On the Practicability of Full-Duplex Relaying in OFDM Systems

Ujjwal Prakash Dhamala

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo 18.08.2011

Supervisor

Professor Risto Wichman

Instructor

Taneli Riihonen, M.Sc. (Tech.)

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Ujjwal Prakash Dhamala

Title: On the Practicability of Full-Duplex Relaying in OFDM Systems

Date: 18.08.2011 Language: English Number of pages: 4+62

Department of Signal Processing and Acoustics

Professorship: Signal Processing Code: S-88

Supervisor: Professor Risto Wichman

Instructor: Taneli Riihonen, M.Sc. (Tech.)

A full-duplex relay is spectrally more efficient than a half-duplex relay because it uses the full band of available frequencies to receive and transmit signals simultaneously. However, the loop interference arriving at the receiver of the relay due to its own transmission is a major hindrance that must be overcome before the idea of full-duplex relaying can be put to practice. The simple technique of subtractive cancellation alone, in theory, could eliminate the loop interference completely from the received signal. In practice, however, the nonidealities inherent in the actual components within the relay transceivers create less than ideal conditions for the cancellation to work perfectly.

This thesis studies the effect of such nonidealities on the performance of a single-input-single-output (SISO) full-duplex relay. The primary focus is on formulating an analytical framework that helps evaluate the feasibility of such a relay. The outcome illustrates that a number of factors determine whether the idea of a full-duplex relay with subtractive loop interference cancellation can be implemented in practice. As expected, it is necessary to have an analog-to-digital converter (ADC) with a large dynamic range at the receiver to ensure that the incoming signal can be digitized with sufficient accuracy. Another important requirement is to have an excellent transmitter with a very small error vector magnitude (EVM) because the contribution of the unknown random error in the transmitted signal to the loop interference cannot be canceled no matter how accurately the incoming signal is digitized. Moreover, the physical design of the relay must, by itself, be able to provide a certain amount of natural isolation between the transmitting and receiving antennas; otherwise, the part of the loop interference resulting from the transmitter error alone can be sufficient to drown the useful signal beyond recovery.

Keywords: full-duplex relay, loop interference cancellation, transceiver nonidealities, analog-to-digital conversion, quantization, dynamic range, error vector magnitude, OFDM

Acknowledgements

The research presented in this thesis has been carried out at the Department of Signal Processing and Acoustics, Aalto University. I would like to express my gratitude to Prof. Risto Wichman, the supervisor of this thesis, for providing me the opportunity to work as a member of the Signal Processing group in the department – a vibrant research group that participates in SMARAD (Smart Radios and Wireless Research), a centre of excellence nominated by the Academy of Finland. I am very grateful to Prof. Wichman for his guidance, support, and kind encouragement throughout the duration of my research.

I am greatly indebted to Taneli Riihonen, the instructor of this thesis, for the time and effort that he has contributed so generously to help me improve the quality of my research and for his encouraging comments that have kept me motivated to do my best.

Finally, I would like to thank my family and friends for their support and encouragement.

Espoo 18.08.2011

Ujjwal Prakash Dhamala

Contents

Chapter 1

Introduction

Relaying, in communication systems, refers to the idea of using an intermediate node positioned somewhere between primary communicating nodes (the source and the destination) to facilitate the transfer of message-carrying signals between them. The intermediate node is referred to as the relay node, and its job is to receive incoming signals from the source node, do the necessary processing, and retransmit them so that the communication between the primary nodes, which might have otherwise failed, becomes successful.

In the context of wireless communications, the term relaying, more often than not, refers to half-duplex relaying, which means that the relay node, at a given time, either receives from the source node or transmits to the destination node using the full band of frequencies available (and quite possibly a single omnidirectional antenna), but does not do both simultaneously[1] so as to avoid interference.

As good as it may seem, half-duplex relaying suffers from two major drawbacks, both of which arise from the need for allocating distinct time slots for reception and transmission by the relay. First, there is a significant loss of throughput compared to that which would be possible if the relay were not required at all. Second, proper scheduling and time-synchronizing have to be enforced among all the three communicating entities involved, which adds significant complexity to the overall system.

Full-duplex relaying, which is a relatively new idea in the wireless context, is one promising approach that aims at mitigating the drawbacks associated with half-duplex relaying, especially the one concerning the loss of throughput since it also translates to loss of spectral efficiency. Basically, a full-duplex relay refers to the kind of relay which uses the full bandwidth to receive and transmit signals simultaneously, thereby causing no loss in throughput.

However, the benefit offered by full-duplex relaying does not come free of cost. Since the relay both receives from the source node and transmits to the destination node using the same band of frequencies (but separate antennas) at the same time, the interference caused by the relay's transmission on its own

[1] More generally, half-duplex relaying also includes schemes in which the relay receives and transmits simultaneously using different frequencies.

received signal could easily disrupt the communication between the primary nodes. This self-interference, commonly referred to as the loop interference, is the major setback to the appeal of full-duplex relaying.

Signal processing theory, nevertheless, provides a simple yet elegant solution to the problem of loop interference: since it is entirely possible for the relay to store the signal that it has transmitted, it can compute a reasonable estimate of the loop interference reaching its receiving antenna and subtract it from the total received signal in order to have a close replica of just the desired signal arriving from the source node.

1.1 Motivation

The solution just described is merely one way of suppressing the effect of loop interference; nevertheless, it is probably the simplest one around and has an additional advantage that it works entirely in the time domain. In theory, this method allows a perfect cancellation of the loop interference no matter how strong the interference is and hence offers the possibility of infinite isolation between the transmitting and the receiving antennas in the relay. In practice, however, perfect cancellation of the loop interference is next to impossible because the actual components that make the receiver and the transmitter in the relay create less than ideal conditions for the subtractive cancellation technique to work perfectly.

The first nonideality that comes into play is the finite word-length of the quantizer within the analog-to-digital converter at the receiver in the relay. Since any practical quantizer can provide only a finite number of quantization levels, it cannot support an unlimited dynamic range that allows accurate representation of all the signal components, possibly with large power differences, superimposed within the arriving signal. And because the loop interference can be much stronger than the useful signal coming from the source node, the limited dynamic range featured by the quantizer might cause the useful signal to be drowned (in the quantization noise) beyond recovery, which, in turn, leads to complete failure of the subtractive cancellation technique (or any other cancellation technique for that matter).

Even when the useful signal does not get drowned completely, the loss of information due to quantization alone causes the estimate of the channel between the transmitting and the receiving antennas of the relay to be less than perfect. This, in turn, causes the estimate of the loop interference to be imperfect and the subtractive cancellation becomes imperfect as well. This is further aggravated by the fact that the various nonidealities present in the transmitter side add unknown random errors to the signal supposed to be transmitted by the relay, thereby causing the channel estimation error to increase and the effectiveness of subtractive cancellation to decrease.

The motivation for this thesis comes from the lack of studies that sufficiently consider the effect of the various nonidealities inherent in all practical systems on

the performance of full-duplex relays with loop interference cancellation. Most of the literature available on full-duplex relaying, while being of tremendous help in pointing out interesting theoretical ideas applicable to mitigating the effect of loop interference, simply fail to address the aforementioned issues of limited dynamic range and imperfect channel estimation. The research in this thesis aims to help bridge these gaps to some extent while trying to answer the question of practicability – whether or not can the idea of loop interference cancellation, which is essential for the functioning of full-duplex relays, be put to actual practice.

1.2 Brief Survey on Full-Duplex Relaying

The theoretical basis for full-duplex relaying has been long established by studies centered on the information-theoretic aspects (such as capacity bounds) of the classic three-node relay channel [1,2]. While these early studies did not consider specific transmission mediums, their results have later been extended for the three-node wireless relay channel in [3] and for a full-fledged wireless network with multiple relay nodes having full-duplex capability in [4]. Such studies do illustrate the theoretical advantage of full-duplex relaying over half-duplex relaying in terms of spectral efficiency; however they tend to assume ideal conditions of operation and do not give due consideration to the issue of loop interference at the relay receiver.

On a more practical level, studies on full-duplex relaying that are based on actual measurements of the useful signal and the loop interference at the receiving antenna of the relay (e.g., [5,6]) have acknowledged the issue of loop interference and discussed the difficulty of overcoming its effect. Some other studies of similar nature (e.g., [7–9]) have considered the benefits as well as the challenges of deploying full-duplex relays (for coverage extension) specifically in digital television broadcasting (DVB-T/H), which is one of the many application areas of orthogonal frequency division multiplexing (OFDM).

Then, there have been some analytical studies that fully consider the effect of loop interference while demonstrating the benefits offered by full-duplex relays over their half-duplex counterparts. The analysis in [10–12], e.g., demonstrates the rate-interference trade-off between full-duplex and half-duplex modes of operation by comparing the two modes in terms of the achievable end-to-end capacity in a three-node communication system. Furthermore, such studies typically explore possible ways of mitigating the effect of loop interference: [13], e.g., introduces the idea of optimized gain control to minimize the effect of loop interference. The analysis in a more recent study [14] shows that the best strategy (in terms of optimizing the achievable rate) is to switch opportunistically between the two modes of relaying, an idea which is therein referred to as hybrid full-duplex/half-duplex relaying.

The advent of multiple-antenna techniques, also known as multiple-input-multiple-output (MIMO), has been a major driving force for the research in

full-duplex relaying. While the mitigation of the effect of loop interference in single-input-single-output (SISO) full-duplex relays is limited mostly to subtractive cancellation (with possible variations in the technique of estimating the loop interference channel), the additional spatial dimension introduced by MIMO, in turn, gives rise to a new class of techniques for mitigating the effect of loop interference – the class of spatial suppression. Because this class encompasses several possible techniques, one can find the study of such possibilities distributed across various research papers: comparative study of time-domain subtractive cancellation and spatial suppression using, e.g., null-space projection has been conducted in [15, 16], the possibility of using beamforming techniques for spatial suppression of loop interference has been explored, e.g., in [17, 18], and the analysis of a broad range of interference mitigation schemes has been presented in [19].

1.3 Scope and Contributions of this Thesis

The scope of this thesis is limited to analyzing the effect of practical transceiver nonidealities on loop interference cancellation in a SISO full-duplex relay in the context of an OFDM system. The primary focus is on formulating an analytical framework that, on the basis of a quantifiable performance metric, helps evaluate the feasibility of a full-duplex relay based on the subtractive technique of loop interference cancellation.

The novelty of this work lies in the fact that it includes a thorough analysis of the combined effect of limited dynamic range at the relay receiver (due to finite-resolution quantization of the incoming signal) and other nonidealities in the relay transmitter upon the viability of full-duplex operation. The primary contribution of this thesis is a closed-form expression for the signal-to-interference plus noise ratio (SINR) after subtractive loop interference cancellation in a SISO full-duplex relay with transceiver nonidealities. The expression is based on applying the well-known Bussgang's theorem (for nonlinearities with Gaussian inputs) to model the quantizer at the relay receiver as a device that merely introduces additive noise uncorrelated to the input signal. As a secondary contribution, a general method for computing the parameters of such a model for any quantizer with deterministic quantization levels is presented, and its application is demonstrated in determining the exact values of the signal-to-quantization noise ratio (SQNR) for the specific cases of uniform and optimum nonuniform quantizers. Last but not the least, this thesis outlines a systematic way of applying the aforementioned SINR expression to determine the minimum conditions that must be fulfilled in order to ensure that a full-duplex relay with transceiver nonidealities performs satisfactorily in a practical scenario.

1.4 Thesis Report Organization

The rest of this thesis report is organized as follows.

- Chapter 2 gives a general overview of the full-duplex relaying system which forms the basis for the analysis presented in the remaining chapters.
- Chapter 3 presents a discussion on the various nonidealities typically present in the practical components making a full-duplex relay while focusing mainly on the model for the quantizer element that makes it mathematically tractable.
- Chapter 4 uses the results from Chapter 3 to develop a full signal model of the full-duplex relay with transceiver nonidealities and presents the derivation of the aforementioned expression for the SINR after subtractive loop interference cancellation in such a relay.
- Chapter 5 verifies the validity of the SINR expression from Chapter 4 and shows, by means of specific examples, how it can be used to determine the conditions necessary for the full-duplex relay to perform satisfactorily.
- Finally, Chapter 6 presents the concluding remarks and some directions for future work.

Chapter 2

System Overview

This chapter forms the basis on which we develop our analysis in the following chapters. We begin by choosing a specific scenario that depends on having a relay node to ensure successful communication between two primary nodes in the system. We then briefly describe the elements constituting the system while gradually formulating the signal model along the way. Moreover, we discuss the significance of loop interference inherent in full-duplex relays and include it in the signal model of the system under consideration.

2.1 Full-Duplex Relay Link

Let us consider a wireless communication link (see Figure 2.1) operating in a scenario in which the destination node D (e.g., a mobile terminal) is positioned so far from the source node S (e.g., a base station) that it is improbable for it to directly receive, at a useful power level, the signal transmitted by node S. The relay node R is, therefore, positioned between the two nodes such that it can receive the signal from node S and retransmit it at a suitable power level making it possible for node D to receive the relayed signal and successfully decode the message originally transmitted by node S. The relay node is equipped with two, physically separated antennas so as to enable reception from the source and transmission to the destination using the same band of frequencies at the same time, i.e., full-duplex relaying. The dotted line between nodes S and D represents the possibility of a weak channel, if any, connecting those two nodes directly. A similar scenario is analyzed, e.g., in [20].

We base our system on orthogonal frequency division multiplexing (OFDM), which is one efficient implementation of multicarrier modulation (MCM). The basic idea in MCM is to divide a data stream (possibly with high information rate) into a number of substreams and transmit them over parallel, ideally orthogonal, subchannels obtained by dividing the available transmission bandwidth in frequency. The number of substreams is chosen such that the bandwidth occupied by each resulting subchannel is less than the coherence bandwidth of the original channel, thereby ensuring that each subchannel experiences a relatively flat fading and the transmission over each of them undergoes minimal

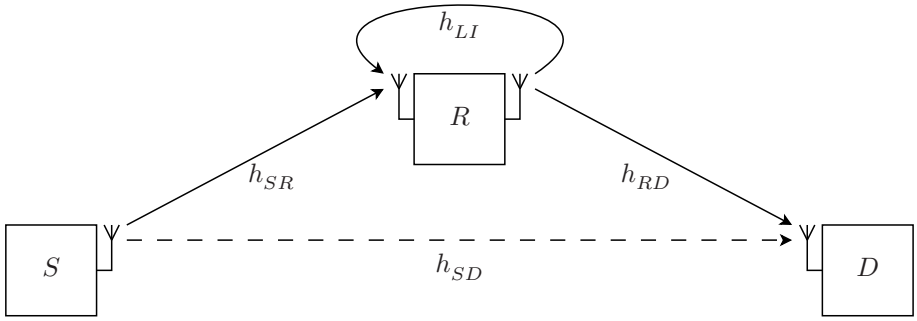


Figure 2.1: A two-hop wireless link with a full-duplex relay. The structures of nodes S, D, and R are further illustrated in Figure 2.2, Figure 2.3, and Figure 2.4, respectively. Symbols h_{SR}, h_{RD}, h_{LI}, and h_{SD} represent multipath channels in the time domain.

intersymbol interference (ISI) [21].

We make a reasonable assumption that all channels in our system vary slow enough in time allowing them to be considered as time-invariant over the duration of receiving one OFDM symbol. Similar assumption is made in the analysis of OFDM-based relaying systems in general (see, e.g., [22]). If we then analyze the operation of the communication link in the time domain on a symbol-to-symbol basis, we can treat the response of each channel over the OFDM symbol period to be that of a linear time-invariant system.

Let $x_S(t)$ be the continuous-time signal transmitted by the source corresponding to one OFDM symbol, and $h_{SR}(t)$ be the impulse response of the source-to-relay channel over the symbol duration. Then the signal reaching the relay, excluding additive noise for the time being, is given by

$$s_R(t) = h_{SR}(t) * x_S(t), \tag{2.1}$$

where $*$ denotes the convolution operation. This signal, after being processed and retransmitted by the relay, should ideally be

$$x_R(t) = x_S(t - \tau_R), \tag{2.2}$$

where τ_R is the processing delay incurred by the relay. The signal finally arriving at the destination via the relay-to-destination channel with impulse response $h_{RD}(t)$ takes the form

$$s_D(t) = h_{RD}(t) * x_R(t). \tag{2.3}$$

Depending upon its position, the destination node D may also directly receive, in addition to the signal transmitted by the relay node R, the original signal from the source node S through the source-to-destination channel $h_{SD}(t)$. In this case, the signal reaching the destination will be the superposition of the signals from the source and the relay:

$$\begin{aligned} s_D(t) &= h_{SD}(t) * x_S(t) + h_{RD}(t) * x_R(t) \\ &= h_{SD}(t) * x_S(t) + h_{RD}(t) * x_S(t - \tau_R), \end{aligned} \tag{2.4}$$

where the last step is the result of applying (2.2).

This superposition of two copies of the same signal with different delays arriving at the receiver through different wireless channels will not pose a threat to the correct functioning of the receiver as long as the total delay spread of the so formed composite channel, including the relay processing delay τ_R, is kept smaller than the duration of the cyclic prefix used by the OFDM modulator. In fact, as long as this condition holds true, the signal from the source coming along the weak channel h_{SD} will appear to the destination receiver to be merely an additional multipath component of the total incoming signal, thereby allowing the demodulator to function correctly. In other words, the destination node does not need to know at all of the existence of two separate transmitting entities in order to correctly recover the transmitted information from the received signal.

The foregoing analysis takes for granted that the relay is able to regenerate and transmit a perfect (delayed, but otherwise perfect) replica of the signal $x_S(t)$ originally transmitted by the source. However, such is not always the case in practice. Owing to the full-duplex operation of the relay, its receiving antenna, at any given time, receives not only the signal transmitted by the source, but also a second unwanted signal occupying the same frequency band as the first, transmitted by the relay itself and meant to be received by the destination. This amounts to the signal received by the relay being a composite of two signals, one desired and another undesired, of which the latter may be significantly stronger because of the physical proximity of the transmitting and the receiving antennas in the relay. This undesired signal, as mentioned in Chapter 1, is usually referred to as loop interference, and the channel through which it arrives is termed accordingly as the loop interference channel. Unless the relay is able to somehow estimate and cancel out the loop interference component from its received signal, the relayed signal will not be a close enough replica of the original signal from the source.

Let $h_{LI}(t)$ be the impulse response of the loop interference channel. Then the loop interference at the relay receiver will be

$$i_R(t) = h_{LI}(t) * x_R(t). \tag{2.5}$$

Having defined the noiseless versions of all signals arriving at the relay receiver, the composite signal received by the relay can now be finally written as the superposition between the desired signal $s_R(t)$ given in (2.1), the loop interference $i_R(t)$ in (2.5), and the additive white Gaussian noise $w_R(t)$ at the relay receiver:

$$\begin{aligned} y_R(t) &= s_R(t) + i_R(t) + w_R(t) \\ &= h_{SR}(t) * x_S(t) + h_{LI}(t) * x_R(t) + w_R(t). \end{aligned} \tag{2.6}$$

Before delving into the possibility of approximating and removing the loop interference $i_R(t)$ from the signal $y_R(t)$, it is useful to look into the structures of the communicating nodes in our system (see Figure 2.1). As we move on to the next section that briefly describes the structure of each of the nodes S, R

and D, let us close this section by updating the expression in (2.4) for the signal received by the destination node to include the additive white Gaussian noise $w_D(t)$ at the destination receiver:

$$y_D(t) = h_{SD}(t) * x_S(t) + h_{RD}(t) * x_S(t - \tau_R) + w_D(t). \tag{2.7}$$

2.2 Structures of Communicating Nodes

Let us begin this discussion regarding the structures of communicating entities in our system with the first node, i.e., the source node S. This node, as shown in Figure 2.2, can be roughly considered to be consisting of three units: the data stream provider, the OFDM modulator, and the transmitter front-end. The first unit, as its name suggests, is responsible for providing a continuous stream of data bits representing whatever information is intended to be communicated by the source to the destination. For the purpose of this thesis, we need not consider the details of how the bit stream is formed; mere assumption that a continuous stream somehow gets there suffices. For the sake of completeness, let us also include the functionality of forward-error-correction (FEC) coding within this unit. The OFDM modulator then systematically generates a sequence of complex symbols from the incoming bit stream. The characteristics of these OFDM symbols are directly relevant for the purpose of this thesis; therefore, we will consider, to some extent, the details of OFDM in Section 2.4. The complex symbols coming from the OFDM modulator are used by the transmitter front-end to modulate a continuous-time high-frequency carrier signal which is finally amplified to a suitable power level and transmitted with an antenna.

The destination node (see Figure 2.3) is essentially an inverse structure of the source node. The receiver front-end takes the radio frequency signal impinging upon its antenna and recovers, in the ideal case, the exact OFDM symbols that were used to modulate the carrier signal in the transmitter front-end of the source node. We will consider the structures of these transceiver front-ends in Section 2.5. The complex OFDM symbols recovered by the receiver front-end are then used by the OFDM demodulator (which performs an inverse operation of that carried out by the OFDM modulator) to produce an exact copy of the bit

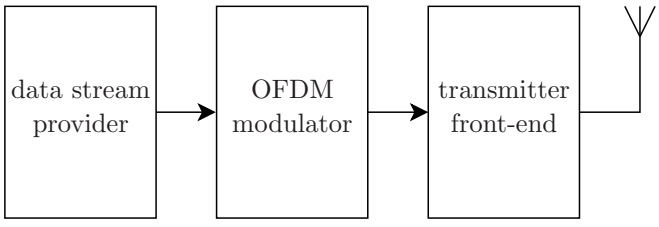


Figure 2.2: Structure of the source node S. The structures of the OFDM modulator and the transmitter front-end are further illustrated in Figure 2.5 and Figure 2.7, respectively.

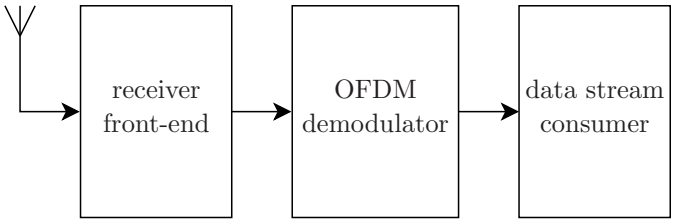


Figure 2.3: Structure of the destination node D. The structures of the OFDM demodulator and the receiver front-end are further illustrated in Figure 2.6 and Figure 2.8, respectively.

stream originally supplied by the data stream provider in the source node. The resulting bit stream is finally fed to the data stream consumer, e.g., the speech decoder in a mobile telephone set, the details of which we need not consider in this study. Reciprocal to the case with the data stream provider in the source node, the data stream consumer is assumed to include the functionality of necessary FEC decoding.

The relay node (see Figure 2.4), at its receiving and transmitting ends, contains front-end units whose structures and functions are similar to those in the destination and the source nodes. The receiver front-end in this case recovers the complex symbols corresponding to the composite signal $y_R(t)$ in (2.6), and not the desired signal $s_R(t)$ in (2.1). It is then the responsibility of the loop interference canceler to approximate and remove the contribution of the unwanted loop interference from the symbols recovered by the receiver front-end before passing them on to the transmitter front-end for retransmission.

The simplest possible approach, at least in the theoretical sense, to removing the contribution of the loop interference from the received signal is the technique of subtractive cancellation in the time domain, which can be expressed in our case as

$$\begin{aligned} \hat{s}_R[n] &= y_R[n] - \hat{i}_R[n] \\ &= y_R[n] - \hat{h}_{LI}[n] * x_R[n], \end{aligned} \tag{2.8}$$

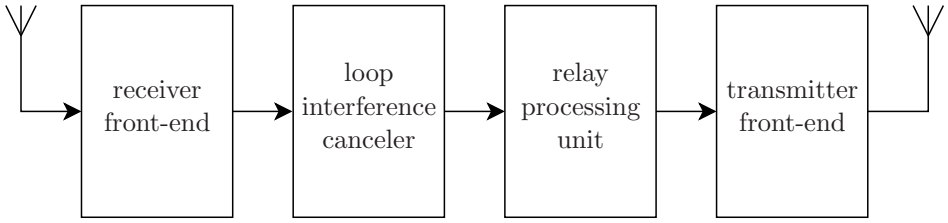


Figure 2.4: Structure of the relay node R. The structures of the transmitter front-end and the receiver front-end are further illustrated in Figure 2.7 and Figure 2.8, respectively.

where $\hat{h}_{LI}[n]$ is an estimate of the loop interference channel $h_{LI}[n]$ obtained from some channel estimation technique, $\hat{i}_R[n]$ is the approximation of the loop interference computed from the estimate $\hat{h}_{LI}[n]$ and the transmitted signal $x_R[n]$ known perfectly to the interference canceler, and $\hat{s}_R[n]$ is the resulting approximation of the interference-free desired signal $s_R[n]$. It should be noted that we have expressed all signals here in the discrete-time domain (with n as the time index) because channel estimation algorithms, as well as other functions to be performed by the relay, almost invariably require that the signals be processed digitally. As we will later see in Section 2.5, the conversion of the continuous-time received signal to the discrete-time domain for digital processing, and the conversion of a digitally processed signal to the continuous-time domain for transmission, are accomplished within the receiver and transmitter front-ends, respectively.

The relay processing unit, which appears before the transmitter front-end in Figure 2.4, is an entity whose structure and function depend largely upon the chosen relaying scheme. In the next section, we attempt to give a brief overview of the possible relaying schemes, along with the functions that the relay processing unit needs to perform in each case.

2.3 Relaying Schemes

Relaying schemes can be broadly classified into two categories: decode-and-forward (DF) relaying and amplify-and-forward (AF) relaying. DF relaying, also known as regenerative relaying, is defined as the relaying operation carried out by relays which first recover the exact bit stream modulating the OFDM signal received from the source, then regenerate the corresponding OFDM symbols again to retransmit them. On the other hand, the definition of AF relaying, or non-regenerative relaying, encompasses the operation of all relays which simply amplify and retransmit their received signals, with the possibility of some linear processing before transmission.

By the definition of DF relaying, the function of the relay processing unit in DF relays in general[1] is at least the aggregate of the functions of the OFDM demodulator and the OFDM modulator, including also forward-error-correction (FEC) decoding and encoding in between. The need for symbol demodulation and modulation causes the relay processing delay to be at least one OFDM symbol. This amounts to the total delay spread of the composite channel, formed by the direct source-to-destination channel and the indirect source-to-relay-to-destination channel, being greater than the duration of the cyclic prefix in an OFDM symbol. Consequently, the signal transmitted by the source that reaches the destination along the direct channel appears not to be an additional multipath component of the total received signal, but rather to be an unwanted interference. As a result, the destination receiver can function reasonably well only if this interference, i.e., the signal coming along the direct channel, is weak

[1] A discussion on some specific DF relaying protocols can be found in [23].

enough. This restriction limits the usability of DF relaying to situations where the source is sufficiently far away from the destination to avoid significant interference to the relayed signal arriving at the destination receiver.

AF relays, on the other hand, are less prone to suffer from the aforementioned drawback because they do not necessarily undertake demodulation and modulation before retransmission and are, therefore, able to keep the relay processing delay within the OFDM cyclic prefix in most cases. This also implies that the complexity of the relay processing unit in an AF relay is generally smaller than that in a DF relay. In the simplest theoretical case, an AF relay essentially does no more than amplifying and retransmitting whatever it receives; therefore, the relay processing unit in this case need not exist at all since it is always in the final stage of the transmitter front-end where the signal is amplified to the desired transmit power.

However, in the practical case, an AF relay usually needs to do more than simply scale and retransmit its received signal. Generally speaking, any relay that performs linear processing on the received signal without undergoing FEC decoding and encoding, so as to, e.g., compensate for known channel imperfections, is categorized under AF relays. Consequently, the actual complexity of the relay processing unit depends upon the specific processing that is required from the relay. Depending upon the nature of the specific requirement, one of either time-domain processing or frequency-domain processing becomes favorable over the other, even mandatory in some cases.

If frequency-domain processing is required, then the relay processing unit needs to have an OFDM demodulator in order to separate the incoming signal into its frequency components, and also an OFDM modulator to combine back the frequency components after processing. An example of frequency-domain processing in AF relays can be found in [24], where the relay dynamically allocates non-uniform gains to subcarriers, based on the knowledge of channel state information, while keeping the total transmit power constant. One can see that the requirement of demodulating at least one complete OFDM symbol before processing amounts to the relay processing delay being larger than the OFDM cyclic prefix, thus bringing into effect the same limitation as with DF relays described earlier.

Time-domain processing, on the other hand, does not require the received signal to be demodulated before processing. In most situations, where the complexity of the required signal processing is not very high, this permits the relay processing delay to be kept smaller than the duration of the OFDM cyclic prefix. In some cases, it is even possible to emulate frequency-domain processing by implementing time-domain finite impulse response (FIR) filters having a smaller number of taps than the cyclic prefix length. An example of such can be found in [25], where the relay is equipped with a filter that performs suitable phase rotation of the subcarriers in the relayed signal so as to ensure a significant coherent combining gain with the direct signal from the source also possibly arriving at the destination.

2.4 OFDM Modulator and Demodulator

The block diagram of a typical OFDM modulator is shown in Figure 2.5. The input bit stream from the data stream provider is mapped into a sequence of independent complex symbols using the quadrature amplitude modulation (QAM) scheme. The resulting complex symbol stream is then fed to a serial-to-parallel converter which generates successive sets of N parallel complex symbols, $X[k]$ for $k = 0, 1, \ldots, N-1$, where N is the number of substreams chosen for transmission over parallel subcarriers. This also implies that the N symbols represent the discrete frequency components of the composite OFDM symbol intended to be eventually transmitted over the available spectrum. The set of these N complex symbols are then processed by the IDFT block, which, as the name suggests, computes the inverse discrete Fourier transform of the discrete frequency components to generate N time-domain complex samples:

$$x[n] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X[k] e^{j2\pi kn/N}, \quad 0 \leq n \leq N-1. \tag{2.9}$$

This set of N time-domain complex samples obtained from the IDFT operation is said to constitute one OFDM symbol. The cyclic prefix is then added to the OFDM symbol and the resulting samples are ordered by the parallel-to-serial converter to obtain a sequence of $N + N_{\text{cp}}$ time samples, where N_{cp} is the length of the cyclic prefix in samples. The sequence of complex samples thus obtained is then fed to the transmitter front-end for transmission.

Figure 2.6 shows the block diagram of an OFDM demodulator, which essentially performs the inverse operation of the OFDM modulator just described. From the incoming stream of complex samples provided by the receiver front-end (to be discussed in Section 2.5), the first block in the OFDM demodulator groups the samples into sequences of length $N + N_{\text{cp}}$, then removes from each group the first N_{cp} samples representing the cyclic prefix, and provides in parallel the resulting N complex numbers constituting one OFDM symbol to the DFT block. The DFT block computes the discrete Fourier transform of the input set of complex numbers to generate the discrete frequency components present in

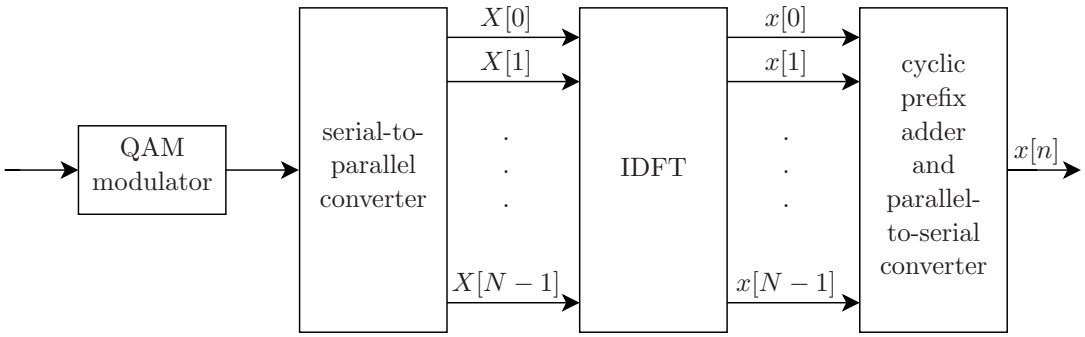


Figure 2.5: OFDM modulator.

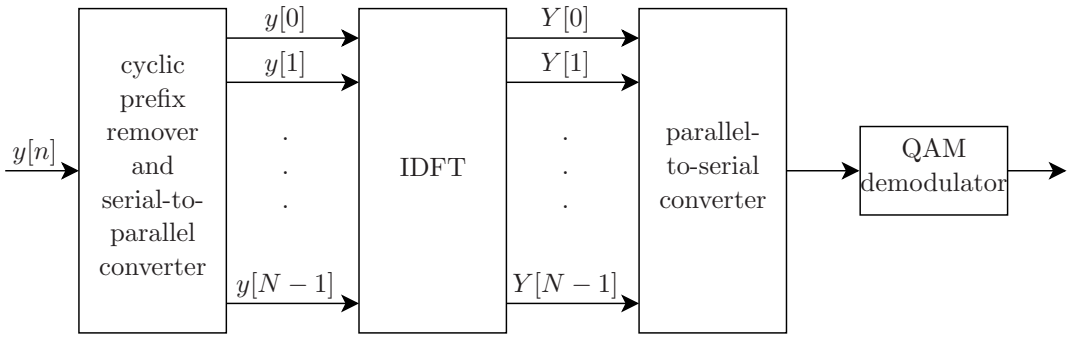


Figure 2.6: OFDM demodulator.

the OFDM symbol. Since each subchannel in OFDM is a flat fading channel as explained earlier in the beginning of this chapter, the discrete frequency components computed by the DFT are merely scaled (and noise-added) versions of the original discrete frequency components generated at the OFDM modulator. The outputs from the DFT block are then converted from parallel to serial and fed to the QAM demodulator which recovers the original data stream.

From (2.9), one can see that each time sample $x[n]$ in an OFDM symbol is the result of phase-rotating and adding N independent complex numbers $X[k], k = 0, 1, \ldots, N-1$. For sufficiently large values of N, the central limit theorem holds [26, §7.3], and each $x[n]$ converges to a zero-mean, complex-valued Gaussian random variable. This Gaussian approximation is fairly accurate for practical OFDM systems having $N \geq 128$. In fact, it has been shown in [27] that the asymptotic convergence of OFDM samples to Gaussianity holds not only for uncoded OFDM systems, but also for many coded OFDM systems and those with unequal power allocation across subcarriers. The significance of this special property of OFDM samples in the time domain will be further explored in the next chapter.

2.5 Transmitter and Receiver Front-ends

The block digram of a typical direct-conversion transmitter front-end is shown in Figure 2.7. The incoming sequence of complex time samples obtained from the OFDM modulator is first broken down into two sequences, one carrying the real part of each sample and the other carrying the imaginary part. These sequences are converted by digital-to-analog converters (DACs) to two continuous-time signals, which are then individually passed through pulse-shaping transmit filters to obtain signals $x_I(t)$ and $x_Q(t)$, referred to as the in-phase component and the quadrature component of the signal to be transmitted, respectively. These signals are then mixed with high frequency carrier signals $\cos(2\pi f_c t)$ and $\sin(2\pi f_c t)$, where f_c is the carrier frequency, and the results are combined as shown in the diagram. The single continuous-time signal $x(t)$ thus obtained is amplified by a power amplifier and eventually transmitted by a suitable antenna.

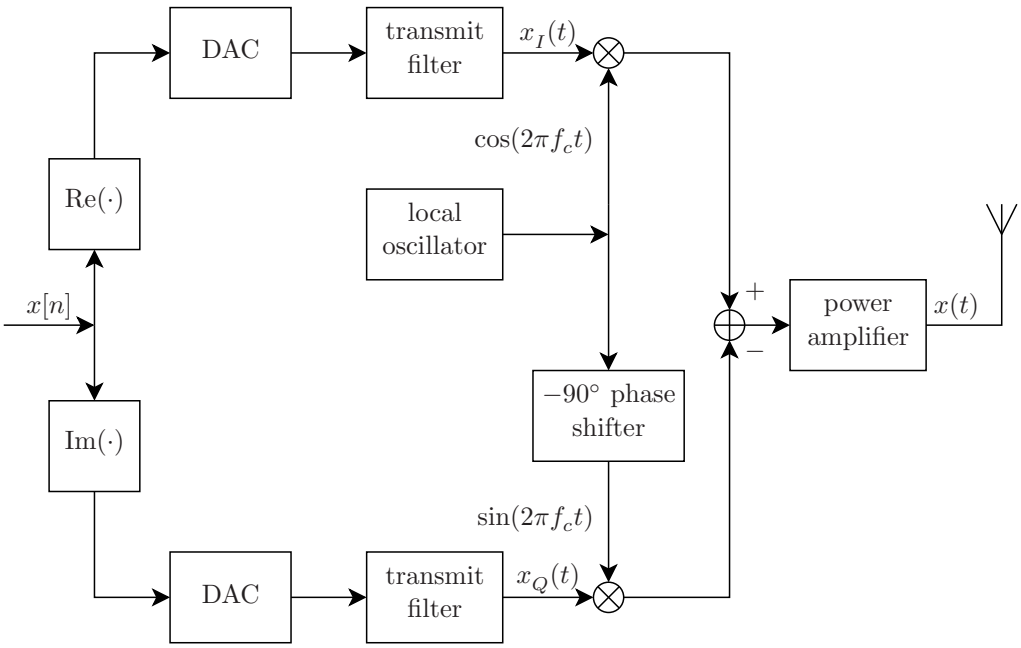


Figure 2.7: Direct-conversion transmitter front-end.

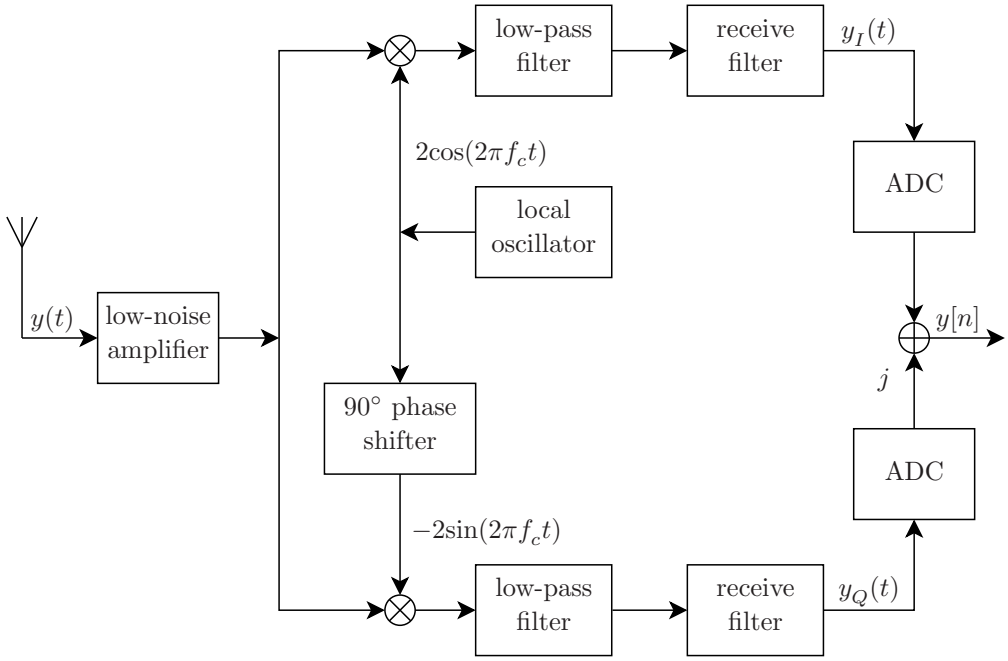


Figure 2.8: Direct-conversion receiver front-end.

The transmitted signal $x(t)$ can be written as

$$\begin{aligned} x(t) &= x_I(t)\cos(2\pi f_c t) - x_Q(t)\sin(2\pi f_c t) \\ &= \mathrm{Re}\{\tilde{x}(t)e^{j2\pi f_c t}\}, \end{aligned} \tag{2.10}$$

where $\tilde{x}(t) = x_I(t) + jx_Q(t)$ is referred to as the complex envelope of the transmitted signal $x(t)$. Whereas $x(t)$ is a real band-pass signal, $\tilde{x}(t)$ is a complex low-pass signal and is said to be the equivalent base-band representation of $x(t)$ [28].

The block diagram of the direct-conversion receiver front-end is shown in Figure 2.8. The continuous-time band-pass signal $y(t)$ received by the antenna is first amplified with a low-noise amplifier to bring its amplitude up to the level required for further processing. It is then separated into the in-phase and quadrature components by mixing it with local oscillator signals $2\cos(2\pi f_c t)$ and $-2\sin(2\pi f_c t)$, and individually passing the resulting signals first through low-pass filters, then through receive filters, as shown in the diagram. The receive filters[2] are designed such that they complement the action of the pulse-shaping transmit filters used in the transmitter so as to maximize the received signal-to-noise ratio (SNR). The continuous-time signals $y_I(t)$ and $y_Q(t)$ thus obtained are then converted into discrete-time signals by separate analog-to-digital converters (ADCs) and the results combined to obtain a single sequence $y[n]$ of complex numbers, to be processed by the OFDM demodulator in the destination node (or by the interference canceler in the relay node).

[2]We do not include the effects of the transmit filter and the receive filter on the signal model because they do not modify the fundamental assumption on which our analysis in the upcoming chapter is based – the assumption that each OFDM symbol transmitted or received by the relay has a Gaussian distribution.

Chapter 3

Transceiver Nonidealities

This chapter deals with the analytical characterization of nonidealities that are inherent in the practical components building the transceiver front-ends in the full-duplex relaying system presented in Chapter 2. We consider those nonidealities in particular that directly affect the performance limits of loop interference cancellation carried out by the relay node. On the receiver side, we focus our attention solely on quantization since it is inextricably tied to our study of the impact of limited dynamic range offered by practical analog-to-digital converters (ADCs) on the limits of loop interference cancellation. We begin by analyzing the signal distortion due to quantization in general and then demonstrate the applicability of the analysis by comparing the degree of distortion among two well-known quantization schemes. Finally, we also consider the effect of various nonidealities in the transmitter side and attempt to quantify it as a combined error vector magnitude (EVM).

Our analysis of the quantization process is based on the observation detailed in Section 2.4 that each time-domain sample from an OFDM signal is a complex-valued Gaussian random variable. It follows from the observation that either of the two real-valued sequences obtained by sampling the in-phase and quadrature components of the continuous-time OFDM signal is a real-valued Gaussian process that can safely be considered as independent of the other. To keep our analysis simple, we begin with the scalar quantization of samples from a single real-valued Gaussian process, and later show how the results can be applied in a straightforward manner to a complex-valued process formed by the combination of two such processes.

3.1 Analog-to-Digital Conversion at the Receiver

Analog-to-digital conversion is a two-step process that involves sampling an analog signal at discrete time instants followed by assigning finite-precision values to the samples so that they can be processed with a digital processor. The second step of this process, referred to as quantization, deserves special attention in the study of loop interference cancellation in full-duplex relays because it is precisely what enforces a limited dynamic range to the relay input.

Quantization is the process of mapping an infinite number of values falling in a subset of real numbers in the continuous domain to a finite number of values from a smaller subset of real numbers in the discrete domain. It is an irreversible process and it inevitably leads to loss of information. It is also a nonlinear transformation which makes it difficult to express analytically in the general case. However, when the input values come from a stationary Gaussian process, it is possible to define the input-output relation of the scalar quantizer, like any memoryless nonlinearity, with a compact, closed-form expression [29]

$$z[n] = \mathrm{T}(y[n]) = \alpha y[n] + d[n], \tag{3.1}$$

where $\mathrm{T}(\cdot)$ is the memoryless nonlinear transformation characterizing the quantizer, $y[n]$ and $z[n]$ are its input and output, respectively, and $d[n]$ is a signal uncorrelated with $y[n]$, i.e., $\mathbb{E}\{y[n+m]d[n]\} = 0$, with $\mathbb{E}(\cdot)$ being the statistical expectation operator. The scaling factor α is a constant given by

$$\alpha = \frac{\mathbb{E}\{yz\}}{\mathbb{E}\{y^2\}} = \frac{1}{\sigma_y^2} \int_{-\infty}^{\infty} yz f_Y(y) \mathrm{d}y, \tag{3.2}$$

where[1] σ_y^2 is the variance of the zero-mean Gaussian random variable $y[n]$ and hence its average power as well, and $f_Y(\cdot)$ is the probability density function of the same [26, Eq. (4.47)]:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{y^2}{2\sigma_y^2}}, \qquad -\infty < y < \infty,\ \sigma_y > 0. \tag{3.3}$$

This representation of the quantization nonlinearity, which is based on Bussgang's theorem [30], allows the output $z[n]$ to be treated as a superposition of the input $y[n]$ (scaled by a constant α but otherwise unmodified) and a new signal $d[n]$ quantifying the unwanted distortion arising from quantization. Furthermore, the fact that $d[n]$ is uncorrelated with $y[n]$ allows us to equate the average power of the output $z[n]$ to the sum of the average powers of its constituents, i.e.,

$$\mathbb{E}\{z^2\} = \alpha^2 \mathbb{E}\{y^2\} + \mathbb{E}\{d^2\}. \tag{3.4}$$

The expression above allows us to compute the average power of the distortion $d[n]$ generated by the quantizer once we have obtained the average powers of the input $y[n]$ and the output $z[n]$. To make things simpler, let us introduce a constant β, defined to be the power gain of the nonlinearity, i.e., the ratio between the average output power and the average input power:

$$\beta = \frac{\mathbb{E}\{z^2\}}{\mathbb{E}\{y^2\}} = \frac{1}{\sigma_y^2} \int_{-\infty}^{\infty} z^2 f_Y(y) \mathrm{d}y. \tag{3.5}$$

[1]For convenience of notation, we choose to drop the time index n from signals when they appear within statistical expectations as in (3.2), (3.4), and (3.5) since our analysis of the quantizer is based on the assumption that the input (and by extension, the output as well) comes from a stationary process whose statistical properties do not vary with time.

Using this definition of β and rearranging (3.4), we get

$$\mathbb{E}\{d^2\} = \mathbb{E}\{z^2\} - \alpha^2\mathbb{E}\{y^2\} = (\beta - \alpha^2)\mathbb{E}\{y^2\}. \tag{3.6}$$

Having expressed the average distortion power in terms of the average input power, we can now conveniently formulate the signal-to-quantization noise ratio (SQNR) for the quantizer nonlinearity, in the same way as done in, e.g., [31] and [32], as

$$\gamma_Q = \frac{\mathbb{E}\{(\alpha y)^2\}}{\mathbb{E}\{d^2\}} = \frac{\alpha^2\mathbb{E}\{y^2\}}{(\beta - \alpha^2)\mathbb{E}\{y^2\}} = \frac{1}{\frac{\beta}{\alpha^2} - 1}. \tag{3.7}$$

Let us now proceed to find the analytical expressions for the constants α and β starting from (3.2) and (3.5). If a quantizer is said to have a resolution of b bits, it features a maximum of $L = 2^b$ quantization levels, i.e., it maps each input value to one among the L quantization levels. In order to do so, it divides the entire range of the possible input values into L fixed intervals (quantization bins) and assigns, to each interval, a single output value (quantization level) for all input values within that interval.

In the general case, let us assume that the lth ($l = 1, 2, 3, \ldots, L$) quantization bin is bounded by (y_l, y_{l+1}) and the corresponding quantization level for that bin is z_l. To be precise, the quantizer output z becomes z_l whenever the input y falls within the interval (y_l, y_{l+1}). Since the L quantization bins are supposed to cover the entire range of input values, we can rewrite (3.2) as

$$\begin{aligned}\alpha &= \frac{1}{\sigma_y^2}\sum_{l=1}^{L}\left(z_l \int_{y_l}^{y_{l+1}} y f_Y(y)\mathrm{d}y\right) \\ &= \frac{1}{\sigma_y^2}\sum_{l=1}^{L}\left(z_l \int_{y_l}^{y_{l+1}} y \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{y^2}{2\sigma_y^2}}\mathrm{d}y\right),\end{aligned} \tag{3.8}$$

where the last step is the result of substituting $f_Y(y)$ from (3.3). We can simplify the integral appearing in each term of the sum by changing the integration variable from y to $\xi = \frac{y^2}{2\sigma_y^2}$. Then, the limits of integration y_l and y_{l+1} get transformed into $\frac{y_l^2}{2\sigma_y^2}$ and $\frac{y_{l+1}^2}{2\sigma_y^2}$, respectively, the differential $\mathrm{d}y$ becomes $\frac{\sigma_y^2}{y}\mathrm{d}\xi$, and we get

$$\begin{aligned}\int_{y_l}^{y_{l+1}} y \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{y^2}{2\sigma_y^2}}\mathrm{d}y &= \int_{\frac{y_l^2}{2\sigma_y^2}}^{\frac{y_{l+1}^2}{2\sigma_y^2}} \frac{\sigma_y}{\sqrt{2\pi}} e^{-\xi}\mathrm{d}\xi \\ &= -\frac{\sigma_y}{\sqrt{2\pi}} e^{-\xi}\Bigg|_{\frac{y_l^2}{2\sigma_y^2}}^{\frac{y_{l+1}^2}{2\sigma_y^2}} \\ &= \frac{\sigma_y}{\sqrt{2\pi}}\left(e^{-\frac{y_l^2}{2\sigma_y^2}} - e^{-\frac{y_{l+1}^2}{2\sigma_y^2}}\right).\end{aligned}$$

Using this result in (3.8), we get

$$\alpha = \frac{1}{\sqrt{2\pi\sigma_y^2}} \sum_{l=1}^{L} z_l \left(e^{-\frac{y_l^2}{2\sigma_y^2}} - e^{-\frac{y_{l+1}^2}{2\sigma_y^2}} \right). \tag{3.9}$$

Likewise, the expression for β in (3.5) becomes

$$\beta = \frac{1}{\sigma_y^2} \sum_{l=1}^{L} \left(z_l^2 \int_{y_l}^{y_{l+1}} \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{y^2}{2\sigma_y^2}} \mathrm{d}y \right), \tag{3.10}$$

where the integral within the sum can be simplified, as done previously, by changing the integration variable from y to $\zeta = \frac{y}{\sigma_y}$. Then, $\mathrm{d}y = \sigma_y \mathrm{d}\zeta$, and we get

$$\begin{aligned} \int_{y_l}^{y_{l+1}} \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{y^2}{2\sigma_y^2}} \mathrm{d}y &= \frac{1}{\sqrt{2\pi}} \int_{\frac{y_l}{\sigma_y}}^{\frac{y_{l+1}}{\sigma_y}} e^{-\frac{\zeta^2}{2}} \mathrm{d}\zeta \\ &= \frac{1}{\sqrt{2\pi}} \left(\int_{\frac{y_l}{\sigma_y}}^{\infty} e^{-\frac{\zeta^2}{2}} \mathrm{d}\zeta - \int_{\frac{y_{l+1}}{\sigma_y}}^{\infty} e^{-\frac{\zeta^2}{2}} \mathrm{d}\zeta \right) \\ &= \mathcal{Q}\left(\frac{y_l}{\sigma_y} \right) - \mathcal{Q}\left(\frac{y_{l+1}}{\sigma_y} \right), \end{aligned} \tag{3.11}$$

where

$$\mathcal{Q}(y) = \frac{1}{\sqrt{2\pi}} \int_{y}^{\infty} e^{-\frac{\zeta^2}{2}} \mathrm{d}\zeta \tag{3.12}$$

is the Gaussian $\mathcal{Q}$-function [26, Eq. (4.52)], a compact expression for the complement of the cumulative distribution function of a standard (zero mean, unit variance) Gaussian random variable. Using the result from (3.11) in (3.10), we get

$$\beta = \frac{1}{\sigma_y^2} \sum_{l=1}^{L} z_l^2 \left(\mathcal{Q}\left(\frac{y_l}{\sigma_y} \right) - \mathcal{Q}\left(\frac{y_{l+1}}{\sigma_y} \right) \right). \tag{3.13}$$

It should be noted that the results in (3.9) and (3.13) are general enough to hold for any quantization scheme with L quantization levels as long as the input samples have the Gaussian distribution. As an example of a specific case, let us find out what these results look like when uniform quantization is applied.

3.1.1 Uniform Quantization

In uniform quantization, all quantization bins have equal width and the quantization level for each bin is chosen to be a certain fixed point within the bin's

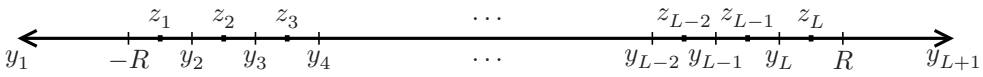


Figure 3.1: Uniform quantization.

interval, preferably the centroid of all possible input values falling within that interval. For the sake of simplicity, we consider the scheme (see Figure 3.1) where successive quantization levels coincide with the mid-points of the corresponding quantization bins. As the quantization bins have uniform width, the quantization levels in this scheme get uniformly spaced as well.

For any practical quantizer, the amplitudes of the input samples that can be quantized without clipping need to be restricted within a finite interval. In our analysis of the uniform quantizer, we consider the non-clipping interval to be $(-R, R)$ under the assumption that the input samples are known to take both positive and negative values and have a mean value of zero. Any input sample having an amplitude falling outside $(-R, R)$ is treated as if its amplitude were, depending upon its sign, either $-R$ or R, and the quantization level from the first or the last bin is obtained at the output. In other words, any sample lying outside the allowed range is first clipped to the maximum allowed amplitude and then quantized as any other sample lying within the allowed range.

For a non-clipping range of $2R$ and a total of L quantization levels, the width of each quantization bin becomes $\Delta = \frac{2R}{L}$. The lower and upper thresholds y_l and y_{l+1} of the quantization bins, and the corresponding quantization levels z_l (see Figure 3.1) can then be expressed as

$$y_1 \to -\infty, \tag{3.14a}$$

$$y_2 = -R + \Delta = -R + \frac{2R}{L} = \left(\frac{2-L}{L}\right)R, \tag{3.14b}$$

$$y_3 = -R + 2\Delta = -R + 2\left(\frac{2R}{L}\right) = \left(\frac{4-L}{L}\right)R, \tag{3.14c}$$

$$\vdots$$

$$y_l = -R + (l-1)\Delta = -R + (l-1)\left(\frac{2R}{L}\right) = \left(\frac{2l-2-L}{L}\right)R, \tag{3.14d}$$

$$\vdots$$

$$y_L = -R + (L-1)\Delta = -R + (L-1)\left(\frac{2R}{L}\right) = \left(\frac{L-2}{L}\right)R, \tag{3.14e}$$

$$y_{L+1} \to \infty, \tag{3.14f}$$

and,

$$z_1 = -R + \frac{\Delta}{2} = -R + \frac{1}{2}\left(\frac{2R}{L}\right) = \left(\frac{1-L}{L}\right)R, \tag{3.15a}$$

$$z_2 = y_2 + \frac{\Delta}{2} = \left(\frac{2-L}{L}\right)R + \frac{1}{2}\left(\frac{2R}{L}\right) = \left(\frac{3-L}{L}\right)R, \tag{3.15b}$$

$$\vdots$$

$$z_l = y_l + \frac{\Delta}{2} = \left(\frac{2l-2-L}{L}\right)R + \frac{1}{2}\left(\frac{2R}{L}\right) = \left(\frac{2l-1-L}{L}\right)R, \tag{3.15c}$$

$$\vdots$$

$$z_L = y_L + \frac{\Delta}{2} = \left(\frac{L-2}{L}\right)R + \frac{1}{2}\left(\frac{2R}{L}\right) = \left(\frac{L-1}{L}\right)R. \tag{3.15d}$$

Using the expressions from (3.14) and (3.15) in (3.9), we arrive at the final expression for α for a uniform quantizer:

$$\begin{aligned}
\alpha &= \frac{R}{\sqrt{2\pi\sigma_y^2}}\left(\frac{1-L}{L}\right)\left(0 - e^{-\frac{\left(\frac{2-L}{L}\right)^2 R^2}{2\sigma_y^2}}\right) \\
&\quad + \frac{R}{\sqrt{2\pi\sigma_y^2}}\sum_{l=2}^{L-1}\left(\frac{2l-1-L}{L}\right)\left(e^{-\frac{\left(\frac{2l-2-L}{L}\right)^2 R^2}{2\sigma_y^2}} - e^{-\frac{\left(\frac{2l-L}{L}\right)^2 R^2}{2\sigma_y^2}}\right) \\
&\quad + \frac{R}{\sqrt{2\pi\sigma_y^2}}\left(\frac{L-1}{L}\right)\left(e^{-\frac{\left(\frac{L-2}{L}\right)^2 R^2}{2\sigma_y^2}} - 0\right) \\
&= \frac{\mu}{\sqrt{2\pi}}\left(\frac{2L-2}{L}\right)e^{-\frac{1}{2}\left(\frac{L-2}{L}\right)^2\mu^2} \\
&\quad + \frac{\mu}{\sqrt{2\pi}}\sum_{l=2}^{L-1}\left(\frac{2l-1-L}{L}\right)\left(e^{-\frac{1}{2}\left(\frac{2l-2-L}{L}\right)^2\mu^2} - e^{-\frac{1}{2}\left(\frac{2l-L}{L}\right)^2\mu^2}\right),
\end{aligned} \tag{3.16}$$

where μ is the clipping margin defined as

$$\mu = \frac{R}{\sigma_y}, \tag{3.17}$$

i.e., the ratio between the maximum non-clipping input level R and the standard deviation σ_y of the zero-mean Gaussian distributed input samples.

Likewise, using the expressions from (3.14) and (3.15) in (3.13), we obtain the final expression for β for a uniform quantizer:

$$\begin{aligned}
\beta &= \frac{R^2}{\sigma_y^2}\left(\frac{1-L}{L}\right)^2\left(1 - \mathcal{Q}\left(\frac{\left(\frac{2-L}{L}\right)R}{\sigma_y}\right)\right) \\
&\quad + \frac{R^2}{\sigma_y^2}\sum_{l=2}^{L-1}\left(\frac{2l-1-L}{L}\right)^2\left(\mathcal{Q}\left(\frac{\left(\frac{2l-2-L}{L}\right)R}{\sigma_y}\right) - \mathcal{Q}\left(\frac{\left(\frac{2l-L}{L}\right)R}{\sigma_y}\right)\right) \\
&\quad + \frac{R^2}{\sigma_y^2}\left(\frac{L-1}{L}\right)^2\left(\mathcal{Q}\left(\frac{\left(\frac{L-2}{L}\right)R}{\sigma_y}\right) - 0\right) \\
&= \mu^2\left(\frac{1-L}{L}\right)^2\left(1 - \mathcal{Q}\left(\frac{2-L}{L}\mu\right) + \mathcal{Q}\left(\frac{L-2}{L}\mu\right)\right) \\
&\quad + \mu^2\sum_{l=2}^{L-1}\left(\frac{2l-1-L}{L}\right)^2\left(\mathcal{Q}\left(\frac{2l-2-L}{L}\mu\right) - \mathcal{Q}\left(\frac{2l-L}{L}\mu\right)\right).
\end{aligned} \tag{3.18}$$

For a given uniform quantizer with L quantization levels (or, equivalently, that with a resolution of $b = \log_2 L$ bits), it can be seen from (3.16) and (3.18) that the values of parameters α and β, and, in turn, from (3.7) that the signal-to-quantization noise ratio γ_Q, depend solely upon the clipping margin μ. The significance of this observation lies in the fact that it makes it possible to optimize the performance of a given uniform quantizer simply by adjusting the clipping margin, which merely requires scaling the average power σ_y^2 of the input samples to a suitable level since the non-clipping range $(-R, R)$ of the quantizer remains fixed.

Clipping Margin and Clipping Probability

Before proceeding to the details of the relationship between the clipping margin and the signal-to-quantization noise, let us first look briefly at a more direct effect of the clipping margin: the effect on the probability that a clipping event occurs. It is easy to see that the larger the range of input values supported by the quantizer without clipping, or the smaller the fluctuation of input samples around their mean value zero, the lower is the probability that a clipping event occurs. This observation, combined with the definition of clipping margin from (3.17), implies that higher values of clipping margin result in lower values of clipping probability. To verify that this is indeed the case, let us compute the probability P_c that a clipping event occurs:

$$\begin{aligned} P_c &= \text{Prob}(|y| > R) \\ &= 1 - \int_{-R}^{R} f_Y(y)\mathrm{d}y \\ &= 1 - \frac{1}{\sqrt{2\pi\sigma_y^2}} \int_{-R}^{R} e^{-\frac{y^2}{2\sigma^2}}\mathrm{d}y, \end{aligned}$$

where the last step is the result of substituting $f_Y(y)$ from (3.3). We can further simplify the expression above by changing the integration variable from y to $\zeta = \frac{y}{\sigma_y}$. Doing so transforms the limits of integration $-R$ and R into $-\frac{R}{\sigma_y} = -\mu$ and $\frac{R}{\sigma_y} = \mu$, respectively, and the differential $\mathrm{d}y$ into $\sigma_y \mathrm{d}\zeta$. Then, the last equation becomes

$$\begin{aligned} P_c &= 1 - \frac{1}{\sqrt{2\pi}} \int_{-\mu}^{\mu} e^{-\frac{\zeta^2}{2}}\mathrm{d}\zeta \\ &= 1 - \frac{1}{\sqrt{2\pi}} \left(\int_{-\mu}^{\infty} e^{-\frac{\zeta^2}{2}}\mathrm{d}\zeta - \int_{\mu}^{\infty} e^{-\frac{\zeta^2}{2}}\mathrm{d}\zeta \right) \\ &= 1 - \Big(\mathcal{Q}(-\mu) - \mathcal{Q}(\mu) \Big) \\ &= 2\mathcal{Q}(\mu), \end{aligned} \tag{3.19}$$

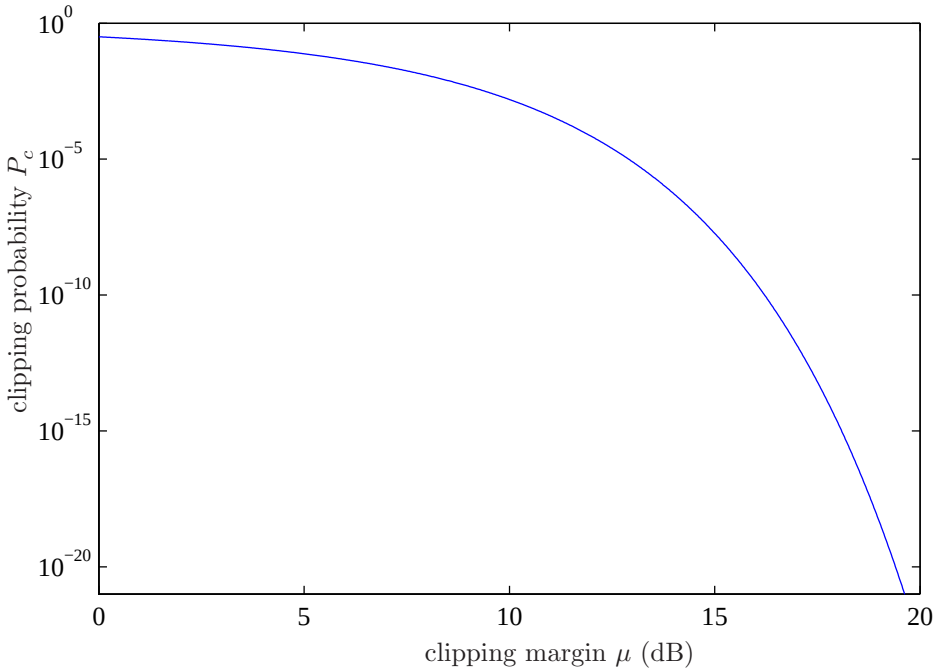


Figure 3.2: Effect of the clipping margin on the clipping probability.

where $\mathcal{Q}(\cdot)$ is the Gaussian $\mathcal{Q}$-function defined in (3.12), and the last step is the result of applying the identity $\mathcal{Q}(y) = 1 - \mathcal{Q}(-y)$ [26, Eq. (4.53)]. The above relation is illustrated graphically in Figure 3.2, which completes the verification of our earlier statement that an increase in the clipping margin amounts to a decrease in the clipping probability.

Clipping Margin and SQNR

We are now in a position that allows us to take a closer look at the relationship between the clipping margin and the signal-to-quantization noise ratio of a uniform quantizer. To do so, we first compute the signal-to-quantization noise ratio γ_Q analytically as a function of the clipping margin μ by using the results of (3.16) and (3.18) in (3.7). To ascertain that our analytical results are correct, we then simulate the behavior of the uniform quantizer at a number of closely spaced values of the clipping margin and compute the signal-to-quantization ratio from the quantizer's input and output data.

Figure 3.3 illustrates the results obtained from both approaches for three different resolutions of the quantizer: 10, 12 and 14 bits. In each case, it can be seen that increasing the clipping margin, starting from a small initial value, leads to a steep rise in the signal-to-quantization ratio (SQNR). However, when the clipping margin reaches a certain value, increasing it further does not lead to an increase in the SQNR, but instead results in a gradual fall in the SQNR from its maximum. This can be explained by the fact that increasing the clipping margin in the beginning leads to a significant decrease in the number of clipping

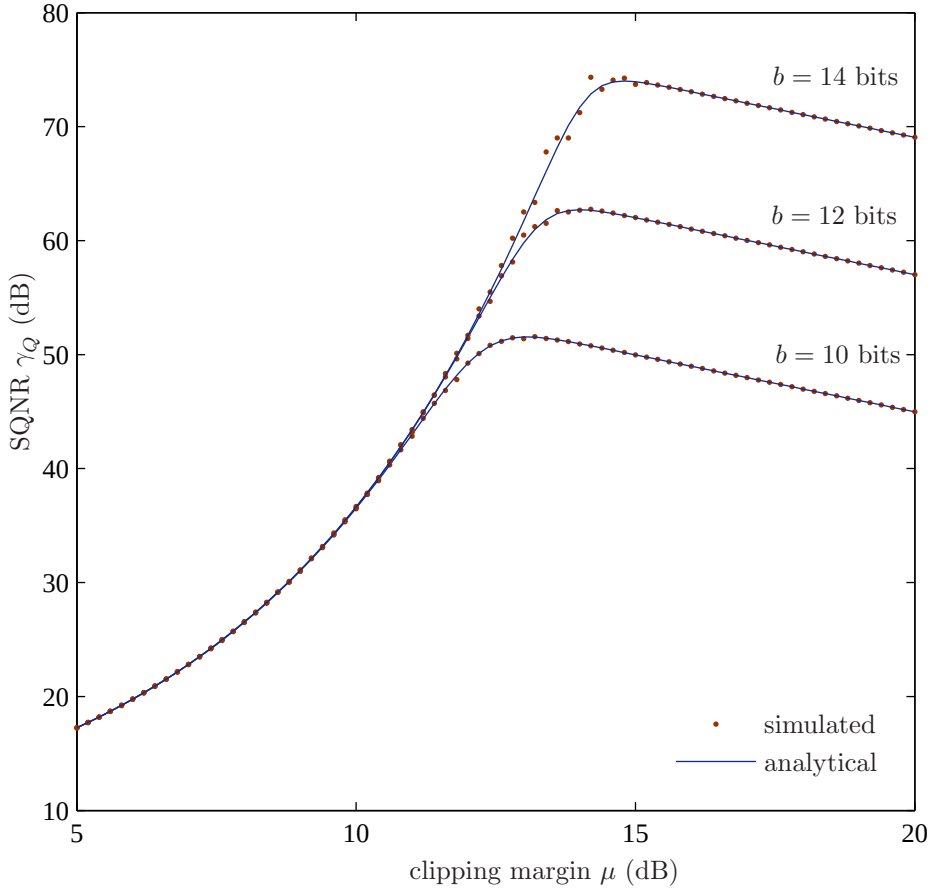


Figure 3.3: Effect of the clipping margin on the signal-to-quantization noise ratio of a uniform quantizer.

Table 3.1: Optimum clipping margin μ_{opt} and the corresponding signal-to-quantization noise ratio $\gamma_{Q,\text{max}}$ for various resolutions b of the uniform quantizer. The column P_c gives the theoretical clipping probability at the optimum clipping margin μ_{opt}, but it could as well be read independent of the quantizer resolution b.

b (bits)	μ_{opt} (dB)	$\gamma_{Q,\text{max}}$ (dB)	$P_c(\mu_{\text{opt}})$	b (bits)	μ_{opt} (dB)	$\gamma_{Q,\text{max}}$ (dB)	$P_c(\mu_{\text{opt}})$
3	7.40	14.10	1.91×10^{-2}	12	14.00	62.71	5.32×10^{-7}
4	8.57	19.33	7.33×10^{-3}	13	14.42	68.35	1.46×10^{-7}
5	9.57	24.55	2.61×10^{-3}	14	14.79	74.01	3.97×10^{-8}
6	10.45	29.82	8.68×10^{-4}	15	15.15	79.70	1.07×10^{-8}
7	11.22	35.17	2.73×10^{-4}	16	15.47	85.40	2.88×10^{-9}
8	11.90	40.57	8.23×10^{-5}	17	15.77	91.13	7.85×10^{-10}
9	12.51	46.03	2.40×10^{-5}	18	16.07	96.88	2.04×10^{-10}
10	13.06	51.55	6.85×10^{-6}	19	16.34	102.65	5.30×10^{-11}
11	13.55	57.11	1.92×10^{-6}	20	16.60	108.47	1.39×10^{-11}

events, thereby resulting in a steep rise in the SQNR. When the clipping margin reaches its optimum value, however, the clipping probability will already have become so small (see Table 3.1) that attempting to reduce it further will not improve the SQNR but rather degrade it because then it will become more likely that the input samples fall mostly within the inner quantization bins, thereby wasting a significant portion of the dynamic range offered by the quantizer.

We can infer from the foregoing analysis that the average noise power due to a quantizer is in fact the combined effect of two distinct phenomena, namely, clipping and quantization, and that the optimum value of the clipping margin is one which rightly balances the interplay between those two phenomena. When the quantizer operates in the region where the clipping margin is too low, the major source of distortion is the occurrence of clipping events, whereas when it operates in the region where the clipping margin is too high, quantization is the primary if not the only source of distortion. The large difference between the slopes of the SQNR curve in the two regions separated by the optimum clipping margin (see Figure 3.3) implies that the impact of clipping on the SQNR is far more detrimental than that of quantization. It is, therefore, useful to avoid clipping even at the cost of having less dynamic range for quantization.

Having said that there exists an optimum value of the clipping margin that results in the SQNR being maximized, we must also emphasize that it is not possible to derive a closed-form expression for the optimum clipping margin. This is because the expression for the SQNR as a function of the clipping margin, which can be obtained by combining results from (3.7), (3.16) and (3.18), is such that the zero of its derivative does not have an explicit, closed-form solution. We must, therefore, resort to numerical maximization of the SQNR expression to compute the optimum clipping margin. Table 3.1 lists the results of such numerical maximization (obtained using Matlab^{TM}) for quantizers having resolutions of 3 to 20 bits.

3.1.2 Optimum Nonuniform Quantization

Owing to the bell-shaped probability density curve of the Gaussian distribution, it is more probable that the value taken by a stationary Gaussian process at a given time instant is closer to the mean of the random process than that it is farther from the mean. This implies that if the process is fed to a scalar quantizer, the incoming samples tend to fall far more frequently among the inner quantization bins than the outer ones. It then follows by intuition that a quantizer with uniformly spaced quantization levels, no matter how optimum its clipping margin is, can most certainly not quantize the incoming Gaussian process such that the overall quantization error becomes as small as possible. It also follows that there has to be a different form of quantization scheme that utilizes the knowledge of the nonuniform probability density of the incoming random process in order to minimize the overall quantization error.

In their classic papers, Max [33] and Lloyd [34] have independently addressed the issue of devising a practical quantizer that is optimum in the sense of min-

imizing the mean-squared quantization error given that the probability density of the incoming random process is known. For a quantizer with a given number of finite quantization levels, both works have produced identical sets of sufficient conditions that minimize the mean-squared quantization error. In literature, the quantizer satisfying these conditions is commonly referred to as the Lloyd-Max quantizer.

The conditions essential to the Lloyd-Max quantizer can be expressed in the form of simultaneous equations as [33]

$$\int_{y_l}^{y_{l+1}} (y - z_l) f_Y(y) \mathrm{d}y = 0, \qquad l = 1, 2, \ldots, L \tag{3.20}$$

and,

$$y_{l+1} = \frac{1}{2}(z_l + z_{l+1}), \qquad l = 1, 2, \ldots, L-1 \tag{3.21}$$

where the symbols carry their usual meanings, i.e., L is the number of quantization levels, y_l and y_{l+1} are the lower and the upper thresholds of the lth quantization bin, z_l is the corresponding output value for that bin, and $f_Y(y)$ is the probability density function of the input. Also, by definition, $y_1 \to -\infty$ and $y_{L+1} \to \infty$, which are the same as in uniform quantization. Equations (3.20) and (3.21) tell us that the optimum output level for each quantization bin is the centroid of the area under the curve of the probability density function of the input between the thresholds of that bin and that the border separating two successive bins lies exactly half-way between their output levels.

The problem of finding the exact thresholds and the output levels for the quantization bins of an optimum quantizer is then limited to solving the $2L - 1$ simultaneous equations in (3.20) and (3.21) for all y_l and z_l. However, one can see that these simultaneous equations do not typically form a set simple enough to provide closed-form solutions for all y_l and z_l. In fact, if $f_Y(y)$ is the Gaussian probability density function as in our case, then closed-form solutions do not exist whenever $L > 2$ and we must rely on iterative numerical techniques to solve those equations.

In order to iteratively solve (3.20) and (3.21), Max [33] has suggested that we begin by choosing a value for z_1 and, with $y_1 \to -\infty$, compute y_2 from (3.20) by using a numerical root-finding procedure on a digital computer. Using the newly found value of y_2 and the original value of z_1, we can then determine the value of z_2 from (3.21). With y_2 and z_2, we can repeat the same two-step procedure of solving (3.20) followed by (3.21) to find y_3 and z_3, then y_4 and z_4, and so on until we have found y_L and z_L. The final step then is to use these values of y_L and z_L, and that of y_{L+1}, which is ∞ by definition, to evaluate the left side of (3.20) and check whether the result is really zero as it should be. If that is not the case, we adjust the current value of z_1 with an appropriate increment or decrement depending upon the result of the last evaluation and then repeat the whole process as many times as required for the result of the final step to be

reasonably close to zero. The values for all y_l and z_l that we have at this point will be the optimum ones.

As simple as this algorithm sounds, ensuring that it works as intended can be quite tricky, especially when the number of quantization levels L is high. According to Bucklew and Gallagher [35], the convergence of the algorithm is highly dependent on the initial choice of z_1. In the same paper, they have suggested two methods based on a companding technique suggested earlier by Smith [36], namely, the g-approximation method and the λ-approximation method, both of which provide a good initial guess for z_1 that speeds up the convergence of the algorithm considerably. Since then, researchers have come up with more efficient techniques for obtaining the initial guess, as well as those for updating the value as the algorithm proceeds (see, e.g., [37] and [38]).

In our implementation of Max's algorithm, we apply the g-approximation method because of its relative simplicity. In this method, the companding function $g(y)$ is defined as

$$g(y) = 1 - \mathcal{Q}\left(\frac{y}{\sqrt{3}}\right),$$

and the initial estimate for the first output level z_1 (or any output level z_l for that matter) is obtained by using the inverse of the companding function as

$$z_l = g^{-1}(\hat{z}_l),$$

where the inverse of the companding function is given by

$$g^{-1}(z) = \sqrt{3}\mathcal{Q}^{-1}(1 - z),$$

and $\hat{z}_l$ is a crude pre-estimate of z_l defined as

$$\hat{z}_l = \frac{2l - 1}{2L}, \qquad l = 1, 2, \ldots, L.$$

Once the implementation of Max's algorithm is in place, we fix the value of L (or, equivalently, the quantizer's resolution $b = \log_2 L$) and compute the values of y_l for $l = 1, 2, \ldots, L + 1$ and z_l for $l = 1, 2, \ldots, L$. We insert these computed values in (3.9) and (3.13) to obtain the values for α and β, which we eventually apply to (3.7) and obtain the value of the signal-to-quantization noise ratio for the optimum nonuniform quantizer.

Table 3.2 lists the numerical values of the signal-to-quantization noise ratio thus obtained for the optimum quantizers of various resolutions and compares them with the maximum values of the signal-to-quantization noise ratio that uniform quantizers with the same resolutions can provide. Cursory observation tells us that the gain Δ_{γ_Q} that an optimum quantizer provides over its uniform counterpart keeps increasing as the quantizer resolution increases. However, it would probably be wise to assume that this gain, in practice, can most certainly not increase beyond some saturation value.

Table 3.2: Signal-to-quantization noise ratio $\gamma_{Q,\text{opt}}$ for various resolutions b of the optimum nonuniform quantizer. Additional columns $\gamma_{Q,\text{max,ufm}}$ and Δ_{γ_Q} show the maximum SQNR achievable from a uniform quantizer of the same resolution as the optimum one and the difference in SQNR between the two quantization schemes, respectively.

b (bits)	$\gamma_{Q,\text{opt}}$ (dB)	$\gamma_{Q,\text{max,ufm}}$ (dB)	Δ_{γ_Q} (dB)	b (bits)	$\gamma_{Q,\text{opt}}$ (dB)	$\gamma_{Q,\text{max,ufm}}$ (dB)	Δ_{γ_Q} (dB)
3	14.46	14.10	0.36	12	67.90	62.71	5.19
4	20.18	19.33	0.85	13	73.92	68.35	5.57
5	26.00	24.55	1.45	14	79.94	74.01	5.93
6	31.91	29.82	2.09	15	85.96	79.70	6.26
7	37.86	35.17	2.69	16	91.98	85.40	6.58
8	43.85	40.57	3.28	17	98.00	91.13	6.87
9	49.86	46.03	3.83	18	104.03	96.88	7.15
10	55.87	51.55	4.32	19	109.95	102.65	7.30
11	61.88	57.11	4.77	20	115.87	108.47	7.40

3.1.3 Generalization to Complex-Valued Input

So far in the derivation of the signal-to-quantization noise ratio, we have considered a real-valued stationary Gaussian process at the input of the quantizer. However, as we shall soon see, the results can be easily generalized to the case when the input signal is a complex-valued stationary Gaussian process having independent real and imaginary components.

Let us consider Figure 3.4(a) showing a system that has a complex valued input process $y[n]$ with independent real and imaginary components $y_R[n]$ and $y_I[n]$. These independent components are processed by separate but identical quantizer nonlinearities, each characterized by the real-valued transformation $\mathrm{T}(\cdot)$. The parameters α and β for these nonlinearities, as defined in (3.2) and (3.5), are given by

$$\alpha_R = \frac{\mathbb{E}\{y_R z_R\}}{\mathbb{E}\{y_R^2\}} = \frac{\mathbb{E}\{y_I z_I\}}{\mathbb{E}\{y_I^2\}} = \alpha_I, \tag{3.22}$$

and,

$$\beta_R = \frac{\mathbb{E}\{z_R^2\}}{\mathbb{E}\{y_R^2\}} = \frac{\mathbb{E}\{z_I^2\}}{\mathbb{E}\{y_R^2\}} = \beta_I, \tag{3.23}$$

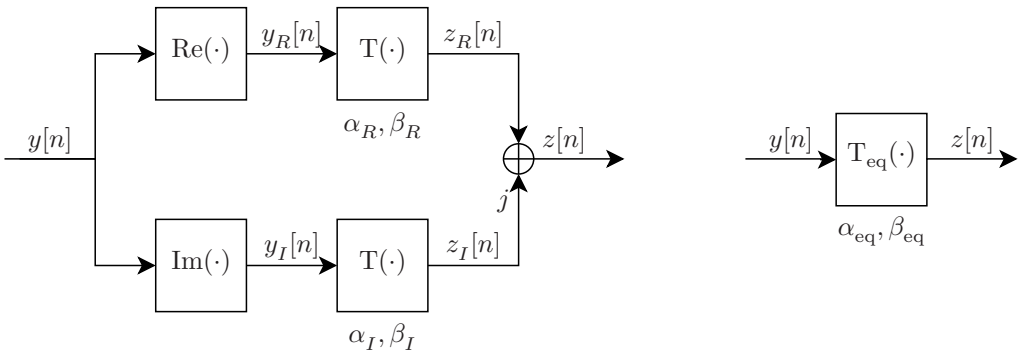
where the expectations $\mathbb{E}\{\cdot\}$ involving only the real components and the corresponding ones involving only the imaginary components are considered to be equal, assuming identical statistical properties of the real and the imaginary components of the input y.

Using the same transformation model as in (3.1), we can then express the outputs of the two quantizers as

$$z_R[n] = \mathrm{T}(y_R[n]) = \alpha_R y_R[n] + d_R[n], \tag{3.24}$$

and,

$$z_I[n] = \mathrm{T}(y_I[n]) = \alpha_I y_I[n] + d_I[n], \tag{3.25}$$



(a) (b)

Figure 3.4: Quantizers for complex-valued input: (a) a conceptual system with separate but identical transformations for the real and imaginary components, (b) a compact representation with a single transformation that is equivalent to the system in (a).

where $d_R[n]$ and $d_I[n]$ are the distortions generated by the nonlinearities transforming the real and the imaginary components of the input $y[n]$, respectively. The final output of the system is obtained by combining the quantities in (3.24) and (3.25) as

$$\begin{aligned} z[n] &= z_R[n] + jz_I[n] \\ &= \big(\alpha_R y_R[n] + d_R[n]\big) + j\big(\alpha_I y_I[n] + d_I[n]\big) \\ &= \alpha_R(y_R[n] + y_I[n]) + (d_R[n] + jd_I[n]) \qquad \because \alpha_R = \alpha_I \\ &= \alpha_R y[n] + d[n]. \end{aligned} \tag{3.26}$$

If we consider Figure 3.4(b) to be a system that is equivalent to that in Figure 3.4(a), then $\mathrm{T}_{\mathrm{eq}}(\cdot)$ represents a nonlinear transformation with a complex-valued stationary Gaussian process $y[n]$ at its input and a complex-valued stationary process $z[n]$ at its output. Then, we can write

$$z[n] = \mathrm{T}_{\mathrm{eq}}(y[n]) = \alpha_{\mathrm{eq}} y[n] + d[n], \tag{3.27}$$

where $d[n]$ is the complex-valued distortion generated by the nonlinearity $\mathrm{T}_{\mathrm{eq}}(\cdot)$. Similar to that in Section 3.1, $d[n]$ is uncorrelated with $y[n]$, and we can write

$$\begin{aligned} \mathbb{E}\{|d|^2\} &= \mathbb{E}\{|z|^2\} - \alpha_{\mathrm{eq}}^2 \mathbb{E}\{|y|^2\} \\ &= \big(\beta_{\mathrm{eq}} - \alpha_{\mathrm{eq}}^2\big)\mathbb{E}\{|y|^2\} \end{aligned} \tag{3.28}$$

where β_{eq}, by definition of the β parameter, is the ratio of the average output

power to the average input power:

$$\begin{aligned}\beta_{\text{eq}} &= \frac{\mathbb{E}\{|z|^2\}}{\mathbb{E}\{|y|^2\}} \\ &= \frac{\mathbb{E}\{z_R^2\} + \mathbb{E}\{z_I^2\}}{\mathbb{E}\{y_R^2\} + \mathbb{E}\{y_I^2\}} \\ &= \frac{\mathbb{E}\{z_R^2\}}{\mathbb{E}\{y_R^2\}} = \frac{\mathbb{E}\{z_I^2\}}{\mathbb{E}\{y_I^2\}},\end{aligned} \tag{3.29}$$

where the second and the third steps are the result of applying the assumptions that the real and the imaginary components of the input are independent of each other and that they have identical statistical properties. Finally, by comparing (3.26) and (3.27), we get

$$\alpha_{\text{eq}} = \alpha_R = \alpha_I, \tag{3.30}$$

and by comparing (3.23) and (3.29), we get

$$\beta_{\text{eq}} = \beta_R = \beta_I. \tag{3.31}$$

Now, the signal-to-quantization noise ratio offered by the equivalent system in Figure 3.4(b) is given by

$$\begin{aligned}\gamma_Q &= \frac{\mathbb{E}\{|\alpha_{\text{eq}} y|^2\}}{\mathbb{E}\{|d|^2\}} \\ &= \frac{\alpha_{\text{eq}}^2 \mathbb{E}\{|y|^2\}}{\big(\beta_{\text{eq}} - \alpha_{\text{eq}}^2\big)\mathbb{E}\{|y|^2\}} \\ &= \frac{1}{\frac{\beta_{\text{eq}}}{\alpha_{\text{eq}}^2} - 1},\end{aligned} \tag{3.32}$$

where (3.28) has been used to substitute for $\mathbb{E}\{|d|^2\}$.

Making use of the observation that this ratio depends only upon the parameters α and β, as was the case in (3.7) for the real-valued input process, and the results $\alpha_{\text{eq}} = \alpha_R = \alpha_I$ and $\beta_{\text{eq}} = \beta_R = \beta_I$, we can conclude that the expressions and the numerical values of the signal-to-quantization noise ratio obtained earlier for uniform and non-uniform optimum quantizers with real-valued Gaussian inputs hold as well for complex-valued Gaussian inputs with independent real and imaginary components.

3.2 Transmitter Nonidealities and the EVM

Having analyzed the effect of quantization on the received signal, we now turn our attention to the nonidealities that can affect the performance of the transmitter in our relaying system. This does not mean that the receiver does not suffer from nonidealities other than quantization; it just means that we are keeping our analysis focused on what interests us the most (i.e., the limited dynamic

range at the receiver due to quantization) by treating the remaining nonidealities as if they contribute merely to raise the noise level at the receiver. Along similar lines, we intend to quantify the combined effect of all transmitter nonidealities on the overall signal model with a single parameter, the benefit of which becomes apparent in Chapter 4.

Some of the most commonly studied transmitter nonidealities that impair the performance of OFDM systems include phase noise, I/Q imbalance, and power amplifier nonlinearity [39]. Phase noise refers to the mismatch of the fluctuation of phase between the oscillators in the transmitter and the receiver circuits and it affects the performance of OFDM systems in two distinct ways [40]. The first effect, known as the common phase error (CPE), causes the entire constellation of the received symbols to be rotated by a fixed amount with respect to the ideal constellation for the specific modulation scheme (e.g., PSK or QAM) used. This effect is not particularly troublesome as it is already addressed by channel estimation and equalization that invert the phase rotation caused by the wireless channel. The second effect, known as the intercarrier interference (ICI), is more severe since it represents the loss of orthogonality among the subcarriers due to the addition of random errors to the modulated subcarriers. A more comprehensive study of the effects of phase noise on the performance of OFDM systems can be found in [41].

For the direct-conversion transmitter (as well as the receiver) described in Chapter 2 to work perfectly, the complex analog oscillator, i.e., the theoretical combination of the oscillator and the phase-shifter that generates the carrier signals to be modulated by the in-phase (I) and the quadrature-phase (Q) components of the complex baseband signal, must be able to provide equal amplitude carriers with a phase difference of exactly 90°. Such a perfect oscillator cannot be realized in practice due to the limited accuracy of analog components, and the mismatch that becomes inevitable between the I and the Q components constituting the transmitted signal is referred to as I/Q imbalance. The effect of I/Q imbalance is that it significantly reduces the (infinite) image rejection capability of the direct-conversion architecture [39], resulting in the formation of unwanted signal components at frequencies other than those of the actual subcarriers.

The power amplifier nonlinearity is a major factor that severely limits the performance of an OFDM transmitter. Because each time-domain OFDM symbol is in fact the superposition of a typically large number of independent complex random variables (refer to Section 2.4 for details), the transmitted signal has a very large dynamic range. This has a significant impact on the design of the power amplifier at the transmitter because the amplifier has to have a large input power backoff (the ratio between the largest input level that keeps the output linear and the average input level) in order to ensure that its output remains sufficiently linear. Such a requirement is difficult to fulfill in practice because it demands a huge waste of power. In order to keep the power efficiency reasonable while operating the power amplifier in its linear region, it is, there-

fore, necessary to reduce the peak-to-average power ratio (PAPR) of the signal prior to amplification. [42] presents an analytical study of the effect of power amplifier nonlinearity on the performance of an amplify-and-forward (AF) relay link.

One simple way to reduce the PAPR is to clip an oversampled version of the baseband OFDM signal and then apply low-pass filtering to the clipped signal so as to remove the high frequency nonlinear distortion caused by clipping [43]. Despite the relative simplicity of this method, significant reduction of the PAPR can be achieved through repeated clipping and filtering without considerably increasing the out-of-band power in the OFDM signal [44]. However, clipping causes a loss in signal fidelity which cannot be recovered by the receiver unless a specialized forward-error-coding (FEC) with a high degree of redundancy (certainly not desirable) is applied to the sequence of symbols prior to modulation.

The nonidealities discussed above do not constitute the entire set of the possible sources of imperfections in the transmitter; nevertheless, they make a strong subset and we leave out the discussion of any other nonideality. In any case, what we are after is a way of quantifying the combined effect of all the transmitter nonidealities with a single parameter. One such entity, which is extensively used as a figure of merit for assessing transmitter performance, is the error vector magnitude (EVM). A detailed analysis of the effect on the EVM due to phase noise and I/Q imbalance has been presented in [45] and that due to clipping in [46].

In the simplest possible terms, the EVM is defined as the magnitude of the difference between the ideally expected value of a demodulated symbol (in the complex plane) and the measured value of the actual received symbol, expressed as a fraction of the magnitude of the expected value. Mathematically,

$$\text{EVM} = \frac{|S_{\text{ideal}} - S_{\text{actual}}|}{|S_{\text{ideal}}|}, \tag{3.33}$$

where S_{ideal} and S_{actual}, respectively, represent the ideally expected (i.e., the same as that meant to be transmitted) and the measured values of the symbol in question S and the difference between the two (i.e., $S_{\text{ideal}} - S_{\text{actual}}$) is referred to as the error vector. It should be stressed that both of these complex values should first be normalized to the same scale.

Since the transmitted signal is composed of a sequence of symbols, each of which comes from the constellation formed by a specific set of symbols, a more accurate metric would be the root-mean-squared value of those given by the expression above for all the possible symbols. Such a metric is referred to as the RMS EVM and it is given by [47]

$$\epsilon = \sqrt{\frac{\frac{1}{M}\sum_{m=1}^{M} |S_{\text{ideal}}[m] - S_{\text{actual}}[m]|^2}{\frac{1}{M}\sum_{m=1}^{M} |S_{\text{ideal}}[m]|^2}}, \tag{3.34}$$

where M denotes the constellation size, $S_{\text{ideal}}[m]$ and $S_{\text{actual}}[m]$ represent the normalized versions of the value meant to transmitted and the value that is measured for the mth symbol, respectively, and ϵ is the notation that we will use for the RMS EVM.

In practice, however, the constellation size M in the expression above is replaced by the length $M'(>> M)$ of the symbol sequence actually transmitted in order to measure the EVM. Then, by denoting the error signal corresponding to the sequence of the error vectors (i.e., $S_{\text{ideal}}[m] - S_{\text{actual}}[m]$ for $m = 1, 2, \ldots, M'$) as $v[n]$ and the signal corresponding to the sequence of the symbols meant to be transmitted (i.e., $S_{\text{ideal}}[m]$ for $m = 1, 2, \ldots, M'$) as $x[n]$, we can rewrite the ratio above in the form of a statistical measure as

$$\epsilon = \sqrt{\frac{\mathbb{E}\{|v|^2\}}{\mathbb{E}\{|x|^2\}}}, \tag{3.35}$$

which is how we will express the RMS EVM in the following chapters.

Chapter 4

Full-Duplex Relay with Nonidealities

In this chapter, we develop a complete signal model of the full-duplex relay node in our two-hop wireless link (refer to Figure 2.1) taking also into account the contributions due to the imperfections discussed in Chapter 3. Moreover, we discuss the degradation of loop interference cancellation caused by the inability of estimation algorithms to obtain accurate information on the loop interference channel when the observed signal is distorted by the aforementioned imperfections. Finally, we formulate the signal-to-interference plus noise ratio (SINR) in the relay node after loop interference cancellation.

4.1 Signal Model

Starting with the definition in (2.6) for the signal $y_R(t)$ received by the relay and making use of the representation in (3.1) for the quantizer nonlinearity, we can express the output from the analog-to-digital converter in the receiver front-end of the relay as

$$\begin{aligned} z_R[n] &= \alpha \left(\sqrt{G_Q} \Big(s_R[n] + i_R[n] + w_R[n] \Big) \right) + d[n] \\ &= \alpha \, y_R[n] + d[n], \end{aligned} \tag{4.1}$$

where α is the scaling factor of the quantizer nonlinearity (defined for real-valued signals in (3.2) and generalized to complex-valued signals in Section 3.1.3); $d[n]$ is the distortion noise due to the nonlinearity; G_Q is the power gain of a low-noise amplifier with an automatic gain control (AGC) mechanism that scales the signal prior to quantization to a level suitable for the optimum operation of the quantizer; $s_R[n]$, $i_R[n]$, and $w_R[n]$ are discrete-time representations of the desired signal from the source node, the loop interference in the relay node, and the additive white Gaussian noise at the relay receiver, respectively; and $y_R[n]$ is a compact discrete-time representation of the overall signal at the relay receiver just before quantization.

It is useful to recall from Section 3.1 that the distortion $d[n]$ due to the quantizer is uncorrelated with the input $y_R[n]$ and that its average power, as in (3.28), can be expressed in terms of the average input power and the parameters α and β as

$$\mathbb{E}\{|d|^2\} = \left(\beta - \alpha^2\right) \mathbb{E}\{|y_R|^2\}. \tag{4.2}$$

One should note that the validity of (4.1) and (4.2) depends on an implicit assumption that $y_R[n]$ comes from a stationary complex Gaussian process, which, in turn, requires that the three components of $y_R[n]$, namely, $s_R[n]$, $i_R[n]$, and $w_R[n]$, themselves come from separate complex Gaussian processes that are all stationary and uncorrelated with one another. Under this assumption, the average power of $y_R[n]$ is given by

$$\mathbb{E}\{|y_R|^2\} = G_Q \left(\mathbb{E}\{|s_R|^2\} + \mathbb{E}\{|i_R|^2\} + \mathbb{E}\{|w_R|^2\}\right),$$

and the average distortion power in (4.2) becomes

$$\mathbb{E}\{|d|^2\} = G_Q \left(\beta - \alpha^2\right) \left(\mathbb{E}\{|s_R|^2\} + \mathbb{E}\{|i_R|^2\} + \mathbb{E}\{|w_R|^2\}\right). \tag{4.3}$$

4.2 SINR at the Output of the ADC

One can see from (4.1) that $s_R[n]$ is the only desired or "signal" component in $z_R[n]$ and that the signals $i_R[n]$, $w_R[n]$, and $d[n]$ constitute the undesired or "interference plus noise" component. Because $d[n]$ is uncorrelated with $y_R[n] = s_R[n] + i_R[n] + w_R[n]$ and the components of $y_R[n]$ are uncorrelated with one another, the signal-to-noise plus interference ratio (SINR) at the output of the analog-to-digital converter (ADC) in the relay is thus given by

$$\begin{aligned}
\gamma &= \frac{\alpha^2 G_Q \mathbb{E}\{|s_R|^2\}}{\alpha^2 G_Q \left(\mathbb{E}\{|i_R|^2\} + \mathbb{E}\{|w_R|^2\}\right) + \mathbb{E}\{|d|^2\}} \\
&= \frac{\alpha^2 G_Q \mathbb{E}\{|s_R|^2\}}{G_Q \left(\beta - \alpha^2\right) \mathbb{E}\{|s_R|^2\} + \beta G_Q \left(\mathbb{E}\{|i_R|^2\} + \mathbb{E}\{|w_R|^2\}\right)} \\
&= \frac{1}{\left(\frac{\beta}{\alpha^2} - 1\right) + \frac{\beta}{\alpha^2}\left(\frac{1}{\mathrm{SIR}} + \frac{1}{\mathrm{SNR}}\right)},
\end{aligned} \tag{4.4}$$

where the intermediate step is the result of substituting $\mathbb{E}\{|d|^2\}$ from (4.3) and rearranging the denominator to bring similar terms together, and the last step is the result of dividing both the numerator and the denominator by $\alpha^2 G_Q \mathbb{E}\{|s_R|^2\}$ and replacing $\frac{\mathbb{E}\{|s_R|^2\}}{\mathbb{E}\{|i_R|^2\}}$ and $\frac{\mathbb{E}\{|s_R|^2\}}{\mathbb{E}\{|w_R|^2\}}$ with more common terms – the signal-to-interference ratio (SIR) and the signal-to-noise ratio (SNR), respectively.

4.3 Estimation and Cancellation of Loop Interference

Similar to its continuous-time counterpart in (2.5), the discrete-time representation $i_R[n]$ of the loop interference in the relay can be expressed as the response of the loop interference channel to the signal transmitted by the relay:

$$\begin{aligned} i_R[n] &= h_{LI}[n] * \Big(x_R[n] + v_R[n] \Big) \\ &= \sum_{m=0}^{M-1} x_R[n-m]\, h_{LI}[m] + \sum_{m=0}^{M-1} v_R[n-m]\, h_{LI}[m], \end{aligned} \tag{4.5}$$

where $h_{LI}[n]$ is the discrete-time impulse response of the loop interference channel assumed to have M taps, $x_R[n]$ is the discrete-time representation of the signal ideally meant to be transmitted by the relay, and the new term $v_R[n]$ represents the error signal due to transmitter imperfections in the relay, as discussed in Section 3.2. Assuming, for the sake of simplicity, that $v_R[n]$ is uncorrelated with $x_R[n]$, the average loop interference power can be expressed as

$$\mathbb{E}\{|i_R|^2\} = G_{LI} \Big(\mathbb{E}\{|x_R|^2\} + \mathbb{E}\{|v_R|^2\} \Big), \tag{4.6}$$

where

$$G_{LI} = \sum_{m=0}^{M-1} |h_{LI}[m]|^2 \tag{4.7}$$

denotes the average power gain of the loop interference channel. Using the definition of error vector magnitude (EVM) in (3.35), $\mathbb{E}\{|v_R|^2\}$ may be written as

$$\mathbb{E}\{|v_R|^2\} = \epsilon^2\, \mathbb{E}\{|x_R|^2\}, \tag{4.8}$$

and (4.6) becomes

$$\mathbb{E}\{|i_R|^2\} = G_{LI} \left(1 + \epsilon^2\right) \mathbb{E}\{|x_R|^2\}. \tag{4.9}$$

As discussed in Chapter 2, the loop interference $i_R[n]$ is generally much stronger than the desired signal $s_R[n]$ from the source node because the separation between the transmitting and receiving antennas within the relay is much smaller than that between the transmitting antenna in the source node and the receiving antenna in the relay node. This means that the signal-to-interference ratio (SIR) is typically very small (well below 0 dB), which, in turn, means that the overall signal-to-interference plus noise ratio (SINR) in the relay is even smaller since the expression in (4.4) contains the reciprocal of SIR in its denominator. Therefore, it is necessary to cancel out the effect of the loop interference $i_R[n]$ from the received signal, and doing so requires estimating the loop interference channel $h_R[n]$ as already pointed out in Section 2.2.

Fortunately, the fact that $i_R[n]$ is much stronger than $s_R[n]$ turns out to be desirable in this context since it facilitates the estimation of $h_R[n]$ by allowing $z_R[n]$ to be modeled simply as an observation of $i_R[n]$ perturbed by a much

weaker additive noise component, which is some linear combination of $s_R[n]$, $v_R[n]$, $w_R[n]$, and $d[n]$. Moreover, because $x_R[n]$ (from which $i_R[n]$ originates) is known completely to the estimation unit, the entire signal transmitted by the relay essentially becomes the "pilot" signal, and one can expect $h_R[n]$ to be identified reasonably well.

Substituting $i_R[n]$ in (4.1) with the expression in (4.5), we get

$$z_R[n] = \alpha \sqrt{G_Q} \sum_{m=0}^{M-1} x_R[n-m]\, h_{LI}[m] + u_R[n], \tag{4.10}$$

where

$$u_R[n] = \alpha \sqrt{G_Q} \left(s_R[n] + \sum_{m=0}^{M-1} v_R[n-m]\, h_{LI}[m] + w_R[n] \right) + d[n] \tag{4.11}$$

is a compact representation for all signal components present in $z_R[n]$ except that corresponding to the output of the M-tap channel $h_{LI}[n]$ due solely to the input $x_R[n]$. In effect, $u_R[n]$, in the context of estimating $h_{LI}[n]$, can be considered to be the "observation noise" that aggregates all extraneous signal components present in $z_R[n]$ observed at the output of the linear system $h_{LI}[n]$ in response to the input $x_R[n]$.

4.3.1 Channel Estimation

Assuming that the channel estimation unit in the loop interference canceler is capable of processing $N\,(\geq M)$ observations at a time and that the channel $h_{LI}[n]$ is varying slowly with respect to the sampling interval, we can express a block of N samples of $z_R[n]$ from (4.10) compactly as

$$\mathbf{z_R} = \mathbf{X_R}\, \mathbf{h_{LI}} + \mathbf{u_R}, \tag{4.12}$$

where

$$\begin{aligned}
\mathbf{z_R} &= \begin{bmatrix} z_R[0] & z_R[1] & z_R[2] & \cdots & z_R[N-1] \end{bmatrix}^T, \\
\mathbf{u_R} &= \begin{bmatrix} u_R[0] & u_R[1] & u_R[2] & \cdots & u_R[N-1] \end{bmatrix}^T, \\
\mathbf{h_{LI}} &= \begin{bmatrix} h_{LI}[0] & h_{LI}[1] & h_{LI}[2] & \cdots & h_{LI}[M-1] \end{bmatrix}^T,
\end{aligned}$$

and

$$\mathbf{X_R} = \alpha \sqrt{G_Q} \begin{bmatrix} x_R[0] & x_R[-1] & x_R[-2] & \cdots & x_R[1-M] \\ x_R[1] & x_R[0] & x_R[-1] & \cdots & x_R[2-M] \\ x_R[2] & x_R[1] & x_R[0] & \cdots & x_R[3-M] \\ \vdots & \vdots & \vdots & & \vdots \\ x_R[N-1] & x_R[N-2] & x_R[N-3] & \cdots & x_R[N-M] \end{bmatrix}. \tag{4.13}$$

The entries in $\mathbf{X_R}$ with negative time indices represent samples towards the end of the previous processing block, i.e., $x_R[-1]$ represents the last sample $x_R[N-1]$

in the previous block, $x_R[-2]$ represents second to the last sample $x_R[N-2]$ in the previous block, and so on.

Channel estimation then involves determining, from the known vector $\mathbf{X_R}$ and the known matrix $\mathbf{z_R}$, the value of the unknown vector $\mathbf{h_R}$ such that it satisfies (4.12) as closely as possible. It should be noted that the elements of $\mathbf{u_R}$ cannot be exactly known: the information available on $\mathbf{u_R}$, if any, can describe (some of) its statistical properties at best. In any case, the estimator of $\mathbf{h_R}$ is some vector-valued function of $\mathbf{X_R}$ and $\mathbf{z_R}$, and the most straightforward way of obtaining the time-evolving estimate of $\mathbf{h_R}$ is to repeatedly evaluate this estimator function with continuously updated values of $\mathbf{X_R}$ and $\mathbf{z_R}$.

In estimation theory, the representation in (4.12) that relates the "observation vector" $\mathbf{z_R}$ to the unknown "parameter vector" $\mathbf{h_R}$ through the known matrix $\mathbf{X_R}$ and the "observation noise" $\mathbf{u_R}$ is commonly referred to as the linear model. For this model, there are quite a few standard estimators available for obtaining the value of the unknown parameter vector, and the better performance of one over others depends upon such factors as the statistical properties of the measurement noise and whether or not some *a priori* information on the parameter vector is available [48]. The details of a number of such estimators, when applied to channel estimation in OFDM systems, can be found, e.g., in [49].

Apart from the factors mentioned above, the applicability of a standard estimator to channel estimation is also determined by its computational complexity, especially when the channel is known to have a long impulse response that varies significantly over time. In such a scenario, it is better – even essential sometimes – to have a recursive implementation that, instead of updating $\mathbf{z_R}$ and $\mathbf{X_R}$ and evaluating the whole estimator function every time a new sample becomes available, somehow computes only the incremental value of the estimator corresponding to the new sample and then adds it to the previous estimate to get the new estimate. The good thing about such recursive algorithms is that, in most cases, they allow the resulting estimators to be implemented directly as adaptive filters, some examples of which can be found in [50].

Rather than delving into the details of specific estimators, we keep the discussion in this section general enough to hold good for any estimator that can be derived from the linear model in (4.12). The reason for doing so becomes apparent in Section 4.4, where we intend to express the SINR in the relay after loop interference cancellation in a concise closed form that does not contain elements specific to any estimator but is, nevertheless, sufficiently parameterized to allow convenient evaluation in case a particular estimator comes into the picture.

Before we close this section, we will take a brief look at the standard notation in basic estimation theory that is useful for the upcoming sections. It is a common practice to denote an estimator for an unknown parameter vector $\boldsymbol{\theta}$ (as well as the estimate obtained from this estimator) by the symbol $\hat{\boldsymbol{\theta}}$ and the corresponding estimation error vector, i.e., the difference between the true value and the estimated value of the parameter, by $\tilde{\boldsymbol{\theta}}$. Then, we have

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}. \tag{4.14}$$

The estimation error covariance matrix, whose elements give the covariances among the elements of the estimation error vector $\tilde{\theta}$, is usually denoted by $\mathbf{C}_{\tilde{\boldsymbol{\theta}}}$ and computed as

$$\mathbf{C}_{\tilde{\boldsymbol{\theta}}} = \mathbb{E}\{\tilde{\boldsymbol{\theta}}\,\tilde{\boldsymbol{\theta}}^H\}, \tag{4.15}$$

where the superscript H denotes the Hermitian transpose.

4.3.2 Time-Domain Subtractive Cancellation

Assuming that we have an estimator for the loop interference channel $h_{LI}[n]$, we are now ready to take a closer look at the details of the time-domain subtractive technique introduced in Section 2.2 for canceling the loop interference $i_R[n]$. With $\hat{\mathbf{h}}_{\mathbf{LI}} = \begin{bmatrix}\hat{h}_{LI}[0] & \hat{h}_{LI}[1] & \cdots & \hat{h}_{LI}[M-1]\end{bmatrix}^T$ as the available estimate of the loop interference channel, and with the signal $x_R[n]$ transmitted by the relay perfectly known to the canceler, we can obtain an estimate of the loop interference $i_R[n]$ as

$$\hat{i}_R[n] = \hat{h}_{LI}[n] * x_R[n] = \sum_{m=0}^{M-1} x_R[n-m]\,\hat{h}_{LI}[m].$$

When we have the estimate $\hat{i}_R[n]$, subtractive cancellation of the loop interference involves nothing more than subtracting an appropriately scaled version of $\hat{i}_R[n]$ from the signal $z_R[n]$ in (4.10). The resulting signal after such cancellation is given by

$$\begin{aligned}
\xi_R[n] &= z_R[n] - \alpha\sqrt{G_Q}\,\hat{i}_R[n] \\
&= \alpha\sqrt{G_Q}\Bigg(s_R[n] + \sum_{m=0}^{M-1} x_R[n-m]\,\tilde{h}_{LI}[m] \\
&\qquad\qquad + \sum_{m=0}^{M-1} v_R[n-m]\,h_{LI}[m] + w_R[n]\Bigg) + d[n] \\
&= \alpha\sqrt{G_Q}\Big(s_R[n] + \tilde{i}_R[n] + e_R[n] + w_R[n]\Big) + d[n],
\end{aligned} \tag{4.16}$$

where $\tilde{h}_{LI}[m] = h_{LI}[m] - \hat{h}_{LI}[m]$, for $m = 1, 2, \ldots, M-1$, are the elements of the channel estimation error vector $\tilde{\mathbf{h}}_{\mathbf{LI}}$,

$$\tilde{i}_R[n] = \sum_{m=0}^{M-1} x_R[n-m]\,\tilde{h}_{LI}[m] = \tilde{h}_{LI}[n] * x_R[n] \tag{4.17}$$

is the residual loop interference in $\xi_R[n]$ attributable to the so-called residual loop interference channel $\tilde{h}_{LI}[n]$ corresponding to the error in the estimation of $h_{LI}[n]$, and

$$e_R[n] = \sum_{m=0}^{M-1} v_R[n-m]\,h_{LI}[m] = h_{LI}[n] * v_R[n] \tag{4.18}$$

is the residual loop interference in $\xi_R[n]$ due to $v_R[n]$, the unknown error in the signal transmitted by the relay.

One should note that the degradation in signal quality caused by $\tilde{i}_R[n]$ depends on how close the estimate $\hat{h}_{LI}[n]$ of the loop interference channel is to its true value $h_{LI}[n]$ and, therefore, gets smaller as the estimate gets better, whereas the degradation caused by $e_R[n]$ depends on how large the deviation $v_R[n]$ of the signal transmitted by the relay is from its ideal value $x_R[n]$, not on the accuracy of $\hat{h}_{LI}[n]$. Since $v_R[n]$ is unknown, $e_R[n]$ cannot be canceled from $z_R[n]$ even in the hypothetical case where $h_{LI}[n]$ has somehow been perfectly identified.

4.4 SINR after Loop Interference Cancellation

By identifying the desired and undesired signal components in (4.16), we can express the signal-to-interference plus noise ratio (SINR) after loop interference cancellation as

$$\gamma = \frac{\alpha^2 G_Q \, \mathbb{E}\{|s_R|^2\}}{\alpha^2 G_Q \Big(\mathbb{E}\{|\tilde{i}_R|^2\} + \mathbb{E}\{|e_R|^2\} + \mathbb{E}\{|w_R|^2\} \Big) + \mathbb{E}\{|d|^2\}}, \tag{4.19}$$

where $\mathbb{E}\{|e_R|^2\}$ can be obtained from (4.18) by using (4.7), (4.8), and (4.9) as

$$\begin{aligned} \mathbb{E}\{|e_R|^2\} &= \sum_{m=0}^{M-1} |h_{LI}[m]|^2 \, \mathbb{E}\{|v_R|^2\} \\ &= \frac{\epsilon^2}{1+\epsilon^2} \, \mathbb{E}\{|i_R|^2\} \end{aligned} \tag{4.20}$$

and $\mathbb{E}\{|\tilde{i}_R|^2\}$ from (4.17) and (4.9) as

$$\begin{aligned} \mathbb{E}\{|\tilde{i}_R|^2\} &= \sum_{m=0}^{M-1} |\tilde{h}_{LI}[m]|^2 \, \mathbb{E}\{|x_R|^2\} \\ &= \frac{\tilde{G}_{LI}}{G_{LI} \left(1+\epsilon^2\right)} \, \mathbb{E}\{|i_R|^2\}, \end{aligned} \tag{4.21}$$

with

$$\tilde{G}_{LI} = \sum_{m=0}^{M-1} |\tilde{h}_{LI}[m]|^2 \tag{4.22}$$

denoting the average power gain of the residual loop interference channel $\tilde{h}_{LI}[n]$.

Using the expressions in (4.3), (4.20), and (4.21), we can rewrite the denominator of the ratio in (4.19) in terms of the average powers, $\mathbb{E}\{|s_R|^2\}$, $\mathbb{E}\{|i_R|^2\}$, and $\mathbb{E}\{|w_R|^2\}$, of the three components of the signal $y_R[n]$ at the relay receiver as

$$\begin{aligned} \text{dmtr.}(\gamma) = \alpha^2 G_Q \left(\frac{\tilde{G}_{LI}}{G_{LI} \left(1+\epsilon^2\right)} \, \mathbb{E}\{|i_R|^2\} + \frac{\epsilon^2}{1+\epsilon^2} \, \mathbb{E}\{|i_R|^2\} + \mathbb{E}\{|w_R|^2\} \right) \\ + \left(\beta - \alpha^2\right) G_Q \Big(\mathbb{E}\{|s_R|^2\} + \mathbb{E}\{|i_R|^2\} + \mathbb{E}\{|w_R|^2\} \Big) \end{aligned}$$

$$
\begin{aligned}
&= \left(\beta - \alpha^2\right) G_Q \, \mathbb{E}\{|s_R|^2\} + \beta \, G_Q \, \mathbb{E}\{|w_R|^2\} \\
&\quad + \left(\beta + \alpha^2 \left(\frac{\tilde{G}_{LI}}{G_{LI}\left(1+\epsilon^2\right)} + \frac{\epsilon^2}{1+\epsilon^2} - 1\right)\right) G_Q \, \mathbb{E}\{|i_R|^2\} \\
&= \left(\beta - \alpha^2\right) G_Q \, \mathbb{E}\{|s_R|^2\} + \beta \, G_Q \, \mathbb{E}\{|w_R|^2\} \\
&\quad + \left(\beta - \frac{\alpha^2}{1+\epsilon^2}\left(1 - \frac{\tilde{G}_{LI}}{G_{LI}}\right)\right) G_Q \, \mathbb{E}\{|i_R|^2\}.
\end{aligned}
$$

If we substitute the expression above for the denominator in (4.19), then divide the numerator as well as the denominator of the result by $\alpha^2 \, G_Q \, \mathbb{E}\{|s_R|^2\}$, and finally replace $\frac{\mathbb{E}\{|s_R|^2\}}{\mathbb{E}\{|i_R|^2\}}$ by SIR and $\frac{\mathbb{E}\{|s_R|^2\}}{\mathbb{E}\{|w_R|^2\}}$ by SNR, the expression for the SINR in the relay after loop interference cancellation becomes

$$
\gamma = \frac{1}{\left(\frac{\beta}{\alpha^2} - 1\right) + \left(\frac{\beta}{\alpha^2} - \rho\right) \frac{1}{\mathrm{SIR}} + \frac{\beta}{\alpha^2} \frac{1}{\mathrm{SNR}}}, \tag{4.23}
$$

where

$$
\rho = \frac{1}{1+\epsilon^2}\left(1 - \frac{\tilde{G}_{LI}}{G_{LI}}\right). \tag{4.24}
$$

Comparison of (4.23) with (4.4) tells us that the subtractive cancellation of loop interference indeed amounts to an improvement in the SINR as the coefficient $\frac{\beta}{\alpha^2}$ of the term $\frac{1}{\mathrm{SIR}}$ in the denominator of the ratio gets reduced by the quantity ρ defined in (4.24). The larger the value of ρ, the greater is the improvement in the SINR.

One can see from (4.24) that the value of ρ and, in turn, the degree of improvement in the SINR due to loop interference cancellation depends upon two factors: the accuracy of the chosen channel estimation technique, which is quantified by the residual loop interference channel gain $\tilde{G}_{LI}$, and the magnitude of the transmitter side imperfections, which is quantified by the EVM ϵ. What cannot be seen from (4.24) is the fact that those seemingly distinct factors are not entirely independent: the EVM has a considerable impact on the accuracy with which the loop interference channel can be estimated as it contributes to the "observation noise" in the signal model defined by (4.10) and (4.11). In the most desirable but unrealistic case where the relay transmitter is perfect (i.e., $\epsilon = 0$) and the estimate of the loop interference channel is accurate (i.e., $\tilde{G}_{LI} = 0$), ρ becomes 1, which is the largest it can be, and the SINR attains its maximum possible value for the given set of α, β, SIR, and SNR. In all realistic cases (where $\epsilon > 0$ and $\tilde{G}_{LI} < 0$), however, ρ remains less than 1 and the SINR remains sub-optimum.

Even in the hypothetical case where ρ is equal to 1, one should note that the coefficient of the term $\frac{1}{\mathrm{SIR}}$ in the denominator of the SINR in (4.23) does not vanish altogether but takes a value given by $\frac{\beta}{\alpha^2} - 1$, which, by the definitions of α and β, gets increasingly closer to zero as the quantizer resolution grows but never reaches zero. This quantity, in a sense, represents the residual loop interference

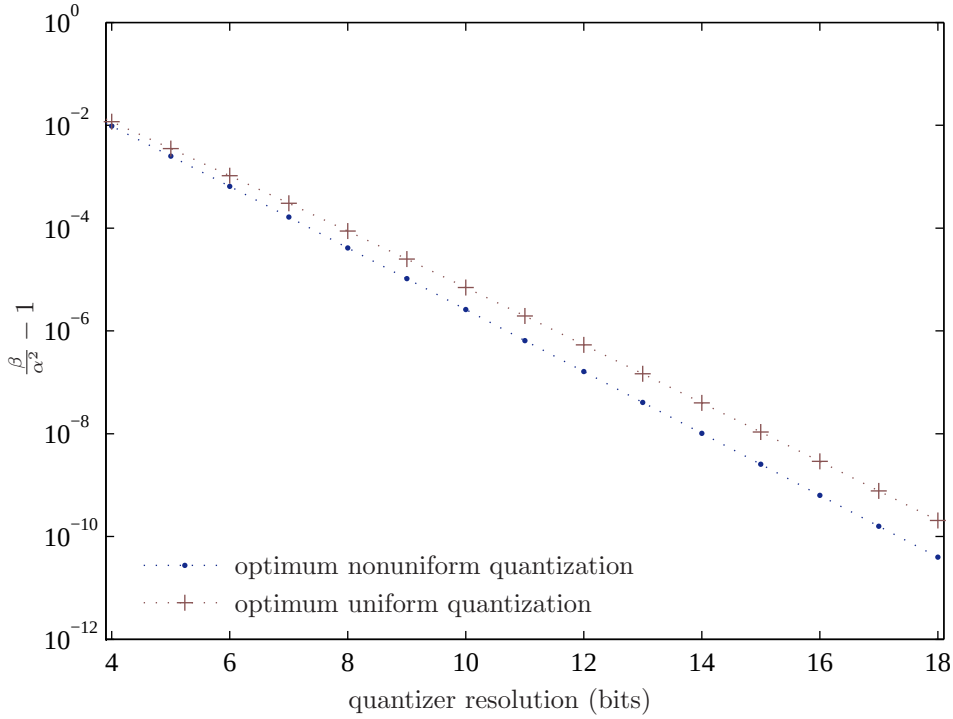


Figure 4.1: Effect of the quantizer resolution on the residual loop interference power expressed as a fraction of the desired signal power in an ideal system.

power expressed as a fraction of the desired signal power and tells us that even the slightest degradation of signal quality caused by quantization alone in an otherwise perfect system is sufficient to rule out the possibility of completely canceling the effect of loop interference. Figure 4.1 plots the exact values of $\frac{\beta}{\alpha^2} - 1$ for a number of quantizer resolutions, and it includes both schemes of quantization discussed in Chapter 3: uniform and optimum nonuniform. For the uniform quantizer, the clipping margin is chosen such that the signal-to-quantization noise ratio gets maximized (refer to Section 3.1.1 for more details).

An Example with the Best Linear Unbiased Estimator

Up to this point in this chapter, we have purposefully kept our discussion regarding the estimation and cancellation of the loop interference as general as possible and accordingly derived an expression for the SINR after cancellation that holds good regardless of a specific estimation technique. Before we conclude this chapter, however, we will consider, as an example, a widely applied standard estimation technique and see how the general SINR expression in (4.23) turns out for this specific case.

Let us once again consider the representations in (4.10) and (4.12) for the signal in the relay node prior to loop interference cancellation. If we make a typical (and reasonable) assumption that the observation noise $u_R[n]$ is white,

then $\mathbf{u_R}$ is a zero-mean random vector with elements that are uncorrelated with one another, and, with no restriction on the true probability density function of $\mathbf{u_R}$, the best linear unbiased estimator[1] for the unknown parameter vector $\mathbf{h_{LI}}$ that satisfies the model in (4.12) is given by [48]

$$\hat{\mathbf{h}}_{\mathbf{LI}\,(\mathrm{BLUE})} = \Big(\mathbf{X}_{\mathbf{R}}^{H}\,\mathbf{C}_{\mathbf{u}}^{-1}\,\mathbf{X}_{\mathbf{R}}\Big)^{-1}\,\mathbf{X}_{\mathbf{R}}^{H}\,\mathbf{C}_{\mathbf{u}}^{-1}\,\mathbf{z}_{\mathbf{R}}, \tag{4.25}$$

where $\mathbf{C_u} = \mathbb{E}\{\mathbf{u_R}\,\mathbf{u_R^H}\}$ is the covariance matrix of the zero-mean noise vector $\mathbf{u_R}$. For this estimator, the estimation error vector defined in (4.14) becomes

$$\tilde{\mathbf{h}}_{\mathbf{LI}} = \mathbf{h_{LI}} - \hat{\mathbf{h}}_{\mathbf{LI}\,(\mathrm{BLUE})},$$

which is zero on average, and the estimation error covariance matrix defined in (4.15) becomes [48]

$$\mathbf{C}_{\tilde{\mathbf{h}}_{\mathbf{LI}}} = \mathbb{E}\{\tilde{\mathbf{h}}_{\mathbf{LI}}\,\tilde{\mathbf{h}}_{\mathbf{LI}}^{H}\} = \Big(\mathbf{X}_{\mathbf{R}}^{H}\,\mathbf{C}_{\mathbf{u}}^{-1}\,\mathbf{X}_{\mathbf{R}}\Big)^{-1}. \tag{4.26}$$

The diagonal elements of $\mathbf{C}_{\tilde{\mathbf{h}}_{\mathbf{LI}}}$ give the estimation error variances for the individual channel taps.

Since the elements of $\mathbf{u_R}$ are assumed to be uncorrelated with one another, $\mathbf{C_u}$ is a diagonal matrix. If we further assume that the elements of $\mathbf{u_R}$ have identical second-order statistics, the noise covariance matrix takes the form $\mathbf{C_u} = \sigma_u^2\,\mathbf{I_N}$, where $\mathbf{I_N}$ is an $N \times N$ identity matrix and σ_u^2 is the variance of each element in the zero-mean noise vector $\mathbf{u_R}$. Then, (4.25) and (4.26) get reduced to

$$\hat{\mathbf{h}}_{\mathbf{LI}\,(\mathrm{BLUE})} = \Big(\mathbf{X}_{\mathbf{R}}^{H}\,\mathbf{X}_{\mathbf{R}}\Big)^{-1}\,\mathbf{X}_{\mathbf{R}}^{H}\,\mathbf{z}_{\mathbf{R}} \tag{4.27}$$

and

$$\mathbf{C}_{\tilde{\mathbf{h}}_{\mathbf{LI}}} = \sigma_u^2\Big(\mathbf{X}_{\mathbf{R}}^{H}\,\mathbf{X}_{\mathbf{R}}\Big)^{-1}, \tag{4.28}$$

respectively.

Let us now investigate the extent to which the cancellation of loop interference is possible if we use $\hat{\mathbf{h}}_{\mathbf{LI}\,(\mathrm{BLUE})}$ to estimate the loop interference channel. To do so, we first determine the expression, specific to this case, for the average residual loop interference gain $\tilde{G}_{LI}$ defined in (4.22). We then substitute the resulting expression for $\tilde{G}_{LI}$ in (4.24) and see what ρ becomes. As we discussed earlier, the closer is the value of ρ to 1, the better is the extent of loop interference cancellation.

When $\hat{\mathbf{h}}_{\mathbf{LI}} = \hat{\mathbf{h}}_{\mathbf{LI}\,(\mathrm{BLUE})}$, the average residual loop interference channel gain in (4.22) is given by

$$\tilde{G}_{LI} = \mathrm{tr}\Big(\mathbf{C}_{\tilde{\mathbf{h}}_{\mathbf{LI}}}\Big), \tag{4.29}$$

[1]The best linear unbiased estimator $\hat{\boldsymbol{\theta}}_{(\mathrm{BLUE})}$ for an unknown parameter $\boldsymbol{\theta}$ is unbiased, which means that $\mathbb{E}\{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{(\mathrm{BLUE})}\} = \mathbf{0}$, and it has the minimum error variance among all unbiased estimators that can be expressed as linear transformations of the observation vector.

the trace of the (loop interference channel) estimation error covariance matrix $\mathbf{C}_{\tilde{\mathbf{h}}_{\mathbf{LI}}}$ in (4.28).

According to (4.28), $\mathbf{C}_{\tilde{\mathbf{h}}_{\mathbf{LI}}}$ can be obtained by inverting the matrix $\mathbf{X}_{\mathbf{R}}^H \mathbf{X}_{\mathbf{R}}$ and scaling the result by σ_u^2. Taking into consideration the fact that the samples of $x_R[n]$ should typically be independent of one another, one can see from (4.13) that $\mathbf{X}_{\mathbf{R}}^H \mathbf{X}_{\mathbf{R}}$ is, on average, given by

$$\begin{aligned}\mathbb{E}\Big\{\mathbf{X}_{\mathbf{R}}^H \mathbf{X}_{\mathbf{R}}\Big\} &= N\, G_Q\, \alpha^2\, \mathbb{E}\{|x_R|^2\}\, \mathbf{I}_{\mathbf{M}} \\ &= \frac{N\, G_Q\, \alpha^2\, \mathbb{E}\{|i_R|^2\}}{G_{LI}\left(1+\epsilon^2\right)}\, \mathbf{I}_{\mathbf{M}},\end{aligned}$$

where $\mathbf{I}_{\mathbf{M}}$ is an $M \times M$ identity matrix and the second step is the result of using (4.9) to write $\mathbb{E}\{|x_R|^2\}$ in terms of $\mathbb{E}\{|i_R|^2\}$. The inverse of $\mathbf{X}_{\mathbf{R}}^H \mathbf{X}_{\mathbf{R}}$ is thus given by

$$\mathbb{E}\left\{\left(\mathbf{X}_{\mathbf{R}}^H \mathbf{X}_{\mathbf{R}}\right)^{-1}\right\} = \frac{G_{LI}\left(1+\epsilon^2\right)}{N\, G_Q\, \alpha^2\, \mathbb{E}\{|i_R|^2\}}\, \mathbf{I}_{\mathbf{M}}. \tag{4.30}$$

On account of the entities in (4.11), we can write down the variance σ_u^2 of each sample from the white noise process $u_R[n]$ (again, on average) as

$$\sigma_u^2 = \alpha^2\, G_Q \left(\mathbb{E}\{|s_R|^2\} + \sum_{m=0}^{M-1} |h_{LI}[m]|^2\, \mathbb{E}\{|v_R|^2\} + \mathbb{E}\{|w_R|^2\}\right) + \mathbb{E}\{|d|^2\},$$

which, after substituting $\mathbb{E}\{|d|^2\}$ from (4.3) and $\sum_{m=0}^{M-1} |h_{LI}[m]|^2\, \mathbb{E}\{|v_R|^2\}$ using (4.7), (4.8), and (4.9), becomes

$$\begin{aligned}\sigma_u^2 &= \alpha^2\, G_Q \left(\mathbb{E}\{|s_R|^2\} + \frac{\epsilon^2}{1+\epsilon^2}\mathbb{E}\{|i_R|^2\} + \mathbb{E}\{|w_R|^2\}\right) \\ &\quad + \left(\beta - \alpha^2\right) G_Q \left(\mathbb{E}\{|s_R|^2\} + \mathbb{E}\{|i_R|^2\} + \mathbb{E}\{|w_R|^2\}\right) \\ &= \beta\, G_Q\, \mathbb{E}\{|s_R|^2\} + \left(\beta - \frac{\alpha^2}{1+\epsilon^2}\right) G_Q\, \mathbb{E}\{|i_R|^2\} + \beta\, G_Q\, \mathbb{E}\{|w_R|^2\}. \end{aligned} \tag{4.31}$$

Using the expressions in (4.30) and (4.31) to evaluate $\mathbf{C}_{\tilde{\mathbf{h}}_{\mathbf{LI}}}$ in (4.28) and applying the result to (4.29), we get

$$\tilde{G}_{LI} = \frac{M\left(1+\epsilon^2\right) G_{LI}}{N} \left(\frac{\beta}{\alpha^2}\, \frac{\mathbb{E}\{|s_R|^2\}}{\mathbb{E}\{|i_R|^2\}} + \left(\frac{\beta}{\alpha^2} - \frac{1}{1+\epsilon^2}\right) + \frac{\beta}{\alpha^2}\, \frac{\mathbb{E}\{|w_R|^2\}}{\mathbb{E}\{|i_R|^2\}}\right),$$

since $\mathrm{tr}\big(\mathbf{I}_{\mathbf{M}}\big) = M$. Then, ρ in (4.24) becomes

$$\begin{aligned}\rho &= \frac{1}{1+\epsilon^2} - \frac{1}{1+\epsilon^2}\, \frac{\tilde{G}_{LI}}{G_{LI}} \\ &= \frac{1}{1+\epsilon^2} - \frac{M}{N}\left(\frac{\beta}{\alpha^2} - \frac{1}{1+\epsilon^2} + \frac{\beta}{\alpha^2}\, \frac{\mathbb{E}\{|s_R|^2\} + \mathbb{E}\{|w_R|^2\}}{\mathbb{E}\{|i_R|^2\}}\right). \end{aligned} \tag{4.32}$$

Since β is greater than α^2, the quantity within the parentheses in the expression above is positive and ρ is less than $\frac{1}{1+\epsilon^2}$. The parameters α and β remain constant for a given quantizer, and so does the EVM ϵ, more or less, for a chosen transmitter. The largest that ρ in (4.32) can become, while still restricted to values smaller than $\frac{1}{1+\epsilon^2}$, thus depends mainly upon the choice of two other factors. The first is the ratio between the assumed length M of the loop interference channel and the number of observation samples N processed at a time; the second is the sum of the average powers of the desired signal $s_R[n]$ and the additive noise $w_R[n]$ at the receiver, relative to the average power of the loop interference $i_R[n]$. The lower the value of either of these ratios, the greater is the value of ρ, and the better is the SINR. This is not at all unexpected because a larger number of samples available for processing at a time or a stronger loop interference received by the relay should indeed result in a more accurate estimation and cancellation of the loop interference.

Chapter 5

Discussion

We begin this chapter by verifying through simulations that our expression in (4.23) for the signal-to-interference plus noise ratio (SINR) in a full-duplex relay after subtractive loop interference cancellation holds true. Then, by choosing the SINR as the metric for evaluating whether the relay performs satisfactorily in a given scenario, we use the aforementioned expression to numerically determine the acceptable range for the other parameters of interest such as the error vector magnitude (EVM) for the transmitter, the signal-to-interference ratio (SIR) and the signal-to-noise ratio (SNR) at the relay input, and the quantizer resolution in the receiver, one at a time. This will eventually help us develop a systematic way of determining the overall criteria that needs to be fulfilled to ensure that the full-duplex relay operates successfully in a given practical scenario.

5.1 Verification of the SINR Expression

Let us once again consider the expression in (4.23) for the SINR after loop interference cancellation, repeated here for convenience:

$$\gamma = \frac{1}{\left(\frac{\beta}{\alpha^2} - 1\right) + \left(\frac{\beta}{\alpha^2} - \rho\right)\frac{1}{\mathrm{SIR}} + \frac{\beta}{\alpha^2}\frac{1}{\mathrm{SNR}}}, \tag{5.1}$$

where

$$\rho = \frac{1}{1+\epsilon^2}\left(1 - \frac{\tilde{G}_{LI}}{G_{LI}}\right). \tag{5.2}$$

Because the number of parameters upon which the SINR depends is apparently large, it becomes rather difficult to vary all the parameters of interest, each over its practical range, within one huge simulation and present all the results at one place. It is, therefore, preferable that we break this simulation down into a number of smaller ones and examine the validity of the expression by varying only one parameter at a time while keeping all the others constant. We can, nevertheless, choose to group the results from several such simulations into composite plots if doing so makes the presentation more informative.

Since the values of the parameters α and β are determined by the resolution of the quantizer and the chosen quantization scheme, one way of verifying

the dependence of the SINR upon these parameters is to select a quantization scheme, preferably the uniform quantization scheme owing to the simplicity of its use in simulations, and vary the resolution of the quantizer across different simulation runs. However, since α and β depend also upon the clipping margin μ in this case (refer to Section 3.1.1 for more details), it is logical that we hold the clipping margin constant at the value that maximizes the signal-to-quantization noise ratio (see Table 3.1) during an entire simulation run, and then repeat the same for different resolutions.

The parameter ρ is a bit trickier to handle because, as (5.2) says, its value depends further upon the values of two other parameters: the EVM and the residual loop interference channel gain expressed as a fraction of the original channel gain before cancellation. For each simulation run, therefore, we fix the value of the EVM and vary the fractional residual loop interference so as to vary ρ in effect. We then repeat the same for different values of the EVM to verify the dependence of the SINR on both of these parameters.

Figure 5.1 shows the results of varying the aforementioned parameters one at a time while keeping the others constant. It is a composite plot showing the results of six different simulation runs (represented by dotted trails) compared with their analytical counterparts (represented by solid lines) given by (5.1). In each run, the fractional residual interference channel gain (i.e., $\tilde{G}_{LI}$ expressed as a fraction of G_{LI}) is gradually varied from $-60\,\text{dB}$ to $-10\,\text{dB}$ and the resulting SINR is plotted while keeping all the other parameters constant. Across all runs, the SIR is held constant at $-10\,\text{dB}$ (indicating stronger loop interference than the desired signal at the relay input) and the SNR at $30\,\text{dB}$ (a reasonable practical value). The six cases result from setting the EVM at three different levels (1%, 2%, and 5%) one at a time, each repeated for two different resolutions of the quantizer (8 and 12 bits). In all cases, the overlapping of the solid lines with the dotted trails indicates that the analytical results agree with those obtained from the simulations, thereby verifying the validity of the expression in (5.1) at least partially.

An important observation that is also illustrated by this composite plot is that the higher the value of the EVM, the lower is the improvement in the SINR that comes with an increased resolution of the quantizer. This is because the EVM, which represents the strength of the unknown portion of the loop interference that cannot be canceled no matter how precise the digital representation of the incoming signal in the relay is, becomes increasingly dominating. This is well demonstrated in the figure by how close the two lines corresponding to the 8-bit and 12-bit quantizers get when the EVM is set to 5%.

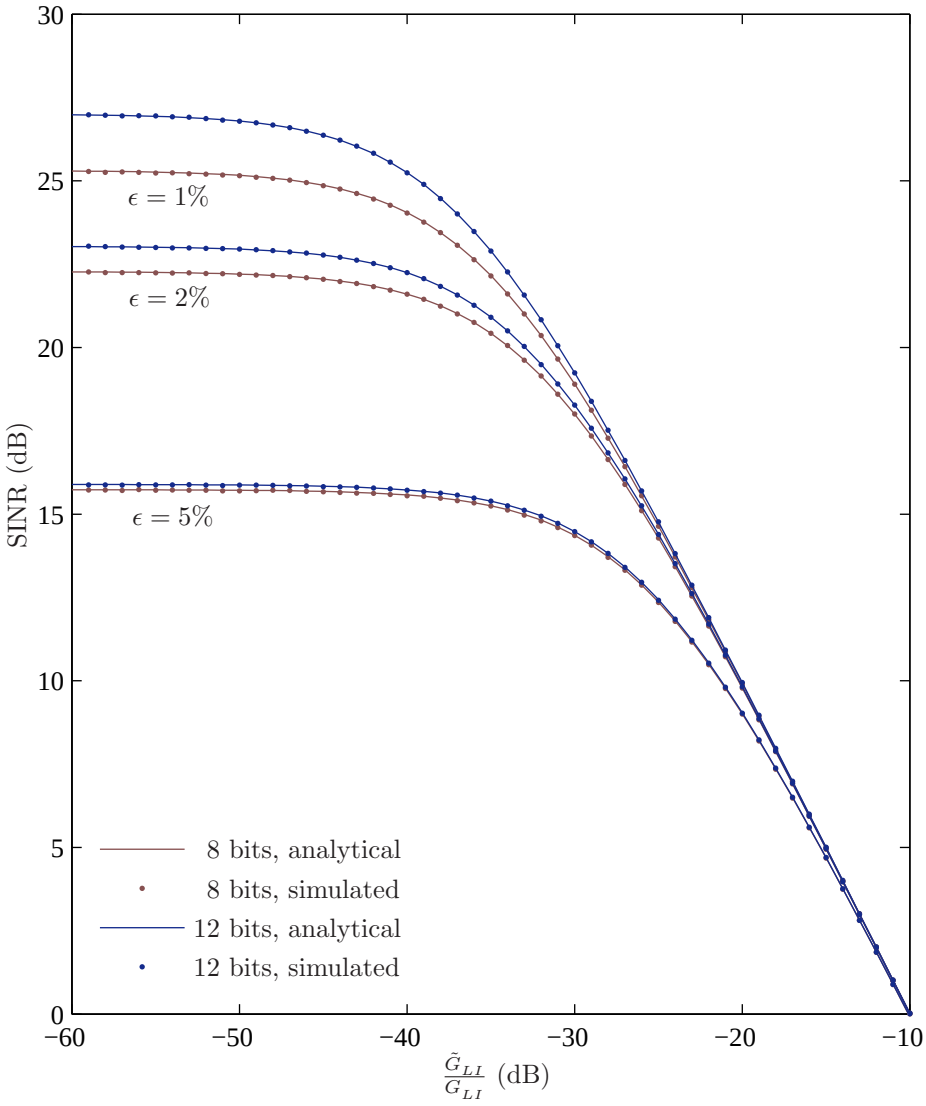


Figure 5.1: Effect of the residual loop interference channel gain, the transmitter EVM, and the quantizer resolution on the signal-to-interference plus noise ratio after loop interference cancellation. In all cases, the signal-to-interference ratio is held constant at $-10\,\mathrm{dB}$ and the signal-to-noise ratio at $30\,\mathrm{dB}$.

To verify the dependence of the SINR in (5.1) upon the two remaining parameters, namely, the SIR and the SNR at the relay input, we take a similar approach as in the previous experiment. This time, we hold the SNR constant during each simulation run and vary the input SIR over a wide range of values so as to examine its impact on the SINR after loop interference cancellation. We then repeat the same for different values of the SNR (30dB and 40dB) and again for different resolutions of the quantizer (8, 10, and 12 bits); however, we fix the EVM and the fractional residual loop interference channel gain constant (at 1% and -60 dB, respectively) throughout all runs in this experiment.

Figure 5.2 illustrates the results obtained from this experiment along with those obtained analytically from (5.1). This time too, all the solid lines overlap with the dotted trails, thereby completing the verification of the validity of (5.1) when viewed together with the results of the previous experiment. And because the EVM is again the dominating parameter, a quantizer with a higher resolution does not necessarily bring about a significant improvement in the SINR as demonstrated by the proximity of the lines corresponding to 10 and 12-bit quantizers. Lastly, Figure 5.2 also demonstrates the well-understood observation that the received SNR places a limit on the maximum achievable SINR, no matter how favorable the values of the remaining parameters are.

5.2 Criteria for Successful Relay Operation

Now that the validity of the closed-form SINR expression in (5.1) has been established, let us see how it can be applied to determine the criteria necessary for ensuring that a full-duplex relay with subtractive loop interference cancellation performs satisfactorily in practical scenarios. This requires, as mentioned in the beginning of this chapter, choosing the SINR to be the performance metric and assigning a minimum threshold that the SINR must attain in order to label the performance as acceptable. Then, all that needs to be done is solving (5.1) for the parameter of interest by substituting γ with the chosen threshold and the remaining parameters with values that are appropriate for the given scenario.

Example 1: Maximum Tolerable Relay Transmit Power

Let us consider a scenario in which the distance of the relay from the base station (i.e., the source node) is such that the average received SNR is, say, 30 dB. This is the case when, e.g., the transmit power of the base station is 40 W (i.e., 46 dBm), the attenuation due to propagation between the base station and the relay is 120 dB, and the noise level at the relay receiver is -104 dBm. Let us assume that the physical design of the relay and the surrounding infrastructure offer a certain amount of isolation between the transmitting and the receiving antennas of the relay. Let us also assume that the EVM for the relay transmitter is known to be 0.1% (indicating an excellent transmitter), and that the loop interference channel can be estimated well enough to keep the value of $\tilde{G}_{LI}$ at 60 dB below that of G_{LI}.

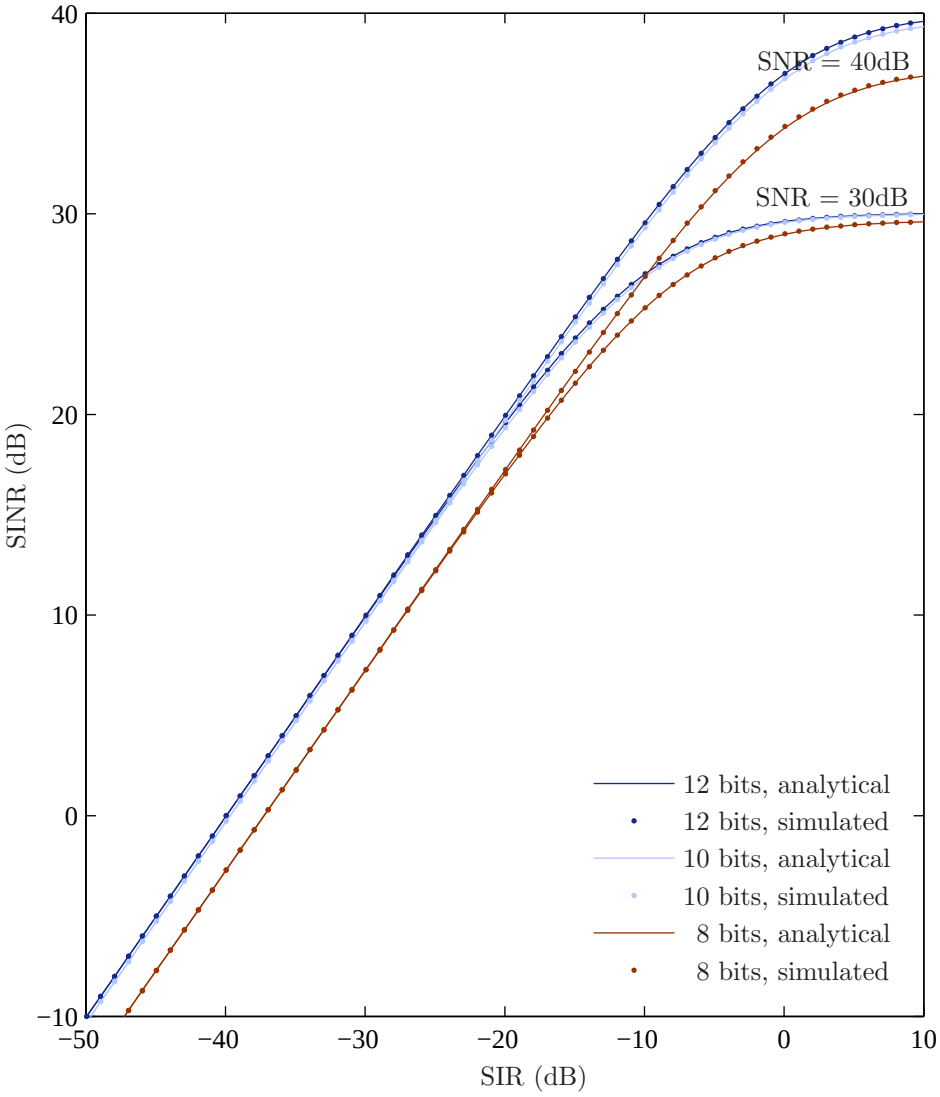


Figure 5.2: Effect of the signal-to-interference ratio, the signal-to-noise ratio, and the quantizer resolution on the signal-to-interference plus noise ratio after subtractive loop interference cancellation. In all cases, the transmitter EVM is held constant at 1% and the fractional residual loop interference channel gain at −60 dB.

Then, one might be interested in determining the maximum relay transmit power that can safely be used for a given amount of isolation existing between the two antennas of the relay while ensuring that the SINR after subtractive loop interference cancellation does not fall below, say, 20 dB (which is 10 dB less than the received SNR). To accomplish this by using (5.1), we first express the SIR as

$$\mathrm{SIR} = \frac{P_S\, G_{SR}}{P_R\, G_{LI}}, \tag{5.3}$$

where P_S is the transmit power of the base station, G_{SR} is the average power gain of the base station-to-relay channel, P_R is the transmit power of the relay, and G_{LI} is the average power gain of the loop interference channel. Then, by substituting the expression above in (5.1) and rearranging the terms to isolate the relay transmit power P_R, we get

$$P_R \leq \frac{1 + \frac{1}{\gamma_{\mathrm{th}}} - \frac{\beta}{\alpha^2}\left(1 + \frac{1}{\mathrm{SNR}}\right)}{\left(\frac{\beta}{\alpha^2} - \rho\right)\frac{G_{LI}}{P_S\, G_{SR}}}, \tag{5.4}$$

where γ_{th} is the SINR threshold of acceptable performance.

The values of the parameters required for evaluating the expression above can be obtained as follows.

- α and β should be obtained by evaluating (3.9) and (3.13) for the desired quantization scheme (and resolution). It should be stressed that the values of α and β do not depend on the actual signal voltage at the relay receiver as it is scaled to the value that is optimum for the chosen quantization scheme by the low-noise amplifier (with automatic gain control) prior to quantization (refer to Section 4.1 for details on the signal model).

- The average loop interference channel gain G_{LI} is given simply by the reciprocal of the isolation present between the transmitting and the receiving antennas of the relay.

- The values of the remaining parameters can be derived from the scenario description given above as

$$\begin{aligned}
\gamma_{\mathrm{th}} &= 20\,\mathrm{dB} = 100,\\
\mathrm{SNR} &= 30\,\mathrm{dB} = 1000,\\
\rho &= \frac{1}{1+\epsilon^2}\left(1 - \frac{\tilde{G}_{LI}}{G_{LI}}\right)\\
&= \frac{1}{1+0.001^2}\left(1 - 10^{-6}\right) = 0.999998,\\
P_S &= 40\,\mathrm{W},\ \text{and}\\
G_{SR} &= -120\,\mathrm{dB} = 10^{-12}.
\end{aligned}$$

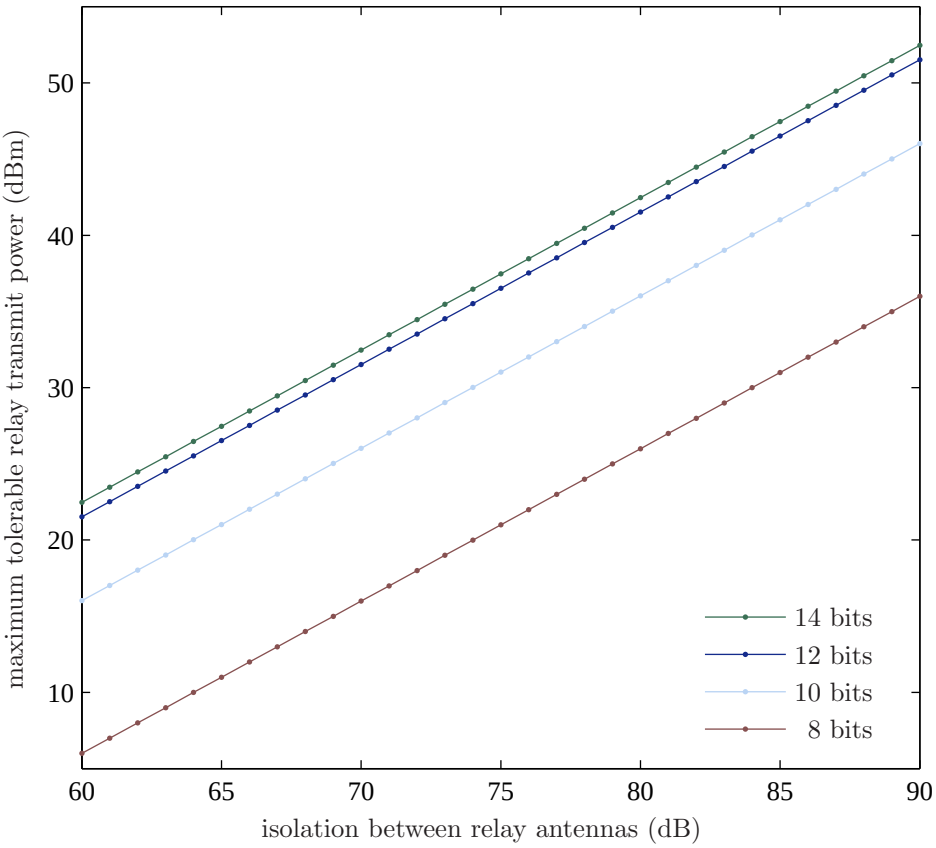


Figure 5.3: Effect of the amount of isolation present between the relay antennas and the quantizer resolution upon the maximum transmit power that can be used while maintaining the SINR threshold of 20 dB. In all cases, the source transmit power is held constant at 40 W, the source-to-relay channel gain at −120 dB, the transmitter EVM at 0.1%, and the fractional residual loop interference channel gain at −60 dB.

Figure 5.3 shows the results obtained by evaluating the expression in (5.4) (with the parameter values as listed above) for the amount of isolation between the two antennas in the relay ranging from 60 dB to 90 dB. The multiple plots in the figure represent different resolutions of the uniform quantizer, each operating at its optimum clipping margin. The plots tell us that the maximum tolerable relay transmit power increases at a constant rate with an increase in the isolation between the two antennas in the relay and that this rate of increase is more or less independent of the quantizer resolution. However, as the quantizer resolution goes higher and higher, the improvement brought about the increased quantizer resolution becomes smaller and smaller. This is because, as mentioned earlier, the part of the transmitted signal that is quantified by the EVM is unknown to the loop interference canceler within the relay and, therefore, cannot be canceled

no matter how accurately the incoming signal is digitized.

Example 2: Minimum Isolation Required Between Relay Antennas

As our second example, let us consider a scenario very similar to the one in the previous example except that, this time, we are interested in determining the minimum amount of isolation that must exist in between the transmitting and the receiving antennas of the relay in order to perform satisfactorily for a given relay transmit power. This might be of interest when, for example, it has been established that the relay needs to transmit with a certain power in order to ensure coverage to a certain region and the system designer needs to determine and enforce the minimum amount of isolation necessary between the two antennas in the relay so as to keep the SINR after loop interference cancellation at or above the minimum threshold of acceptable performance.

Following a process similar to that in Example 1, we arrive at the following expression for the minimum isolation required between the two antennas in the relay:

$$\frac{1}{G_{LI}} \geq \frac{\left(\frac{\beta}{\alpha^2} - \rho\right) \frac{P_R}{P_S G_{SR}}}{1 + \frac{1}{\gamma_{\text{th}}} - \frac{\beta}{\alpha^2}\left(1 + \frac{1}{\text{SNR}}\right)}. \tag{5.5}$$

Using the same parameters values as in the previous example, (5.5) gives us the results presented in Figure 5.4. The plots on the figure illustrate one simple observation that a higher amount of isolation is required between the transmitting and the receiving antennas of the relay if it has to transmit at a higher power level. Also, because of the same reason as in the previous example, having a quantizer of a higher resolution does not always significantly ease the requirement on the minimum isolation necessary for acceptable performance.

Example 3: Minimum SIR Required at the Relay Input

As our final example, let us consider a scenario where the only known parameter is the transmit power of the base station, and based on this information, one has to determine the optimum location and the transmit power of the relay. The proper choice of both of these parameters is crucial as they jointly define the coverage of the relay. Besides coverage, the location of the relay also has a definite impact on the path loss suffered by the useful signal coming from the base station and hence the received SNR at the relay input. The second parameter, i.e., the transmit power of the relay, together with the information provided by the first parameter and the information on the amount of isolation that can be enforced between the two antennas of the relay, determines the SIR at the relay input. Then, for ascertaining the optimum location and the transmit power of the relay, one first needs to determine, as a function of the received SNR, the minimum SIR at the relay input required to ensure that the SINR after loop interference cancellation exceeds the minimum threshold of acceptable performance.

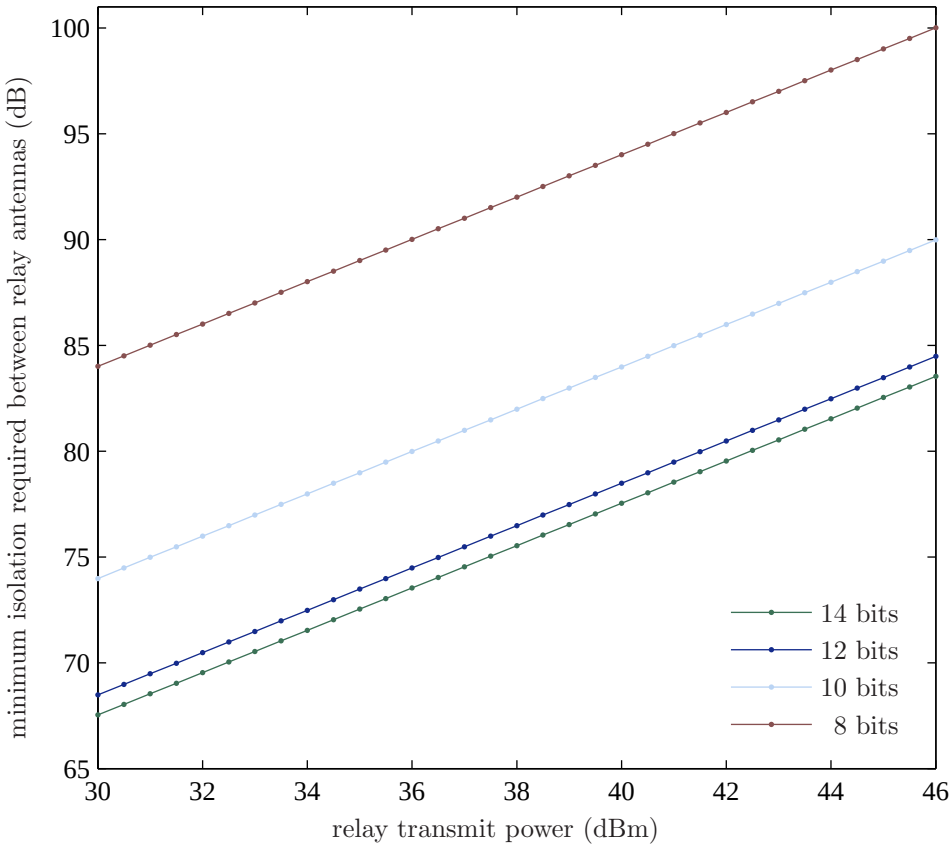


Figure 5.4: Effect of the relay transmit power and the quantizer resolution on the amount of minimum isolation required between the relay antennas so as to maintain the SINR threshold of 20 dB. In all cases, the source transmit power is held constant at 40 W, the source-to-relay channel gain at −120 dB, the transmitter EVM at 0.1%, and the fractional residual loop interference channel gain at −60 dB.

Rearranging the terms in (5.1) so as to isolate the SIR, we arrive at

$$\text{SIR} \geq \frac{\frac{\beta}{\alpha^2} - \rho}{1 + \frac{1}{\gamma_{\text{th}}} - \frac{\beta}{\alpha^2}\left(1 + \frac{1}{\text{SNR}}\right)}. \tag{5.6}$$

Using the same values for α, β, ρ, and γ_{th} as in Example 1, the expression above gives us the results presented in Figure 5.5. It can be seen from each plot that the range of the SIR that can be tolerated after loop interference cancellation improves when the input SNR increases and that this improvement is the most remarkable for the input SNR values that are close to the minimum SINR threshold γ_{th} (which, in this example, has been set to be 20 dB).

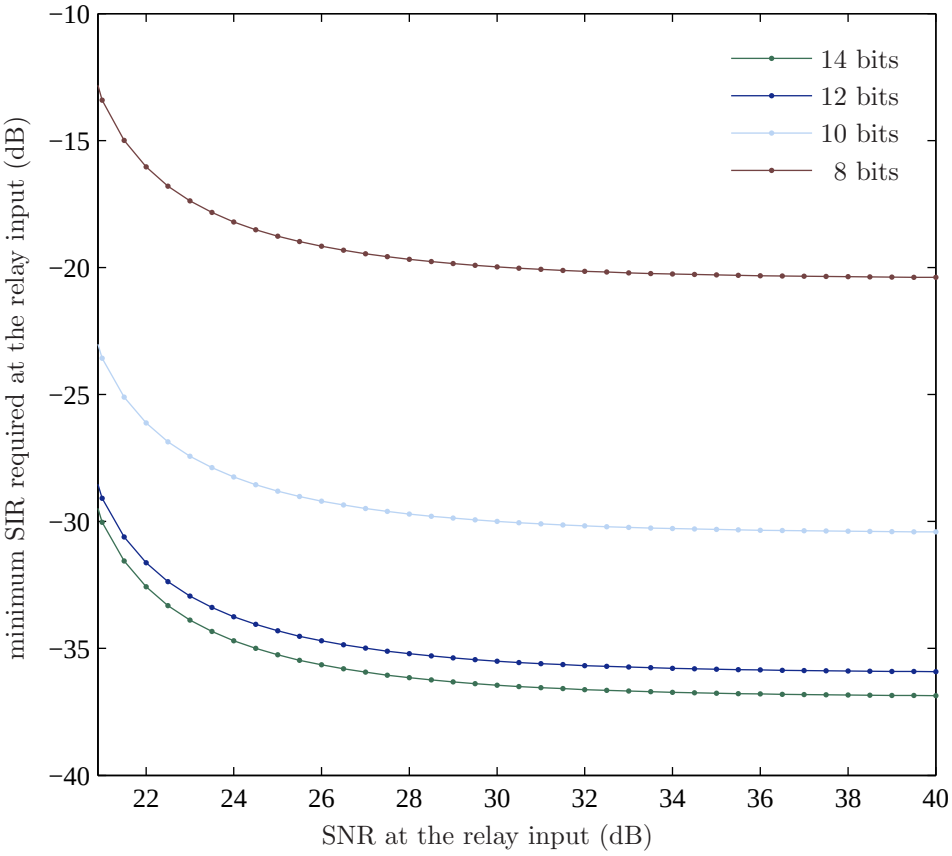


Figure 5.5: Effect of the received SNR and the quantizer resolution on the minimum SIR that is required at the relay input in order to maintain the SINR threshold of 20 dB. In all cases, the transmitter EVM is held constant at 0.1% and the fractional residual loop interference channel gain at −60 dB.

Chapter 6

Conclusion

The results presented in this thesis show that the idea of full-duplex relaying in OFDM systems can indeed be put to practice provided that certain criteria are fulfilled. As expected, it is necessary to have a quantizer with a good resolution (and hence a large dynamic range) at the relay receiver so as to ensure that the incoming signal, which is a superposition of a weak useful signal and a much stronger loop interference, can be digitized with sufficient accuracy. Another important requirement is to have the relay equipped with an excellent transmitter characterized with a very small EVM figure. This is because the error in the transmitted signal is unknown to the processing unit within the relay and, therefore, its contribution to the loop interference cannot be canceled no matter how accurate the digital representation of the incoming signal is. Moreover, prior to applying any loop interference cancellation, the physical design of the relay (along with the surrounding infrastructure) must, by itself, be able to provide a certain amount of natural isolation between its transmitting and receiving antennas; otherwise, the part of the loop interference contributed by the transmitter error alone can be sufficient to drown the useful signal, thereby rendering any further processing fruitless. With the framework developed in this thesis, it becomes easy to analyze the connection between all these aspects of practical system design.

Moving forward, an interesting extension to this work could be a more detailed study of the effects of the most prominent if not all transmitter nonidealities on the signal model so as to be able to better parameterize the SINR expression as opposed to using the single EVM parameter to represent all of them. Further extension would then be to look for possible ways to compensate the effect of each nonideality as this would potentially improve the achievable SINR and make the full-duplex relay even more feasible. Another interesting area to which this research could be extended is the area of multiple-input-multiple-output (MIMO) full-duplex relays briefly introduced in Section 1.2 of this thesis.

Bibliography

[1] E. Van Der Meulen, "Three-Terminal Communication Channels," *Advances in Applied Probability*, pp. 120–154, 1971.

[2] T. Cover and A. Gamal, "Capacity Theorems for the Relay Channel," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, 1979.

[3] A. Host-Madsen and J. Zhang, "Capacity Bounds and Power Allocation for Wireless Relay Channels," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2020–2040, 2005.

[4] V. Cadambe and S. Jafar, "Degrees of Freedom of Wireless Networks with Relays, Feedback, Cooperation, and Full Duplex Operation," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2334–2344, 2009.

[5] W. Slingsby and J. McGeehan, "Antenna Isolation Measurements for On-Frequency Radio Repeaters," in *Proceedings of the 9th IET International Conference on Antennas and Propagation*, pp. 239–243, 1995.

[6] C. Anderson, S. Krishnamoorthy, C. Ranson, T. Lemon, W. Newhall, T. Kummetz, and J. Reed, "Antenna Isolation, Wideband Multipath Propagation Measurements, and Interference Mitigation for On-Frequency Repeaters," in *Proceedings of the IEEE Southeast Conference*, pp. 110–114, 2004.

[7] H. Hamazumi, K. Imamura, N. Iai, K. Shibuya, and M. Sasaki, "A Study of a Loop Interference Canceller for the Relay Stations in an SFN for Digital Terrestrial Broadcasting," in *Proceedings of the IEEE Global Telecommunications Conference*, pp. 167–171, 2000.

[8] K. Salehian, M. Guillet, B. Caron, and A. Kennedy, "On-Channel Repeater for Digital Television Broadcasting Service," *IEEE Transactions on Broadcasting*, vol. 48, no. 2, pp. 97–102, 2002.

[9] K. Nasr, J. Cosmas, M. Bard, and J. Gledhill, "Performance of an Echo Canceller and Channel Estimator for On-Channel Repeaters in DVB-T/H Networks," *IEEE Transactions on Broadcasting*, vol. 53, no. 3, pp. 609–618, 2007.

[10] T. Riihonen, S. Werner, and R. Wichman, "Comparison of Full-Duplex and Half-Duplex modes with a Fixed Amplify-and-Forward Relay," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, 2009.

[11] T. Riihonen, S. Werner, R. Wichman, and Z. Eduardo, "On the Feasibility of Full-Duplex Relaying in the Presence of Loop Interference," in *Proceedings of the 10th IEEE Workshop on Signal Processing Advances in Wireless Communications*, pp. 275–279, 2009.

[12] T. Riihonen, S. Werner, and R. Wichman, "Rate-Interference Trade-off Between Duplex Modes in Decode-and-Forward Relaying," in *Proceedings of the 21st IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, pp. 690–695, 2010.

[13] T. Riihonen, S. Werner, and R. Wichman, "Optimized Gain Control for Single-Frequency Relaying with Loop Interference," *IEEE Transactions on Wireless Communications*, vol. 8, no. 6, pp. 2801–2806, 2009.

[14] T. Riihonen, S. Werner, and R. Wichman, "Hybrid Full-Duplex/Half-Duplex Relaying with Transmit Power Adaptation," *IEEE Transactions on Wireless Communications*, 2011 (in press).

[15] T. Riihonen, S. Werner, and R. Wichman, "Spatial Loop Interference Suppression in Full-Duplex MIMO Relays," in *Proceedings of the 43rd Asilomar Conference on Signals, Systems and Computers*, November 2009.

[16] T. Riihonen, S. Werner, and R. Wichman, "Residual Self-Interference in Full-Duplex MIMO Relays after Null-Space Projection and Cancellation," in *Proceedings of the 44th Asilomar Conference on Signals, Systems and Computers*, pp. 653 –657, November 2010.

[17] P. Larsson and M. Prytz, "MIMO On-Frequency Repeater with Self-Interference Cancellation and Mitigation," in *Proceedings of the 69th IEEE Vehicular Technology Conference*, pp. 1–5, 2009.

[18] T. Riihonen, A. Balakrishnan, K. Haneda, S. Wyne, S. Werner, and R. Wichman, "Optimal Eigenbeamforming for Suppressing Self-Interference in Full-Duplex MIMO Relays," in *Proceedings of the 45th IEEE Conference on Information Sciences and Systems (CISS)*, 2011.

[19] T. Riihonen, S. Werner, and R. Wichman, "Mitigation of Loopback Self-Interference in Full-Duplex MIMO Relays," *IEEE Transactions on Signal Processing*, 2011 (in press).

[20] T. Riihonen, S. Werner, R. Wichman, and J. Hämäläinen, "Outage Probabilities in Infrastructure-Based Single-Frequency Relay Links," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, April 2009.

[21] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[22] T. Riihonen, K. Haneda, S. Werner, and R. Wichman, "SINR Analysis of Full-Duplex OFDM Repeaters," in *Proceedings of the 20th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 3169–3173, September 2009.

[23] T. Riihonen, R. Wichman, and S. Werner, "Capacity Evaluation of DF Protocols for OFDMA Infrastructure Relay Links," in *Proceedings of the IEEE Global Telecommunications Conference*, December 2009.

[24] I. Hammerström and A. Wittneben, "Power Allocation Schemes for Amplify-and-Forward MIMO-OFDM Relay Links," *IEEE Transactions on Wireless Communications*, vol. 6, pp. 2798–2802, August 2007.

[25] T. Riihonen, S. Werner, J. Cousseau, and R. Wichman, "Design of Co-Phasing Allpass Filters for Full-Duplex OFDM Relays," in *Proceedings of the 42nd Asilomar Conference on Signals, Systems and Computers*, pp. 1030–1034, October 2008.

[26] A. Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*. Prentice Hall, 2008.

[27] S. Wei, D. Goeckel, and P. Kelly, "Convergence of the Complex Envelope of Bandlimited OFDM Signals," *IEEE Transactions on Information Theory*, vol. 56, pp. 4893–4904, October 2010.

[28] S. Haykin, *Communication Systems*. Wiley, 2000.

[29] H. E. Rowe, "Memoryless Nonlinearities with Gaussian Inputs: Elementary Results," *Bell Systems Technical Journal*, vol. 61, no. 7, pp. 1520–1523, 1982.

[30] J. Bussgang, *Crosscorrelation Functions of Amplitude-Distorted Gaussian Signals*. Research Laboratory of Electronics, MIT, 1952.

[31] D. Dardari, "Joint Clip and Quantization Effects Characterization in OFDM Receivers," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, pp. 1741–1748, August 2006.

[32] P. Zillmann, "Relationship Between Two Distortion Measures for Memoryless Nonlinear Systems," *IEEE Signal Processing Letters*, vol. 17, pp. 917–920, November 2010.

[33] J. Max, "Quantizing for Minimum Distortion," *IRE Transactions on Information Theory*, vol. 6, pp. 7–12, March 1960.

[34] S. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137, March 1982.

[35] J. Bucklew and N. Gallagher, Jr., "A Note on the Computation of Optimal Minimum Mean-Square Error Quantizers," *IEEE Transactions on Communications*, vol. 30, pp. 298–301, January 1982.

[36] B. Smith, "Instantaneous Companding of Quantized Signals," *Bell Systems Technical Journal*, vol. 36, no. 3, pp. 653–709, 1957.

[37] F.-S. Lu and G. Wise, "A Further Investigation of Max's Algorithm for Optimum Quantization," *IEEE Transactions on Communications*, vol. 33, pp. 746–750, July 1985.

[38] X. Wu, "On Initialization of Max's Algorithm for Optimum Quantization," *IEEE Transactions on Communications*, vol. 38, pp. 1653–1656, October 1990.

[39] G. Fettweis, M. Löhning, D. Petrovic, M. Windisch, P. Zillmann, and W. Rave, "Dirty RF: A New Paradigm," *International Journal of Wireless Information Networks*, vol. 14, no. 2, pp. 133–148, 2007.

[40] A. Garcia Armada, "Understanding the Effects of Phase Noise in Orthogonal Frequency Division Multiplexing (OFDM)," *IEEE Transactions on Broadcasting*, vol. 47, no. 2, pp. 153–159, 2001.

[41] P. Mathecken, T. Riihonen, S. Werner, and R. Wichman, "Performance Analysis of OFDM with Wiener Phase Noise and Frequency Selective Fading Channel," *IEEE Transactions on Communications*, vol. 59, pp. 1321–1331, May 2011.

[42] T. Riihonen, S. Werner, F. Gregorio, R. Wichman, and J. Hamalainen, "BEP Analysis of OFDM Relay Links with Nonlinear Power Amplifiers," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, 2010.

[43] H. Ochiai and H. Imai, "Performance Analysis of Deliberately Clipped OFDM Signals," *IEEE Transactions on Communications*, vol. 50, pp. 89–101, January 2002.

[44] J. Armstrong, "Peak-to-Average Power Reduction for OFDM by Repeated Clipping and Frequency Domain Filtering," *IET Electronics Letters*, vol. 38, pp. 246–247, February 2002.

[45] A. Georgiadis, "Gain, Phase Imbalance, and Phase Noise Effects on Error Vector Magnitude," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 2, pp. 443–449, 2004.

[46] I. Kotzer, S. Har-Nevo, S. Sodin, and S. Litsyn, "An analytical Approach to the Calculation of EVM in Clipped OFDM Signals," in *Proceedings of the 26th IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, pp. 193–197, 2010.

[47] M. McKinley, K. Remley, M. Myslinski, J. Kenney, D. Schreurs, and B. Nauwelaers, "EVM Calculation for Broadband Modulated Signals," in *64th ARFTG Conference Digest*, pp. 45–52, 2004.

[48] S. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall, 1993.

[49] J.-J. van de Beek, O. Edfors, M. Sandell, S. Wilson, and P. Borjesson, "On Channel Estimation in OFDM Systems," in *Proceedings of the 45th IEEE Vehicular Technology Conference*, vol. 2, pp. 815–819, July 1995.

[50] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 2002.