp.70-78, 2017.

[65] O. Yilmaz, Y. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, and Joachim Sachs, “Analysis Ultra-Reliable and Low-Latency 5G Communication for a Factory Automation Use Case," *In Proc. IEEE Int. Conf. on Comm. (ICC) Workshop*, pp.1190 - 1195, 2015.

[66] H. Ji, W. Kim, and B. Shim, “Pilot-less Sparse Vector Coding for Short Packet Transmission,” *submitted to IEEE Wire. Let.*.

[67] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, 2005.

[68] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, M. Fallgren, “Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Commun. Mag.*, vol.52, no.5, pp.26 - 35, 2014.

[69] Recommendation ITU-R M.2083, “Framework and overall objectives of the future development of IMT for 2020 and beyond", Sep. 2015.

[70] J. W. Choi, B. Shim, and S. Chang, “Downlink Pilot Reduction for Massive MIMO Systems via Compressed Sensing," *IEEE Commun. Letters*, vol. 19, no. 11, pp.1889 - 1892, 2015.

[71] Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli, “Estimation of sparse MIMO channels with common support," *IEEE Trans. Commun.*, vol. 60, no. 12, pp. 3705-3716, Dec. 2012.

[72] J. W. Choi, B. Shim, Y. Ding, B. Rao, and D. I. Kim, “Compressed Sensing for Wireless Communications: Useful Tips and Tricks.," *to appear in IEEE Commun. Surv. Tutor.*

[73] M. Ozdemir and H. Arslan, “Channel estimation for wireless OFDM systems," *IEEE Commun. Surv. Tutor.*, vol.9 no. 2, pp. 18-48, 2007.

[74] 3GPP Technical Report (TR) 45.820, “Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT)".

[75] S. Dmitriy, “Clustering wireless channel impulse responses in angular-delay domain," *IEEE 5th Workshop on Signal Processing Advances in Wireless Commun.*, pp. 253-257, 2004.

[76] F. Antonio, D. J. Love, and R. W. Heath, “Simplified spatial correlation models for clustered MIMO channels with different array configurations," *IEEE Trans. Veh. Technol.*, vol. 56, no.4, pp.1924-1934, 2007.

[77] Z. Gao, L. Dai, B. Shim, and Z. Wang, “Structured compressive sensing-based spatio-temporal joint channel estimation for FDD massive MIMO," *IEEE Trans. on Commun.*, vol. 64, no. 2, pp. 601-617, 2016.

[78] H. Ji, Y. Kim, J. Lee, O. Eko, Y. Nam, Z. Zhang, B. Lee, and B. Shim, “Overview of Full-Dimension MIMO in LTE-Advanced Pro," *IEEE Commun. Mag.*, vol.55, no. 2, pp. 176-184, 2017.

[79] S. Han, C. I, Z. Xu, and C. Rowell, “Large-scale antennna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol.53, no.1, pp. 186-193, 2015.

[80] A. Adhikary, J. Nam, J. Y.Ahn, and G. Caire, “Joint spatial division and multiplexing the large-scale array regime," *IEEE Trans.Inf. Theory,* vol. 59, no.10, pp.6441-6463, 2013.

[81] H. Ji, J. Choi, and B. Shim, “Channel Sparsification Beamforming for Internet-of-Things Systems,” *Proc. of IEEE ICC*, 2017.

[82] T. L. Marzetta, “Non cooperative cellular wireless with unlimited numbers of base station antennas", *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp.3590 -3600, 2010

[83] Y. Kim, H. Ji, J. Lee, Y.H. Nam, B.L. Ng, I. Tzanidis, Y. Li and J. Zhang, “Full Dimension MIMO (FD-MIMO): The Next Evolution of MIMO in LTE Systems," *Wireless Commun. Mag.*, vol. 21, issue 3, 2014

[84] Y. H. Nam, B. L. Ng, Y. Sayana, Y. Li, J. Zhang, Y. Kim and J. Lee, “Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Commun. Mag.*, vol. 51, issue 6, 2014

[85] W. Zhang, J. Xiang, Y.R. Li, Y. Wang, Y. Chen, P. Geng and Z. Lu, “Field Trial and Future Enhancements for TDD Massive MIMO Networks," in *Proc. on 26th Intl. Symp. on Personal, Indoor, and Mobile Radio Comm. (PIMRC) Workshop Advancements in Massive MIMO*, 2015, pp. 1114-1118

[86] 3GPP Technical Reports TR36.873, “Study on 3D channel model for LTE".

[87] C. Lim, T. Yoo, B. Clerckx, , B. Lee and B. Shim, “Recent trends in MU-MIMO,", *IEEE Commun. Mag.*, vol. 51, issue 3, 2014.

[88] N. Jindal, “MIMO broadcast channels with finite-rate feedback." *IEEE Trans. on Inform. Theory*, vol. 52, issue 11, 2006.

[89] 3GPP Technical Reports TR36.897, “Study on Elevation Beamforming/Full-Dimension (FD) MIMO for LTE".

[90] 3GPP Technical Reports TR36.847, “E-UTRA and UTRA; Radio Frequency (RF) requirement background for Active Antenna System (AAS) Base Station (BS)".

[91] D. Tse, and P. Viswanath, Wireless Communication. Cambridge University Press, 2005.

[92] B. Lee, J. Choi, J. Seol, D. Love, and B. Shim, “Antenna grouping based feedback compression for FDD-based massive MIMO systems", *IEEE Trans. on Commun.*, vol. 63, no. 9, pp. 3261-3274, Sept. 2015.

[93] J. Choi, B. Shim, Y. Ding, B. Rao, and D. Kim, “Compressive sensing for wireless communications: useful tips and tricks", submitted to *IEEE Commun. Survey and Tutorials.*

[94] J. Wang, “Support recovery with orthogonal matching pursuit in the presence of noise," *IEEE Trans. Signal Process.*, vol. 63, no. 21, 5868-5877, Nov. 2015.

초 록

5세대 무선통신 시스템의 새로운 기술 혁신은 무인 차량 및 항공기, 스마트 도시 및 공장, 원격 의료 진단 및 수술, 인공 지능 기반 맞춤형 지원과 같은 전례 없는 서비스 및 응용프로그램으로 부상하고 있다. 이러한 새로운 애플리케이션 및 서비스와 관련된 통신 방식은 대기 시간, 에너지 효율성, 신뢰성, 유연성 및 연결 밀도 측면에서 기존 통신과 매우 다르다. 현재의 무선 액세스 방식을 비롯한 종래의 접근법은 이러한 요구 사항을 만족할 수 없기 때문에 최근에 sparse processing과 같은 새로운 접근 방법이 연구되고 있다. 이 새로운 접근 방법은 표본 추출, 감지, 압축, 평가 및 탐지를 포함한 기존의 정보 처리에 대한 효율적인 대체기술로 활용되고 있다. 지난 10년 동안 compressed sensing (CS)기법은 의료영상, 기계학습, 탐지, 컴퓨터 과학, 통계 및 기타 여러 분야에서 빠르게 확산되었다. 또한, 신호의 희소성(sparsity)를 이용하는 CS 기법은 다양한 무선 통신이 연구되었다. 주목할만한 예로는 채널 추정, 간섭 제거, 각도 추정, 및 스펙트럼 감지가 있으며 현재까지 연구는 주어진 신호가 가지고 있는 본래의 희소성에 주목하였으나 본 논문에서는 기존의 접근 방법과 달리 인위적으로 설계된 희소성을 이용하여 통신 시스템의 성능을 향상시키는 방법을 제안한다.

우선 본 논문은 다운링크 전송에서 희소 신호 매핑을 통한 데이터 전송 방법을 제안하며 짧은 패킷 (short packet) 전송에 적합한 CS 접근법을 활용하는 기술을 제안한다. 제안하는 기술인 희소벡터코딩 (sparse vector coding, SVC)은 데이터 정보가 인공적인 희소벡터의 nonzero element의 위치에 매핑하여 전송된 데이터 패킷은 희소벡터의 0이 아닌 위치를 식별함으로 원신호 복원이 가능하다. 분석과 시뮬레이션을 통해 제안하는 SVC 기법의 패킷 오류률은 ultra-reliable and low latency communications (URLLC) 서비스를 지원을 위해 사용되는 채널코딩방식보다 우수

한 성능을 보여준다. 또한, 본 논문은 SVC기술을 다음의 세가지 영역으로 확장하였다. 첫째로, 여러 개의 SVC 기반 패킷을 동일한 자원에 겹치게 전송함으로 상향링크에서 대규모 전송을 지원하는 방법을 제안한다. 중첩된 희소벡터를 다중사용자 CS 디코딩 방식을 사용하여 채널 간섭에 강인한 성능을 제공하고 비직교 다중 접속 (NOMA) 방식과 유사한 성능을 제공한다. 둘째로, SVC 기술이 희소 벡터의 support만을 식별한다는 사실을 이용하여 파일럿 전송이 필요없는 pilotless-SVC 전송 방법을 제안한다. 채널 정보가 없는 경우에도 희소 벡터의 support의 크기는 채널의 크기에 비례하기 때문에 pilot없이 복원이 가능하다. 셋째로, 희소벡터의 support의 크기에 추가 정보를 전송함으로 복원 성능을 향상 시키는 enhanced SVC (E-SVC)를 제안한다. 제안된 E-SVC 전송 방식의 핵심 아디디어는 짧은 패킷을 전송되는 정보를 희소 벡터로 변환하고 정보 복원을 보조하는 추가 정보를 희소 벡터의 크기 (magnitude)로 매핑하는 것이다.

마지막으로, SVC 기술을 파일럿 전송에 활용하는 방법을 제안한다. 특히, 채널 추정을 위해 채널 임펄스 응답의 신호를 희소화하는 프리코딩 기법을 제안한다. 파일럿 신호을 프로코딩 없이 전송되는 기존의 방식과 달리, 제안된 기술에서는 파일럿 신호를 빔포밍하여 전송한다. 제안된 기법은 기지국에서 다중 안테나를 활용하여 채널 응답의 0이 아닌 요소를 최소화하는 시간 영역 희소 프리코딩을 적용하였다. 이를 통해 더 적확한 채널 추정을 가능하며 더 적은 파일럿 오버헤드로 채널 추정이 가능하다.

주요어: 희소인지, 압축 센싱, 5세대 통신, 희소 벡터 코딩, 희소 빙포밍
학번: 2015-30204

감사의 글

시작한 것이 얼마 되지 않은 것 같은데, 벌써 졸업을 앞두고 회사로 돌아갈 시간이 되었습니다. 많은 분들의 도움 없이는 절대 이룰 수 없는 시간이었기에 본 글을 통해 감사의 마음을 전합니다.

우선 박사 학위를 지도해주신 심병효 교수님께 감사의 말씀을 전합니다. 기회를 주시고 무한한 에너지로 갈고 닦는 능력은 저의 부족한 부분을 알게 해주셨습니다. 또한 많은 배려로 좋은 성과를 낼 수 있었습니다. 학위 논문 심사과정에서 위원장을 맡아 주시고, 심사해주신 이재홍 교수님께 감사드립니다. 아울러 바쁘신 와중에도 심사위원을 맡아 주시고 소중한 충고를 해 주신 최성현 교수님, 한양대학교 최준원 교수님께도 깊은 감사를 드립니다. 또한, 회사에서 물심양면으로 도와주고 심사해 주신 이주호 마스터에게도 깊은 감사를 전합니다.

대학원 생활 동안 함께한 연구실 친구들에게도 감사드립니다. 특별히, 병주, 선호와 이제는 졸업했지만 함께한 희진, 계용, 그리고 원준, 진홍, 상태, 루옹, 승환, 구영, 준한에게 깊은 감사를 전합니다. 새로운 후배들인 현규, 광진, 용준, 준원도 앞으로 좋은 성과 이어 나가길 바랍니다. 특히, 함께 고생했던 신원재 교수에게도 깊은 감사를 전합니다.

아울러 파견나오기 까지 많은 도움을 주신 최성호 상무, 김윤선 수석, 한진규 수석을 비롯한 삼성전자 표준research팀 여러분께도 깊은 감사를 전합니다. 특히, 오랜 시간 함께한 이효진님께 감사드립니다. 학위 기간 동안 삼성전자에서 제공해준 시간적, 경제적 지원으로 연구에 매진할 수 있었습니다. 사랑하는 가족들에게 깊은 감사의 마음과 사랑을 전합니다. 믿고 기도, 사랑으로 응원해 주신 아버지, 어머님께 깊은 감사를 전합니다. 아들과 같이 사랑으로 지원해 주시고 믿어 주신 장인어른, 장모님 정말 감사합니다. 응원해 주신 형, 형수님과 사랑하는 조카 선우, 선준에게도 감사의 말을 전합니다.

마지막으로 가장 중요한 저의 사랑하는 아내에게 감사 드립니다. 이해와 헌신, 배려 없이는 아무 것도 이룰 수 없었을 것입니다. 아내의 지원이야 말로 가장 큰 에너지이고 힘이었습니다. 아울러 사랑하는 딸 은우와 많은 시간을 함께 할 수 있어서 감사했습니다.

이 기회가 밑거름이 되어 세상에 도움이 되고 나, 우리, 가족, 그리고 세상이 더 행복해 지도록 노력하겠습니다. 고마운 마음 잊지 않겠습니다. 하나님 감사합니다.