Master's Thesis

Implementing Linear Predictive Coding based on a statistical model for LTE fronthaul

Ghassan Nakkoul
Mohammad Shohag Hassan Kaikobad

Department of Electrical and Information Technology,
Faculty of Engineering, LTH, Lund University, 2016.

Implementing Linear Predictive Coding based on a statistical model for LTE fronthaul

Ghassan Nakkoul
ghassannakkoul@gmail.com

Mohammad Shohag Hassan Kaikobad
hassan_shohag@yahoo.com

Department of Electrical and Information Technology
Lund University

Advisor: Stefan Höst and Yezi Huang

Examiner: Maria Kihl

November 20, 2016

Printed in Sweden
E-huset, Lund, 2016

Popular science article

Due to the growth of number of users and faster communication methods, mobile operators have to use the allocated resources more efficiently to meet the user demands. Like any other systems, mobile communication networks go through series of updates over time. In mobile communication system, these updates are known as "Releases". The transition from 3rd Generation (3G) to 4th Generation (4G) took place with Release 8 in 2008. Many new techniques are introduced in 4G in order to use the available resources more efficiently for improving quality of services (QoSs).

LTE (Long term evolution) or more commonly known as 4G communication system deals with much larger amount of data traffic than any other previous technologies. Hence it is of utmost importance that the operators make use of the allocated bandwidth more efficiently to serve the ever increasing number of users. It is possible for LTE to deal with this large amount of data due to the use of OFDM modulation technique which ensures better quality of communication. In OFDM, there exists multiple blocks of frequency bands stacked together as a whole, which are not related to one another.

The LTE structure is different from any previous systems. In telecommunication systems, there exists a unit which handles all the data traffic to and from the transmitter and the receiver. This module is called the base station. In LTE, the base station is divided into two parts namely the Baseband Unit (BBU) and the Radio Unit (RU), where almost all the data processing takes place at the BBU, and the RU is used as both transmitter and receiver when data is exchanged to and from a mobile device. In recent years, a new type of architecture is proposed, which is called the C-RAN (Cloud Radio Access Network). In C-RAN, the BBU and the RU would be placed at two different locations. Multiple BBUs can be placed together at a single place called the BBU pool, whereas the RUs will be placed in separate places far from the BBU pool and connected via optical fibers. In this structure, RU is known as RRH (Remote Radio Head) as they are separated from the BBU. One main advantage of such a structure is that, only the RRH is placed near the user locality and the BBU can be put at the network operator's vicinity. This also helps in reducing the operating and maintenance cost for the operator in many ways.

Since the LTE imposes with massive amount of data traffic on the fronthaul (almost tenfold of the actual information data after applying error correcting

coding, control signals etc.), it is very important to carry out compression of those data traffic before they are sent from the BBU to the RU. If good compression is carried out, then it becomes possible to accommodate more users, using the available resources. Although analog signals are used to transmit a message from the transmitter to the receiver over a medium, it is always important to convert those analog signal to digital signal to be transmitted from one block to the next block for processing, through the connecting link.

The main purpose of this thesis work is to apply a compression technique which will minimize the number of bits needed to represent each of those data samples transmitted from the BBU to the RRH. The compression technique used in this thesis is to employ a module which will use certain number of previous data samples values to make a prediction of the next data sample. Then this predicted data sample is compared with the actual data sample and their difference is found. The difference between these two samples has a low magnitude, as a result it is possible to use lower number of bits in the digital domain to represent this value, and finally transmitted through the link to the RRH.

At the RRH, the same prediction module is used to utilize these received samples of low magnitude, to make a prediction of the original data samples which are intended to be sent at the first place. In order to make the prediction module to function properly, it is very important to set up the filter values, which are known as the prediction coefficients. These coefficients play the role of successfully predicting data samples which are very similar to the original data samples. These coefficients are calculated by statistical method so that they can be used for any set of random data sample vector in the LTE.

This thesis studies the performance of applying this prediction technique in LTE. In order to identify the efficiency of this applied compression technique, certain parameters are calculated using various simulations, and compared with the set of values as specified by the main researching bodies of the LTE.

It is found that, the applied compression technique works fine in LTE as the simulation results support the validity of the scheme. It also proves that, it is possible to introduce this compression technique as an extension to the upcoming upgrades of the LTE, and this will facilitate accommodating more users with the available infrastructure resources.

Abstract

This thesis studies the application of Linear Predictive coding (LPC) in the downlink of Long Term Evolution (LTE) fronthaul, which comprises of BBU and RRH. This can act as an additional module in the existing system. Today, the transmission of a single complex sample from the BBU to the RRH consumes 30 bits. The research of the thesis is to analyze the application of linear prediction theory in the LTE downlink transmission, which will work as a compression scheme and reduce this 30 bits to lower value, at the same time fulfill the Error Vector Magnitude (EVM) requirement stated in the LTE standards made by 3rd Generation Partnership Project (3GPP).

As 4G-LTE and the upcoming access technologies will deal with large number of data samples in the transmission, it is an advantage if those data samples can be compressed without destroying the information content.

LPC or linear prediction coding has been proved to be a very effective method for speech compression in audio related applications. In this thesis, the same logic of compression is applied on digital data samples of the LTE and the results are analyzed.

It is found that, if LPC is applied properly on the LTE, it is possible to compress data samples efficiently and transmit them from the BBU to the RRH with fewer bits. At the RRH those compressed data samples can be processed and the main information data can be reconstructed, with additional quantization error and noise. This is obvious because LPC is a lossy compression method. A statistical model is established to generate a table of linear prediction filter coefficients which will be present both at the BBU and the RRH, when compression and decompression of data samples are performed.

Entropy is also calculated in order to analyze the achievable compression on an actual error vector after implementing certain compression coding such as Huffman coding. The specific coding technique is left as a scope of future research.

Acknowledgment

We would like to show our deep gratitude to our supervisors Stefan Höst and Yezi Huang for their great support and guidance to overcome the obstacles which we have faced during this work and making this thesis successful. We would also like to thank our thesis examiner Maria Kihl for her comments which helped us to improve the content of the this thesis. In this work we have tried to maintain authenticity in every respect.

I (Mohammad Shohag Hassan Kaikobad) would like to begin by thanking Almighty Allah. Then I would like to thank Swedish Institute (SI) for giving me the opportunity to study at Lund University and complete my Master Degree in Wireless Communications. Without their support it would be very tough to pursue with this program. I would also like to thank my family and friends for their continuous support during my studies here at Lund University.

I (Ghassan Nakkoul) would like to thank my family and my friends for their support encouragement and trust.

Finally, we would like to thank each other for being cooperative and supportive during the entire period of this thesis work.

Acronym

3GPP 3rd Generation Partnership Project

4G 4th Generation

BBU Baseband Unit

CAPEX capital expenditure

CP Cyclic Prefix

CPRI Common Public Radio Access

C-RAN Cloud Radio Access Network

eNodeB Evolved Node B

EVM Error Vector Magnitude

FDD Frequency Division Duplex

FFT Fast Fourier Transformer

FIR Finite-Impulse Response

IC Integrated Circuits

IFFT Inverse Fast Fourier Transformer

IIR Infinite-Impulse Response

IMT International Mobile Telecommunications

IMT-A International Mobile Telecommunications-Advanced

ISI Inter-symbol interference

ITU-R International Telecommunication Union - Radio Communication Sector

LPC Linear Predictive coding

LTE Long Term Evolution

LTE-Advanced Long Term Evolution-Advanced

MIMO Multiple-Input Multiple-Output

OFDM Orthogonal Frequency Division Multiplexing

OPEX operating expense

PRB Physical Resource Block RB

PRBs Physical Resource Blocks RB

QAM Quadrature Amplitude Modulation

RE Resource Element

RF Radio Frequency

RRH Remote Radio Head

RU Radio Unit

SFN Single Frequency Network

TDD Time Division Duplex

UEs User Equipments

UTP Unshielded Twisted pair

WP5D Working Party 5D

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

In this chapter, a comparison of the difference in the architecture is presented for the existing Long Term Evolution (LTE) and the upcoming Cloud Radio Access Network (C-RAN) fronthaul systems in the mobile networks [10]. Based on the upcoming C-RAN architecture, new challenges in terms of the demand of higher data traffic is needed to be tackled. Compression methods can be a solution which can be understood from this chapter. This chapter is constructed to explain the need of applying compression in the LTE which helps to handle large amount of data.

1.1 Background

The curve of rapid advancement of telecommunication systems with time is very steep in nature. This is due to the fact that the demand for faster communication methods is growing day by day. The application of wireless based communication systems is increasing in every possible fields. With the implementation of LTE, the amount of data traffic has grown tremendously in recent years. If this increasing number of data traffic is compared with the exchange of information in the early years of the telecommunication, then it is easily observed that today the amount of information has increased by couple of hundred times. As a result, it is always desirable to utilize the available resources to accommodate the increasing users' demands. The application of compression schemes is a very good option when large amount of data is to be transported from a source to a destination. It is also important that when compression is applied, the content of received information should not get distorted and the receiver should be able to recover the original message with minimum error. So it is essential to keep in mind that by applying compression, it also serves the purpose of using the maximum of the avaiable resources to accommodate the increasing amount of data traffic.

The compression scheme as applied in this thesis work is in the LTE downlink. The LTE base station unit known as Evolved Node B (eNodeB) is mainly divided into two main sections, named the Baseband Unit (BBU) and the Radio Unit (RU). Normally both the BBU and RU are physically situated closely as shown in Figure 1.1. Whereas in a C-RAN, the BBU and the RU are separated in two different geographical locations, where the radio unit is called Remote

Radio Head (RRH) as shown in Figure 1.2.

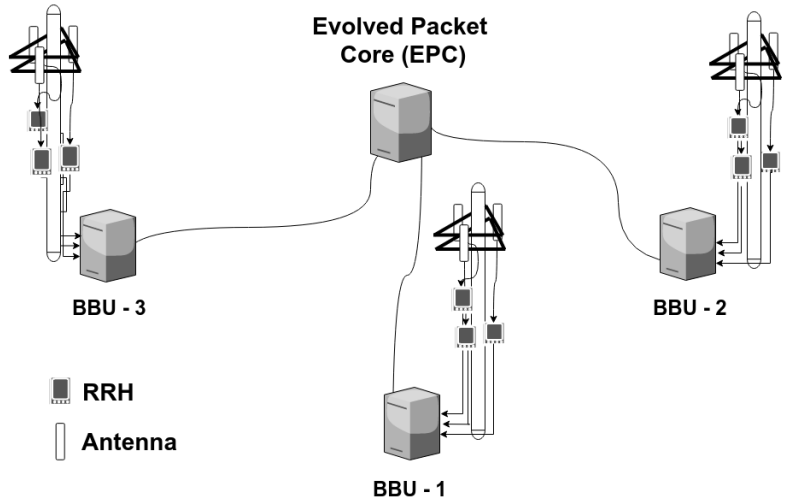


Figure 1.1: The traditional LTE fronthaul structure.

The main processing unit BBU can be located at the operators-controlled premises whereas the RRH units are placed in different locations to improve the system efficiency, meeting users demand, minimize the complexity by using simpler radio equipment at the network edge, facilitating easier maintenance and cheaper operating cost [4].

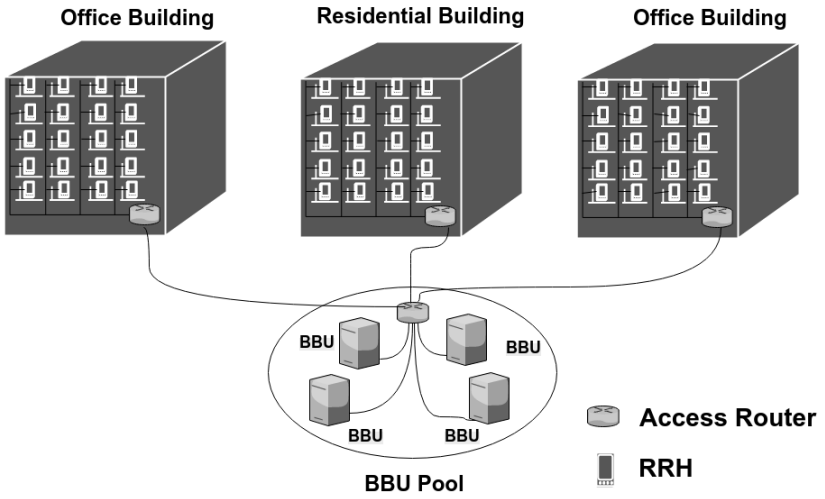


Figure 1.2: New LTE fronthaul structure i.e., C-RAN.

The use of RRH in the C-RAN architecture as shown in the Figure 1.2, has some advantages over the current LTE fronthaul architecture in Figure 1.1. To meet the requirement of higher data rate and the concept of always connected,

increasing the densification of the RRHs is a must. Hence it seems that the capital expenditure (CAPEX) is higher with increasing the number of RRH, however it is much less than deploying a full eNodeB. On the other hand, the main operating expense (OPEX) is also lower in the C-RAN. The new structure makes the network more flexible when maintenance of the network is taken into consideration. This is because in the new structure, each of the RRH functions independently. As a result if it is required to do some maintenance on a specific RRH the rest of the RRHs are not disturbed.

Multiple BBUs can be aggregated in one common location called BBU pool, which is more convenient from an operator's point of view since it reduces effort in terms of efficient controlling and maintenance. Some of the BBUs with no load can be turned off which reduces the operating cost. Cooperative radio technique and cooperative multipoints are also used at the BBU pool to reduce interference between different radio communication systems and increase system performance [4]. The main challenges for C-RAN is to efficiently handle the increase of the data traffic per user, maintaining low-latency and accurate synchronization between BBU and RRH. The combination of these two elements in the eNodeB is commonly known as LTE fronthaul. Optical fibers are the best option to meet these requirements today using the Common Public Radio Access (CPRI) protocol. The issue with CPRI is that it represent each data sample with 30 bits. However the traffic increases tremendously when large number of small cells are introduced in combination with Multiple-Input Multiple-Output (MIMO) and beamforming.

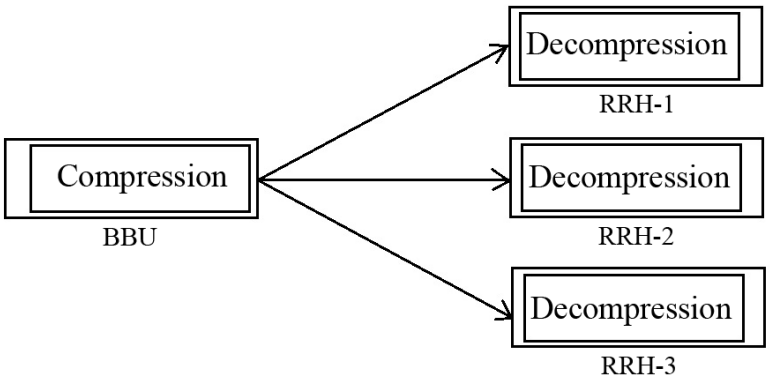


Figure 1.3: LTE fronthaul in C-RAN.

In this thesis, a compression scheme is proposed to be used in the BBU for the downlink which will use Linear Predictive coding (LPC) algorithm to minimize the number of bits that are used to quantize the samples in each of the Orthogonal Frequency Division Multiplexing (OFDM) signal. By lowering the number of bits used per OFDM time samples, it is possible to meet the demand of higher number of users in a densely populated environment using the current protocol namely known as CPRI. On the other hand, Unshielded Twisted pair (UTP) cabling can also be used to meet user demand in low density areas. LPC

theory has provided good support for various type of digital signal processing such as speech compression, computer generation of speech waveforms and also most recently it is used for blind identification of noisy communication channel [16].

LPC is able to provide an effective data compression in the LTE as found in the research of this thesis. The good compression ratio has facilitated lower order quantization level for successful transmission of signal from the BBU to the RRH in the downlink as shown in Figure 1.3, and as a result more users can be served with the same CPRI connection.

1.2 Problem Statement

In this thesis, it is tried to implement and optimize a compression technique that has been implemented in speech coding applications but not in LTE. This technique can compress the data which is going to be sent from the BBU to RRH in the downlink to save infrastructure resources. In this way the total available resources can handle more data traffic by using the proposed compression technique. A statistical model is implemented to derive the compression technique parameters and also to check the validity of the system by using those parameters.

Data compression is always a desirable option when large amount of data are transmitted from a source to a destination. In recent years, due to massive demand for high speed data communication, the 4^{th} Generation (4G) telecommunication system standard is developed by 3^{rd} Generation Partnership Project (3GPP). Today, although the transmission of information is becoming mostly radio based, traditional copper wire and optical fiber links are still being used alongside. Like any other medium, communication over air is also limited for use in terms of bandwidth. As a result, regulatory body such as International Telecommunication Union - Radio Communication Sector (ITU-R) has to control the allocation of bandwidth for different types of communication systems to operate side by side without causing noticable interference to one another and to ensure better user experience and facilitate innovative services [15].

The use of OFDM in LTE means that a large number of signals is to be sent in parallel at a time. This requires not only large amount of bandwidth but also higher number of quantization bits to define each of the OFDM samples with different amplitude which can reduce the amount of the quantization error. This can ensure minimum Error Vector Magnitude (EVM) stated in the standard recommended by ITU-R Working Party 5D (WP5D) and other regional standardization bodies for International Mobile Telecommunications (IMT) and International Mobile Telecommunications-Advanced (IMT-A) or commonly known as LTE and Long Term Evolution-Advanced (LTE-Advanced) [3]. In our thesis, we have focused on data compression by reducing the number of quantization bits required for each of the the OFDM samples to be transmitted from the BBU to the RRH in the downlink. The complex-valued baseband LTE signals contain I-part and Q-part. Today, quantization of 15 bits is used to represent each I-part and Q-part of the complex signal [6]. Quantization is applied on the OFDM

signals using a less granulated scale.

Moreover Linear Predictive coding (LPC) has been implemented to get rid of the oversampling drawback in the LTE signal. By doing so, the dynamic range of the amplitude for the transmitted signal can be reduced. This reduces redundancy without introducing any undesired degradation in the signal. Furthermore, in C-RAN large number of RRHs are connected with the BBU pool. This compression scheme will also be able to reduce the number of optical fibers connecting them and hence reduce the cost of the system implementation. Simulations have been carried out using MATLAB for different numbers of quantization bits and corresponding EVM values are plotted for comparison.

1.3 Thesis outline

The following parts of this thesis are divided into four chapters. In Chapter 2, the basics of OFDM and its application in LTE are discussed. The thesis mainly deals with three bandwidths, which are 10, 15 and 20 MHz. The frame structure of LTE for each of these bands are also presented along with the modulation scheme used in the reference channels as stated in the LTE standards. It is shown the difference of reference channels in terms of the size of Fast Fourier Transformer (FFT) and the total number of samples in the frame. Brief idea of quantization and dequantization is also presented, as signals transmitted over the LTE fronthaul link are in digital form. In the last section, method of calculating EVM and entropy is shown, which will help in the later chapters to compare between results of the thesis.

Chapter 3 introduces the main technique of linear prediction theory which is a crucial part of this work. A short history of linear prediction estimation is given to make it interesting for readers. As the idea of linear prediction theory began with the invention of the Wiener filters during 1940's, it is of importance to show how such a filter works. The implemented structure used in this thesis is explained. At the end, method of calculating the filter coefficients are shown.

Chapter 4 is dedicated to present all the simulations to compare between the outcomes of applying the LPC on LTE for different spectrum size. Plots of EVM and entropy are shown for different size of filter, number of quantization level and for different reference channels that are used. The EVM and entropy plots are shown for both the calculated and statistical model of filter coefficients.

Finally, in Chapter 5 the results are compared to show that compression of the samples using LPC can actually help to reduce the redundancy, which will be able to accomodate more users within the available physical links between the BBU and RRH. Possible future works related to this topic are also discussed.

Chapter 2

LTE Fronthaul system architecture

In this chapter, the LTE frame structure is discussed along with the idea of OFDM symbols which are used in LTE. The importance of Cyclic Prefix (CP) in LTE is also discussed as it is a very important feature of OFDM symbols. As the transmitted data samples are quantized when they are sent from the BBU to the RRH and then dequantization is applied at the RRH, the idea of quantization and dequantization is presented. They play a major role in the proposed scheme of LPC in LTE system. And finally, in order to compare between the results obtained in the thesis, it is also important to recap the idea of EVM and entropy.

2.1 Application of OFDM in LTE

OFDM is a multi-carrier transmission technique which is based on using large number of orthogonal narrowband subcarriers to achieve higher data rates. All subcarriers are packed next to each other in the frequency domain and the use of a rectangular pulse in the time domain ensures a sinc-shaped spectrum in the frequency domain for each of the subcarriers. Figure 2.1 shows the subcarriers in the frequency domain with frequency spacing equal to $\Delta f = \frac{1}{T_u}$, where T_u is the OFDM symbol time. The frequency spacing in LTE system is 15 kHz according to 3GPP standardization. The OFDM signal is represented as:

$$x(t) = \sum_{k=0}^{N_c-1} x_k(t) = \sum_{k=0}^{N_c-1} a_k^{(m)} \cdot e^{j2\pi k\Delta f t} \tag{2.1}$$

where x_k is the subcarrier, $a_k^{(m)}$ is modulation symbol carried on the subcarrier k and N_c represent the number of modulators. The modulators convert the Quadrature Amplitude Modulation (QAM) amplitudes into a QAM signal, at the output of the Inverse Fast Fourier Transformer (IFFT) block.

Each of those subcarriers x_k are mutually orthogonal over the time interval $mT_u \leq t < (m+1)T_u$ and can be represented by the following equation[3]:

$$\int_{mT_u}^{(m+1)T_u} x_{k_1}(t) \cdot x_{k_2}^* dt = \int_{mT_u}^{(m+1)T_u} a_{k1} \cdot a_{k2}^* e^{j2\pi k1\Delta f t} \cdot e^{-j2\pi k2\Delta f t} dt = 0 \tag{2.2}$$

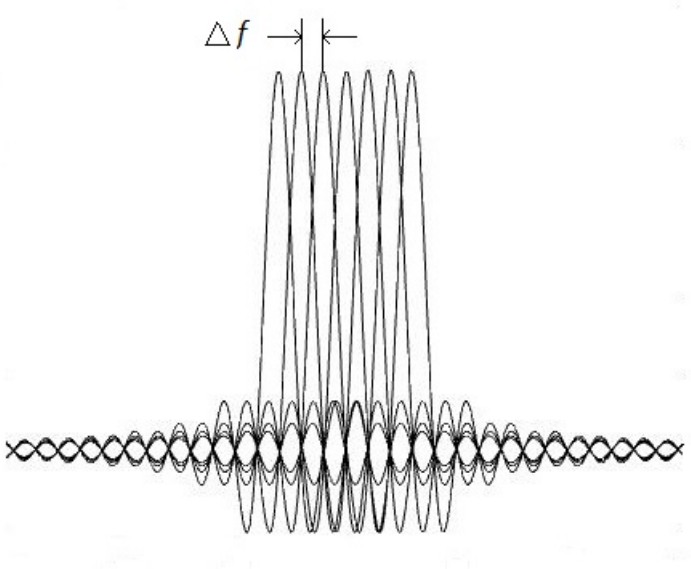


Figure 2.1: OFDM subcarriers in frequency domain.

A simplified structure of OFDM system is shown in Figure 2.2. At the transmitter side a serial to parallel converter is used to convert the serial samples into parallel sets of samples. Then they are placed at the input of the IFFT which is implemented to convert the samples from frequency-domain to time-domain.

A CP is then introduced at the beginning of each OFDM symbol which will be discussed in Section 2.2. All the symbols at the output of the transmitter are sent in sequence as they are in the time domain. Whereas at the receiver side the CP is removed and then FFT is implemented to convert the received signals from time-domain to frequency-domain. The blocks in the receiver in Figure 2.2, reverse the process that has been applied at the transmitter.

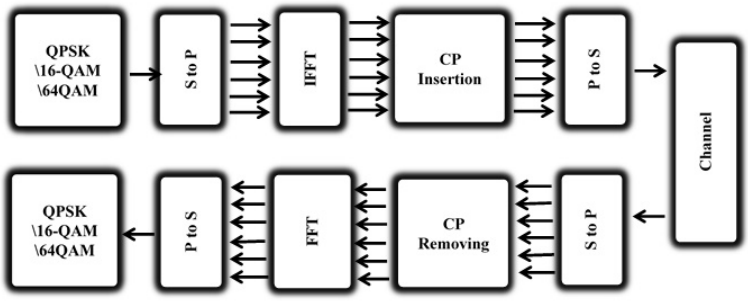


Figure 2.2: OFDM system structure.

The number of inputs to the IFFT and FFT is power of 2 for efficient implementation of the FFT and IFFT structures [3]. As it is shown in the Table 2.1, the size of the IFFT for 20 MHz, 15 MHz, 10 MHz and 5 MHz is 2048, 1536, 1024 and 512 respectively [5].

2.2 Cyclic Prefix (CP)

The orthogonality of the subcarriers will be corrupted due to time dispersion in the radio channel. In the time-domain, inter-symbol interference (ISI) happens when multiple copies of the same signal arrive at the receiver at different time intervals from multipath channels, after getting scattered by objects in the transmission medium. It is a problem for the receiver when making a decision of the received symbol because it causes over-lapping symbols on one another.

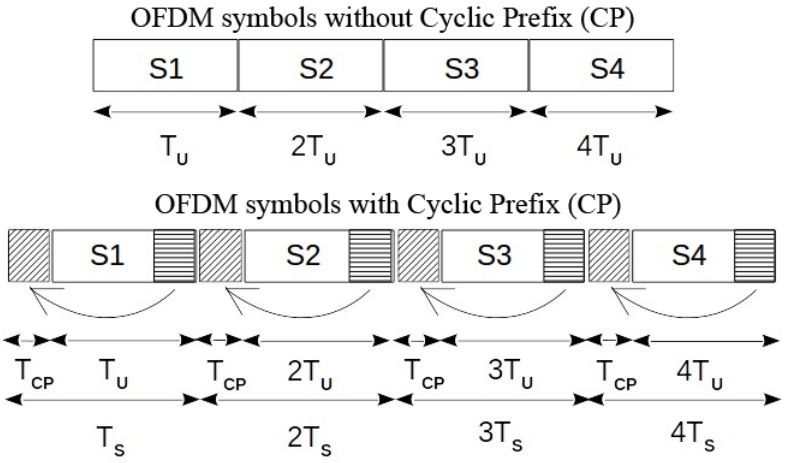


Figure 2.3: Cyclic Prefix insertion.

The orthogonality between the subcarriers in OFDM is not only because of frequency domain separation but also due to specific frequency domain structure of each subcarrier. This means that, even if the channel is constant over a bandwith of the main lobe of each subcarriers, the side lobes might be corrupted due to the frequency selectivity of the radio channel and also the orthogonality between them can be lost due to the inter-subcarrier interference [3]. Hence the receiver cannot fully separate the desired information from the corrupted signal.

In order to overcome the problem of ISI, a CP is inserted at the beginning of each OFDM symbol. The length of CP is chosen such that it is not shorter than the maximum delay of the multipath channel and is able to guard each symbol suffering from corruption due to ISI by the previous symbols after they reach the receiver. The last part of an OFDM symbol is copied and inserted at the beginning of that symbol.

If the duration of an OFDM symbol is T_u seconds, and the CP duration is T_{CP} seconds, the total duration T_s of the transmit signal is equal to summation of T_u and T_{CP}, i.e.,

$$T_s = T_u + T_{CP} \tag{2.3}$$

After the OFDM symbol is received at the receiver, the CP is removed. The rest of the samples are taken for demodulating the OFDM symbol.

In the LTE standards, the CP can be of two types. Normal CP has a length of $4.7\mu sec$, whereas extended CP is of $16.7\mu sec$. Normal CP is used in small cell environments to reduce the CP overhead. Extended CP is used in large cell environments for better protection against extreme time dispersion. They can also be used in Single Frequency Network (SFN) [3].

2.3 The frame structure for different LTE bandwidth

3GPP has defined an LTE frame structure in order to maintain and synchronize the communications in the air-interface between the eNodeB and the User Equipments (UEs)[12].

In LTE there are two modes, Frequency Division Duplex (FDD) and Time Division Duplex (TDD). In this thesis work only FDD mode is studied. FDD requires two separated communication bands: one for uplink and another for downlink with sufficient guard band between them, so they do not interfere with each other.

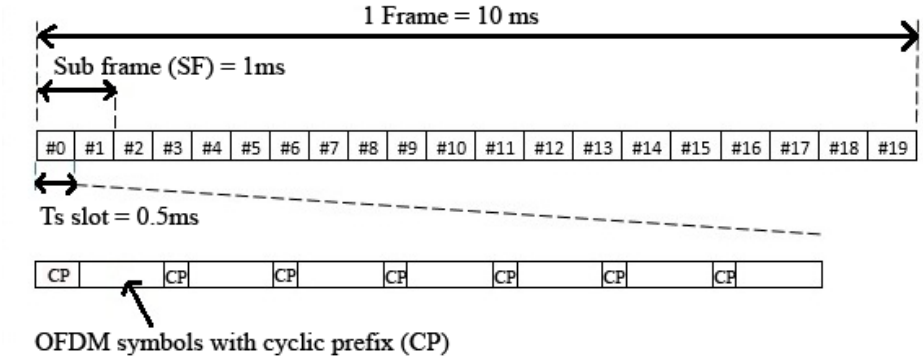


Figure 2.4: LTE FDD mode downlink frame structure.

In the downlink of the LTE FDD mode, the length of the LTE frame is 10 ms and each frame is divided into 10 subframes. Each subframe has a length of 1 ms and it contains two slots, each one has a duration of 0.5 ms. Each slot contains 7 OFDM symbols for normal CP length and 6 OFDM symbols with extended CP.

Figure 2.4 illustrates the LTE frame structure. The extended CP has longer time length than the normal CP to increase the protection against dispersion and ISI, but it reduces the number of symbols from 7 to 6 to maintain the 0.5 ms length of the slot. Table 2.1 shows the time length and number of samples in one OFDM symbol in time domain for normal and extended CP.

For the case of 20 MHz band with modulation scheme of 64-QAM as stated in the standard [5], $T_u = 66.7\mu s\ (2048 \cdot T_s)$ and $T_{cp} = 160 \cdot T_s \simeq 5.1\mu s$ for the first symbol in the slot and the time duration for each of the other 6 CP in the slot is $T_{cp} = 144 \cdot T_s \simeq 4.7\mu s$ where $T_s = 1/(15000 \cdot 2048)$.

As a result, the time duration of one slot for the mentioned frame is $T_{slot} = 15360 \cdot T_s$ and simply we can get the duration of one frame by multiplying by 20. The number of samples per CP for different bandwidths are shown in the Table 2.1.

Table 2.1: LTE Downlink physical layer parameters.

$T_XBW(Btx)$	**1.25 MHz**	**2.5 MHz**	**5 MHz**	**10 MHz**	**15 MHz**	**20 MHz**
Number of $PRB(N_{PRB})$	6	12	25	50	75	100
FFT size (N_{FFT})	128	256	512	1024	1536	2048
Sampling frequency	1.92 MHz	3.84 MHz	7.68 MHz	15.36 MHz	23.04 MHz	30.72 MHz
$(f_s = 15\,\text{kHz} \cdot N_{FFT})$	$(1/2 \cdot 3.84\text{MHz})$		$(2 \cdot 3.84\text{MHz})$	$(4 \cdot 3.84\text{MHz})$	$(6 \cdot 3.84\text{MHz})$	$(8 \cdot 3.84\text{MHz})$
Subcarriers/$PRB(N_{SC})$	12					
OFDM symbols (N_{CP}) (CP length)	7/6 (Short/Long CP) $5.21\mu s/16.67\mu s$					
Modulation	QPSK, 16-QAM, 64-QAM					
Number of Samples per slot	960	1920	3840	7680	11520	15360
Number of samples in CP first symbol/rest of symbols per slot	10/9	20/18	40/36	80/72	120/108	160/144

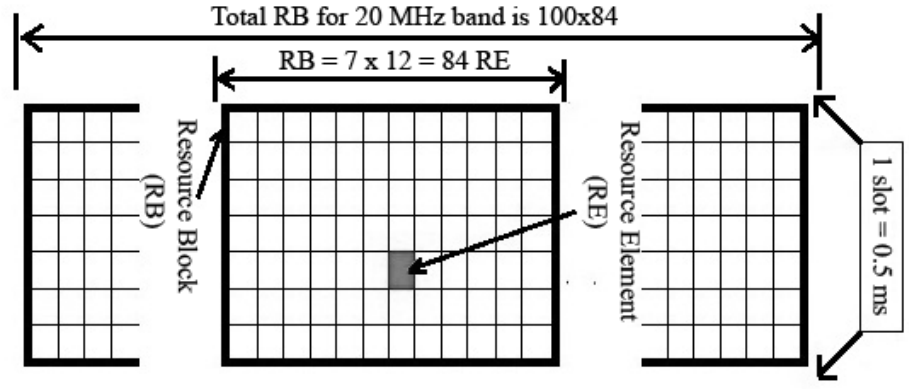


Figure 2.5: Resource Block in LTE FDD downlink frame.

In the frequency domain, 12 subcarriers with frequency spacing of 15 kHz, results in a bandwidth allocation of $12 \cdot 15$= 180 kHz. One Physical Resource Block RB (PRB) is represented by a two-dimensional grid: 7 symbols (1 slot) in time and 12 subcarriers in frequency. The number of PRB varies depending on the used bandwidth. Each PRB contains number of Resource Element (RE) which is equal to $12 \cdot 7 = 84$ RE as shown in Figure 2.5.

2.4 Quantization and Dequantization

As in any telecommunication system, the transmitted signal in the wireless channel is in analog form (unquantized). In the fronthaul the signal transmission is done in digital form. The samples are represented with finite number of bits. this process called quantization.

Quantization starts with finding the maximum and minimum values of the data vector and then based on the step size, the number of bits needed to represent the samples is chosen. Then the data samples are mapped on the quantization levels. The crucial part about this process is to find the balance between acceptable level of accuracy to be able to recover the sent data samples and the minimum number of bits needed to represent them.

When the quantization process is applied at the transmitter side, a reversed process called dequantization must be applied at the receiver side to revert the received quantized values into the original data samples.

During the dequantization process at the receiver, an error is introduced which is known as the quantization error. This error is due to the fact that it is not possible to revert the received sample to the same original value. The difference between the sent data sample and received one is called the quantization error and the received data sample can be described as the following:

$$x_q = x + e_q \Leftrightarrow e_q = x_q - x \tag{2.4}$$

where x_q and x are the quantized received sample and the sent sample respectively.

The error value will never increase the value of the step size which is denoted as Δ. In the uniform quantizer, the mean square quantization error is given by:

$$MSE_q = \frac{\Delta^2}{12} \tag{2.5}$$

The mean square quantization error in Equation (2.5) will be taken in consideration later to evaluate the system performance by adding it while calculating the filter coefficients [14].

2.5 The modulation scheme in LTE and the impact of quantization

In order to obtain higher data rates in the downlink for a specific bandwidth, a higher modulation order can be used. As a result, the number of symbols in the signal constellation increases. So more information bits can be sent by each modulation symbol.

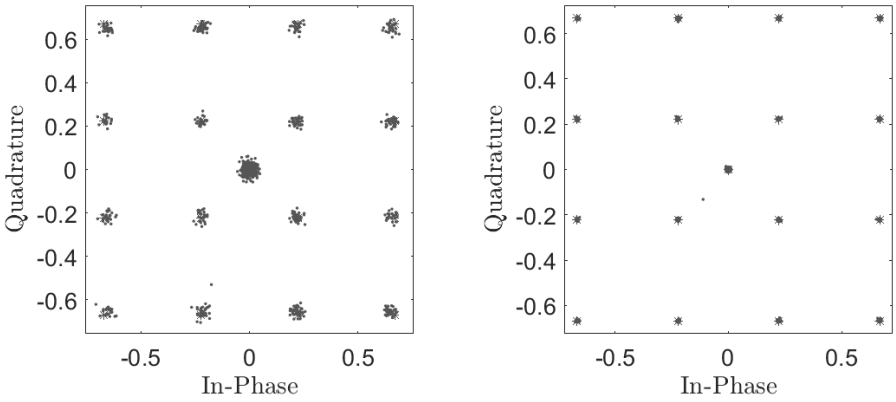
The four modulation schemes that are implemented in LTE are QPSK, 16-QAM, 64-QAM and 256-QAM. For the case of QPSK, only 2 bits can be sent in one modulation interval and total number of symbols is 4.

In 16-QAM there are in total 16 signals in the constellation and each signal is represented by 4 bits. In 64-QAM, each signal constellation is represented by 6 bits. In 256-QAM, each signal constellation is represented by 8 bits.

After performing FFT on the OFDM symbols, they are converted to QAM symbols in the frequency domain. The plot of the samples in the frequency domain forms the signal constellation as shown in Figure 2.6 and Figure 2.7. Each sample of the OFDM symbol is represented with one dot on the signal constellation map. Moreover, the data samples before being sent is converted to time domain as explained earlier. The quantization is applied and as a result quantization error is introduced and deviation from the original signal position on the constellation is observed due to this error.

On the other hand, the higher the quantization level, the lower error is produced. Figure 2.6 and Figure 2.7 illustrate the effect of quantization on the received samples where the red dots represent one transmitted OFDM symbol and the green dots represent the received samples after being dequnantized at the receiver.

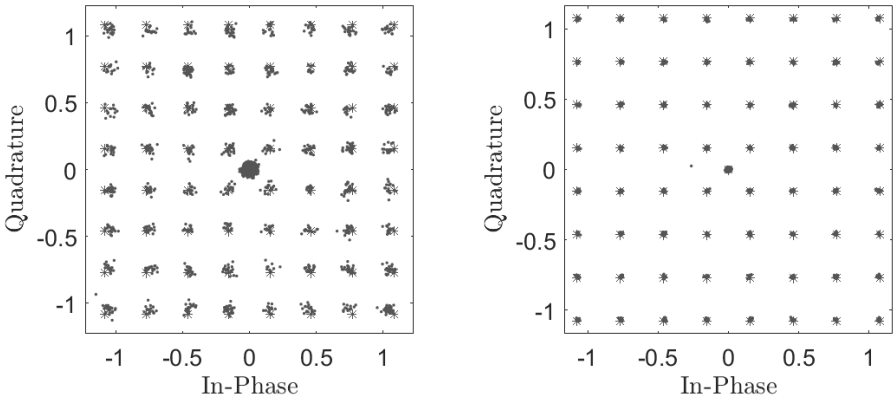
It is obvious that the amount of the error for 64 quantization level is much higher than 256 quantization level. This is because when 256 quatization levels are used, each sample is quantized with 8 bits and the whole range of data sample is divided into 256 levels. This means that the same range of data is divided into smaller steps when 256 quantization levels is applied. As a consequence, a higher accuracy is obtained in the receiver when the data samples is recovered.



(a) 16-QAM samples with quantization level=64.

(b) 16-QAM samples with quantization level=256.

Figure 2.6: 16-QAM signal Constellation.



(a) 64-QAM samples with quantization level=64.

(b) 64-QAM samples with quantization level=256.

Figure 2.7: 64-QAM signal Constellation.

In LTE, the reference channels are defined to describe the communication channel specification which includes the modulation scheme, the bandwidth, and the number of antennas. Some of the parameters of those reference channels are listed in the Table 2.2.

Table 2.2: LTE Downlink Reference channel parameters.

Reference channel Number	Number of resource block	Bandwidth	Modulation Scheme
R.7	50	10 MHz	64-QAM
R.8	75	15 MHz	64-QAM
R.9	100	20 MHz	64-QAM

2.6 Error Vector Magnitude (EVM)

Signals that travel through any physical medium experience distortion due to Radio Frequency (RF) imperfections present in the physical layer. One important way to measure the difference between the desired signal and the signal actually transmitted is the Error Vector Magnitude (EVM) calculation. EVM shows the error between the original data vector and the regenerated data vector at the receiver using linear prediction. EVM calculation is only done between the actual information data samples without taking into consideration of the Cyclic Prefix (CP). The CP is first removed from all the symbols in the LTE frame. The EVM is used as a guidance to check if the compression scheme meets the standarzation requirements or not. It also serves a good indication for which modulation schemes and code rates are to be used [15]. The following equation defines how EVM is calculated over the main data.

$$EVM = \sqrt{\frac{\sum\limits_{V \in T_m} \mid z'(v) - i(v) \mid^2}{\mid T_m \mid \cdot P_0}} \tag{2.6}$$

where, $i(v)$ = Ideal desired signal
$z'(v)$ = Actual transmitted signal
P_0 = Average power of the ideal signal
$\mid T_m \mid$ = Measurement duration (duration of 1 subframe in downlink)

It is always important to synchronize time and frequency of the main and transmitted signal before EVM is calculated. The cyclic-prefix must be removed and FFT should be carried out on the received time domain samples before the EVM is measured. The duration for measuring the EVM in LTE is at least 2 slots for the downlink and 1 slot for uplink. The average EVM value recommended in LTE must be less than 17.5% for QPSK, 12.5% for 16QAM and 8% for 64QAM in the downlink [5].

2.7 Compression of the Error Vector transmitted from the BBU to RRH

The entropy in a broad sense is a measure of uncertainty. According to information theory, entropy is a good tool to measure the amount of information content in a message [13]. If the probability of occurrence is low for certain incidence, then high entropy is obtained which means that, higher number of bits are needed to express it and vice versa. In data communication, often redundancy is present which requires extra number of bits to be used for storage or transmission of data. Entropy is used in data compression to indicate the minimum number of bits required to represent certain information after optimal encoding is done.

If p_a is the probability distribution of certain random variable X, then entropy is calculated as:

$$H(X) = -\sum_{a=1}^{i} p_a log(p_a) \tag{2.7}$$

In order to calculate the entropy, the probability density function (PDF) of the sent error vector is calculated. By using the individual probabilities of each of the unique samples, the entropy is calculated using the Equation 2.7. The unit of entropy in communication theory is bits/sample.

Huffman coding is one of the strong coding methods that can give high compression gain and eliminate redundancy. This can be done by representing the samples with high probability by shorter code length and vice versa. As a result the output of the encoder can be defined using fewer bits.

In this thesis, main compression is done using LPC. LPC is a lossy compression method. Therefore the entropy is calculated to show that if lossless compression method such as Huffman coding is applied on top of LPC, more compression gain can be achieved.

Chapter 3
Linear Prediction Coding (LPC)

The main idea of this thesis is based on the linear prediction coding (LPC). So it is of utter importance to present an overview of LPC as well as its history briefly. In the first section, a brief history is given about linear prediction filtering. In the following sections, LPC is discussed in detail and also the basics of prediction filter is explained. Furthermore, the method of calculating the filter coefficients are shown by using LPC. The implemented structure of LPC is discussed too. And at the end of this chapter, the method of calculating the filter coefficients to establish a statistical model is also presented.

3.1 Theory of Linear Prediction: History and Background

Linear prediction was successfully used to synthesize speech signal at the end of 1960's at AT&T Bell Labs. It is believed that, the idea of linear prediction existed since the beginning of 1940's when Norbert Wiener was working on a mathematical model to separate signal from noise. During World War 2, Wiener also applied the idea of linear prediction on the radar system and anti-aircraft guns to predict and eliminate enemy planes which changes position over time. Researchers from Japan, Fumitada Itakura and Shuzo Saito also developed statistical model of linear prediction for estimating speech spectral density using a maximum likelyhood method during the 1960's [1].

At the beginning, when speech synthesizing was done using LPC, it had a bit rate of 2.4 kbps and the quality of the speech was not good in the sense that it was distorted. Bishnu Atal and his colleagues, a team of researchers at AT&T Bell Labs, tried to minimize the distortion and produce high quality speech with low bitrate. They succeeded to produce high quality speech with a bit rate of 9.6 kbps by using Multipulse LPC. The idea of multipulse LPC is to use the output of the speech synthesizer and compare it with the original speech and make an error vector which is weighted and fed into a multipulse excitation generator. The multipulse excitation generator creates a sequence of pulses which are again used by the speech synthesizer to minimize the weight of the error [1].

Until then, it was not possible to apply LPC in the telecommunication system for synthesizing speech, because LPC required a lot of real-time calculation to be made and the limitation of computational power acted as a barrier. Over the years, as Integrated Circuits (IC) were developed, the drastic development of

handling large amount of calculation with the increased computational power made it possible to apply LPC for higher quality speech synthesizing, which is later applied in cellular systems for mobile communication.

3.2 Discrete-Time Filters

Any Discrete-Time filter is built from *delay units*, *multipliers* and *adders*. The filter implementation uses previous components to calculate the output of the filter based on the input. There are mainly 2 types of the filters known as Finite-Impulse Response (FIR) and Infinite-Impulse Response (IIR) filters.

3.2.1 FIR Filter

The output of the FIR filter is computed based on the combination of weighted past or present samples and the relation between the input and the output can be represented as a *convolution sum*. Hence the output can be expressed as the following:

$$y(n) = \sum_{K=1}^{M} h(k)x(n-K) \tag{3.1}$$

3.2.2 IIR Filter

The output of the IIR filter is computed based on the current input and past samples of the output which is fed into the filter as a feedback. The output of the IIR can be expressed as the following:

$$y(n) = x(n) + \sum_{K=1}^{M} \alpha_k y(n-k) \tag{3.2}$$

where $x(n)$ represent the current input, $y(n)$ represent the current output and $y(n-1), y(n-2)....y(n-k)$ represent the past output samples which are calculated based on filter coefficients $\alpha_1, \alpha_2...\alpha_k$ [7].

3.3 LPC Theory

Linear prediction is a significant process in the signal processing field and can be implemented as a powerful tool in LTE. The idea behind linear prediction is based on using prediction filter to predict the next sample. This can be done by summing the multiplication of each of the filter coefficients with the previous data samples. The number of coefficients and number of previous samples depends on the size of the filter.

The predicted signal $y(t)$ is compared to the desired signal $x(t)$ and the difference between them is called the prediction error signal $e(t)$, as shown in Equation 3.3. The structure of the transmitter is shown in Figure 3.1.

$$e(n) = x(n) - y(n) \tag{3.3}$$

It is also possible to build a receiver that can recover the desired signal from the prediction error signal $e(t)$. At the receiver side, the received error signal $e(t)$ is added to the predicted signal $\hat{y}(t)$ to reconstruct the expected desired signal $\hat{x}(t)$.

The reconstructing process is not optimal because the signal at the output of the receiver $\hat{x}(t)$ is not identical to the desired signal $x(t)$. The reason of not being identical is because the predicted signal contains prediction error. However with good filter coefficients and sufficient filter size, a sub-optimum result can be obtained.

On the other hand, the most important outcome of this structure is compression gain. This is obtained due to the fact that the amplitude of the sent prediction error signal $e(t)$ has lower amplitude comparing to the desired signal. Hence, less number of bits is needed to represent the data sample.

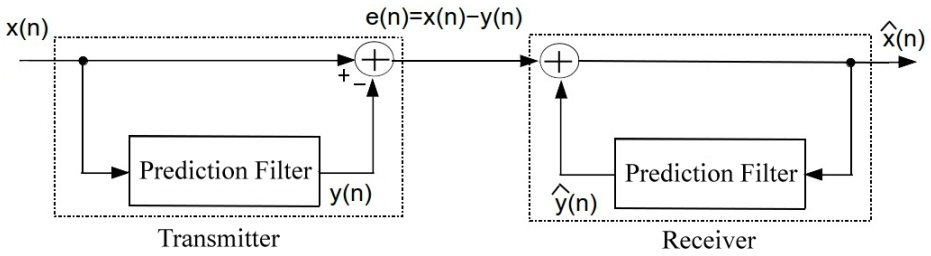


Figure 3.1: Prediction Filter.

The prediction filter function is to estimate the next sample based on the knowledge of the previous samples in the time domain. To study its functionality, let $H = [h_1, h_2, ...h_k]^T$ be the filter coefficients and $x_n, x_{n-1}, x_{n-2}, ...x_N$ represent the discrete-time input signals to the filter. It is assumed here that the initial state of the filter memory is zero and the discrete-time input signal $x(n)$ is *wide-sense stationary* [9].

If we assume that the first filter input is x_{n-1} then the output of the filter can be assumed to be x_n and can be written as:

$$y_n = h_1 \cdot x_{n-1}$$

The value x_{n-1} is stored in the filter memory and the output of the filter y_n is compared to the input x_n to get the error vector. Furthermore when the next input signal x_n is processed in the filter then the output signal x_{n+1} is estimated based on the previous two samples x_{n-1} and x_n, which is written as:

$$y_{n+1} = h_1 \cdot x_n + h_2 \cdot x_{n-1}$$

with applying the same method we can estimate the future sample based on the previous samples for the rest of the other input signals. The number of the

stored samples in the filter depends on the size of filter k. The filter flashes the older sample when the filter memory fills up.

A good filter estimation at the output requires that the filter coefficient to be calculated as correct as possible and a memory to store the k previous samples. The calculation of the predicted samples that are presented in the previous formulas can be written as:

$$\begin{bmatrix} x_{n-1} & 0 & 0 & \cdots & 0 \\ x_n & x_{n-1} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ x_{m-1} & x_{m-2} & \cdots & \cdots & x_{m-k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N & x_{N-1} & \cdots & \ddots & \vdots \\ 0 & x_N & x_{N-1} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & x_N \end{bmatrix}_{[(N+k)\times k]} \cdot \begin{bmatrix} h_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ h_k \end{bmatrix} = \begin{bmatrix} y_n \\ y_{n+1} \\ \vdots \\ y_m \\ \vdots \\ \vdots \\ y_N \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{[(N+k)\times 1]} \tag{3.4}$$

It should be noted here that Equation 3.4 is based on the assumption that, with the right filter coefficients the output of the filter is equal to the next input sample which means for example:

$$y_n = x_n \tag{3.5}$$

If the left matrix of Equation 3.4 is indicated as A and the right hand side of Equation 3.4 as B then it can be written:

$$A \cdot H = B \tag{3.6}$$

and the filter parameters H can be calculated by the following mathematical expression:

$$A^* AH = A^* B \tag{3.7}$$

$$H = (A^* A)^{-1} A^* B \tag{3.8}$$

$$H = A^+ B \tag{3.9}$$

Where A^+ is known as Moore-Penrose pseudo inverse of matrix A [8]. Moreover the implementation of this operation has been done in Matlab using the backslash operator $\backslash$, which solves system of linear equations $A \cdot x = B$ with the operation $x = A \backslash B$. This makes calculating filter coefficients mathematically easy to accomplish.

Previous derivation for the filter coefficients was based on the knowledge of the input signals, which is not the case in reality. However we used it to study and evaluate the system validity and verify that the system can recover the desired signal at the output of the receiver.

3.4 The method of calculating the filter coefficients:

In the previous Section 3.3, the LPC theory and the error prediction signal at output of the transmitter are explained. It is clear that, the smaller the error signal are, the higher compression gain is obtained.

In this section, the minimization of the prediction error based on the statistical criteria is presented. It is consider that the input $x(n)$ assumed to be wide-sense stationary with zero mean. And as a result this apply to the error signal vector $e(n)$ [7][9]. The most common statistical criteria to calculate the filter coefficients is the mean-square value of the error signal and this is mathematically expressed as:

$$\xi = E[e^2(n)] \tag{3.10}$$

the optimal solution is to have $\xi = 0$

On the other hand, the output of the filter in the time domain can be expressed as:

$$y(n) = X^T \cdot H \tag{3.11}$$

where
$X = [x_{n-1}, x_{n-2}....]^T$ and
$H = [h_1, h_2, ...h_k]^T$
In this section the filter coefficients are derived from Equation 3.3 and 3.10:

$$\xi = E\Big[\big(x(n) - y(n)\big)^2\Big] \tag{3.12}$$

The accuracy of $y(n)$ which is the estimation of the sent signal will depend on the size of the filter and can be represented as in (3.11) and as a result the error can be written as:

$$e(n) = x(n) - X^T H \tag{3.13}$$

and we can say that,

$$\xi = e^T \cdot e \tag{3.14}$$

the result of substituting (3.13) in we get:

$$\xi = E[(x - X^T \cdot H)^T \cdot (x - X^T \cdot H)] \tag{3.15}$$

hence,

$$\begin{aligned}\xi &= E[(x^T - H^T \cdot X)(x - X^T \cdot H)] \\ &= E[x^T \cdot x] + E[-x^T \cdot X^T \cdot H - H^T \cdot X \cdot x + H^T \cdot XX^T \cdot H] \\ &= E[x^2] - 2 \cdot H^T \cdot E[X \cdot x] + E[H^T \cdot X \cdot X^T \cdot H]\end{aligned}$$

$$\xi = E[x^2] - 2 \cdot H^T \cdot E[X \cdot x] + H^T \cdot E[X \cdot X^T] \cdot H \tag{3.16}$$

where we can express:

$$E[X \cdot x] = E\left[\begin{bmatrix} x_{n-1} \\ \vdots \\ x_{n-k} \end{bmatrix} \cdot x_n\right] = E\left[\begin{bmatrix} x_{n-1} \cdot x_n \\ \vdots \\ x_{n-k} \cdot x_n \end{bmatrix}\right] \tag{3.17}$$

$E[X \cdot x]$ is called as cross-correlation vector and it is indicated as r_x with size of $k \cdot 1$.

On the other hand, we can express the mean-square error as:

$$E[X \cdot X^T] = E\left[\begin{bmatrix} x_{n-1} \\ \vdots \\ x_{n-k} \end{bmatrix} . \begin{bmatrix} x_{n-1} \cdots x_{n-k} \end{bmatrix}\right]$$

$$= E\begin{bmatrix} x_{n-1} \cdot x_{n-1} & x_{n-1} \cdot x_{n-2} & \cdots & x_{n-1} \cdot x_{n-k} \\ x_{n-2} \cdot x_{n-1} & x_{n-2} \cdot x_{n-2} & \cdots & x_{n-2} \cdot x_{n-k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-1} \cdot x_{n-k} & x_{n-2} \cdot x_{n-k} & \cdots & x_{n-k} \cdot x_{n-k} \end{bmatrix}$$

which can be written as:

$$E[X \cdot X^T] = \begin{bmatrix} r_x(0) & r_x(1) & \cdots & r_x(k-1) \\ r_x(1) & r_x(0) & \cdots & r_x(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(k-1) & r_x(k-2) & \cdots & r_x(0) \end{bmatrix} \tag{3.18}$$

we call the matrix in (3.18) the auto-correlation matrix and indicated as Λ_x and with size $k \times k$ and as a result the mean-square error can be expressed as:

$$\xi = \sigma_x^2 - 2r_x^T \cdot H + H^T \cdot \Lambda_x \cdot H \tag{3.19}$$

The mean-square error achieve the minimum value when the derivatives with respect to all the values of H are zeros concurrently.

$$\frac{\partial \xi}{\partial H} = -2r_x + 2\Lambda_x \cdot H = 0 \tag{3.20}$$

$$H = \Lambda_x^{-1} \cdot r_x \tag{3.21}$$

H is a column vector with a size of $k \times 1$ [7][9]. The elements of H are the filter coefficients which are important to predict the samples at the output of the filter.

In the LTE system, the previous derivation is used to calculate the filter coefficients H. It can be noticed that Λ is a Toeplitz matrix and composed of k number of columns. It is possible to derived from the first column. The content of the first column are k input data samples in the LTE baseband signal after removing the CP.

The vector r_x is cross-correlation vector, which is similar to the first column in the correlation matrix. However it starts from the next sample which is $r_x(1)$. This derivation gives a straight forward method to calculate the filter coefficients that is used for the compression of the samples in LTE system.

Chapter 4

The Implementation of Linear Predictive coding (LPC) in LTE

In this chapter, the compression technique of linear predictive coding which is popular in speech coding systems is implemented in LTE. The main objective is to compress data samples of OFDM symbols in the BBU before transmitting them to the RRH in the downlink. In 20 MHz band only 1200 sub-carries are used in the middle of each OFDM symbol and rest of the other sub-carriers are zeros. This create an oversampling rate of $\frac{2048}{1200} \approx 1.7$ [11]. This oversampling leads to a redundancy in the time domain. As a result the estimation of the predicted sample has higher precision due to the correlation between each samples.

The statistical derivation of the filter coefficients is the main goal of this work. The idea behind this derivation is based on the statistical characteristic of the samples in the time domain. The goal is to derive a code book which contains the filter coefficients. The code book will be present both at the transmitter and the receiver module. The parameters of the code book are the reference channel, size of the filter and number of bits used at the quantizer. The filter coefficients are selected from the code book according to these parameters.

4.1 Implemented structure of LPC

The implemented structure of the LPC in our thesis is using a FIR filter as a part of it. The components which are used to do a linear estimation in the BBU and reconstructing the original signal at the RRH are shown in Figure 4.1. In the BBU, the linear predictive filter is using the filter coefficients, which are calculated in the previous section. The same coefficients are also stored at the RRH in the linear filter. At the initial stage, the filter contents both at the transmitter and the receiver are initialized to be zeros. During the transient period at the beginning of every OFDM symbol, The original data samples are passed into the filter at the transmitter side.The same samples are also transmitted and stored inside the filter at the receiver, in order to train the linear prediction filter at the receiver to use these samples for prediction purpose after the transient period. The receiver put them at the beginning of a new symbol without doing any filtering.

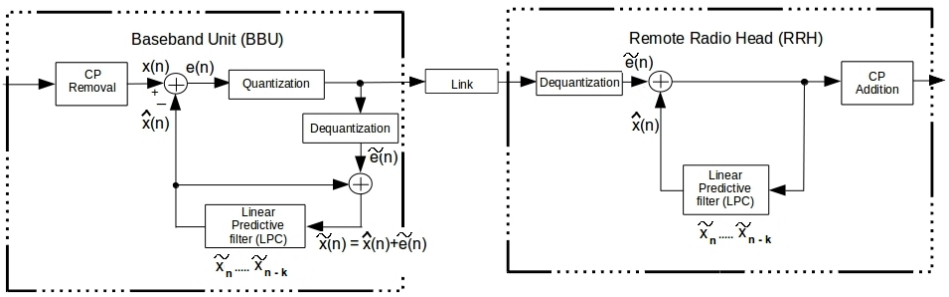


Figure 4.1: LPC structure.

After the transient period, the output of the filter is a discrete value of the next predicted sample $x(n)$ which is compared with the main sample $x(n)$ by the comparator. The output of the comparator is the difference between the main sample and the predicted sample. This difference between the two samples is denoted as estimated error sample. The function of the quantizer is to quantize the the error sample and then it is accumulated in the transmitted sample vector.

One copy of the quantized error sample is transmitted to the RRH and another copy is passed through the dequantizer and then added with the current predicted sample to get the same estimation at the receiver by reducing the weight of the error sample.

The transient period at the beginning of the prediction process, produces error with large amplitude. The amplitude of the error decreases and comes to a lower level as the process continues. This means we can use a quantizer with lower quantization level.

At the RRH, reversed operation is carried out to reconstruct the main samples by using the received error samples and added with the predicted sample. The filter is fed with the previous mentioned summation.

The issue of large amplitude of error samples at the initial transient stage also exist at the RRH and is discussed in a later chapter.

4.2 Statistical filter parameter for LPC implementation

The time domain waveform of LTE signal has a Gaussian distribution with zero mean and variance $\sigma^2 = 1$. As known, each LTE frame compose of 140 OFDM symbols. The distribution for each of the OFDM symbols is approximately the same. The statistical characteristic of the symbols permits to derive a common set of filter coefficients. These filter coefficients can be used to estimate the data samples with minimum prediction error.

The implementation of this idea has been accomplished by transforming each symbol in the time domain to frequency domain to get the power spectral density for each symbol in one LTE frame. Figure 4.2 shows one OFDM symbol in the frequency domain for the reference channel R.9, which means that the symbol has 2048 QAM samples. Then the mean of the power spectral for all the

symbols in multiple frames is calculated and the resulted calculation can be seen in Figure 4.3.

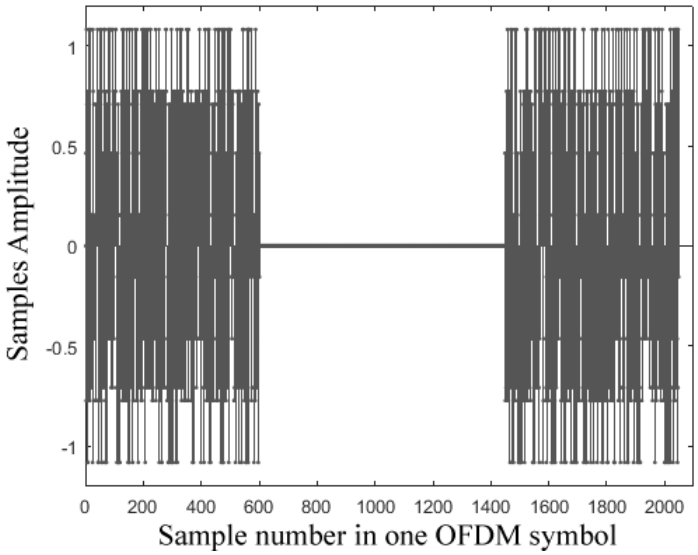


Figure 4.2: The spectrum of one OFDM symbol in the frequency domain.

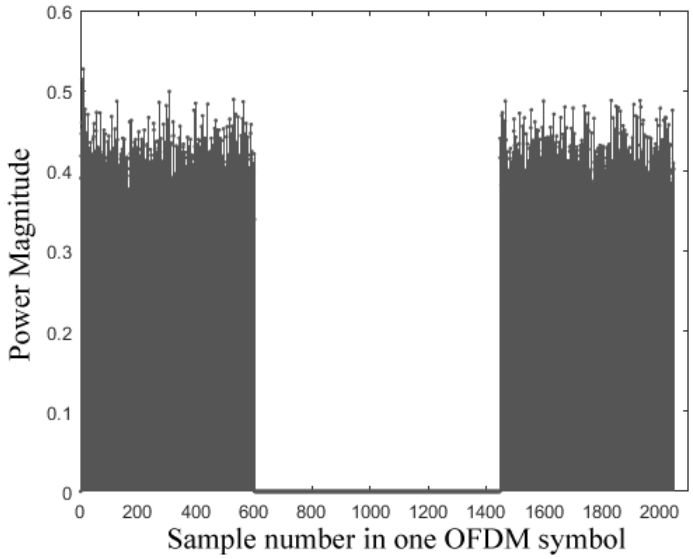


Figure 4.3: The mean spectrum of all OFDM symbols in one LTE frame.

Moreover, to have higher accuracy in the error estimation, the quantization error that has been introduced in the implemented structure is added and its

value depends on the number of bits used to quantize the error as discussed earlier in Section 2.4. The mean level of the spectrum is calculated excluding the zero samples in the middle. Figure 4.4 shows the spectrum of the mean OFDM symbols taking the quantization error in consideration. The resulted spectrum is transformed into the time domain and then filter coefficients are calculated using the Equation 3.21. The benefits of this work is to have a code book for the filter coefficients that can be selected based on the size of the filter, level of quantization used and the reference channel in use.

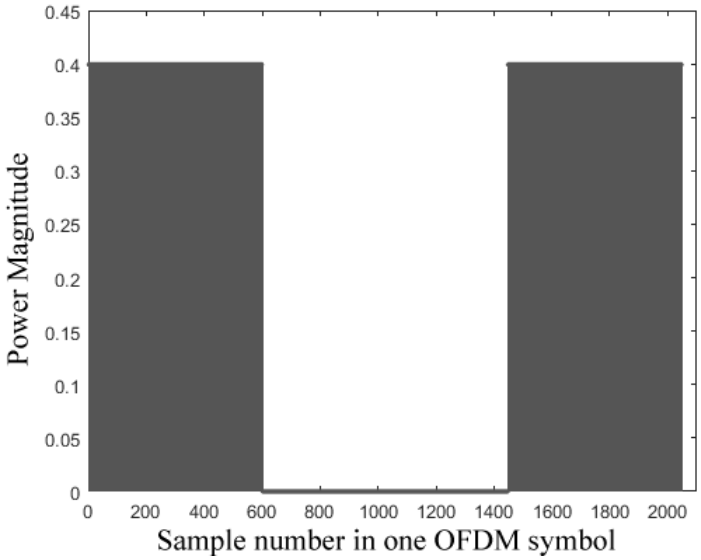


Figure 4.4: The spectrum of all OFDM Symbols including the qunatization error.

Chapter 5
Simulations

In the previous two chapters, the main idea of the linear prediction coding is introduced and the structure of the implemented linear prediction filter in LTE is presented. It is also shown how the filter coefficients are calculated. In this chapter, focus will be given on two simulation models to evaluate LPC for both, a known and unknown set of input data samples. For each of the models two types of experiments are performed. The first is to check the accuracy of the estimation by studying the Error Vector Magnitude (EVM) against the number of bits used in the quantizer for different size of filters and for different LTE bands. The reason is to study the impact of changing the quantiztion levels on the accuracy of the prediction. The second experiment is to study the relation between entropy against the size of the filter for different quantization level. The reason to do this experiment is to study the amount of redundancy after introducing LPC. In addition to check the redundancy for different size of filter, and observe the minimum amount of bits needed to represent one data sample for the same number of bits N used at the quantizer.

In a real world scenario, data samples in an LTE frame are not known. So a statistical model of the linear prediction estimation is more useful than a predetermined set of filter coefficients for known data input.

5.1 Simulation model

In this section the simulation model is presented for two cases. In the first model, linear predictive coding is implemented based on filter coefficients calculated for a known set of data input. In the second model, the linear predictive coding is implemented based on statistically derived filter coefficients for an unknown set data input.

Figure 5.1 shows the simulation model for the filter coefficients which are calculated based on known data input. In this model, random bits of data are generated and then placed in one LTE frame. The LTE frame is generated using the LTE toolbox provided in MATLAB to simulate the structure of an LTE frame. The reference channels (R.7 for 10 MHz, R.8 for 15 MHz and R.9 for 20 MHz) as stated in the LTE standards by 3GPP [5] are used by calling the function `lteRMCDLTool` from the LTE toolbox. The function uses the randomly generated data and load them into the LTE frame. The output of the LTE Toolbox is used as

an input to our proposed scheme to perform the compression at the transmitter side. The output of the LTE toolbox is a vector of complex time-domain samples of one LTE frame. The structure of an LTE frame is discussed earlier in Chapter 2. The number of complex time domain samples are different for each of the bands. In total, there are 140 OFDM symbols in one frame. Each of the OFDM symbols contain cyclic prefix which must be removed before the symbols are passed through the linear prediction filter. The reason for removal of cyclic prefix is to reduce the redundancy of transmitting samples of the cyclic prefix. Since samples between BBU and RRH are transmitted through physical links such as optic fiber or ethernet cable, there is no need for transmitting the cyclic prefixes which are normally used as a protection to avoid overlapping of OFDM symbols when air medium is used for communication.

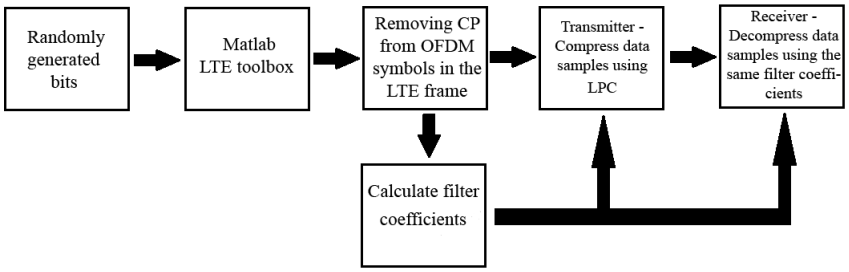


Figure 5.1: Simulation model for deterministic filter coefficients.

After removing the CP, the calculation of the filter coefficients has been performed according to Equation 3.9 in Section 3.3. The filter coefficients are fed into the transmitter and the receiver. This simulation model is used to check the validity of the prediction process in the transmitter and receiver when the filter coefficients are calculated based on known data input as explained in Section 3.3.

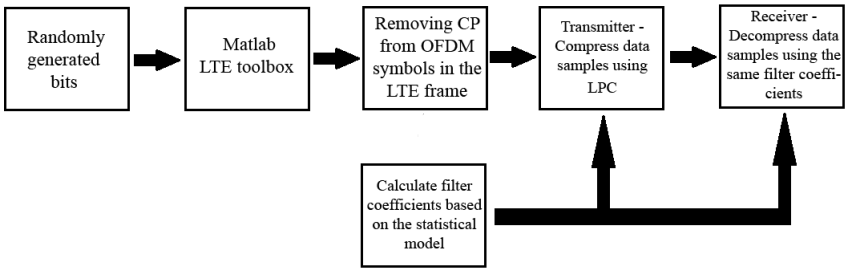
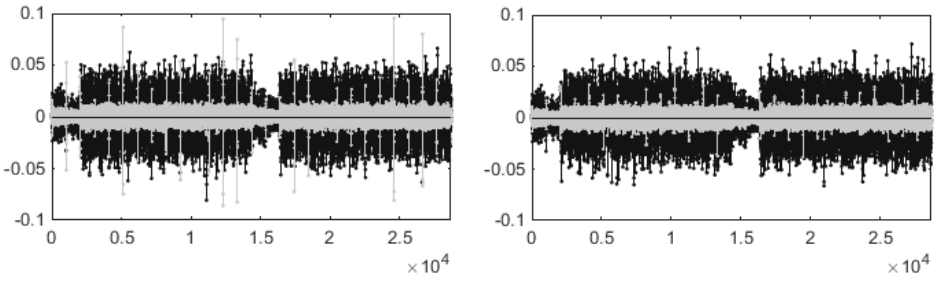


Figure 5.2: Simulation model for statistically calculated filter coefficients.

Figure 5.2 shows the second simulation model that has been implemented. It has almost the same structure of the first model, however the filter coefficients are calculated statistically as explained in Section 4.2. In this model, the filter

coefficients are calculated based on the three factors, type of reference channel, size of the filter and number of bits used in the quantizer. When the receiver and transmitter uses the same parameters, they will choose the same filter coefficients from the stored code book and use it.

5.2 Technical issue between the consecutive OFDM Symbols and the mitigation method



(a) LTE subframe with spikes. **(b)** LTE subframe without spikes.

Figure 5.3: 2 Subframes of 10 MHz band for the calculated filter parameters.

The OFDM symbols are constructed block-wise, and therefore they are assumed to be uncorrelated. This means that, when linear predictive coding is done over samples of a particular OFDM symbol, quantized error with large amplitude will be generated at the beginning of every consecutive OFDM symbols, as shown in Figure 5.3a [2]. It is a problem for the quantizer, because due to these spikes, the quantization range has to be large enough to take them into concideration. As a result the quantization level, using the same number of bits in the quantizer, has to be large. A bigger qunatization error is expected as the gap between the quantization level is getting larger.

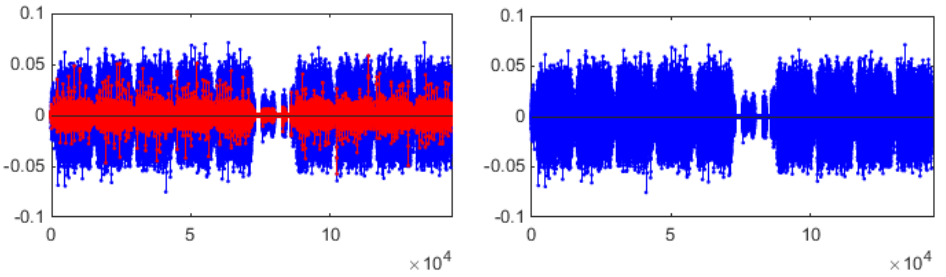
Most of the discrete error samples have low amplitude. So if the quantizer is designed having large gap between each of the levels, then there will occur higher quantization error. As a result, at the RRH the main samples will not be properly regenerated and it will produce a high EVM. So it is very important to remove these spikes from the quantized error samples before they are transmitted to the RRH.

There are many ways to deal with these spikes. In this thesis, a very simple technique is adopted to solve the problem of the spikes. Since the spikes appear repeatedly at the beginning of each of the symbols, so the index values of the position of the spikes are identified and they are replaced with the main samples of low amplitude as shown in Figure 5.3b. This technique reduces both the complexity of the system and the quantization error. Upon receiving the main samples which are replaced for the spikes, the filter at the RRH uses these samples to keep running the recovering process.

5.3 Implementing LPC based on generated filter parameters from known input data

In this section, the filter parameters have been derived based on known input data, as explained in Section 2.4. Moreover both the original, transmitted and the recovered signals are illustrated in the figures in this section. In addition the EVM and entropy calculations have been performed as will be described later.

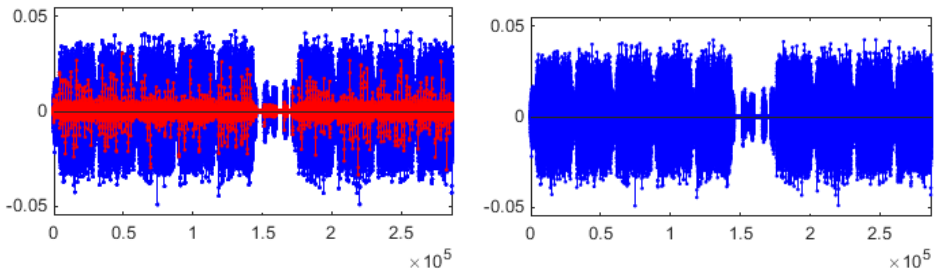
Figures 5.4, 5.5 and 5.6 show LTE frames for 10, 15 and 20 MHz band respectively. For the 10 MHz band, the size of FFT is 1024, and the total number of samples in a frame is 153600. After removing the CP of each symbol, the remaining number of samples in the LTE frame is 14360. For the 15 and 20 MHz bands, the size of FFT is 1536 and 2048 respectively according to the standard of the LTE by 3GPP [5]. But according to MATLAB, the size of FFT is the same for both 15 and 20 MHz bands, which is 2048. As a result, the total numbers of samples for both these bands are 307200 and after removing the CP the length of the LTE frame vector becomes 286720.



(a) — the transmitted signal
— the original signal at the BBU.

(b) Original samples recovered at the RRH.

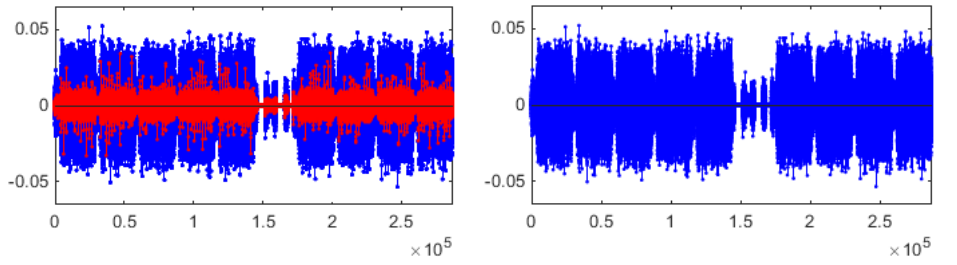
Figure 5.4: LTE frame for 10 MHz band.



(a) — the transmitted signal
— the original signal at the BBU.

(b) Original samples recovered at the RRH.

Figure 5.5: LTE frame for 15 MHz band.



(a) the transmitted signal; the original signal at the BBU.

(b) Original samples recovered at the RRH.

Figure 5.6: LTE frame for 20 MHz band.

In figures 5.4, 5.5 and 5.6, the blue samples at the BBU represents the original time-domain samples of the OFDM symbols and the red samples are the samples which are compressed using LPC. They are transmitted over the link to the RRH. It is quite clear that the implemented LPC achieves significant amount of compression by noticing that the amplitude of the compressed samples is much lower than the original samples. Lower the amplitude of a sample, less number of bits are needed to represent each sample. On the other hand, the samples which are recovered at the RRH are compared with the original samples at the BBU. The accuracy of this recovery will be presented in the next section by studying the EVM.

5.3.1 The simulation of the EVM against the number of bits used in the quantizer

Figure 5.7, 5.8 and 5.9 shows the Error Vector Magnitude (EVM) for 10, 15 and 20 MHz bands respectively with calculated filter parameters. The vertical axis shows the EVM in percentage versus the horizontal axis showing the number of bits (N) used in the quantizer.

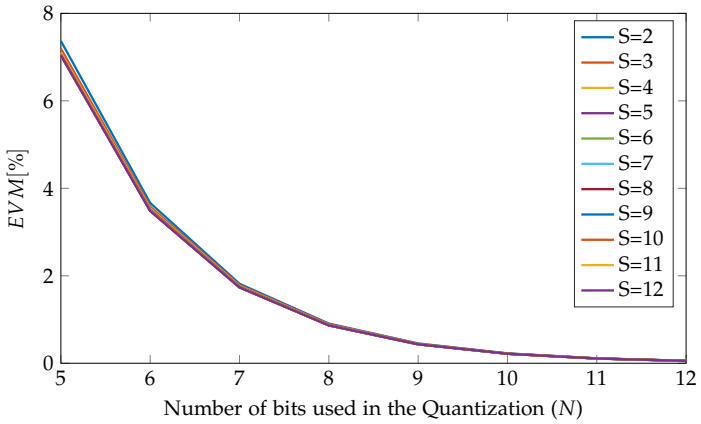


Figure 5.7: EVM for 10 MHz with calculated filter parameters.

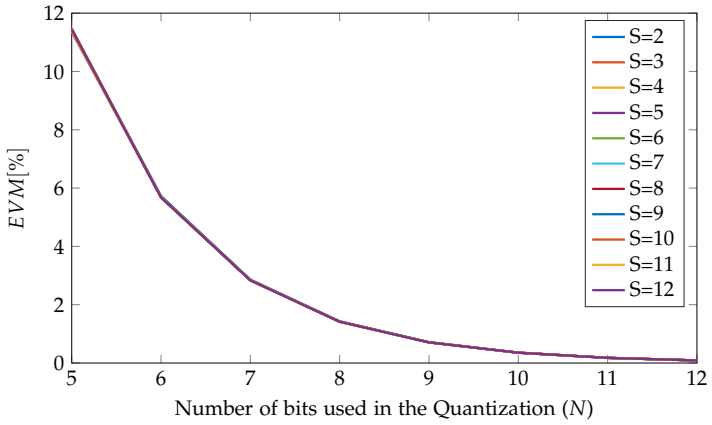


Figure 5.8: EVM for 15 MHz with calculated filter parameters.

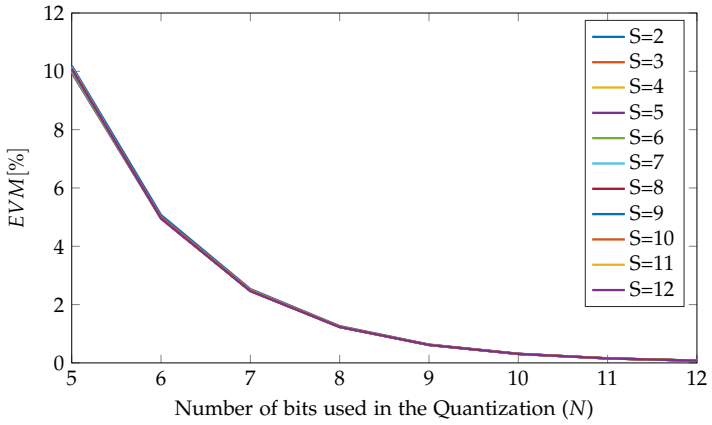


Figure 5.9: EVM for 20 MHz with calculated filter parameters.

For 10 MHz band in Figure 5.7, with a filter size of 2 to 12, when 5 bits are used in the quantizer, that is there are 32 levels of quantization, then the EVM is found to be a little more than 7%. According to the 3GPP standards of LTE, for 64-QAM the maximum EVM should be below 8% [5].

In this case, for the 10 MHz band, the EVM found for $N = 5$ bits, is within the limit stated in the standards. It is always desirable to keep the EVM value as low as possible so that there is minimum error between the transmitted and the received signal in the BBU to the RRH respectively. As a result, it is important to increase the number of bits in the quantizer to reduce the EVM value. When N is increased, the EVM value decreases exponentially.

Similar behavior can be observed for 15 and 20 MHz bands in Figure 5.8 and 5.9 respectively. For the 15 MHz band the EVM is slightly more than 11% when N=5 bits. And for 20 MHz band, it is approximately 10%. So for 15 and 20 MHz bands, a quantizer with 5 bits cannot be used as it exceeds the value stated in the LTE standard. When $N = 6$ bits, the EVM is 5.7% and 5% for 15 and 20 MHz respectively and further decreases with increasing the number of levels in the

quantizer.

5.3.2 Entropy calculation against the filter size for different quantization level for calculated filter coefficients based on known input data

The entropy is an indicator to show how much compression can be done on a data vector. It represents how much redundancy can be removed from a data vector after some sort of coding i.e. Huffman coding, is carried out for data compression. Even though no such coding is done in this thesis, only the entropy is calculated for all the three bands to show that by applying such coding, the number of quantization bits can be reduced to a safe level that will not affect much to the recovered signal at the RRH.

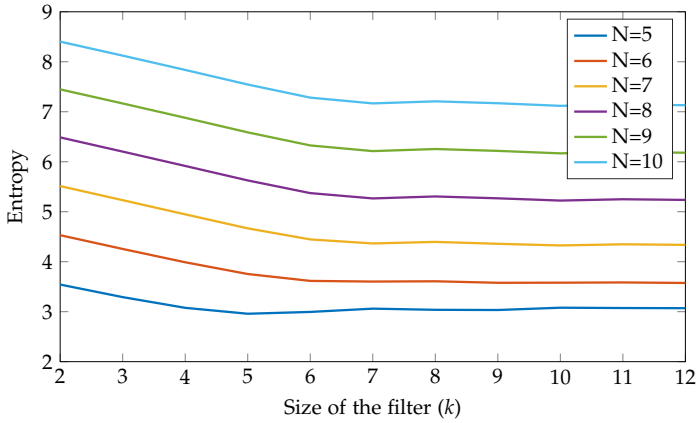


Figure 5.10: Entropy for 10 MHz with calculated filter parameters with different filter size.

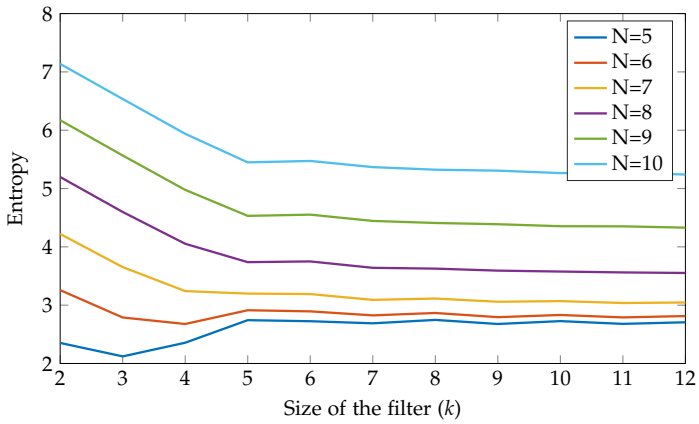


Figure 5.11: Entropy for 15 MHz with calculated filter parameters with different filter size.

Figure 5.10, 5.11 and 5.12 shows the EVM versus size of the filter for 10, 15 and 20 MHz bands, respectively for different levels of quantization (N). It is

found that, for 10 MHz band as shown in Figure 5.10, the behavior of the graphs of entropy is the same for all N values from 6 to 10. There is approximately 1 bit gap in the entropy value between each successive N.

If 7 bits is used in the quantizer, then the entropy level out at 4.4 bits/sample when the size of the filter have a value of 6 and higher.

Regarding the 15 MHz band in Figure 5.11, it is observed that the behavior of the entropy graphs for $N = 7$ bits and more are similar to each other. Whereas for N= 5 and 6 bits, it is showing an abnormal behavior in the entropy when the filter size is less than 5.

An in depth investigation has been done to study the entropy values for these graphs which have different characteristic and the others. This abnormal behavior is due to the low number of bits used in the quantizer. This leads to an excess error, which makes the dequantization inaccurate. Hence this error propagates and leads to poor estimation.

As the filter size increases, the estimation by the filter tends to get worse. And as a result, the transmitted signal to the RRH, which represents the error vector will have a higher amplitude due to bad estimation and the validity of this system will be unreliable.

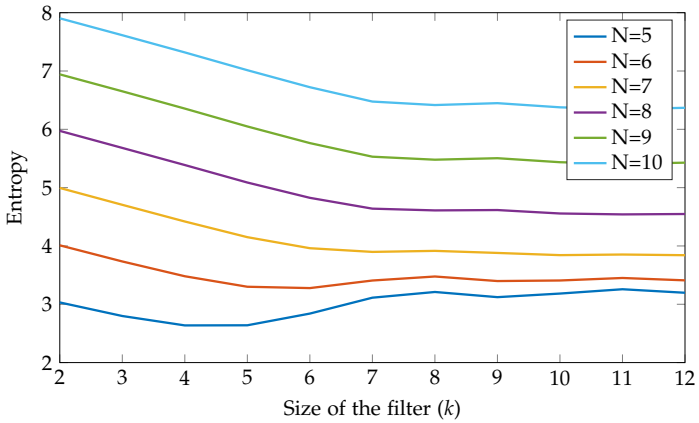


Figure 5.12: Entropy for 20 MHz with calculated filter parameters with different filter size.

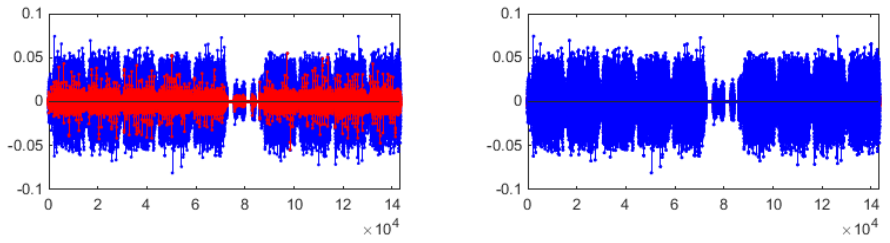
For more details about this, the reader is suggested to check the Appendix to follow the figures that shows the accuracy of the estimation for different number of bits used in the quantizer with varying size of the filter, where three subframes of the transmitted signal along with the original signal are plotted. From those figures for the case of $N = 5$ bits, it can understood that when the size of the filter is low i.e. 2 and 3, is acceptable. But as the size of the filter increases, the prediction is getting worse and it loses its validity. This is because, when the number of bits in the quantizer is kept small but the size of the filter is increased, the input of the filter contains more number of error values as can be seen from Figure 4.1. Hence, the magnitude of each of the error values produced will be larger. And as a result, the entropy values which are obtained, are not reliable any more.

On the other hand, for higher number of bits used in the quantizer, the estimation is getting better with increasing the size of the filter. Similar behavior has been observed regarding 20 MHz band in Figure 5.12. For the case of N=7 bits, the entropy value levels out at 3.8 bit/sample when the size of the filter is 6 or more, which means that 7 bits can be reduced to 3.8 bit/sample, if some sort of compression coding is performed as mentioned earlier.

5.4 Implementing LPC based on statistically generated filter parameters

In this section the filter coefficients that has been derived based on the derivation in section 3.6 using the statistical model is used to evaluated the accuracy of the statistical model to reproduce the transmitted signal at the receiver. In addition, the EVM has been calculated against the number of bits which has been used at the quantizer and entropy against the size of the filter.

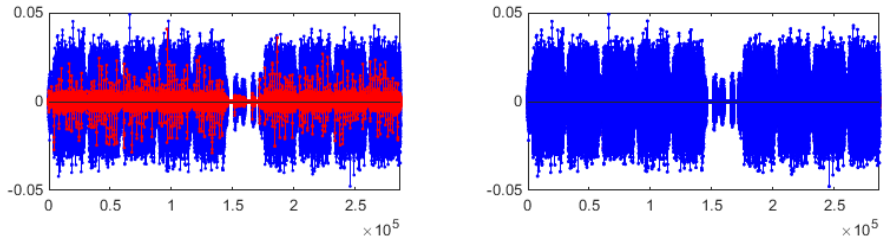
Figures 5.13, 5.14 and 5.15 show the LTE frame for 10, 15 and 20 MHz band respectively with filter size of 10 and 8 bits are used.



(a) — the transmitted signal
— the original signal at the BBU.

(b) Original samples regenerated at the RRH.

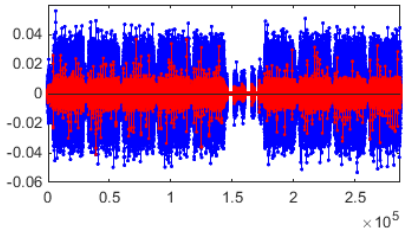
Figure 5.13: LTE frame for 10 MHz band for statistically calculated filter parameters.



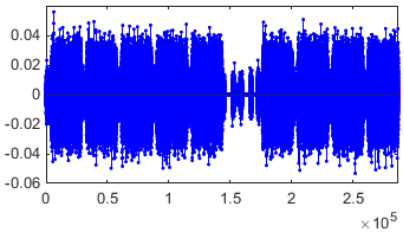
(a) — the transmitted signal
— the original signal at the BBU.

(b) Original samples regenerated at the RRH.

Figure 5.14: LTE frame for 15 MHz band for statistically calculated filter parameters.



(a) — the transmitted signal
— the original signal at the BBU.



(b) Original samples regenerated at the RRH.

Figure 5.15: LTE frame for 20 MHz band for statistically calculated filter parameters.

As described earlier in section 4.2, the filter parameters are calculated based on the statistical model. It is clear that the compression of the original signal is taking effect. This can be observed by checking the red samples which are the transmitted signal comparing to the RRH and the blue original signal. EVM calculation in the next section gives more details about the accuracy of this compression.

5.4.1 The simulation of the EVM against the quantization level based on the statistical model

The EVM based on statistically calculated filter coefficients for the bands 10 MHz, 15 MHz and 20 MHz is shown Figure 5.16, 5.17 and 5.17 respectively.

Figure 5.16 shows the EVM against the number of bits used in quantizing the transmitted samples for different size of filters for 10 MHz band. It is clear that the performance for different size of filters in term of the size of the EVM is close to be identical, specifically when high number of bits are used in the quantizer. The EVM start with approximated value of 7% and decrease exponentially to reach approximately 1% or less when 8 bits or more is used in the quantizer.

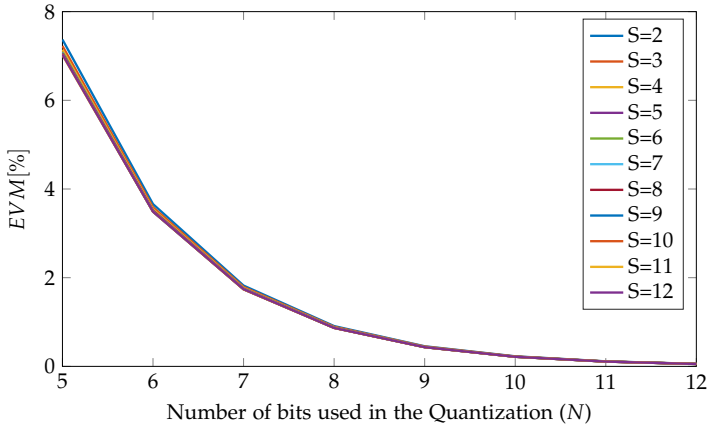


Figure 5.16: EVM for 10 MHz with statistical filter parameters.

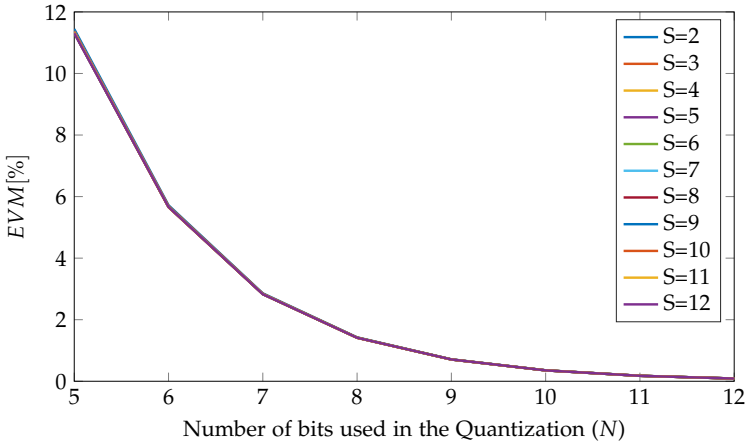


Figure 5.17: EVM for 15 MHz with statistical filter parameters.

Figures 5.17 and 5.18 shows the EVM against the number of bits used in quantizing the transmitted samples for different size of filters for 15 MHz and 20 MHz band. The EVM have the same exponential decrease in its value against the number of bits used in quantization but it starts from 11% for 15 MHz and 10% for 20 MHz. It reaches approximately 1% or less when 8 bits or more is used in the quantizer.

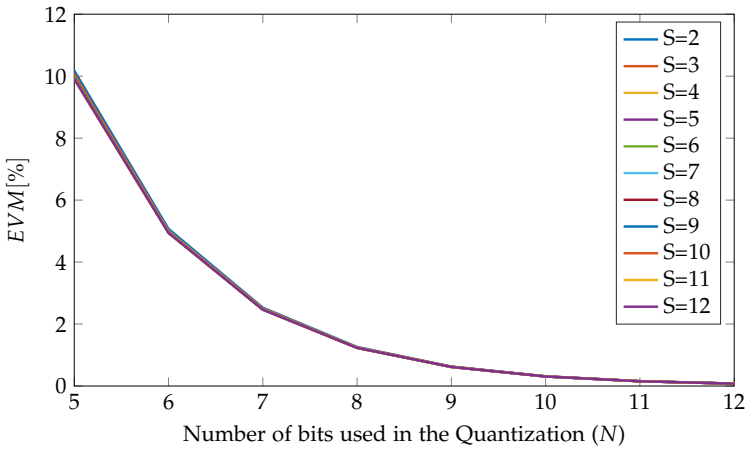


Figure 5.18: EVM for 20 MHz with statistical filter parameters.

5.4.2 Entropy calculation against the filter size for different quantization level for statistically calculated filter coefficients

The calculation of entropy against the size of the filter based on the statistical model is performed for the 10, 15 and 20 MHz bands. Figure 5.19, 5.20 and 5.21 show the entropy for previously mentioned bands respectively.

Figure 5.19 shows the entropy calculation results of 10 MHz band for different number of bits used in the quantizer against the size of the filter. It can be seen that, the entropy has the similar behavior. And it is noticed that for $N = 7$ bits, the entropy value levels out at 3.6 bit/sample when size of the filter is 6 and more.

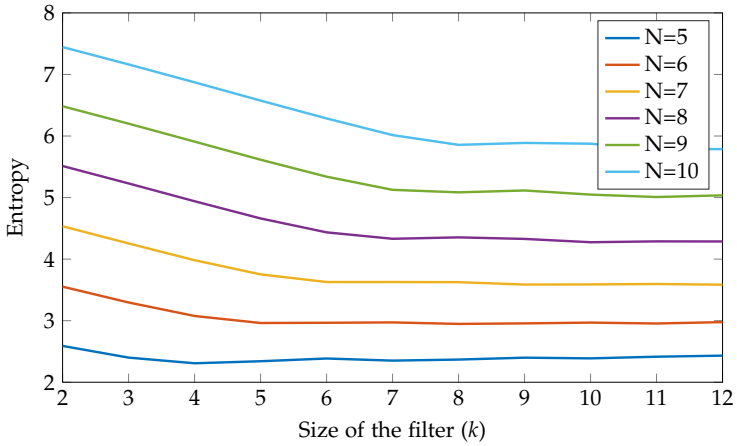


Figure 5.19: Entropy for 10 MHz with statistical filter parameters with different filter size.

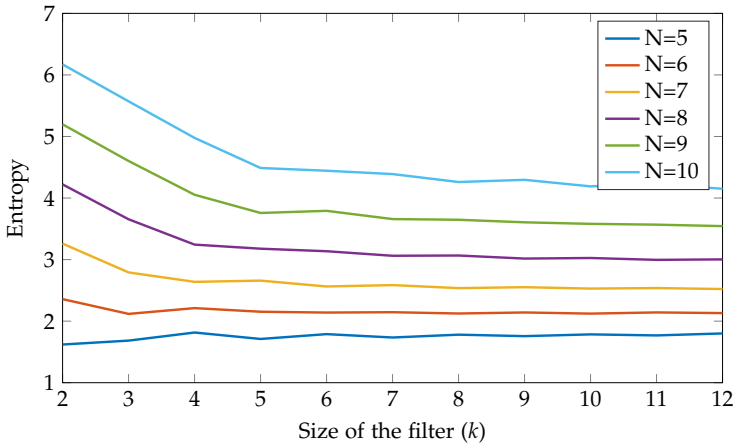


Figure 5.20: Entropy for 15 MHz with statistical filter parameters with different filter size.

Regarding 15 MHz band, it is clear that the entropy curves has similar tendency for $N = 7$ bits and more, and for lower number N values, it has abnormal pattern and the reason is has been already discussed in Section 4.2.2. It can be seen from Figure 5.20 that for $N = 7$ bits, the entropy value saturates at 2.5 bit/sample.

Regarding 20 MHz band which is illustrated in Figure 5.21, it can be seen that when $N = 7$ bits, the entropy value saturates at 3.3 bits/sample.

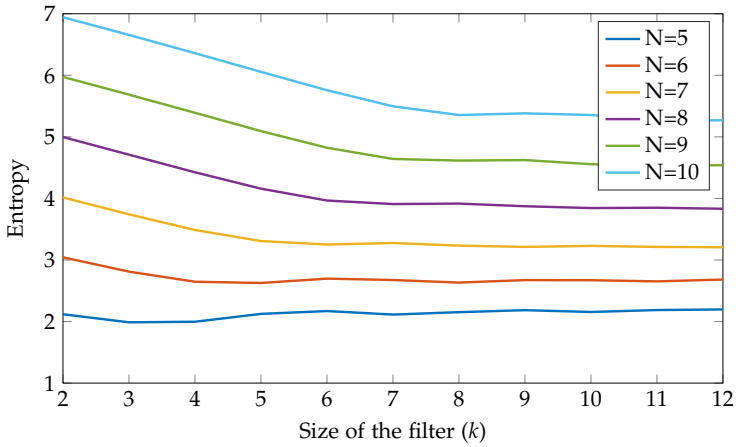


Figure 5.21: Entropy for 20 MHz with statistical filter parameters with different filter size.

5.5 The impact of using only the real part of the data signal

The simulations in this work has been performed only on the real values of the main signal to reduce the required time to run the simulations. To investigate the effect of not taking into consideration the imaginary values, we ran one simulation for 10 MHz, 64-QAM with filter coefficients calculated based on a known data input taking in consideration both the real and the imaginary part. It is found that, by excluding the imaginary part does not affect much the EVM calculation as shown in Figure 5.22 comparing with Figure 5.7.

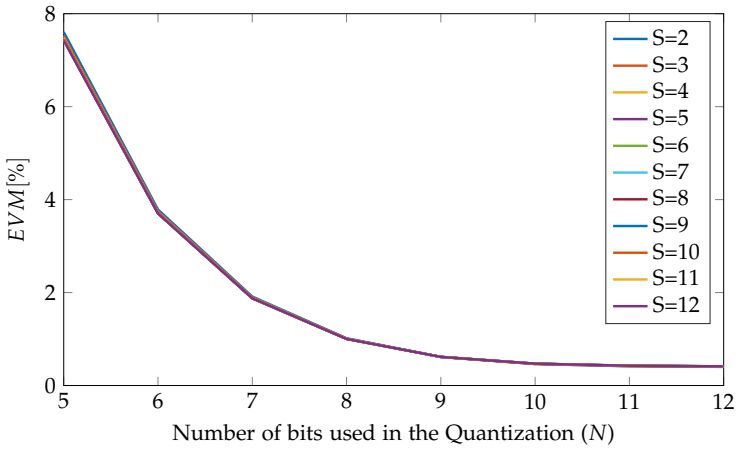


Figure 5.22: EVM for 10 MHz with complex signal for statistical filter parameters.

Chapter 6
Conclusions

In this chapter, a comparison is presented based on the simulation results from the previous chapter. The main purpose of this thesis work is summarized and some possible future research work is suggested.

6.1 Summary

The purpose of this thesis work was to compress the data between the BBU and the RRH in the LTE fronthaul to tackle the main issue of the massive increase in data traffic. This challenge will be one of the major obstacles when massive MIMO and beamforming is introduced in the upcoming network access technologies. The main idea behind this work is to introduce the linear predictive coding to compress the data so that less number of bits can be used to represent each transmitted sample from the BBU, and to be recovered with a minimum EVM value at the RRH meeting the 3GPP standard.

To be able to study the accuracy of linear predictive coding, the EVM is calculated versus for different number of bits used in the quantizer with different size of filter. The entropy is also studied to investigate the possibility of additional compression when Huffman or similar type of coding is applied.

In the previous chapter, the relation between EVM and the number of bits in the quatizer has been analyzed for two different scenarios. In the first scenario, the filter coefficients are calculated for a known data input. For the second scenario, the filter coefficients are statistically calculated. And it has been found that the EVM results for both the scenarios are similar for each of the studied bands (10, 15 and 20 MHz) with different size of filters. This can be interpreted as, the EVM is dependent on the number of bits used in the quantizer. When 7 bits are used in the quantizer, then the EVM is around 2%, 2.8% and 2.5% for 10, 15 and 20 MHz bands respectively as tabulated in Table 6.1.

On the other hand, the relation between the entropy and the size of the filter has been analyzed for the same previous two scenarios as mentioned earlier. When the number of bits is 7 with a filter size of 7, the minimum entropy value achieved was 4.2 bits/samples, 3.05 bits/sample and 3.85 bit/samples for 10, 15 and 20 MHz band with filter coefficients calculated based on known data input. However, it is found that for statistically calculated filter coefficients entropy value were 3.62 bits/sample, 2.58 bits/sample and 3.27 bits/sample for 10, 15

Table 6.1: Simulation results for different reference channels with $N = 7$ bits used in the quantizer.

Reference channel Number	**EVM** [%]	**Entropy (known input data)** [*bit/sample*]	**Entropy (unknown input data)** [*bit/sample*]
R.7	2	4.2	3.62
R.8	2.8	3.05	2.58
R.9	2.5	3.85	3.27

and 20 MHz band respectively. It is quite clear that the statistical model is better than the other scenario. The reason is that the calculation of the filter coefficients for the case of known data input does not take into consideration the effect of abnormal behavior at the beginning of each symbol due to the uncorrelation between them. Whereas for the statistical model the filter parameters are calculated by considering the mean of power spectral density of all the symbols in the frame which reduces the effect of the uncorrelation at the beginning of each symbol.

It should be noted here that introducing the proposed compression scheme of linear predictive coding in LTE does not violate the 3GPP standards and it can be used as an adaptation at the output of the BBU and the input of the RRH in the current LTE system. The derived filter coefficients in this work has achieved a tremendous compression gain. The compression leads to a representation of a complex sample using 2*3.27 = 6.54 bits/sample on average for 20 MHz. Whereas in the current LTE systems, each complex sample is represented by 30 bits. This makes the implemented compression scheme attractive for future telecommunication systems and interesting to be tested on the existing systems.

6.2 Further research work

During the work on linear predictive coding in LTE fronthaul, a lot of issues have been identified and they can be open scopes for future research, considering that the application of LPC is new in this field. Some of the open issues that can be studied are discussed in this section.

One of the important research topic can be finding another solution to solve the uncorrelation issue at the beginning of each of the predicted symbols, study the prediction efficiency and compare it with this method applied in this work. Also to check the impact of the new method on the entropy especially for low number of bits used in the quantizer to achieve higher compression.

Another interesting research issue can be to apply LPC on multiple antenna system and study the outcomes with respect to the EVM and entropy. On the other hand, the goal of this work focused on the downlink in fronthaul. Another crucial research can be to study the validity in the uplink. Further research topic can be to choose appropriate coding methods to achieve higher level of compression.

References

[1] B. S. Atal. The history of linear prediction. *IEEE Signal Processing Magazine*, 23(2):154–161, March 2006.

[2] M.M. da Silva. *Cable and Wireless Networks: Theory and Practice*. CRC Press, 2016.

[3] Erik Dahlman, Stefan Parkvall, and Johan Skold. *4G: LTE/LTE-advanced for mobile broadband*. Academic press, 2013.

[4] A. de la Oliva, J. A. Hernandez, D. Larrabeiti, and A. Azcorra. An overview of the cpri specification and its application to c-ran-based lte scenarios. *IEEE Communications Magazine*, 54(2):152–159, February 2016.

[5] TS ETSI. 136 101 v13. 2.1 (3gpp ts 36.101 v13. 2.1 release 13),". *LTE: Evolved Universal Terrestrial Radio Access (E-UTRA)*.

[6] B. Guo, W. Cao, A. Tao, and D. Samardzija. Lte/lte-a signal compression on the cpri interface. *Bell Labs Technical Journal*, 18(2):117–133, Sept 2013.

[7] Simon S Haykin. *Modern filters*. Macmillan Coll Division, 1989.

[8] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[9] Tokunbo Ogunfunmi and Madihally Narasimha. *Principles of speech coding*. CRC Press, 2010.

[10] M. Peng, C. Wang, V. Lau, and H. V. Poor. Fronthaul-constrained cloud radio access networks: insights and challenges. *IEEE Wireless Communications*, 22(2):152–160, April 2015.

[11] Leonardo Ramalho, Maria Nilma Fonseca, Aldebaro Klautau, Chenguang Lu, Miguel Berg, Elmar Trojer, and Stefan Höst. An lpc-based fronthaul compression scheme. *Submitted to IEEE Communications Magazine*, 2016.

[12] Stefania Sesia, Issam Toufik, and Matthew Baker. *LTE-the UMTS long term evolution*. Wiley Online Library, 2015.

[13] CE Shannon and W Weaver. A mathematical model of communication urbana, 1949.

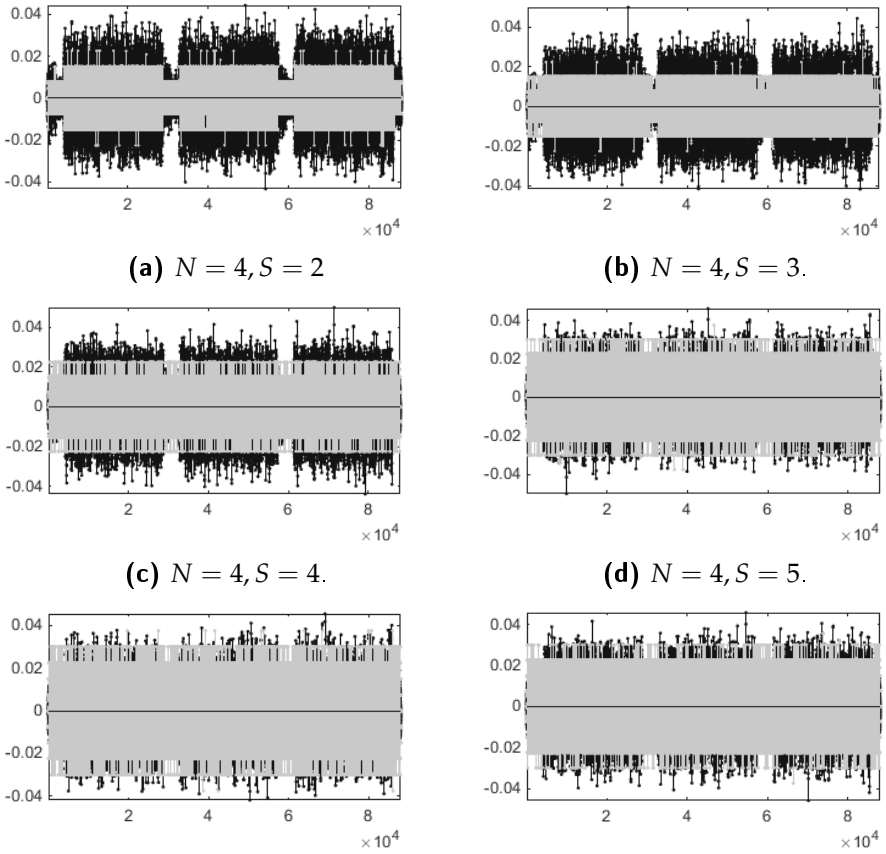
[14] Yun Q Shi and Huifang Sun. *Image and video compression for multimedia engineering: Fundamentals, algorithms, and standards*. CRC press, 1999.

[15] Sesia Stefania, Toufik Issam, and Baker Matthew. Lte-the umts long term evolution: from theory to practice. *A John Wiley and Sons, Ltd*, 6:136–144, 2009.

[16] PP Vaidyanathan. *The theory of linear prediction*, volume 2. Morgan & Claypool Publishers, 2007.

Appendix:

The error estimation when LPC implemented with four bits used in the quantization



(a) $N = 4, S = 2$

(b) $N = 4, S = 3$.

(c) $N = 4, S = 4$.

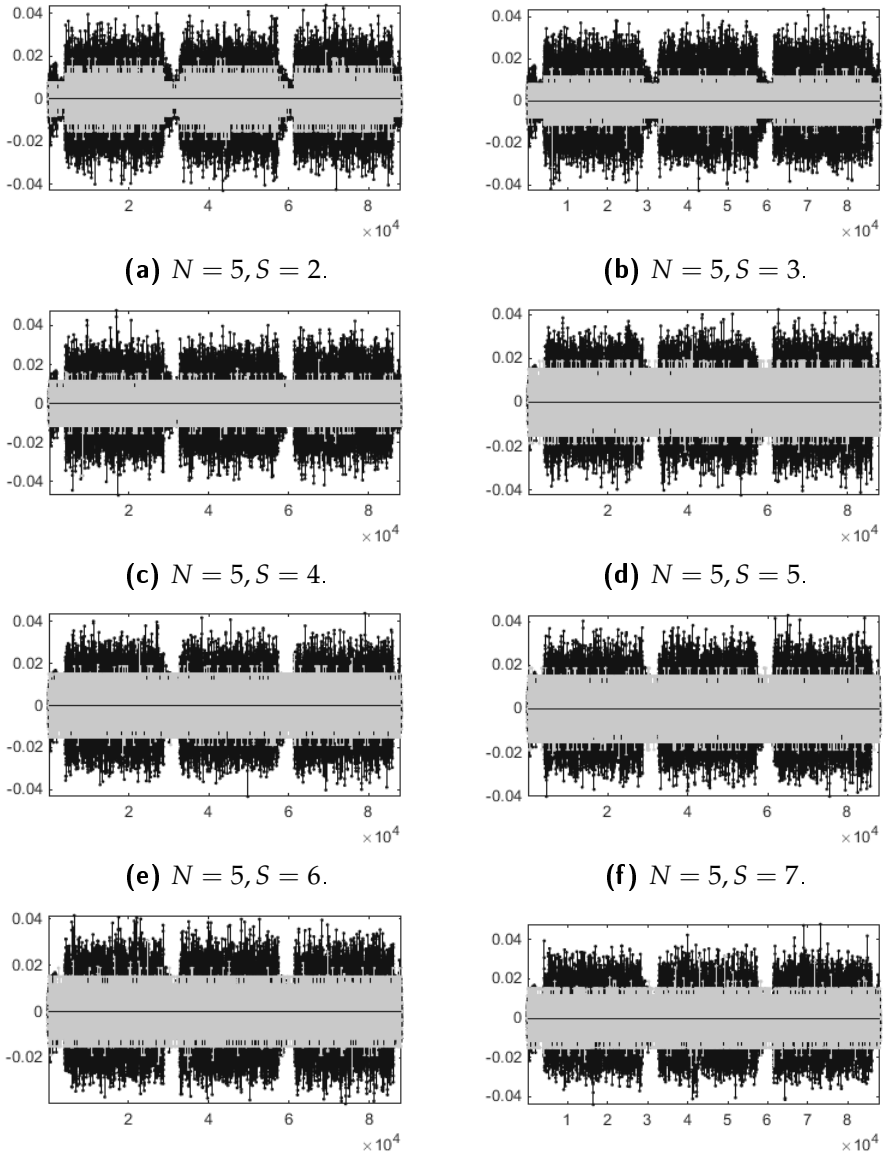
(d) $N = 4, S = 5$.

(e) $N = 4, S = 6$.

(f) $N = 4, S = 9$.

Figure 6.1: 3 subframes of 15 MHz band for $N = 4$ bits with different size of filter.

The error estimation when LPC implemented with five bits used in the quantization



(a) $N = 5, S = 2$. **(b)** $N = 5, S = 3$.

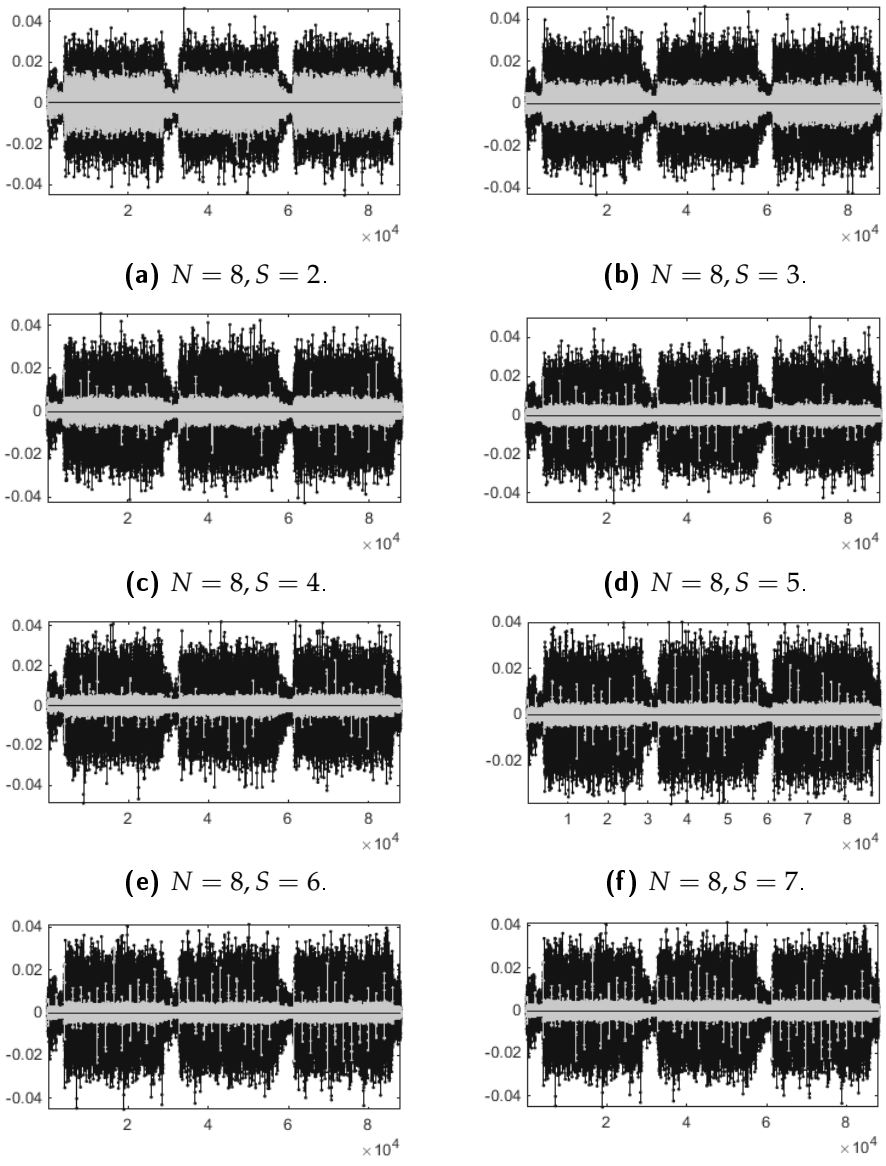
(c) $N = 5, S = 4$. **(d)** $N = 5, S = 5$.

(e) $N = 5, S = 6$. **(f)** $N = 5, S = 7$.

(g) $N = 5, S = 9$. **(h)** $N = 5, S = 11$.

Figure 6.2: 3 subframes of 15 MHz band for $N = 5$ bits with different size of filter.

The error estimation when LPC implemented with eight bits used in the quantization



(a) $N = 8, S = 2$.

(b) $N = 8, S = 3$.

(c) $N = 8, S = 4$.

(d) $N = 8, S = 5$.

(e) $N = 8, S = 6$.

(f) $N = 8, S = 7$.

(g) $N = 8, S = 9$.

(h) $N = 8, S = 11$.

Figure 6.3: 3 subframes of 15 MHz band for $N = 8$ bits with different size of filter.

LUND
UNIVERSITY

Series of Master's theses
Department of Electrical and Information Technology
LU/LTH-EIT 2016-552 http://www.eit.lth.se