Space-Time Fronthaul Compression of Complex Baseband Uplink LTE Signals

Jinseok Choi†, Brian L. Evans† and Alan Gatherer*
†Wireless Networking and Communications Group, The University of Texas at Austin
Email: jinseokchoi89@utexas.edu, bevans@ece.utexas.edu
*Huawei Technologies, Plano, Texas
Email: alan.gatherer@huawei.com

***Abstract*—In this paper, we propose space-time fronthaul compression of baseband uplink LTE signals for cellular networks, in which baseband units (BBUs) support remote radio heads (RRHs) through fronthaul links. In particular, we assume massive antenna arrays in which the number of antennas in a RRH is much larger than the number of active users. The proposed method consists of two phases: dimensionality reduction phase and individual quantization phase. The key idea of the first phase is to apply principal component analysis (PCA). It performs low-rank approximation of a matrix — composed of received signals — by exploiting the correlation of the received signals across space and time. In the second phase, our method individually quantizes the dimensionality-reduced signal by applying transform coding with bit allocation to reduce the number of quantization bits. An LTE link-level simulator provides numerical results which show that the method achieves up to 8× compression ratio for the uplink with 64 antennas and 4 active users, along with improvement in communication system performance as a result of denoising.**

***Index Terms*—LTE uplink, space-time compression, correlation, PCA, dimension reduction, denoising, transform coding, bit allocation**

I. Introduction

To reduce capital and operating expenses while supporting exponential growth in mobile data traffic, some cellular network deployments separate remote radio heads (RRHs) from baseband units (BBUs). In this deployment, one BBU can support RRHs on multiple basestations. Fronthaul links connect RRHs to BBUs and may be optical fiber, Gigabit ethernet or microwave links. A cloud radio access network [1] is an extreme example of such a network.

Separating RRHs and BBUs over fronthaul links requires substantial transport network resources and corresponding investment for the link structure. Current link standards of common public radio interface [2] and open radio interface [3] have insufficient capacity to support the ever increasing data rate. This stresses the importance of lowering transmit data rate of the fronthaul links.

Time domain compression methods for baseband long term evolution (LTE) uplink signals have been proposed to reduce data rate. The following methods result in error vector magnitude (EVM) of about 2%. A low latency compression algorithm in [4] first resamples an LTE signal to remove spectral redundancies, then converts it to a block floating point representation and finally quantizes the I/Q samples in each block with a non-linear quantizer. This method achieved 3× rate reduction with about 2% EVM. In [5] and [6], similar approaches to [4] were proposed with dithering and modified block scaling respectively. These methods reported a compression gain of 3× with 1.5% EVM and 3.3× with 2% EVM. In [7], a Lloyd-Max scalar quantizer designed for a zero-mean Gaussian distribution combined with noise-shaped feedback coding gave 3× compression ratio for both uplink and downlink with less than 2% EVM. Since scalar quantization is not capable of exploiting time correlations in LTE signals, a multi-stage vector quantization based compression was proposed in [8] to exploit the time correlation in the signals. This method achieved 4× compression ratio for uplink with approximate 2% EVM.

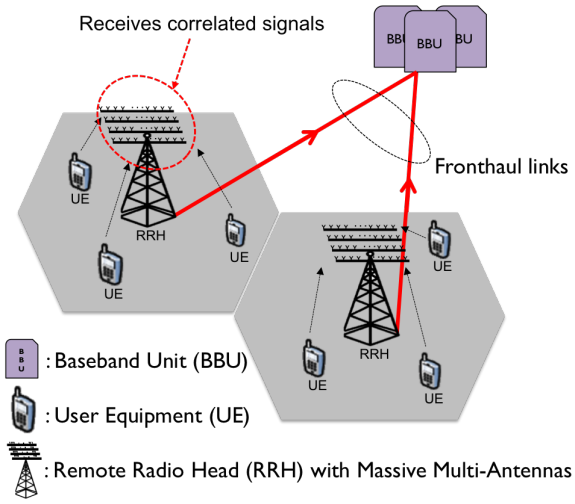


Fig. 1. RRHs with massive multi-antennas connected to BBUs via fronthaul links receive correlated signals from active UEs, and the number of active UEs is far less than the number of receiving antennas at the RRH.

Spatial domain compression methods have been proposed as well as time domain methods. [9] developed compress-and-forward schemes with joint decompression and decoding for the linear Wyner cellular uplink channel with single-antenna terminals, and [10] presented estimate-compress-forward strategies. Distributed compression methods were developed by using distributed Wyner-Ziv coding in [11], and by solving distributed Karush-Kuhn-Tucker conditions

J. Choi and B. L. Evans were supported by gift funding from Huawei

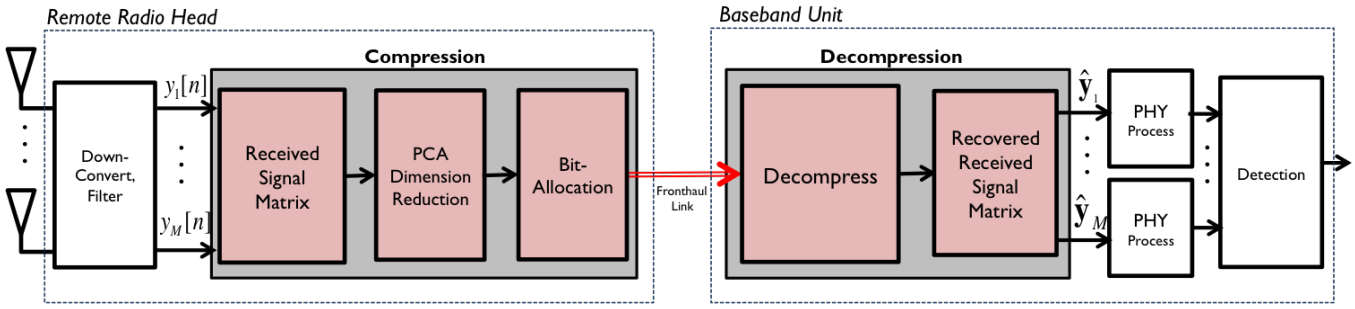


Fig. 2. Space-time fronthaul compression applied between the remote radio head and the baseband unit for uplink LTE signals.

to exploit the correlation of the received signals in [12]. While the proposed strategy in [12] is based on single-layer transmission and compression, [13] studied a layered transmission and compression method to effectively handle the existing uncertainties in the quality of backhaul links. The spatial domain compression methods [9]–[13] do not report LTE EVM results.

In contrast, we aim to jointly exploit both temporal and spatial correlation of received signals to achieve higher compression gain. In particular, we focus on the case in which the receiving antennas outnumber the active user equipment (UE), as large-scale antenna systems have been studied for millimeter wave mobile systems [14] and massive multiple-input multiple-output wireless communications [15]. In millimeter wave cellular communications, for example, a basestation operating at 28 or 38 GHz with a large number of antennas can support UEs up to 200m outdoors [16] — i.e. basestations need to be densely deployed, and consequently the number of the active UEs will be far less than the number of the receiving antennas as shown in Fig. 1.

In such a network, we propose a space-time fronthaul compression of complex baseband uplink time domain LTE signals, which exploits the correlation of received signals across space and time. The proposed method consists of two phases which are dimensionality reduction and individual quantization. The key feature of the dimensionality reduction phase is to perform low-rank approximation, applying principal component analysis (PCA) [17] to leverage the correlation structure across space and time. Since low-rank approximation is often used for denoising image or video signals [18], this method can obtain denoising gain. In the quantization phase, the dimensionality-reduced signals are quantized by using transform coding, which allocates a different number of quantization bits to each sequence of the signals.

II. System Model

In LTE uplink, symbols generated by each UE are precoded using discrete Fourier transform (DFT), and the output of the DFT is mapped by an inverse-DFT to produce an orthogonal frequency-division multiplexing (OFDM) symbol in time domain. Before the OFDM symbol is transmitted, a cyclic prefix (CP) is appended to the symbol. After the transmission of OFDM frames from different UEs, antennas in the RRH receive the time domain signals with noise and interference.

We consider fronthaul compression of this time domain signal for LTE uplink network, which consists of massive multi-antenna RRHs and multiple single-antenna UEs, i.e. each RRH covers each cell, and M antennas in the RRH receive signals from active UEs in the cell. The received signal at m^{th} antenna y_m is

$$y_m[n] = \sum_u x_u[n] * h_{m,u}[n] + w_m[n] \tag{1}$$
$$m \in \{1, 2, \cdots, M\},\ n \in \{0, 1, 2, \cdots\}$$

where $(*)$ represents convolution, x_u is an OFDM symbol of u^{th} user, $h_{m,u}$ is a channel response of u^{th} user to m^{th} antenna and w_m is additive white Gaussian noise (AWGN) of m^{th} antenna. From (1), we build a matrix of received signals $\mathbf{Y} \in \mathbb{C}^{N \times M}$, where N is the number of samples taken for each compression and M is the number of antennas at RRH. The matrix $\mathbf{Y}$ is shown as

$$\mathbf{Y} = \begin{bmatrix} y_1[0] & y_2[0] & \cdots & y_M[0] \\ y_1[1] & y_2[1] & \cdots & y_M[1] \\ \vdots & \vdots & \ddots & \vdots \\ y_1[N-1] & y_2[N-1] & \cdots & y_M[N-1] \end{bmatrix} \tag{2}$$

and each column in $\mathbf{Y}$ is denoted as

$$\mathbf{y}_i = \begin{bmatrix} y_i[0] & y_i[1] & \cdots & y_i[N-1] \end{bmatrix}^\intercal.$$

Since we consider that the columns $\mathbf{y}_i$, $i \in \{1, 2, \cdots, M\}$ are highly correlated, we apply low-rank approximation to $\mathbf{Y}$. In other words, we model the matrix $\mathbf{Y}$ based on (1) as

$$\mathbf{Y} = \mathbf{Y}_0 + \mathbf{E} \tag{3}$$

where $\mathbf{Y}_0 \in \mathbb{C}^{N \times M}$ denotes a low-rank matrix without noise, which includes information of x and h, and $\mathbf{E} \in \mathbb{C}^{N \times M}$ indicates a complex Gaussian noise matrix.

After the compression based on the model in (3), the compressed signals are sent to the BBU via the fronthaul link. Then, the signals are decompressed to recover $\mathbf{Y}$ and decoded at the BBU in the reverse order of the compression process. Fig. 2 shows the uplink compression process at the RRH, and decompression and OFDM decoding process at the BBU.

III. Proposed Compression Method

In this section, we first explore a low-rank approximation approach using PCA to reduce the dimension of the matrix $\mathbf{Y}$, which results in fewer signal samples to be sent over the fronthaul link. Then we discuss a possible performance gain from the denoising effect of low-rank approximation. Finally, we apply transform coding with bit allocation to achieve additional compression.

A. Low-Rank Approximation with PCA

Low-rank approximation reduces the dimension of the matrix $\mathbf{Y}$ which, in turn, reduces the number of signal samples to be sent. Assuming that $\mathbf{Y_0}$ in (3) is a rank-L matrix of dimensions $N \times M$, and $N \gg M > L$ without loss of generality, the low-rank approximation of $\mathbf{Y}$ is

$$\bar{\mathbf{Y}} = \underset{\text{rank}(\hat{\mathbf{Y}})=L}{\text{argmin}} \|\mathbf{Y} - \hat{\mathbf{Y}}\|_F \tag{4}$$

where the norm $\|\cdot\|_F$ represents Frobenius norm, and the optimal solution for (4) is

$$\bar{\mathbf{Y}} = \mathbf{U}_L \mathbf{\Sigma}_L \mathbf{V}_L^H \tag{5}$$

with

$$\mathbf{U}_L = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_L \end{bmatrix}$$
$$\mathbf{V}_L = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_L \end{bmatrix}$$
$$\mathbf{\Sigma}_L = diag\begin{bmatrix} \sigma_1 & \sigma_2 & \cdots & \sigma_L \end{bmatrix}$$

where $(\cdot)^H$, $\mathbf{u}_i \in \mathbb{C}^N$, $\mathbf{v}_i \in \mathbb{C}^M$ and σ_i are conjugate transpose, left eigenvectors, right eigenvectors and singular values. Since various methods of determining the true rank of a noisy matrix have been proposed [19], [20], we assume that the rank L is known.

Using PCA, the proposed method reduces the dimension of $\mathbf{Y}$ and obtains the low-rank approximation of $\mathbf{Y}$ at the BBU. In this work, we perform PCA without mean-centering. Our method finds the matrix of the first L principal components $\mathbf{V}_L$—equivalent to L eigenvectors—using singular value decomposition (SVD). Then, we transform the matrix $\mathbf{Y}$ by multiplying $\mathbf{V}_L$. This transformation maps the received signal vector $\mathbf{y}_i$ from an original space of M variables to a new space of L variables which are uncorrelated over the dataset. The transformed matrix $\mathbf{P}_L \in \mathbb{C}^{N\times L}$ is represented as

$$\mathbf{P}_L = \mathbf{Y}\mathbf{V}_L = \mathbf{U}_L\mathbf{\Sigma}_L \tag{6}$$

where $\mathbf{p}_i$, $i \in \{1, 2, \cdots, L\}$ denotes i^{th} column of $\mathbf{P}_L$, and is a decorrelated data vector.

After the linear transformation, the samples of the matrix $\mathbf{V}_L$ and $\mathbf{P}_L$, instead of $\mathbf{Y}$ are quantized and sent to the BBU via the fronthaul link, and the approximated low-rank matrix $\bar{\mathbf{Y}}$ is obtained at the BBU as follows:

$$\bar{\mathbf{Y}} = \mathbf{P}_L\mathbf{V}_L^H = \mathbf{U}_L\mathbf{\Sigma}_L\mathbf{V}_L^H. \tag{7}$$

Thus, the number of samples to be sent via link becomes $ML + NL$, and the compression ratio for the number of samples achieved by dimension reduction is

$$CR_{\text{DR}} = \frac{MN}{L(M+N)}. \tag{8}$$

We now discuss the denoising performance of low-rank approximation represented in (7). An analytic interpretation is as follows. The rank-L matrix $\mathbf{Y}_0$ has all desired signal components lying on the L-dimensional subspace, so eliminating the $(M-L)$-dimensional subspace only removes noise components that span the subspace of dimension $(M-L)$. To quantify the denoising performance, we rewrite $\bar{\mathbf{Y}}$ based on the signal matrix model in (3) as

$$\bar{\mathbf{Y}} = \mathbf{Y}_0 + \mathbf{\Delta} \tag{9}$$

where $\mathbf{\Delta}$ is the error matrix after low-rank approximation. The matrix $\mathbf{\Delta}$ contains the error from both residual noise and signal information loss due to low-rank approximation. The matrix $\bar{\mathbf{Y}}$ is considered as a denoised matrix, and the denoising SNR gain is defined as $G = \frac{\|\mathbf{E}\|_F}{\|\mathbf{\Delta}\|_F}$, which is the ratio of the total power of the noise before and after the low-rank approximation. The gain G is derived [21] as

$$G = \sqrt{\frac{\sum_{i=1}^{M}\lambda_i(\mathbf{E})}{\sum_{i=1}^{L}\lambda_i(\mathbf{E})}} \tag{10}$$

where $\lambda_i(\mathbf{E})$ is i^{th} eigenvalue of $\mathbf{E}$. The signal to noise ratio (SNR) gain is represented as $20\log_{10}G$ in dB. For the case in which eigenvalues λ_i are all equal, the denoising gain becomes $\sqrt{\frac{M}{L}}$; accordingly, the SNR gain in dB for this case is

$$G_{\text{dB}} = 10\log_{10}\frac{M}{L}. \tag{11}$$

B. Transform Coding with Bit Allocation

Here we discuss a more efficient way of quantizing the matrix $\mathbf{P}_L$ and $\mathbf{V}_L$. This quantization is applied to each real and imaginary part of the matrix $\mathbf{P}_L$ and $\mathbf{V}_L$, and quantizers are considered to be uniform quantizers. For the sake of brevity, we consider the matrix $\mathbf{P}_L$ as either its real components $\text{Re}(\mathbf{P}_L)$ or imaginary components $\text{Im}(\mathbf{P}_L)$, and same applies to $\mathbf{V}_L$ hereinafter.

Since the matrix $\mathbf{P}_L$ is a linearly transformed matrix of $\mathbf{Y}$ by its principal components matrix $\mathbf{V}_L$, we can individually quantize each transform variable $\mathbf{p}_i$ and its corresponding principal component $\mathbf{v}_i$, in order to use less number of quantization bits for additional compression. Such method is called transform coding. The problem here is to determine the number of bits for each quantizer, which achieves a minimum total quantization error for a target compression ratio. Our method solves a bit allocation problem for the individual quantization to optimize the overall coder performance given a bit budget for a target compression ratio.

To solve the bit allocation problem, we aim to minimize weighted overall distortion measure D of quantizing $\mathbf{P}_L$ for a given bit budget B, with respect to $\mathbf{b} = \begin{bmatrix} b_1 & b_2 & \cdots & b_L \end{bmatrix}$. b_i represents the number of quantization bits for $\mathbf{p}_i$. We use eigenvalues λ_i as weights for the distortion measure D. In

other words, we can find $\mathbf{b}$ which minimizes $D(\mathbf{b})$ by solving the cost function, such that

$$\text{minimize } D(\mathbf{b}) = \sum_{i=1}^{L} \lambda_i W_i(b_i) \tag{12}$$

$$\text{subject to } \textstyle\sum_{i=1}^{L} b_i = B,\ b_i \in \mathbb{Z}_+$$

where $\mathbb{Z}_+$ indicates nonnegative integers and $W_i(b_i)$ represents the mean-squared error incurred in quantizing the elements of $\mathbf{p}_i$ with b_i bits. Assuming P_i is a random variable for the samples in $\mathbf{p}_i$, we can approximate $W_i(b_i)$ [22].

$$W_i(b_i) \approx \text{var}(P_i) h_i 2^{-2b_i} \tag{13}$$

Here, $\text{var}(P_i)$ is the variance of P_i and the constant h_i is determined by the pdf $f_{\bar{\text{P}}_i}(p)$ of the normalized random variable $\bar{P}_i = P_i/\sqrt{\text{var}(P_i)}$ as

$$h_i = \frac{1}{12}\left\{\int_{-\infty}^{\infty} [f_{\bar{\text{P}}_i}(p)]^{1/3}\, \text{d}p\right\}^3. \tag{14}$$

We can find the approximate $\text{var}(P_i)$ and h_i using an empirical distribution of the samples in $\mathbf{p}_i$.

A greedy algorithm modified from the algorithm in [22] can solve (12). This modified greedy algorithm increases the number of quantization bits b_i by one for the quantizer with maximum weighted mean-squared error $\lambda_i W_i(b_i)$ in each iteration, until it allocates all B bits. This bit allocation algorithm is shown in Algorithm 1.

Since the principal component $\mathbf{v}_i$ also has the same weight λ_i as its corresponding transformed variable $\mathbf{p}_i$, our method assigns the same b_i bits to the quantizer of $\mathbf{v}_i$. Fig. 3 shows the overall compression process using PCA and transform coding with bit allocation. After the quantization, the quantized $\mathbf{p}_i$ and $\mathbf{v}_i$ are transmitted to the BBU via the fronthaul link. Regarding the total number of bits required to quantize $\mathbf{p}_i$ and $\mathbf{v}_i$, the proposed quantization method reduces the number of the aggregate quantization bits from $(M+N)L\,b_{\text{SD}}$ to $(M+N)\sum_{i=1}^{L} b_i$, where b_{SD} denotes the number of standard quantization bits. Therefore, the compression gain of transform coding with bit allocation is

$$CR_{\text{BA}} = \frac{L\,b_{\text{SD}}}{\sum_{i=1}^{L} b_i}. \tag{15}$$

Without considering the number of additional bits to be sent for quantization side information (QSI), the overall compression ratio regarding both dimensionality reduction and bit allocation becomes $CR_{\text{DR}}\,CR_{\text{BA}}$.

Algorithm 1 Greedy Algorithm for Bit Allocation

1: Initialize $b_i = 0$ for all $i \in \{1, 2, \cdots, L\}$
2: Input B, λ_i, W_i, $i \in \{1, 2, \cdots, L\}$
3: **while** $\sum_{i=1}^{L} b_i < B$ **do**
4: Find $k = \text{argmax}_{i\in\{1,2,\cdots,L\}} \lambda_i W_i(b_i)$
5: $b_k = b_k + 1$
6: **end while**
7: Return $\mathbf{b} = [b_1, b_2, \cdots, b_L]$

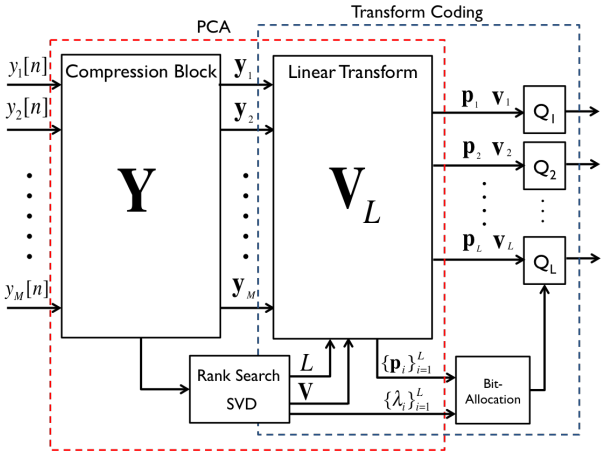


Fig. 3. Space-time fronthaul compression using PCA and transform coding with bit allocation.

The side information includes quantization ranges for each $\mathbf{p}_i$ and $\mathbf{v}_i$, quantization bit b_i information and rank information. The information of quantization ranges can be transmitted with $2\,L\,b_{\text{SD}}$ bits by representing each range with b_{SD} bits. The information for all b_i, $i \in \{1, 2, \cdots, L\}$ requires $L \log_2 b_{\text{SD}}$ bits as the maximum possible value of b_i is b_{SD}, and the rank information can be represented with $\log_2 M$ bits. Thus, the final compression ratio is derived as

$$CR_{\text{DRBA}} = \frac{MN\,b_{\text{SD}}}{(M+N)\sum_{i=1}^{L} b_i + b_{\text{QSI}}} \tag{16}$$

with

$$b_{\text{QSI}} = L(2b_{\text{SD}} + \log_2 b_{\text{SD}}) + \log_2 M$$

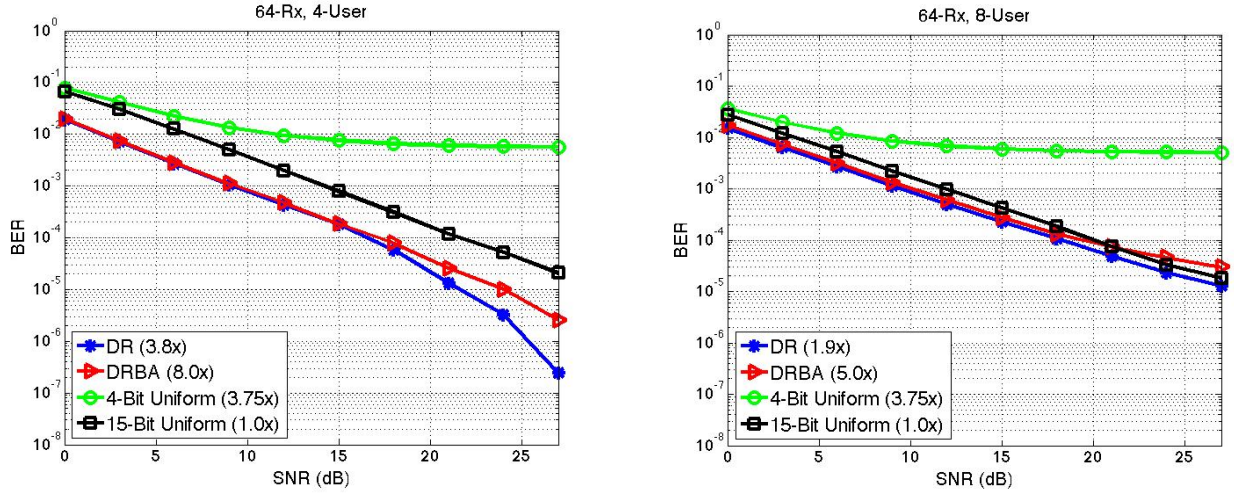
where b_{QSI} is the number of bits for QSI. (16) is the worst case compression ratio as it assumes that there is no QSI to be transmitted for the non-compression case. Since b_{QSI} is almost proportional to the rank L, and N is much larger than L from the assumption; $N \gg M > L$, b_{QSI} becomes negligible compared to the other terms in (16). In this case, the overall compression ratio becomes $CR_{\text{DRBA}} \approx CR_{\text{DR}}\,CR_{\text{BA}}$.

IV. Simulation Results

In this paper, we evaluate the performance of the proposed compression method by using an LTE uplink link-level simulator, and using two performance metrics. First, the simulator records uncoded bit error rates (BERs) to measure the impact of the compression method. Second, it measures EVM, which

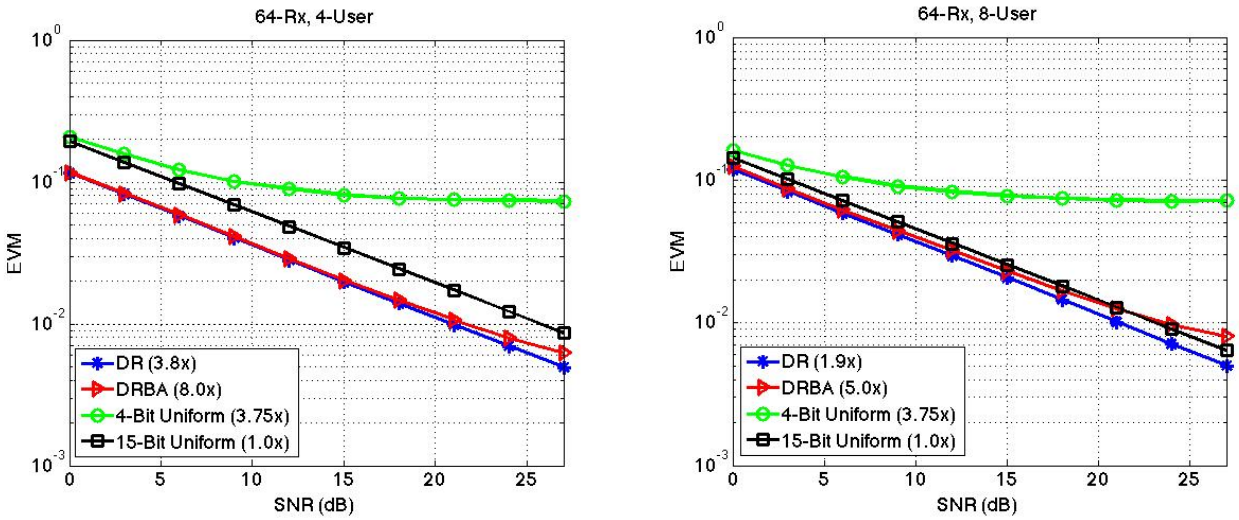
TABLE I
EVM requirements

Modulation Scheme	Required EVM [%]
QPSK	17.5 %
16QAM	12.5 %
64QAM	8 %
256QAM	3.5 %



(a) (b)

Fig. 4. The graphs show uncoded bit-error rate (BER) for compression with dimension reduction (DR), DR with Bit Allocation (DRBA) and 4-bit uniform quantization, and for no compression with 15-bit uniform quantization in (a) the network with 64 receiving antennas and 4 active users, and (b) the network with 64 receiving antennas and 8 active users. All active users transmitted 64-QAM.



(a) (b)

Fig. 5. The graphs show error-vector magnitude (EVM) for compression with dimension reduction (DR), DR with Bit Allocation (DRBA) and 4-bit uniform quantization, and for no compression with 15-bit uniform quantization in (a) the network with 64 receiving antennas and 4 active users, and (b) the network with 64 receiving antennas and 8 active users. All active users transmitted 64-QAM.

shows the difference between the ideal symbols and the decoded symbols after equalization. EVM requirements for QPSK, 16-QAM, 64-QAM and 256-QAM modulations are shown in the Table I [23].

Pedestrian A channel model is considered with 10 MHz LTE, which supports 50 resource blocks. The exponential correlation model [24] is used to generate antenna correlation assuming uniform linear array antennas. 64-QAM modulation is applied for symbol mapping. We simulate the proposed method in two cases: (a) 64 receiving antennas with 4 users, and (b) 64 receiving antennas with 8 users. Equally allocating the resource blocks to each user, overall 48 resource blocks are occupied with user signals for both cases. The number of compression samples N is 1096 (1024 IDFT outputs plus CP) and M is 64, so $\mathbf{Y}$ is the matrix of dimensions 1096×64. The true ranks of the matrix $\mathbf{Y}_0$ for each case are 16 and 32 ($L = 16, 32$) respectively, and b_{SD} is 15.

The simulation shows the BER and EVM curves of the 4 following cases: dimension reduction without bit allocation (DR), dimension reduction with bit allocation (DRBA), 4-bit uniform quantizer and 15-bit uniform quantizer. The 4-bit uniform quantizer case is a simple compression case, in which the number of quantization bits for I/Q samples is reduced to 4 bits without other compression procedures. The 15-bit uniform quantizer case is equivalent to non-compression case as we consider the standard quantization bit $b_{\text{SD}} = 15$. The number

of bits specified is the number of quantization bits per real or imaginary part of the samples. The compression ratios shown in the graphs are drawn based on (16) to take into account all required bits that are transmitted via the fronthaul link.

Fig. 4 shows BER curves for cases (a) and (b). In Fig. 4(a), DR achieves $3.8\times$ compression with 6 dB SNR gain compared to the 15-bit uniform quantizer case. This corresponds to (11) as $G_{dB} = 10\log_{10}4 \approx 6$ (dB) with $M = 64$ and $L = 16$. Accomplishing an additional $2.1\times$ rate reduction from the quantization, DRBA attains $8.0\times$ compression. Except for the very high SNR region, DRBA's SNR gain is similar to DR. The 4-bit uniform quantizer results in significant BER increase compared to the non-compression case with $3.75\times$ compression ratio, which is lower than the compression ratios of both DR and DRBA.

The simulation results in Fig. 4(b) show that DR achieves $1.9\times$ compression with 3 dB SNR gain compared to the 15-bit uniform quantizer curve. This also corresponds to (11) as $G_{dB} = 10\log_{10}2 \approx 3$ (dB) with $M = 64$ and $L = 32$. Although this compression ratio is relatively small as more active users increases the rank L, it can escalate up to $5.0\times$ due to the additional $2.7\times$ rate reduction from the quantization in DRBA. Except for the very high SNR region in which the DRBA curve shows slight increase of BER compared to the non-compression case, DRBA's SNR gain is similar to DR. The 4-bit uniform quantizer shows serious BER increment compared to the non-compression case, achieving $3.75\times$ compression which is lower than the compression ratio of DRBA.

EVM curves for cases (a) and (b) are shown in Fig. 5. DR curve presents about 6 dB and 3 dB SNR gain respectively, compared to the 15-bit quantizer case. These gains also correspond to (11) with given M and L values. Besides the very high SNR region, DRBA shows similar EVM curves with DR, achieving higher compression ratios — $8.0\times$ in case (a) and $5.0\times$ in case (b). Since DR and DRBA show improvements in EVM due to the denoising effect, both methods meet the EVM requirement for 64-QAM with about 1.5% to 3% *EVM improvement* in case (a) and 0.5% to 1% *EVM improvement* in case (b) for a typical LTE SNR range, whereas the 4-bit uniform quantizer results in catastrophic EVM increase.

V. Conclusion

In this paper, we have proposed a space-time fronthaul compression method of uplink LTE signals for the case, in which the number of receiving antennas at the RRH outnumbers the number of active users in the cell. Since neighboring antennas at the RRH receive spatially correlated signals, the propose method leverages both spatial and temporal correlation of the received signals. PCA performs low-rank approximation by exploiting the correlation structure across space and time to reduce dimensionality of the signals. Transform coding with bit allocation is applied for additional compression by individually quantizing the dimensionality-reduced signals with less number of bits. Via numerical results, we demonstrate the validity of the proposed compression method; $8.0\times$ compression is achievable with 6 dB SNR gain in EVM compared to the non-compression in the case of 64 receiving antennas and 4 users. Thus, the application of the proposed compression method can significantly lower transport data rate while simultaneously improving EVM performance.

References

[1] C. Mobile, "C-RAN: the road towards green RAN," *White Paper*, vol. 2, 2011.

[2] CPRI Specification V6.0 (2013-08-30) , "Common Public Radio Interface (CPRI); Interface Specification, 2013."

[3] G. ETSI, "001: Open Radio Equipment Interface (ORI)," *Requirements for Open Radio equipment Interface (Release 3)*.

[4] B. Guo, W. Cao, A. Tao, and D. Samardzija, "CPRI compression transport for LTE and LTE-A signal in C-RAN," in *in Proc. Int. ICST Conf. on Comm. and Networking in China*. IEEE, 2012, pp. 843–849.

[5] D. Samardzija, J. Pastalan, M. MacDonald, S. Walker, and R. Valenzuela, "Compressed transport of baseband signals in radio access networks," *IEEE Trans. on Wireless Comm.*, vol. 11, no. 9, pp. 3216–3225, 2012.

[6] Y. Ren, Y. Wang, G. Xu, and Q. Huang, "A compression method for LTE-A signals transported in radio access networks," in *Proc. IEEE Int. Conf. on Telecomm.*, 2014, pp. 293–297.

[7] K. F. Nieman, and B. L. Evans, "Time-domain compression of complex-baseband LTE signals for cloud radio access networks," in *Proc. IEEE Global Conf. on Signal and Info. Processing*, Dec. 2013, pp. 1198–1201.

[8] Hongbo Si, Boon Loon Ng, Md. Saifur Rahman and Jianzhong Zhang, "A Vector Quantization Based Compression Algorithm for CPRI Link," in *Proc. IEEE Global Comm. Conf. 2015*, December.

[9] A. Sanderovich, O. Somekh, H. V. Poor, and S. Shamai, "Uplink macro diversity of limited backhaul cellular network," *IEEE Trans. on Info. Theory*, vol. 55, no. 8, pp. 3457–3478, 2009.

[10] J. Kang, O. Simeone, J. Kang, and S. S. Shitz, "Joint signal and channel state information compression for the backhaul of uplink network MIMO systems," *IEEE Trans. on Wireless Communications*, vol. 13, no. 3, pp. 1555–1567, 2014.

[11] A. D. Coso and S. Simoens, "Distributed compression for MIMO coordinated networks with a backhaul constraint," *IEEE Trans. on Wireless Communications*, vol. 8, no. 9, pp. 4698–4709, 2009.

[12] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. on Veh. Tech.*, vol. 62, no. 2, pp. 692–703, 2013.

[13] O. Simeone, O. Somekh, E. Erkip, H. V. Poor, and S. Shamai, "Robust communication via decentralized processing with unreliable backhaul links," *IEEE Trans. on Info. Theory*, vol. 57, no. 7, pp. 4187–4201, 2011.

[14] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Comm. Mag.*, vol. 49, no. 6, pp. 101–107, 2011.

[15] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: benefits and challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 742–758, 2014.

[16] A. Bleicher, "The 5G phone future [News]," *IEEE Spectrum*, vol. 50, no. 7, pp. 15–16, 2013.

[17] I. Jolliffe, *Principal component analysis*. Wiley, 2002.

[18] L. Zhang, S. Vaddadi, H. Jin, and S. K. Nayar, "Multiple view image denoising," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1542–1549.

[19] S. J. Qin and R. Dunia, "Determining the number of principal components for best reconstruction," *Journal of Process Control*, vol. 10, no. 2, pp. 245–250, 2000.

[20] S. Kritchman and B. Nadler, "Determining the number of components in a factor model from limited noisy data," *Chemometrics and Intelligent Lab. Systems*, vol. 94, no. 1, pp. 19–32, 2008.

[21] H. M. Nguyen, X. Peng, M. N. Do, and Z.-P. Liang, "Denoising MR spectroscopic imaging data with low-rank approximations," *IEEE Trans. on Biomedical Eng.*, vol. 60, no. 1, pp. 78–89, 2013.

[22] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Springer 2012 (originally published 1992).

[23] Evolved Universal Terrestrial Radio Access V12.8.0 (2015-7), "Base station radio transmission and reception," 2015.

[24] S. L. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *Comm. Letters, IEEE*, vol. 5, no. 9, pp. 369–371, 2001.