Scalable Compressive Sensing-Based Multi-User Detection Scheme for Internet-of-Things Applications

Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan
[tyliu, woody, ciao, andyw]@access.ee.ntu.edu.tw

Abstract—Rapid growth in the number of users (largely sensor nodes) of most Internet-of-Things (IoT) applications is widely expected in recent years. 3G or 4G protocols are designed for regular Human-to-Human communication purposes instead of low-rate, Machine-to-Machine (M2M) communication, which is the basis of IoT. Comparatively, traditional spread-spectrum protocols such as Code-Division Multiple Access (CDMA) are considered applicable for the majority of IoT applications. However, they are unsuitable to be utilized in the expected massive-IoT scenarios. The traditional Multi-User Detection (MUD) approach used in these protocols has poor scalability. When the number of users reaches hundreds, the complexity and hardware cost of traditional MUD become impractically high. In this paper, we employ the concept of Multiple Measurement Vector Compressive Sensing (MMV-CS) to exploit the feature of user sparsity in spread-spectrum-based IoT applications. By reformulating the CS detection model, a memory reduction of approximately 9,400 times can be achieved. Along with other hardware cost-down, the proposed simplified structure of the reconstruction model also allows a faster detection speed (~10x to 1000x improvement) while a satisfactory level of Bit-Error-Rate (BER) is still maintained.

Keywords—Internet-of-Things; multi-user detection; user activity sparsity; compressive sensing; multiple measurement vector

I. INTRODUCTION

Compared to existing mobile electronic devices, devices that constitute the vision of Internet-of-Things (IoT) can and will be an even larger industry, reaching the scale of trillions or beyond [1]. These devices, however, are generally simple, cheap and application-specific, represented by various sensors with low-rate wireless transmission modules and little data processing capability [2]. Given the prospect of IoT topology depicted in Fig. 1 where numerous such devices rely on a single gateway or aggregated node for further data processing and/or analysis, it is prominent to improve the scalability of the Multi-User Detection (MUD) scheme with a reasonable hardware cost.

The feature of user activity sparsity commonly seen in many Machine-to-Machine (M2M) communication-based multi-user IoT applications is the key to improving the scalability of MUD. Due to the nature of certain IoT applications (such as in geological monitoring), only a small percentage of sensor nodes would be actively transmitting data at a time. As the traditional matched filter-based MUD approach needs to allocate a unique bipolar spread-spectrum code to each registered user [3], it is unsuitable for such applications. Since the allocated codes have to be highly mutually-uncorrelated, the traditional MUD method would cause impractical code length when users scale up, leading to high hardware usage and computational complexity.

[4] introduces the idea of Single Measurement Vector Compressive Sensing (SMV-CS) which exploits user activity sparsity. Hence the IoT MUD problem can be readily coped with a much smaller number of shorter (thus cheaper) spreading codes. From another perspective, when resource (number of highly-uncorrelated codes and the corresponding hardware) is limited, SMV-CS-MUD is able to successfully detect active transmissions. Under the same circumstances, matched filter-based MUD would obviously fail [5]. However, when the number of users further soars, the SMV formation will lead to poor sampling matrix property and huge matrix size in the CS model. Both the reconstruction speed and hardware cost will thus suffer from user scaling, let alone the detection performance.

In this paper, we propose a Multiple Measurement Vector Compressive Sensing Multi-User Detection (MMV-CS-MUD) scheme to cope with the user scalability problem in massive-IoT scenarios. The following results have been accomplished:

- The sampling matrix size used in CS-MUD is shrunk by nearly \( L^2 \) (\( L \) = frame length ) times, which under our experimental setting, is approximately 9400 times.
- Compared to the SMV-CS-MUD scheme, we improved the detection speed by over 1000 times when the number of users reaches hundreds.
- Near-optimal detection accuracy with full Inter-Symbol Interference and realistic observation signals is achieved.

The remainder of this paper is organized as follows: Section II will introduce the transmission model IoT MUD problems face and the SMV-CS-MUD approach, along with the theoretical CS problem it relates to. Section III will be the proposal and detailed description of the MMV-CS-MUD scheme. In Section IV the experimental results will be presented and analyzed, followed by the conclusion in Section V.

Fig. 1. The typical massive-IoT multi-user detection scenario, where the number of users is high and the user activity probability is significantly low. --

---

This work was supported by the Ministry of Science and Technology under Grant MOST 103-2218-E-002-033.
II. BACKGROUND

A. The Multi-User Transmission Model

The mathematical model of the user transmission process is as illustrated in Fig. 2.

First, a small number of active users from a total of \( K \) users are set to respectively transmit a frame of convolutionally-encoded binary bits, optionally interleaved, then mapped following the Binary Phase-Shift Keying (BPSK) constellation. Inactive users are set to transmit all-zero frames, thus forming an augmented BPSK constellation \( \{0, \pm1\} \). All frames to be transmitted are then spread using bipolar-valued Pseudo-Noise (PN) spreading sequences before they go through their respective wireless channels (modeled as i.i.d. Rayleigh distributed channels with an exponentially-decaying power delay profile where the expected delay is known). The superposition of these signals through channels is considered to be the observation \( y \) the aggregated node or gateway receives:

\[
y = \sum_{k=1}^{K} H_b S_b x_k + n.
\]  

where \( x_k \) is the encoded and modulated frame of the \( k \)-th user. The frame length is denoted by \( L \).

![Fig. 2. Transmission process of signals till receiving.](image)

The corresponding spreading matrix \( S_b \) and the channel matrix \( H_b \) take the following forms:

\[
S_b = \begin{bmatrix}
    s_{k,1} & 0 & 0 & 0 \\
    s_{k,2} & 0 & 0 & 0 \\
    \vdots & \ddots & \ddots & \ddots \\
    0 & s_{k,L_k} & 0 & 0
\end{bmatrix},
\]

\[
H_b = \begin{bmatrix}
    h_{k,1} & 0 & 0 & 0 \\
    h_{k,2} & 0 & 0 & 0 \\
    \vdots & \ddots & \ddots & \ddots \\
    0 & h_{k,L_b} & 0 & 0
\end{bmatrix}.
\]

In the matrices the sequences \( S_b = [s_{k,1}, \ldots, s_{k,N}]^T \) and \( H_b = [h_{k,1}, \ldots, h_{k,L_b}]^T \) respectively denote the spreading sequence allocated to the \( k \)-th user and the frequency selective channel it corresponds to. \( N_b \) and \( L_b \) are the spreading factor and the channel tap number. Note that the channel matrix is in such an indented formation to model the effect of Inter-Symbol Interference (ISI) commonly seen in wireless communication.

In traditional MUD schemes [3], to detect which user is actively transmitting data upon the receipt of signals, the receiver correlates the received superposition of signals with all spread-spectrum codes allocated in a parallel way. Users corresponding to codes that have far-larger-than-zero correlation values are then labeled as active, as illustrated in Fig. 3.

![Fig. 3. Traditional detection process using parallel correlators, or matched filters [3].](image)

When the number of users scales up, the cost of the traditional MUD scheme would quickly turn prohibitive. Since user activity sparsity can usually be expected in massive-IoT applications, it would be ideal if a same set of spread-spectrum codes could be allocated repeatedly. Hence the resource needed to detect large numbers of users will not rise so drastically.


It can generally be assumed in the aforementioned scenario that only a small percentage of users (e.g., user 2 in Fig. 1) would be simultaneously active. This feature of user activity sparsity leaves room for the utilization of compressive sensing, the use of which theory in communication has expanded greatly in recent years. Consider the compression-sampling process below:

\[
y = \Psi \Phi x + n.
\]

\( x_{N \times 1} \) is the original, sparse signal whose entries are largely zeros, \( \Phi \) is the sparsity basis generally determined by physical constraints of the signal (e.g., \( \Phi \) can be wavelets for image signals), and \( \Psi \) is the measurement matrix. Also, \( n \) is the additive noise and \( y_{M \times 1} \) is the observed signal, where \( M < N \) [6]. The usual goal of CS is to solve the problem below:

\[
\min_{x \in \mathbb{R}^N} \|x\|_0 \text{ subject to } y = \Psi \Phi x + n,
\]

where \( \|x\|_0 \) is the \( l_0 \)-norm of signal \( x \). At this point, the multi-user uplink scenario based on the denotations above can be analogically modeled as:

\[
y = Ax + n.
\]

in which \( x_{KL \times 1} = [x_1^T, x_2^T, \ldots, x_K^T]^T \) is the vertically-stacked aggregation of all \( K \) user signals to be transmitted. \( A_{(N,L+L_b-1) \times KL} = [H_1 S_1, H_2 S_2, \ldots, H_K S_K] \) is the horizontally-stacked equivalence of the sampling matrix constituted by \( K \) pairs of spreading matrices and channel matrices.

However, as the CS reconstruction complexity soon becomes prohibitive with the scaling of user number, in [5] the problem denoted by (4) is equally partitioned into \( L/L_{sub} \) sub-problems. The sizes of the signals \( x_{sub} \) to be detected during the iterations are therefore all \( KL_{sub} \times 1 \):

\[
y_{sub} = A_{sub} x_{sub} + n.
\]

The sampling matrix is also partitioned into \( A_{sub} \) to cater for the compromise. This approach uses Group Orthogonal Matching Pursuit (GOMP) [7] for reconstruction of each sub-problem and the activity of recovered frames are determined before they are fed into the sparsity-aware Viterbi Decoder [5].
Such partitioning will cause the absence of ISI effect between sub-problems in transmission model and is therefore unrealistic in the context of real-case communication scenarios. Consequently, an alternative detection scheme that can swiftly recover the transmitted data from the actual received signal is needed. Also, it should ideally maintain the hardware-friendly characteristic of the SMV-CS-MUD approach, which uses the hardware-implementable series of greedy matching pursuit reconstruction algorithms [4]-[5].

III. PROPOSED MULTIPLE MEASUREMENT VECTOR COMPRESSIVE SENSING-BASED MULTI-USER DETECTION (MMV-CS-MUD) SCHEME

The SMV formation needs to sacrifice the accuracy of the transmission model in order to be satisfactory in the sense of computational speed. As the number of users scales up, the comparison of correlation values carried out in the greedy pursuit algorithm will take longer time and cost more hardware, especially memory. Sub-problem sizes in (5) will therefore have to be reduced, causing a greater loss in modeling accuracy.

We propose a CS-MUD scheme based on Multiple Measurement Vector (MMV) formation, instead of the previous SMV formation. Our efforts are entirely concentrated on the receiver end as MUD schemes cannot in any way affect the transmission process introduced in Section II. To better explain our approach, we give the transmission process presented by (4) in its matrix/vector formation, as shown in Fig. 4.

![Fig. 4. Data constitution of received signal y.](image)

\[ a_k = [a_{k,1}, ..., a_{k,M}]^T \] for \( k = 1, ..., K \) is the convolution of \( h_k \) and \( s_k \). In Fig. 4, the first segment of \( y \), \( y_1 = [y_{1,1}, ..., y_{N_S+h-1}]^T \), has most of its entries solely contributed by the first user symbols when \( N_S (\propto K) \) is significantly larger than \( L_h \). This is true in massive-IoT scenarios. The last \( L_h - 1 \) entries of \( y_1 \) have a mixed source consisted of the first symbols and second symbols of the users. For \( y_2 = [y_{N_S+1}, ..., y_{2N_S+L_h-1}]^T \) to \( y_{L-1} \), both the first \( L_h - 1 \) and the last \( L_h - 1 \) entries are mixed in the same way. For \( y_L \), only its first \( L_h - 1 \) entries are mixed. A simple, hardware-saving compensation approach is to apply equation (6) to the segments accordingly as stated above:

\[
y'_1(k) = \begin{cases} 
c_1y_1(k), & 1 \leq k \leq L_h - 1 \\
c_2y_1(k), & N_S + 1 \leq k \leq N_S + L_h - 1, \\
y_1(k), & \text{otherwise}
\end{cases}
\]

obtaining \( y'_1 \) as the \( l \)-th (\( l = 1 \) or \( L \) should be exceptions that only apply partially) segment of the received signal \( y \) after compensation. According to the exponentially-decaying channel power delay profile, \( c_l = 1 / \text{decay}^{L_h - 1} = 1 - c_R \). In the equation above, \( \text{decay} \) is the expected value of the power decay per tap known at the receiver end as CS-MUD schemes assume full knowledge of the channel parameters. In other words, as long as \( N_S > L_h \), the received signal \( y \) can basically be seen as an approximation of vertically-stacked segments. More importantly, these segments can be detected separately for each set of simultaneous symbols from all \( K \) users.

Taking advantage of the observation made above, a new formulation of the detection model can be written as:

\[
\min_{x_{1:R K}} \|x_1\|_0 \text{ subject to } y'_1 = A'x_1 + n.
\]

where columns of \( A' \) are vectors \( a_1, ..., a_K, M = N_S + L_h - 1 \).

The CS reconstruction problems all involve the identical sampling matrix \( A' \) and are essentially results of sequential transmission of symbols along the timeline. We can therefore horizontally stack \( x_1 \) to \( x_L \) on the right side of (7) and \( y'_1 \) to \( y'_L \) on the left side, assuming the channels would not change during the transmission process of a frame. This is reasonable, as the typical frame length does not exceed a few hundred bits and most existing IoT applications are assumed to be low-rate. The approach above leads to an MMV-CS model [8], as illustrated in Fig. 5. In this way, sequentially-received symbols can be detected in one CS problem as frame-by-frame transmission and receipt of signals are expected in most packet-based communication protocols.

![Fig. 5. The Multiple Measurement Vector Compressive Sensing (MMV-CS) model in matrix formation.](image)

The simultaneous sparse approximation problem [9] to be reconstructed in the MMV-CS model above is as follows:

\[
\min_{X_{0:k} \in l} \|X\|_0 \text{ subject to } Y = AX + N,
\]

where \( Y_{M \times L} = [y_1', y_2', ..., y_L'] \) and \( X_{K \times L} = [x_1, x_2, ..., x_L] \).

This problem can be solved using the Simultaneous Orthogonal Matching Pursuit (SOMP) algorithm proposed in [9]. SOMP can be viewed as an MMV-adaptive version of the classic...
Orthogonal Matching Pursuit (OMP) algorithm used in SMV-CS reconstruction problems [10]. The pseudo-code of the SOMP algorithm is given as follows [5], [8]:

Simultaneous Orthogonal Matching Pursuit (SOMP)

\[ A_0 = \emptyset, A_0 = \{1, 2, \ldots, K\}, l = 0, R_0 = Y \]

repeat

\[ l = l + 1 \]

\[ \lambda_i = \arg\max_{i} \sum_{k \in A_{l-1}} \{|A'R_{l-1}x|_k\} \]

\[ A_l = A_{l-1} \cup \{ \lambda_i \}; \Lambda_l = \Lambda_{l-1} - \{ \lambda_i \} \]

\[ x_i = \arg\min_{x \in \Lambda_l} \|Y - A_\lambda x\|_F \] (Frobenius Norm)

\[ R_l = Y - A_{\Lambda_l}x_l \]

until \( l = \max \text{< criteria}_{\text{stop}}, K > \)

Also, as has been stated in Section II, the rows of X should be either all-zero or bipolar-valued in the first place. The SOMP algorithm operates under the condition that the signal to be reconstructed needs to have same support locations for all of its columns [9]. Consequently the block sparsity of the original signal is maintained as suggested in [4].

The signal flow diagram of the MMV-CS-MUD scheme is as illustrated in Fig. 6, where \( Y_0 = \{y_1, y_2, \ldots, y_L\} \) contains segments before the coping or compensation measure for ISI.

![Fig. 6. Signal flow of the proposed MMV-CS-MUD scheme.](image)

After CS reconstruction, the detected result \( X_{\text{est}} \) is demapped according to the augmented BPSK constellation \([0, \pm1]\) using hard decision. Then the same sparsity-aware Viterbi decoder as in SMV-CS-MUD [5] is utilized for frame decoding, whose output is denoted as \( X'_{\text{est}} \).

To summarize, the proposed MMV-CS-MUD scheme uses the original and most accurate transmission model under full ISI impact. By reformulating the CS detection model and reducing the sampling matrix size by a rate of nearly \( L^2 \), the proposed scheme is expected to improve the detection speed and support a significantly larger number of users.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Detection Performance under Varying Noise Interference

The first simulation examines the effect of noise interference, where the probability of user activity, \( p_{\text{active}} \), is set as 0.02 and the value of the Signal-to-Noise Ratio (SNR) of the Additive White Gaussian Noise (AWGN) changes from 0 to 25dB. 50-bit frames from \( K = 128 \) users are first encoded by a standard terminated non-systematic non-recursive [5; 7]s convolutional code with a constraint length of 7. Then the encoded 104-bit frames are spread by bipolar PN sequences with a spreading factor of \( N_2 = K/4 \). The channels are set to be 6-tap i.i.d. Rayleigh distributed with an exponentially-decaying power delay profile. Each value of SNR is tested for 10,000 Monte-Carlo iterations and the computation time is averaged. The result is shown in Fig. 7.

![Fig. 7. Detection performance versus varying signal-to-noise ratio.](image)

In Fig. 7, the detection performance of MMV-CS-MUD is not improved much as SNR rises. This is due to the fact that in the transmission process, the excessively sparse but realistic sampling matrix caused more severe information loss than the more compact but unrealistic sampling matrix in sub-problem partitioning SMV-CS-MUD did. The reduction of noise interference is thus not able to further lower the BER result. Nevertheless, it can be concluded that the detection performance of MMV-CS-MUD has reached a sub-optimal level.

B. Detection Performance under Varying User Density

Next, the effect of user activity sparsity on detection performance is examined and presented in Fig. 8. Each \( p_{\text{active}} \) value from the range \([0.01, 0.02, \ldots, 0.10]\) is set to be verified under \( SNR=20dB \). All other parameters are kept the same as given in the previous experiment.

![Fig. 8. Detection performance versus varying activity probability \( p_{\text{active}} \).](image)
The MUD scheme to be slightly higher than that of SMV-CS-MUD. Still, the performances are considerably close, especially after $p_{active}$ reaches 0.05, indicating the difference becomes less significant for a higher density of actual transmission activities. In other words, a near-optimal detection performance is guaranteed especially when user density rises.

C. Detection Efficiency under Varying User Scale

We proceed to validate the detection efficiency represented by the computational time needed to reconstruct the transmitted signals when the number of users scales up. The value of $K$ is set to rise from 32, 64, 128 to 256. All users have the same activity probability $p_{active} = 0.02$ and the SNR is set as 20dB. Each value of $K$ is tested for 1,000 Monte-Carlo iterations and the computational time is averaged. The simulation result is illustrated in Fig. 9.

![Fig. 9. Comparison of the needed reconstruction time of the proposed MMV-CS-MUD and the sub-problem partitioning SMV-CS-MUD in [5].](image)

As can be seen in Fig. 9, the x-axis stands for the number of total users, while the y-axis is the averaged reconstruction time required for each detection. When the number of users gradually scales up from tens to hundreds, the difference between the detection efficiency of the compared schemes also rises from approximately 10 times to more than 1,000 times. After $K$ reaches 100, the SMV-CS-MUD scheme takes seconds for each detection. Under larger user scale, the proposed scheme is much more efficient in actual IoT MUD systems due to the real-time requirement. Therefore, user scalability is improved.

Both based on OMP-like algorithms, the complexity of the discussed schemes are largely determined by sizes of the sampling matrices. The significantly smaller matrix size in the proposed scheme not only reduced the memory needed, but also helped scaling down the required computation time. When $K = 128$ and $L = 104$, the size of $A'$ in MMV-CS-MUD is $(N_g + L_h - 1) \times K = 37 \times 128 = 4,736$ and the size of $A$ in direct SMV-CS-MUD is $(N_g L + L_h - 1) \times KL = 44,368,896$. The rate of reduction is approximately 9,400.

Combining the results presented above, it is evident to say the proposed MMV-CS-MUD scheme under full ISI impact reached a similar level of detection performance to that of the sub-problem partitioning SMV-CS-MUD scheme with a more inaccurate transmission model. More importantly, by using the MMV-CS formation, the proposed scheme is able to better satisfy the requirement of real-time detection when users increase drastically. In comparison, solving the SMV-CS problem would be unrealistically costly and slow when $K$ is large.

A brief comparison between the proposed detection scheme, the original, direct SMV-CS-MUD scheme and the sub-problem partitioning SMV-CS-MUD scheme is presented in TABLE I.

<table>
<thead>
<tr>
<th>CS-MUD Schemes</th>
<th>Detection Accuracy</th>
<th>Detection Efficiency</th>
<th>Scalability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-size (original)</td>
<td>N/A</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>SMV-CS-MUD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-problem partitioning</td>
<td>Optimal</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>SMV-CS-MUD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed MMV-CS-MUD</td>
<td>Near-Optimal</td>
<td>Good</td>
<td>Good</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS

In this paper, we have proposed an MMV-CS-MUD scheme that is able to improve the scalability of the multi-user detection process in massive-IoT applications. A near-optimal detection performance under full ISI impact is also achieved. Compared to the direct SMV-CS solution, the proposed scheme reduces the required memory size by approximately 9,400 times under the aforementioned setting. The computational speed is also greatly improved, especially when the number of users significantly rises and reaches hundreds.

REFERENCES
