Analysis of Baseband IQ Data Compression Methods for Centralized RAN

Aya Shehata, Matthieu Crussière and Philippe Mary

Univ Rennes, INSA Rennes, CNRS, IETR-UMR 6164, F-35000 Rennes

***Abstract*—Through recent wireless technologies, such as Centralized Radio Access Network, baseband unit and remote radio heads are physically separated and connected using fronthaul links. Compressing complex baseband signal samples prior transmission over fronthaul link is an effective way to satisfy the pressing need to decrease the huge required transported data rates. In this paper, we analyze the existing IQ data compression schemes exploiting time and spectral signal characteristics. We consider compression system evaluation parameters to have a smooth trade-off between required signal quality and complexity performance while achieving an acceptable compression gain. We propose an optimized uniform quantization technique combined with entropy coding achieving non-uniform quantization performance by exploiting signal temporal statistical characteristics with much less computational complexity. We also present a comparison between simulation results analyzing the trade-off between the removal of the signal spectral redundancies and vector quantization in terms of performance and complexity.**

***Index Terms*—CPRI, Uniform and Non-Uniform Quantization, Decimation, Vector Quantization, Compression Ratio.**

I. Introduction

The centralized, cooperative, Cloud Radio Access Network (C-RAN) relying on centralized processing, collaborative radio and real time cloud infrastructure, define the next generation wireless network architecture. Through C-RAN, processing is done in a Baseband Unit (BBU) located in a pool configuration and connected to many Remote Radio Heads (RRHs), where solely radio frequency (RF) units remain [1]. Fronthaul links, which are implemented over electrical or optical based interfaces, are then responsible for the transmission of digitized complex IQ baseband signals [2]. Consequently, the C-RAN strategy leads to an increased amount of traffic on the fronthaul links which hereby become the bottleneck of such an architecture. A natural solution to face this issue is to accommodate the increasing data traffic demand through the installation of more higher bandwidth optical fibers. A less expensive approach is to find ways to decrease the required transmitted data rate over the fronthaul before being transmitted. Hence, baseband IQ data rate compression has gained an increased interest through the last years. Various strategies of data rate compression have recently been investigated, namely taking basis on time-domain samples streams resulting from Orthogonal Frequency Division Multiplexing (OFDM), as used in many of the recent technologies. Therefore, fronthaul link compression techniques commonly found in literature are based on exploiting the temporal and spectral characteristics of the OFDM signal.

In [3] and [4], uniform scalar quantization was used on the oversampled OFDM signal and block scaling has been carried out to compensate for the large dynamic range of signal power. This is especially useful for uplink signals facing large and small scale propagation effects. In [5], non-uniform quantization based on an iterative gradient algorithm was proposed to take advantage of the signal statistical structure over uniform quantization, and dithering is performed across parallel links to reduce compression error. In [6], relatively low complexity compression methods were used, where compression is done by encoding the difference between the current and the previous sample and the algorithm was tested using high oversampling factors. Vector quantization combined with decimation, block scaling before quantization and entropy coding for the quantizer output was adopted in [7]. The aim was to exploit the correlation between samples and led to improved performance however at the expense of substantial increase of computational complexity. In [8] and [9] authors explored the correlation between samples by the well known linear predictive coding, hereby trying to solve the complexity issue introduced by vector quantization. This approach remains however restricted to the uplink signal.

Throughout this paper, we explore the well known lossy and lossless compression methods exploiting the scalar-based and vector-based characteristics of the samples to find a solution which balances the compression performance, complexity and system end to end performance. After a reminder of the main ingredients which can be used to build a global compression scheme, we explore two main strategies at the scalar and vector levels. We first propose to leverage the statistical signal characteristics by using a simple optimized uniform quantizer combined with entropy coding instead of designing a more complex non-uniform quantizer as usually done in the literature. Our proposed approach yields a substantial compression gain and a lower computational complexity.

Secondly, we investigate how correlation between signal samples can be efficiently exploited from a compression goal perspective. Thus, according to the wireless system specifications, we present an illustration in attaining optimum performance in terms of compression, signal quality and complexity between i) removing OFDM signal oversampling overhead using decimation and ii) exploiting this time correlated IQ baseband signals using vector quantization. Differently to what is proposed in [7] where decimation is followed by vector quantization, we rather compare both strategies since the former eliminates the main memory advantage of the latter due to diminishing the signal correlation. A performance/complexity analysis is then led and discussed.

The remaining of this paper is organised as follows. Section II presents the system model. In Section III, compression algorithms are described in details. Numerical analysis are presented in Section IV. Finally, Section V concludes the paper.

II. System Model

Fig. 1 depicts the basic functional blocks of the system used where time domain baseband signal is compressed before being transmitted through the fronthaul link. Throughout this study, the considered modules for compression are decimation, quantization and entropy encoding at the transmitter side and reverse operations are performed at the receiver side.

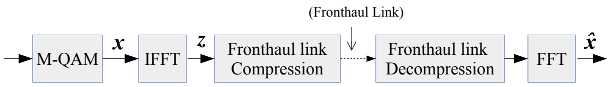

Figure 1: Fronthaul link compression algorithm framework.

Let the signal $\mathbf{x} \in \mathbb{C}^{N_c \times 1}$ be the M-QAM frequency domain symbol vector where N_c is the number of loaded sub-carriers. $\mathbf{z} \in \mathbb{C}^{N_f \times 1}$ is the corresponding time domain symbol vector loaded with $N_f - N_c$ empty subcarriers, i.e. known as guard subcarriers and N_f is the FFT size. $\mathbf{z}$ can be expressed as

$$\mathbf{z} = \mathbf{F}^H \mathbf{M}\mathbf{x} \tag{1}$$

where $\mathbf{F} \in \mathbb{C}^{N_f \times N_f}$ is the DFT matrix and $\mathbf{M} \in \mathbb{C}^{N_f \times N_c}$ is a mapping matrix decomposed into $\frac{N_f - N_c}{2} \times N_c$ null matrices and identity matrix of size N_c as follows

$$\mathbf{M} = \begin{bmatrix} \mathbf{0}_{\frac{N_f - N_c}{2} \times N_c} \\ \mathbf{I}_{N_c} \\ \mathbf{0}_{\frac{N_f - N_c}{2} \times N_c} \end{bmatrix}$$

where $\mathbf{0}_{n \times p}$ and $\mathbf{I}_n$ is the $n \times p$ null matrix and the square identity matrix of size n, respectively. Finally, $\hat{\mathbf{x}}$ is the demodulated symbol vector at the receiver. According to the Central Limit Theorem (CLT), for a sufficiently large IFFT length N_f, the resulting statistical distribution of the amplitude of the real and imaginary components converge to a zero-mean Gaussian distribution. Thus, exploiting OFDM signal statistical distribution is an important consideration in the quantizer design. Practically, the sampling value of the OFDM signal is usually higher than the minimum required by the Nyquist theorem $(N_f > N_c)$. Thus, time correlated baseband signals are expected and compression techniques exploiting this correlation must be considered.

On this basis, Compression Ratio (CR) and Modulation Error Ratio (MER) are the metrics used to evaluate the performance of the compression system, as introduced hereafter.

A. Compression Ratio

The compression ratio is defined as the ratio between the uncompressed size and the compressed one, defined as the following

$$\text{CR} = \frac{R_o}{R} \tag{2}$$

R_0 is the number of uncompressed bits of each I or Q sample, e.g. 15 (bits/sample) according to [2], and R is the number of compressed bits of each I or Q sample.

B. Performance

MER quantifies the ratio between the power of the original signal over the distortion introduced by the quantization in the log domain:

$$\text{MER(dB)} = 10\log_{10}\left(\frac{\mathbb{E}\{|x|^2\}}{\mathbb{E}\{|x - \hat{x}|^2\}}\right) \tag{3}$$

III. Compression Schemes

A. Scalar based Compression Techniques

1) Scalar Uniform Quantization: IQ samples are quantized sample by sample using a quantizer with R_{sq} bit resolution per each complex component. Uniform quantization (UQ) is optimum only for a uniform distributed signal [10]. N quantization levels are simply uniformly distributed in the range $\left[-2^{R_{sq}-1}, \cdots, 2^{R_{sq}-1} - 1\right]$, centered at zero, with total number of $N = 2^{R_{sq}}$ levels. Quantizer performance is evaluated by quantization distortion measured by the Mean Square Error (MSE) between the signal and the chosen quantization level expressed as

$$D = \sum_{i=1}^{N} \int_{t_{i-1}}^{t_i} (x - y_i)^2 p_X(x)\, dx \tag{4}$$

where y_i is the i^{th} quantization level, t_{i-1} and t_i are the decision thresholds of the i^{th} level. The first and last threshold levels t_0 and t_N are set to $-\infty$ and ∞ respectively. MSE in (4) could be decomposed into two effects as following

$$\begin{aligned} D = & \overbrace{\sum_{i=2}^{N-1} \int_{t_{i-1}}^{t_i} (x - y_i)^2 p_X(x)\, dx}^{D_1} \\ & \underbrace{+ \int_{-\infty}^{t_1} (x - y_1)^2 p_X(x)\, dx + \int_{t_{N-1}}^{\infty} (x - y_N)^2 p_X(x)\, dx}_{D_2} \end{aligned} \tag{5}$$

where D_1 is the distortion occurred when an input value lies in the bounded intervals, a.k.a. the granular distortion, and D_2 is the distortion at the first and last unbounded intervals, a.k.a. the overload distortion.

UQ performance could be enhanced by dynamically adapting the input signal level to have a trade-off between granular and overload distortions [9]. The loading factor γ is defined as

$$\gamma = \frac{y_N}{\sigma_x s_x} \tag{6}$$

where y_N is the maximum quantizer's output amplitude, σ_x is the standard deviation of the original random input signal and both are fixed attributes. s_x is the adjustable factor scaling the input signal before quantization to have the optimum quantizer performance. When $R_{sq} \to \infty$, D_1 can be expressed as follows [10]

$$D_1 = \frac{\gamma^2 \sigma^2}{3N^2} \tag{7}$$

and we derive a closed form of D_2 for a zero-mean Gaussian distributed signal in terms of γ as

$$\begin{aligned} D_2 = & \left[\sigma^2(1+\gamma^2) Q\left(\gamma\left(1 - \frac{1}{N}\right)\right)\right] \\ & - \left[\frac{\gamma\sigma^2}{\sqrt{2\pi}}\left(1 + \frac{1}{N}\right) e^{-\frac{\gamma^2}{2}\left(1 - \frac{1}{N}\right)^2}\right] \end{aligned} \tag{8}$$

where $Q(u) = (1/\sqrt{2\pi}) \int_u^\infty \exp(-v^2/2) dv$ is the Q-function. According to (8) which consists of a constant and monotonically decreasing functions, overload distortion is decreasing by increasing γ. while, in contrast, granular distortion increases with γ. Thus, as well known, explicit solution for γ trading between granular and overload distortions satisfying

$$\gamma_{opt} = \arg\min_{\gamma} D \tag{9}$$

is impossible and numerical search is used to find the optimum value depending on the input signal variance and the number of quantization levels.

2) Scalar Non-uniform Quantization: Optimum quantizer for non-uniformly distributed data samples requires more quantization levels in the range where samples have a high probability of occurrence. Thus, non-uniform quantizer (NUQ) has non-equal distances between quantization levels according to the distribution of the input signal. According to [11], for minimum mean square error (MMSE) optimality, the decision thresholds for each quantization level are derived by setting the partial derivative of (4) with respect to t_i to zero, which generates

$$t_i = \frac{y_i + y_{i+1}}{2} \tag{10}$$

and by setting the partial derivative of (4) with respect to y_i to zero, each quantization level appears to be the centroid of its decision region, i.e.

$$y_i = \frac{\int_{t_{i-1}}^{t_i} x p_X(x) dx}{\int_{t_{i-1}}^{t_i} p_X(x) dx}. \tag{11}$$

For a zero-mean Gaussian distributed signal, we have

$$y_i = \frac{\sigma}{\sqrt{2\pi}} \frac{\left(e^{-\frac{t_{i-1}^2}{2\sigma^2}} - e^{\frac{-t_i^2}{2\sigma^2}}\right)}{\mathrm{Q}\left(\frac{t_{i-1}}{\sigma}\right) - \mathrm{Q}\left(\frac{t_i}{\sigma}\right)} \tag{12}$$

3) Entropy Coding: Entropy coding is a lossless data compression scheme that utilizes the probability mass function (PMF) of the quantization levels. A large number of bits is used to represent levels with a low probability of occurrence while levels with a high probability of occurrence are represented with a lower number of bits. Huffman coding [13] is the most common practically used entropy coding technique approaching the Shannon's lossless source coding theorem. Thus the average codeword length assigned to the i^{th} quantization level is

$$L_{\text{Huff,i}} = -\log_2 \int_{t_{i-1}}^{t_i} p_X(x) dx \tag{13}$$

For a zero-mean Gaussian distributed signal, we derive it in a closed form as

$$L_{\text{Huff,i}} = -\log_2 \left(Q\left(\frac{t_{i-1}}{\sigma}\right) - Q\left(\frac{t_i}{\sigma}\right)\right) \tag{14}$$

Hence, additional compression is gained by eliminating the potential redundancy from the distribution of quantization levels.

4) Entropy Constrained Quantization (ECQ): it merges quantization with entropy coding [14]. Through ECQ, distortion and codeword lengths are optimized simultaneously by embedding entropy coding into the Lloyd's algorithm. The optimal quantizer is then obtained by minimizing the distortion subject to a constraint on the maximum transmission rate as following

$$\begin{gathered} \min_{t_i, y_i} . D = \sum_{i=1}^{N} \int_{t_{i-1}}^{t_i} (x - y_i)^2 p_X(x) dx \\ \text{Subject to } -\sum_{i=1}^{N} p(y_i) L_i \leq R_{max} \end{gathered} \tag{15}$$

Quantization level are still as in (11) but decision thresholds are updated to

$$t_i = \frac{y_i + y_{i+1}}{2} + \frac{\lambda}{2} \frac{L_{i+1} - L_i}{y_{i+1} - y_i} \tag{16}$$

where λ is the Lagrange multiplier of the rate constraint. This method has a remarkable complexity enhancement compared to using quantization and entropy coding functions separately.

B. Vector based Compression Techniques

Additional compression gain could be achieved by considering the fact that the amplitude of the OFDM signals are oversampled. Two solutions will be used to exploit this spectral redundancy.

1) Spectral Domain Redundancies Removal: OFDM signal is transmitted in a broader spectrum than is strictly required by the Nyquist criterion, i.e. oversampled by a factor L defined as

$$L = \frac{f_s}{f_m} \tag{17}$$

where, f_s is the signal sampling rate and f_m is the system bandwidth. This can be exploited to further increase the compression without degrading the signal quality by removing the redundant spectrum leaving a signal with a lower sampling rate f_{ds} before being transmitted through the fronthaul link.

Decimation is used in the redundancy removal process, first by upsampling the input signal by a factor m. The upsampled signal is then filtered with an FIR low-pass filter with a bandwidth limited to $[-f_{ds}/2, f_{ds}/2]$. Finally, filtered signal is downsampled by factor a k. The decimation factor F is the key parameter for performance evaluation and is defined as

$$F = \frac{f_s}{f_{ds}} = \frac{k}{m} \geq 1 \tag{18}$$

which is limited to the oversampling factor L, i.e. that depends on the amount of redundancy to be removed in the spectrum while retaining the useful data, i.e. $F \leq L$, and the filter length which must be chosen to trade between the spectral shaping requirements and the filter complexity.

2) Vector Quantization: Vector quantization (VQ) differs from the scalar one in exploiting time correlation between samples. VQ is a vector codebook based quantizer. The input is a k-dimensional vector whose components are the IQ samples $\mathbf{x} = [x_1, x_2, \cdots, x_k]$ and mapped to one of the k-dimensional vectors $\mathbf{y}_i$ in the codebook, where $i \in \{1, 2, \cdots, 2^{k.R_{vq}}\}$ and R_{vq} is the number of bits per each scalar sample. According to [12], Lloyd's optimality conditions can be extended to the vector quantizer (VQ) design.

By having a set of training samples extracted from the source, vector quantizer codebook is designed offline by partitioning the k-dimensional space of random vectors $\mathbf{x}$ into N convex quantization cells, which seeds are the output vector $\mathbf{y}$ occurences. An input vector that belongs to a cell is mapped to its seed, i.e. the quantized codeword. VQ is supposed to outperform the performance of scalar quantization by exploiting the correlation occurred by oversampling the signal. Hence, K-consecutive IQ samples are gathered in a vector and presented as the input to the quantizer.

IV. Analysis and Proposal

In this section we first numerically investigate the scalar quantization performance described in Section III-A. A 64-QAM signal is considered, which modulates $N_c = 27260$ sub-carriers out of $N_f = 32768$ IFFT entries per each OFDM symbol, that corresponds to a 8-MHz DVB-T2 frame structure. In order to assess the performance and compare with the idealized quantizer, we consider independent identically distributed (i.i.d) input samples, i.e. no redundancy in the spectral domain. MER of an ideal scalar quantization compression system is $\text{MER}_{\text{ub}} = 2^{2R_{sq}}$ [15] and it is considered as an upper bound.

A. Scalar based analysis

In Fig. 2, MER is presented as a function of the resolution R_{sq}. First, the dashed curves present the proposed optimized uniform quantizer and Lloyd-based non-uniform quantizer. UQ at each resolution is optimized at certain γ_{opt} according to the number of quantization levels and signal variance as shown in Fig. 3 which presents the MER as a function of the loading factor γ from 6 to 10 quantization bits [9]. Non-uniform quantizer has a clear advantage over the optimized

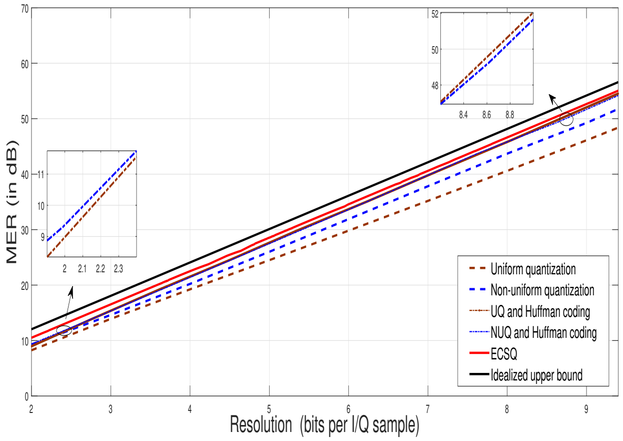

Figure 2: MER as a function of the resolution R_{sq} for i.i.d. complex Gaussian distribution.

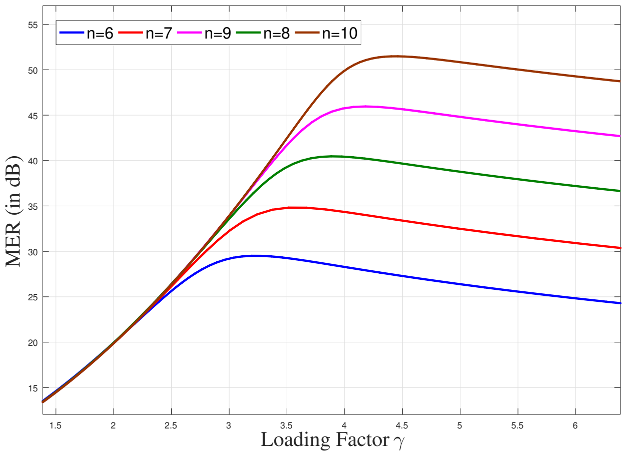

Figure 3: MER versus loading factor γ for different resolution bits.

uniform one and the difference is increasing at higher resolutions as seen on Fig. 2. Using the Huffman coding at the output of both uniform and non-uniform quantizers allows more compression by assigning variable-length codewords to different quantization levels according to their probability of occurrence. As shown in the dashed-dotted lines, UQ and NUQ algorithms with Huffman coding match. As illustrated in the zoomed parts in Fig. 2, for low resolutions NUQ is slightly better than UQ by approximately 0.4 dB at 2 bits. However, UQ becomes better by increasing the number of bits per sample about almost 0.45 dB at 9 bits.

This behavior is due to the increase of optimum loading factor γ_{opt} when increasing the number of bits as shown in (6) and illustrated in Fig. 3. Higher γ leads to a lower scaling factor s_k according to (6), thus, the scaled signal amplitudes at the input of the quantizer diminishes and are quantized to levels near to zero. Huffman algorithm codes these levels with short codeword lengths due to their higher probability of occurrence under Gaussian distribution. Hence, the optimized uniform quantizer combined with entropy coding exploits signal statistical distribution like non-uniform quantization but with a lower computational complexity.

This behavior can be explained by analyzing the asymptotic gap between the entropy of the uniform and non-uniform quantizer outputs derived from the asymptotic distortion-rate expressions for uniform and non-uniform quantizations derived in [10] and [16] respectively. This gap is expressed as the following

Table I: Asymptotic gap between UQ and NUQ with Huffman coding

Resolution bits	2	3	4	5	6	7	8
γ_{opt}	1.80	2.33	2.76	3.21	3.52	3.89	4.09
$R_{UQ}-R_{NUQ}$	0.66	0.29	0.05	-0.17	-0.30	-0.43	-0.52

$$\lim_{R\to\infty} R_{UQ}-R_{NUQ}=\frac{1}{2}\log_2\left(\frac{3\sqrt{3}\pi}{2\gamma^2}\right) \tag{19}$$

Substituting in (19) the optimal γ_{opt} obtained from Fig. 3 for different resolutions, the values of (19) are summarized in Table I, which shows the improvement of UQ over NUQ by increasing number of resolution bits, i.e. more compression gain can be achieved by UQ at the same MER.

Finally, the ECSQ performance can be considered as an upper limit to the compression gain obtained by a scalar quantization scheme and it is approximately 1.6 dB lower than the upper bound MER_{ub}. In the other hand, the MER obtained with UQ followed by Huffman coding is approximately 2.4 dB lower than the upper bound. Thus, according to the system requirements, we can decide whether approximately 1 dB gain obtained by ECSQ over using UQ and Huffman code worth the additional complexity of ECSQ implementation.

B. Vector based analysis

We now consider the fact that the amplitude of the OFDM baseband signals are correlated due to the oversampling as described in Section III-B. This section compares the compression ratios between i) a decimation followed by a scalar uniform quantization and ii) a vector quantization, when the oversampling factor varies. Figs. 4a and 4b illustrate the both techniques.

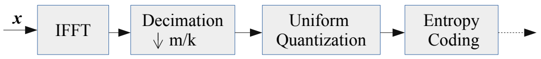

Figure 4a: Decimation followed by scalar quantization.

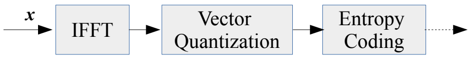

Figure 4b: Vector quantization.

As the oversampling factor L increases, the number of guard zero-padded sub-carriers $N_{Null} = N_f - N_c$ within one OFDM symbol increases. Thus, the ratio $\frac{N_{Null}}{N_f}$ is the oversampling parameter. In Fig. 5, the compression ratio is presented as a function of the oversampling factor represented by N_{Null}/N_f and is computed as CR$=\frac{R_0}{\frac{m}{k}R_{sq}}$ and CR$=\frac{R_0}{R_{vq}}$ for the decimation plus scalar UQ and VQ respectively. Thus, compression ratio increases by increasing the oversampling factor. For the system described in Fig. 4a, the decimation factor F is bounded by the oversampling factor as mentioned in Section III-B1, and more decimation, and hence high compression ratio, can be achieved at high oversampling values.

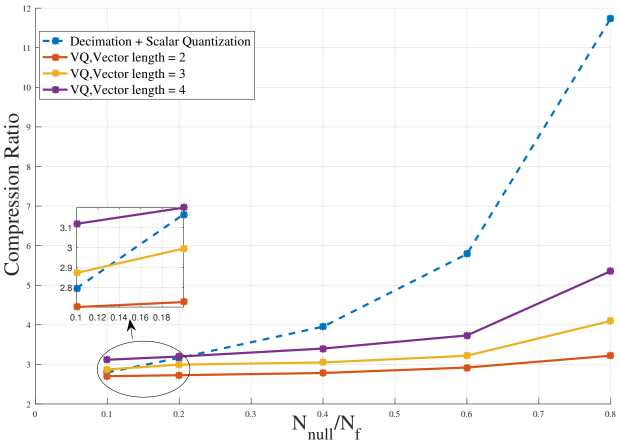

Figure 5: Compression ratio w.r.t. the oversampling factor of two techniques: decimation plus UQ and vector quantization.

For vector quantization implemented using the Lloyd's algorithm, CR increases along with the oversampling factor but less faster than the first solution for a given vector length K. However, additional compression gain can be achieved by increasing the vector dimension K because of the quantizer memory advantage.

In Fig. 5, we assume a fixed MER = 30 dB to compare the compression achieved by the two systems. A Hamming windowed FIR filter is implemented in the decimation process and Table II shows the selected decimation parameters (m,k) and the filter length N_w that achieve an acceptable signal quality and complexity performance tradeoff for different oversampling factors. Simulations results show that the gain obtained by decimation at low oversampling factor is very small due to the small degree of freedom available in that case and VQ with $K=3$ or 4 achieves a better performance. On the other hand, the decimation gain increases significantly by increasing the oversampling factor, achieving about 6 times higher compression gain than the one obtained with vector quantizer with $K=3$ at $L=0.8$.

The computational complexity of the two algorithms is also an important characteristic to be taken into consideration in addition to the compression performance. The complexity of the first algorithm is $\mathcal{O}(N_w \log N_w + 2^{R_{sq}})$, where N_w is the filter length and R_{sq} is the resolution of the quantizer. The complexity of VQ is $\mathcal{O}(2^{KR_{vq}})$ where K is the vector dimension and R_{vq} is the number of bits for each components. VQ has an exponential complexity in the vector length and the resolution while the decimation and UQ technique has a sup-linear complexity in the filter length but exponential with the quantizer resolution. Hence, the above numerical analysis may be used to decide the most suitable compression scheme according to the application. As an illustration, without loss of generality, DVB-T2 or ATSC3.0 broadcasting technologies have empty sub-carriers in approximately $\sim 4/25$ of their spectrum. Thus, for having more compression, VQ with higher vector lengths is preferable, taking into consideration its complexity. While, an LTE like system has about $\sim 2/5$ empty sub-carriers, decimation with UQ can achieve an acceptable compression gain at 30 dB MER with acceptable complexity according to Table II.

Table II: Hamming window FIR filter specifications

N_{null}/N_f	0.1	0.2	0.4	0.6	0.8
Decimation factor (F)	16/15	6/5	3/2	11/5	9/2
Filter length (N_w)	730	470	150	310	310

V. Conclusion

In this paper, we have analyzed various IQ data compression strategies for the data rate limitation on the fronthaul links of a C-RAN architecture. Through our analysis, two main conclusions can be driven. First, we have showed that an optimized scalar uniform quantization algorithm coupled with entropy coding over the resulting scalar codebook efficiently exploits the non-uniform OFDM signal statistical distribution. This proposed strategy achieves a substantial compression gain, about 2.4 dB lower than the idealized scalar quantizer, with much lower complexity compared to the non-uniform quantization approach. Second, we have provided a comparison between decimation and vector quantization in exploiting time-domain correlations. We have showed that a smooth trade-off exists between the required signal quality, compression performance and system complexity based on the choice of suitable parameter values.

Acknowledgement

This work has been funded by BPIfrance for the FUI25 CLOUDCAST project.

References

[1] Chen, Kulin, and Run Duan. "C-RAN the road towards green RAN." China Mobile Research Institute, white paper 2 (2011).
[2] Interface, Common Public Radio. "CPRI Specification V7. 0." Standard Document Specification 1 (2015).
[3] B. Guo, W. Cao, A. Tao and D. Samardzija, "LTE/LTE-A signal compression on the CPRI interface," in Bell Labs Technical Journal, vol. 18, no. 2, pp. 117-133, Sept. 2013.
[4] Peng-ren, Ding, and Zhao Can. "Compressed transport of baseband signals in cloud radio access networks." 9th International Conference on Communications and Networking in China. IEEE, 2014.
[5] D. Samardzija, J. Pastalan, M. MacDonald, S. Walker and R. Valenzuela, "Compressed Transport of Baseband Signals in Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 11,
[6] A. Vosoughi, M. Wu and J. R. Cavallaro, "Baseband signal compression in wireless base stations," 2012 IEEE Global Communications Conference (GLOBECOM), Anaheim, CA, 2012, pp. 4505-4511.
[7] H. Si, B. L. Ng, M. S. Rahman and J. Zhang, "A Novel and Efficient Vector Quantization Based CPRI Compression Algorithm," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 8, pp. 7061-7071, Aug. 2017.
[8] L. Ramalho et al., "An LPC-Based Fronthaul Compression Scheme," *IEEE Communications Letters*, vol. 21, no. 2, pp. 318-321, Feb. 2017.
[9] L. Ramalho, I. Freire, C. Lu, M. Berg and A. Klautau, "Improved LPC-Based Fronthaul Compression With High Rate Adaptation Resolution," *IEEE Communications Letters*, vol. 22, no. 3, pp. 458-461, March 2018.
[10] Gersho, Allen, and Robert M. Gray, *Vector quantization and signal compression*. Vol. 159. Springer Science and Business Media, 2012.
[11] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129-137, March 1982.
[12] Y. Linde, A. Buzo and R. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84-95, January 1980.
[13] D. A. Huffman, "A Method for the Construction of Minimum-Redundancy Codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098-1101, Sept. 1952.
[14] P. A. Chou, T. Lookabaugh and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 1, pp. 31-42, Jan. 1989.
[15] Berger, Toby. "Rate-distortion theory." Wiley Encyclopedia of Telecommunications (2003).
[16] H. Gish and J. Pierce, "Asymptotically efficient quantizing," *IEEE Transactions on Information Theory*, vol. 14, no. 5, pp. 676-683, September 1968.