An Improved CoSaMP Multiuser Detection for Uplink Grant Free NOMA System

Saifullah Adnan¹, Yuli Fu², Jameel Ahmed Bhutto³, Junejo Naveed Ur Rehman⁴, Raja Asif Wagan⁵, Abbas Ghulam⁶

School of Electronics and Information Engineering, South China University of Technology, Guangzhou, China¹,²
School of Control Science and Engineering, South China University of Technology, Guangzhou, China³,⁶
School of Information Science and Engineering, Xiamen University, Xiamen, China⁴
Department of Information Technology, Baluchistan University of Information Technology, Quetta, Pakistan⁵

Abstract—Non-Orthogonal Multiple Access (NOMA) is the most prominent technology that enhances massive connectivity and spectral efficiency in 5G cellular communication. It provides services to the multi-users in time, frequency, and code domain with significant power level. Message Passing Algorithm (MPA) detection in a multi-user uplink grant-free system requires user activity information at the receiver that makes it impractical. To circumvent this problem, (MPA) is combined with Compressed Sensing (CS) based detection which not only detects the user activity but also the signal data. However, the Compressive Sampling Matching pursuit (CoSaMP) algorithm uses Zero Forcing (ZF) detector to estimate the signal but its performance degrades with increment in Signal to Noise Ratio (SNR). Therefore, Minimum Mean Square Error (MMSE) detector in CoSaMP algorithm is deployed in this paper that enhances detection accuracy and BER performance. The simulation results validate that the proposed algorithm attains better performance than MPA and conventional CoSaMP algorithm in high SNR.

Keywords—MMSE; multi-user detection; CoSaMP; NOMA; MPA; SNR

I. INTRODUCTION

In the era of mobile communication, multiple access technology such as Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), Code Division Multiple Access (CDMA), and Orthogonal Frequency Domain Multiple Access (OFDMA) are the most prominent schemes of the conventional Orthogonal Multiple Access (OMA) which are used to distinguish all generations from 1G to 4G [1]. FDMA assigns the frequency bands, CDMA assigns the channel-code, TDMA and Global System for Mobile Communication (GSM) assign time slots to each user in 1G, 2G and 3G communication networks, respectively. These all schemes are employed to assign orthogonal resources to the users for mitigating inter-user interference (IUI) [2]. However, the number of supportable users in OMA are still restricted by the amount of available resources that makes it unsuitable to meet the explosive connectivity of future generation cellular network. The main goal of future 5G network is to achieve 10 times increase in previous technology with the massive connectivity [3]. In order to achieve these requirements, NOMA has been introduced that serves multiple users on the basis of power in time, frequency and code domain [4].

NOMA is to cluster users according to different channel state information of different users, then assign different power to users in the cluster, and finally transmit the information of all users in a cluster in order to improve the performance of the system. High latency and signal overhead problem is occurred in uplink system due to the fact that transmission is controlled by the BS through a request-grant technique and this problem is drastically enhanced for 5G cellular networks. In order to resolve this problem, grant free transmission is employed in uplink NOMA, where users can communicate frequently with system to perform its activity of transmitting data without providing information to the BS. The decoding is performed on the priority based such as the strongest and powerful signal is treated at first by considering other signals as noise. Furthermore, it excludes the most powerful signal from the received signal and this process repeats till the required signal for the intended user is obtained. The most prominent novelty in NOMA system is that it assigns intelligently the transmit power to random users on the base of the difference in channel-conditions thus multiple users can use entire bandwidth simultaneously by sharing the power difference.

In order to continue to meet future communication needs and further improve system capacity and throughput, NOMA technology allows users to share time and frequency resources through power domain or code domain multiplexing [5]. The existing NOMA system technical solutions can be divided into two categories, namely power domain multiplexing [6-8] and code domain multiplexing, including Low Density Spread Spectrum (LDS) [9, 10], Sparse Code Multiple Access [11] and Multi-user Shared Access (MUSA) [12] and Successive-interference cancellation Aenable Multiple Access (SAMA) [6]. LDS-OFDMA is generic solution and it can be extended to aforementioned NOMA schemes.

NOMA has attracted a lot of attention. In order to avoid signaling overhead and access delay caused by access request, in a NOMA cell, the system assigns a unique address code to each user in the cell in advance. We call the NOMA system that assigns the non-orthogonal address code in advance as grant free NOMA system. However, if the user does not send the access request in advance, it will bring a problem that the BS end does not know which user is accessing the network. In addition, in the actual cellular mobile communication system, the proportion of users who send data at the same time is less than 10%. Therefore, the BS end needs to jointly estimate the user status and the data it sends, which is actually a sparse signal estimation problem. In the grant free access mode, all
users are virtual in which users who do not send data are in the sleep state, and users who need to send data will enter the active state. This granting strategy can significantly reduce the transmission delay and signaling load, simplify the physical layer design, reduce the node power consumption and equipment cost, and has a broad application prospect in large-scale machine communication system.

In the grant free NOMA system the active users have sparse characteristics which meet the requirements of signal sparse in the theory of CS. Some scholars have applied the sparse reconstruction algorithm in the theory of compression sensing to the signal detection [13] of uplink grant free NOMA system. The main principle is that of greedy algorithms to continuously identify support subsets (index sets of non-zero elements) and refine them until the best support set is found. If all the supporting elements are found accurately, then the columns of the observation matrix corresponding to the supporting elements can be taken out to transform the original underdetermined system into overdetermined system, and the original signal can be recovered directly by using MMSE or least square conventional estimation scheme estimator. In [14, 15], the algorithm of orthogonal matching pursuit (OMP) is introduced.

In [16], NOMA is investigated for 5G systems to realize the massive connectivity by controllable interference at the cost of slightly increase in complexity. MPA is proposed for MUD which mitigate the interference among multiple users and reduces the receiver complexity[6]. Nevertheless, the conventional MPA requires user activity at receiver that does not exist in grant-free system. In [17], the authors jointly used CS and MPA to detect both user activity and data. In addition, the user activity is detected by CoSaMP algorithm. But, it detects wrong element in the identification step due to the interference from off diagonal elements in measurement matrix. Consequently, this paper improves the detection accuracy that is effected in CoSAMP by employing MMSE detection in the support set selection step of CoSaMP that overcomes the problem of noise enhancement in ZF and lead to choose the user activity more accurately. The simulation results substantiate that the proposed MMSE based CoSaMP detector attain better performance than conventional detectors.

The rest of paper is organized as follows. The System Model is demonstrated in Section II. Details of the CoSaMP based detector are elaborated in Section III. Section IV explains the proposed MMSE-CoSaMP based CS-MPA model. Simulation results are elaborated in Section V. Section VI presents conclusion.

II. SYSTEM MODEL

The grant-free NOMA is a potential technique for communicating the short packet devices as BS does not require grant procedure. It is assumed that the number of potential users can be very large but only a few number of users are active to transmit the data in one time slot. Therefore, the communication between the users is sporadic in nature. In this system, the BS not only detects the users which transmit the information but also decodes the received data. The uplink NOMA system is considered in this scenario for transmitting the information symbols to the BS and users $u$.

The transmitted symbols $s_u$ are selected from the constellation matrix $\mathbb{C}$. The selected members are transmitted and these symbols are modulated on the spreading sequence $s_u$ of length $L$. After that spreading sequence of all users is superimposed in order to transmit over $s_u$ carriers. More specifically, this paper considers the overloaded scenario which allows the massive connectivity in MTC. The number of non-zero elements in each spreading sequence is much larger than the available media $L$. The received signal mathematically is represented.

$$V_l = \sum_{u=1}^{U} G_{l,u} S_{l,u} x_u + z_l,$$

where $G_{l,u}$ is gain of the channel and $S_{l,u}$ is the sequence length of $l_{th}$ Sub-carrier. Hence, the received signal $v = [v_1, v_2, v_3 \ldots , v_I]^T$ at the station can be represented as

$$v = Hx + z$$

where $x = [x_1, x_2, x_3 \ldots , x_I]^T$ is transmitted signal, $H$ is channel matrix, whose elements in $l_{th}$ rows and $u_{th}$ columns equals to the $G_{l,u}$, $S_{l,u}$ while $z = [z_1, z_2, z_3 \ldots , z_I]$ is considered as noise vector succeeding the probability distribution $CN(0, \sigma^2 I_p)$. In LDS-OFDM system, a sparse spreading sequence designed for each user plays an important role in interference cancellation at the receiver. In addition, a unique decoder with the same input vector is developed for the decoding purpose. The selection of non-zero elements position controls the overlapping factor. The good performance of the system can be obtained by covering a large amount of dispersed data. After that non-zero elements from the constellation matrix are opted [18]. The small amount of the data is spread on each sub-carrier, as a result, the superimposed signal will be much smaller than the active users. Hence, equation(1) can be written as

$$V_l = \sum_{u \in L(l)} G_{l,u} S_{l,u} x_u + z_l = \sum_{u \in L(l)} h_{l,u} x_u + z_l,$$

where $L(l)$ is the set of non-zero active users, it can be demonstrated as $L(l) = [u | S_{l,u} \neq 0]$.

Message passing algorithm requires user activity information at the base station. Therefore, CS provide the service of user activity to MPA that makes output of compressed sensing algorithm to become input to the MPA. This algorithm is represented by the factor graph and it is known as Low Density Parity Check (LDPC) codes and this name have been assigned to it because each parity check is only connected to a small number of code word bits [19, 20]. In Fig. 1, the codeword bit is the variable nodes and parity check bit are factor nodes. The transmitted symbols by all
users are variable nodes, the data only can be transmitted, if and only if parity check bits are non-zero when $S_{lu} \neq 0$. The marginal distribution is the product of the received message from the nodes, the $S_{th}$ iteration can be represented as [17].

$$m_{i \rightarrow u}^{(t)}(x_i) \propto \sum_{i \in L(t) \cup u} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \parallel v_i - h_{i,u} x_i \parallel \right\} \prod_{i \in L(t) \cup u} m_{i \rightarrow u}^{(t-1)}(v_i)$$

(4)

$$m_{i \rightarrow u}^{(t)}(v_u) \propto \prod_{i \in L(u) \cup i} m_{i \rightarrow u}^{(t-1)}(x_i)$$

(5)

where $m_{i \rightarrow u}^{T}(x_u)$ represents the communication signal from the factor node to a variable node in the same way from variable node to factor node. After $ST$ iterations, the marginal probability distribution of $v_u$ can be calculated by,

$$(x_u) \propto \prod_{i \in L(u)} m_{i \rightarrow u}^{T}(x_u)$$

(6)

All the estimated symbols with maximum probability distribution are chosen from the constellation matrix. The MPA computational complexity grows exponentially with the maximum number of symbols in spreading sequences on the sub-carriers $\omega$ instead of the total number of transmitted symbols $u$.

Active user’s information at the BS make MPA multiuser detection impractical in the scenario of uplink LDS-OFDM system where users transmit the data randomly without BS scheduling[17]. Therefore, CS has been combined with MPA to resolve this issue and to obtain the user activity due to the sparse nature of the active and non-active users. In CS-MPA, user activity is computed by CS before applying the message passing algorithm. The CoSaMP in CS cannot detect the user activity accurately, due to the interference from the offdiagonal elements in the support set computation. Therefore, this paper enhances the detection performance by replacing the ZF detection step of CoSAMP with MMSE. Further, details of the proposed scheme are elaborated in the Section III.

III. CoSAMP BASED DETECTOR

The Sampling matrix is firmly embedded with the set of sparse signals and sparse signal recovery methods are utilized to recover spare signals from the samples. Sparse signal recovery algorithms are divided into three categories. 1) Greedy Algorithms, 2) Convex relaxation Algorithms, 3) Bayesian Inference algorithm. Greedy algorithms are fast and easy to implement and are particularly used for support recovery. The CoSaMP is the most prominent greedy algorithm that not only converges fast but also has good detection accuracy (Fig. 2).

Properties of CoSaMP: The implementation method determines how the algorithm achieves the required performance.

1) The algorithm requires a satisfying constant in the form of a matrix. Many matrices have this property. As a result, the algorithm can be used with many measuring technologies.

2) Binding the matrix depends on the sampling of the matrix. The algorithm can recover the signal density from a small number of samples.

3) The associated error shows that the algorithm is successful for all signal samples, even the signal is surrounded by noise. As expected, CoSaMP is the best for CS and sparsity.

4) The computational cost of the algorithm depends on the approximate signal that we apply.

![Fig. 1. Factor Graph Representation of MMSE based Receiver.](image1.png)

![Fig. 2. Block Diagram of CoSaMP.](image2.png)
The algorithm is initialized using a trivial estimation of the signal which means that the initial residual value is the entire unknown target signal. Each iteration consists of five main steps.

1) **Identification.** Using the available samples at the current stage, the correlated vector is computed by the algorithm first which is known as a signal proxy. The signals with strong power are estimated.

2) **Support merger:** The number of detected components from the first step are embedded in the set of currently approximate components.

3) **Estimation:** The least-squares’ problem for approximating the estimated signal on the integrated set of components will be resolved using an algorithm.

4) **Pruning:** In the estimated least-square signal, the algorithm retains only the large or high power entries in new approximation.

5) **Residual:** At the final step samples are updated every time, so that the remaining part reproduce and updated which has not been approximated yet.

**IV. PROPOSED ALGORITHM**

Keeping deep sight on the demand of the upcoming generation in the sense of spectral efficiency and massive connectivity, the proposed MMSE-CoSaMP based system mitigates the interference from off-diagonal elements using MMSE detector. In [21], the compressed sampling matching pursuit (CoSaMP) algorithm is used. In this algorithm, multiple elements are selected in each iteration. In addition to the element selection criteria, the atoms selected in each iteration of CoSaMP may be discarded in the next iteration, because there is no problem of active user error stacking. The user activity is estimated with the compressive sensing algorithm, while the data is detected through the message passing algorithm MP. This system is much better than the conventional CoSaMP algorithm as it reduces the noise enhancement issue that occurred in the ZF step of CoSaMP algorithm. The LDS-OFDM uplink grant-free system has been considered in this work which consists of $L$ Sub-carriers for the channel gain keeping the bandwidth constant. However, it is assigned to the users differently. i.e., for any $l = 1, 2, ... L$, we have gain $G_{l,u} = G_u$, where, $u = 1, 2, ... n$ and reference signal $x = x_1, x_2, x_3, ... x_U$ for users of observation matrix $x$ is transmitted first. $x = [x_1, x_2, x_3, ... x_U]$ $L \times U$, the received communication signal $x_t$ for the activity detection can be demonstrated as

$$r_t = \sum_{u=1}^{U} x_u G_u I_u + z_t = xG + z_t$$

where $G=[G_{1,1}, G_{1,2}, ..., G_{1,U}, I_u]$ for $u = 1, 2, 3, ... U$. Logical variable is used to demonstrate weather the user is active or not. $I_u = 1$ if user $u$ is active, $I_u = 0$ when user is inactive, and $z_t \sim \text{CN}(0, \sigma^2 I)$.

Due to the scarcity of user activity $G$, estimation of a sparse signal from a minimal number of samples can be considered as the issue of the signal recovery in CS, in this system, only those models are considered in which the noise factor is involved. The constellation matrix can be designed to follow the RIP with strong probability [22]. Hence, the CS can be used for sparse recovery from which the user activity can be detected by demonstrating the position of non-zero elements. We used MMSE based compressive sampling matching pursuit CoSaMP algorithm in the place of the ZF based CoSaMP, because ZF has noise enhancement. Because of its low complexity and robustness in high SNR, we choose MMSE [23-25] in the estimation part of algorithm.

Algorithm 1 Modified MMSE-CoSaMP Based MUD Detection algorithm

1. **Input:** Sparsity $U$ and constellation matrix $\Phi$, signal matrix $R$ and noise variance $\sigma^2$.
2. **Initialize:** $x=1$, $r_2 = H_2 x_{active} + z_2$, $i = 1$, and supported set $\Omega = \emptyset$, Iterations $T$
3. **Iterations:** run the following from 1 to 5 into loop $n$ cycle ($1 \leq n \leq m$)
4. $\hat{T}_{n-1} = \arg \max (\Phi^\dagger r_{n-1})$ %Compute support set
5. $\hat{T}_n = T_{n-1} \cup \hat{T}_{n-1}$ %Merging support set
6. $T_n = (H^H H + \sigma^2 I_n)^{-1} H^H$ %Filter support set
7. $R = R - (H^H H + \sigma^2 I_n)^{-1} H^H$ %residual update
8. if $|| r_n ||_2 \geq || r_{n-1} ||_2$ then
9. $T_n = T_{n-1}$
10. **end if**
11. location= find ($\hat{G} \neq 0$) and $r_{active} = r(location)$
12. **Data Detection**
13. **Initialize:** $x = 1$, $m_{i \rightarrow u}^{(0)}(r_u) = 1$ and $m_{i \rightarrow u}^{(0)}(x_u) = 0 \forall u, l$
14. **while** $x \leq T_2$ **do**
15. simulating equation……. (6)
16. $m_{i \rightarrow u}^{(t)}(r_u) \propto \prod_{i \in L(u)} m_{i \rightarrow u}^{(t)}(x_i)$
17. **end while**
18. $(x_u) \propto \prod_{i \in L(u)} m_{i \rightarrow u}^{T}(x_u)$ %
19. **Output** $x_u$

The compressive sensing algorithm CoSaMP has capacity to detect the position of the non-zero sparse signal elements accurately as we as it can detect the activity of the user at the same time, another advantage is that, the position of random active users can be detected with higher accuracy, it has better performance as mentioned in [20][21]. The received modulated signal can be determined as bellow.

$$r_2 = H_2 x_{active} + z_2$$ (6)
where $x_{active}$ is the active users, $H_2$ is the same as $H$ the difference is that it contains the information of active users only, which can be obtained from time or frequency domain pilots [26, 27]. The MPA is used for data detection whereas CoSaMP is used to detect the location of active uses. The CS-MPA combines the advantages of both MPA and CS. The proposed algorithm modify CS algorithm by using MMSE detector in CoSaMP in the place of ZF which reduces the interference and copes up the noise enhancement problem.

V. SIMULATIONS AND RESULTS

This section demonstrates the performance comparison of proposed algorithm with the conventional detectors in terms of Bit Error Rate (BER) versus SNR. Parameters used in this paper are provided in the Table I. The length of given signal $L = 40$ and the number of users is $U = 50$ clarifies that the overloading of the system is too high. The Non-zero elements (sparsity) are active users which are selected from the constellation matrix.

Fig. 4 illustrates the BER performance of conventional MPA and CoSaMP with MMSE-CoSaMP under the scenario mentioned in the table. It can be seen from this diagram that the proposed algorithm has achieved better performance. The user activity and the data both are detected with more accuracy. Fig. 3 illustrates the effect of the number of active users and BER performance of both proposed MMSE-CoSaMP based and conventional CoSaMP based systems.

<table>
<thead>
<tr>
<th>PARAMETERS</th>
<th>VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modulation</td>
<td>QPSK</td>
</tr>
<tr>
<td>Orthogonal resources</td>
<td>40</td>
</tr>
<tr>
<td>Users</td>
<td>80</td>
</tr>
<tr>
<td>Active users</td>
<td>4, 6, 8, 10, 12</td>
</tr>
<tr>
<td>SNR</td>
<td>0 to 25</td>
</tr>
<tr>
<td>Channel</td>
<td>Raleigh fading channel</td>
</tr>
<tr>
<td>Noise</td>
<td>AWGN</td>
</tr>
</tbody>
</table>

After the analysis of both algorithms, it is concluded that increasing the number of active users degrade the detection of signal accuracy, but when the results of both proposed and conventional systems are compared it is observed that the proposed system has better performance than the conventional one. Our main focus is to meet the demand of the upcoming generation. The proposed system has better performance so it can be considered as the step towards the betterment of the system. From the Fig. 5, it can be observed that in a conventional system when SNR is 5dB then BER can be noticed on the scale that it lies on $10^{-3}$ while in the proposed system it can be seen that it is near to $10^{-4}$. Simulation results elaborate that proposed MMSE-CoSaMP has much better performance than the conventional CoSaMP and the proposed algorithm is the most suitable candidate for upcoming 5G communication because it attains a BER=$10^{-3}$ for active users.
VI. CONCLUSION

MUD is the most crucial part of uplink grant-free NOMA systems for 5G cellular communication. This paper addresses the problem of MUD at the BS for estimating user activity and data. The massive connectivity in NOMA systems exploits the structured scarcity of user activity due to which CS is combined with MAP that jointly detects user activity data. But, CoSaMP based detection wrongly chooses the signal activity due to the interference from the off-diagonal elements. The proposed algorithm attains better output results than the conventional CoSaMP an algorithm that independently calculates support for each signal and simultaneously updates the same support for many rare signals. Therefore, the proposed algorithm can improve the performance of signal detection in NOMA systems with acceptable complexity and it accomplishes the gain of 5dB than that of the conventional algorithm. This can decrease significantly the latency and remove the signal interference to better extent in 5G wireless communications. This can be deployed in other NOMA schemes.

REFERENCES


