Quantization-Based Watermarking: Methods for Amplitude Scale Estimation, Security, and Linear Filtering Invariance

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. J. T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 12 maart 2007 om 12.30 uur

door

Ivo Dimitrov SHTEREV

Master of Electronic Engineering
Technical University of Sofia at Plovdiv
geboren te Plovdiv, Bulgarije

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. ir. R. L. Lagendijk

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. ir. R. L. Lagendijk	Technische Universiteit Delft, promotor
Prof. dr. ir. A. J. van der Veen	Technische Universiteit Delft
Prof. dr. ir. H. J. Sips	Technische Universiteit Delft
Prof. dr. ir. L. J. van Vliet	Technische Universiteit Delft
Prof. dr. ir. P. H. Hartel	Universiteit Twente
Dr. M. van der Veen	Philips Research Europe
Prof. dr. B. Macq	Université Catholique de Louvain, Belgium

Quantization-Based Watermarking: Methods for Amplitude Scale Estimation, Security, and Linear Filtering Invariance
Shterev, Ivo Dimitrov
Thesis Delft University of Technology - With Ref. - With Summary in Dutch
ISBN/EAN 978-90-9021723-9

Abstract

Watermarking is the process of imperceptibly embedding a message (watermark) into a host signal (audio, video). The resulting signal is called a watermarked signal. The message should introduce only tolerable distortion to the host signal and it should be recoverable by the intended receiver after signal processing operations on the watermarked data.

Watermarking schemes based on quantization theory have emerged as a result of information theoretic analysis. In terms of additive noise attacks, these schemes have proven to perform better than traditional spread spectrum watermarking because they can completely cancel the host signal interference, which makes them invariant to the host signal. The existence of good lattices in high dimensions that can be directly and efficiently implemented has made quantization-based schemes of practical interest.

Quantization (Lattice)-based schemes are vulnerable to amplitude scale and linear filtering attacks, because these attacks introduce mismatch between the encoder and the decoder lattice volumes. Furthermore, these attacks induce a large amount of distortion with respect to the mean squared error, but do not cause significant perceptual degradations. Such operations on watermarked signals are quite common in many applications.

In this thesis we study quantization-based watermarking. We incorporate statistical techniques into quantization-based schemes to build watermarking systems that are robust to amplitude scale and linear filtering attacks. These watermarking systems are applicable in situations where the attacks are unintentional, due to standard signal processing operations. Since traditional quantization-based schemes are not robust to amplitude scale and linear filtering attacks, and due to the frequent presence of these operations in many signal processing applications, we develop amplitude scale estimation procedures for quantization-based watermarking, and construct quantization-based watermarking systems that are robust to linear filtering attacks. The estimation procedures are based on Fourier analysis of the watermarked and attacked signals, and maximum likelihood estimation. The robustness to linear filtering is achieved by applying quantization-based techniques on the amplitudes in the frequency domain.

To develop the estimation procedures we first derive probability density function models of the watermarked and attacked data for general host signals.

We develop a Fourier-based estimation procedure. It exploits the structure in the probability density function of the watermarked data, due to the encoding process. The approach gives accurate results for high watermark-to-noise ratios, for synthetic as well as real signals.

To increase the estimation accuracy for low watermark-to-noise ratios, we develop a maximum likelihood estimation approach. The estimation technique performs well even when there is a mismatch between the probability density function of the host signal and that of the model assumed at the estimator. The estimator also gives accurate results for real host signals (speech, music).

Traditional quantization-based watermarking schemes are not secure, in the sense that an attacker having knowledge of the embedding distortion can reconstruct the decoder and decode (read) the watermark. To increase the security we apply a well known idea, which is the incorporation of subtractive dither into the watermarking system. The dither realization is assumed to be known to the decoder, but not to the attacker. By adding a scaled version of the dither sequence to the watermarked data, we are able to obtain a signal with a distribution that has a clear structure, which we can describe mathematically. Based on that we are able to derive probability density function models and therefore a maximum likelihood estimation procedure for estimation of amplitude scaling factors in the presence of dither. Due to the complexity of the probability expressions, we derive approximations to them under the condition that the dither variance is much smaller than the host signal variance.

For security purposes, we design the dither statistics such that an attacker without having knowledge of the dither realization is not able to decode the watermark. We design the dither using the probability of error of the watermarking scheme as an objective function. It is shown that the uniformly distributed over the base quantization cell dither is sufficient for security purposes. We also show that the security of the system is unaffected by the conditions needed in the approximation of the density models for the estimation of the scaling factor in the presence of subtractive dither.

To reduce the computational complexity of the maximum likelihood estimation approach, we apply a different model of the attack channel, in which the amplitude scaling and additive noise are reversed. While conceptually the two channels are equal, this allows for considerable reduction in estimation complexity. Exploiting this reduction, we show how to jointly estimate the attacker's amplitude scaling factor and noise variance, and apply this technique to synthetic images.

We notice the duality between amplitude scaling and linear filtering, since linear filtering in the time domain can be seen as multiplication in the frequency domain. To improve the robustness to linear filtering, we study an intermediate case where the attack is a multi band amplitude scaling channel, with scaling in the frequency domain. We develop a maximum likelihood procedure for estimating the scaling. We show experimentally that only a few filter coefficients are sufficient for constructing accurate probability density models.

To provide robustness to linear filtering attacks, we implement Rational Dither Modulation in the frequency domain. Due to the finite Fourier transform length, there are errors in the pass-band zone. To reduce these errors we incorporate a windowing operation on each frame. However, this is achieved at the expense of increased distortion of the watermark encoder. To eliminate the distortion due to windowing, we incorporate overlapped windows. The overlapped windows allow for increased watermark payload, at the expense of increased decoding errors in the pass-band zone due to the overlap.

Samenvatting

Watermerken is het proces van het onwaarneembaar inbedden van een boodschap (het watermerk) in een bronsignaal, zoals geluidssignalen, foto's en video. Het resulterende signaal wordt een gewatermerkt signaal genoemd. De ingebedde (of *gecodeerde*) boodschap mag slechts een kleine verstoring in het bronsignaal teweeg brengen maar de boodschap moet wel teruggewonnen kunnen worden uit het gewatermerkte signaal door de watermerkdetector (*decodering*).

Watermerkmethoden gebaseerd op kwantisatietheorie (zogenaamde *QIM* watermerkmethoden) zijn voortgekomen uit informatietheoretische analyse. Deze methoden presteren beter dan traditionele *spread spectrum* watermerkmethoden onder additieve ruisaanvallen omdat ze de interferentie van het bronsignaal volledig teniet doen. Dit maakt QIM methoden invariant voor het bronsignaal. Het bestaan van goede hoogdimensionale roosters (*lattices*) die efficiënt geïmplementeerd kunnen worden, heeft QIM methoden van praktisch belang gemaakt.

Kwantisatie (*lattice*) gebaseerde watermerkmethoden zijn echter kwetsbaar voor aanvallen die de amplitude van het gewatermerkte signaal schalen of die het signaal (lineair) filteren, omdat door deze aanvallen de codeer-en decodeerroosters niet langer overeenstemmen. Bovendien introduceren deze aanvallen een grote gemiddelde kwadratische fout, zonder significante perceptuele verstoringen te veroorzaken. Zulke bewerkingen op gewatermerkte signalen komen echter wel vaak voor in toepassingen.

In dit proefschrift bestuderen we QIM watermerkmethoden. We gebruiken een statistische aanpak om QIM watermerkmethoden te ontwikkelen die robuust zijn voor amplitudeschaling en lineaire filteringbewerkingen. Deze watermerkmethoden zijn toepasbaar in situaties waar de aanvallen onbedoeld zijn, bijvoorbeeld ten gevolge van gebruikelijke signaalbewerkingsoperaties. De ontwikkelde benadering worden schattingsprocedures gebruikt voor de amplitudeschaling en lineaire filteringbewerking die gebaseerd zijn op Fourieranalyse en op een *maximum likelihood* benadering. De robuustheid voor lineair filteren wordt bereikt door QIM watermerkmethoden op amplitudes in het frequentiedomein toe te passen.

Om de schattingsprocedures te ontwikkelen, leiden we eerst kansdichtheidsfuncties af voor de gewatermerkte en aangevallen gewatermerkte signalen.Vervolgens ontwikkelen we een Fourier-gebaseerde schattingsprocedure. Deze methode maakt gebruik de structuur in de kansdichtheidsfunctie van de gewatermerkte data ten gevolge van het coderingsproces. De aanpak geeft nauwkeurige resultaten voor zowel synthetische als realistische muzieksignalen mits de watermerk-ruisverhouding voldoende hoog is.

Om de schattingsnauwkeurigheid voor lage watermerk-ruisverhoudingen te verbeteren, ontwikkelen we een *maximum likelihood* schattingsaanpak. De schattingstechniek presteert ook goed wanneer er een verschil is tussen de kansdichtheidsfunctie van het bronsignaal en

die van het aangenomen model bij de schatter. De schatter geeft ook nauwkeurige resultaten voor realistische signalen zoals muziek.

Traditionele QIM watermerkmethoden zijn niet veilig, in de zin dat een aanvaller met kennis van het inbeddingsproces de door het watermerk gedragen boodschap kan decoderen en daardoor ook verstoren. Om de veiligheid te vergroten, passen we een welbekend idee toe, namelijk de toevoeging van *subtractive dither* in het watermerksysteem. Het gekozen *dither*-signaal wordt verondersteld bekend te zijn bij de decoder, maar niet bij de aanvaller. Door een geschaalde versie van het *dither*-signaal aan de gewatermerkte data toe te voegen, zijn we in staat een signaal te verkrijgen waarvan de kansdichtheidsfunctie een structuur heeft die wederom direct afhankelijk is van de amplitudeschaling. We kunnen wederom een *maximum likelihood* schattingsprocedure af leiden voor het schatten van amplitudeschalingsfactor, nu in de aanwezigheid van *subtractive dither*. Omdat de kansdichtheidsfunctie erg ingewikkeld is, leiden we een benadering af voor het (meest gebruikelijke) geval dat de variantie van de *dither* veel kleiner is dan de variantie van het bronsignaal.

Om maximale veiligheid te verkrijgen kiezen we het gedrag van de *dither* zo dat een aanvaller zonder kennis van het gekozen *dither*-signaal niet in staat is het watermerk te decoderen. We tonen aan dat als de *dither* uniform verdeeld is met een juist gekozen variantie, het watermerk systeem veilig is. We tonen bovendien aan dat de veiligheid van het systeem niet aangetast wordt door de aannamen die nodig zijn voor het benaderen van de kansdichtheidsfunctie voor de schatting van de amplitudeschaalfactor in de aanwezigheid van *subtractive dither*.

Om de rekenlast van de *maximum likelihood* schattingmethode te verlagen, passen we een ander model van het aanvalskanaal toe waarin de amplitudeschaling en additieve ruisterm in volgorde verwisseld zijn. Dit leidt tot een aanzienlijke afname van de schattingscomplexiteit, terwijl de twee modellen conceptueel hetzelfde zijn. Tevens laten we zien dat de amplitudeschalingsfactor en variantie van de additieve ruisterm gezamenlijk geschat kunnen worden. We passen deze methode toe op gewatermerkte beelden.

Het probleem van lineair filteren van gewatermerkte signalen wordt als volgt aangepakt. Er bestaat een dualiteit tussen amplitudeschaling en lineair filteren: lineair filteren in het tijddomein komt tenslotte neer op vermenigvuldigen (schaling) in het frequentiedomein. Om robuustheid voor lineair filteren te verkrijgen, bestuderen we de situatie waarin de aanval een multiband amplitudeschalingskanaal is, dat wil zeggen dat enkele frequentiebanden van het gewatermerkte signaal geschaald worden. We ontwikkelen een *maximum likelihood* schattingmethode voor de schalingsfactoren van de frequentiebanden. Aan de hand van experimenten laten we zien dat slechts een paar filtercoëfficienten voldoende zijn voor het nauwkeurig bepalen van de kansdichtheidsmodellen van de gewatermerkte frequentiebanden.

Als laatste kiezen we een recente variant op de QIM watermerkmethode, namelijk *rational dither modulation*, en passen deze toe in het frequentiedomein. Ook op deze manier wordt de robuustheid voor lineaire filter-aanvallen vergroot. Er treden echter fouten op in de doorlaatband van de toegepaste filters ten gevolge van de eindige lengte van de Fouriertransformatie. Om deze fouten te verminderen, passen we een vensteringsoperatie toe op elk segment van het bronsignaal. Dit leidt echter tot een grotere verstoring van het bronsignaal. Om de verstoring door vensteren te verwijderen, passen we overlappende vensters toe. De overlappende vensters leiden tevens tot een grotere watermerkcapaciteit, echter ten koste van een toename in decodeerfouten in de doorlaatband van de lineaire filters.

Table of Contents

Symbols and Abbreviations

$\boldsymbol{X} \cdot \boldsymbol{Y}$	dot product of vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$
$\lvert \cdot \rvert$	absolute value and determinant
$\lVert \cdot \rVert$	L_2 norm
$\lfloor x \rfloor$	the greatest integer smaller or equal to x
$\bigcup$	union of two events
$AWGN$	additive white Gaussian noise
A	indicator set
$B^n(r)$	n-dimensional ball with radius r
C	capacity
D	dither
DAB	digital audio broadcasting
$D(X,Y)$	distortion between X and Y
DFT	discrete Fourier transform
DWR	document-to-watermark ratio
E_E	error exponent
E_{FA}	error exponent for false alarm
E_M	error exponent for miss
$f_X(x)$	probability density function of X
$g(y_{k-1})$	equivalent to $\left(\frac{1}{L}\sum_{m=k-L}^{k-1} \lvert y_m \rvert^p\right)^{\frac{1}{p}}$, where L is the memory of the system and $p \geq 1$
$G(\omega)$, $H(\omega)$	transfer function of analysis and synthesis filters
$G_n(\Lambda)$	normalized second moment of Λ
G_n^*	normalized second moment of a ball
$h(\tau)$	filter impulse response
I	identity matrix
I_A	indicator function
$\mathbb{I}^+$	the set of positive integers
$I(X,Y)$	mutual information between X and Y
$K_X[\cdot]$	$n \times n$ correlation matrix of X
L_f	filter length (order)
LF	likelihood function
LTI	linear time invariant
MAP	maximum a posteori probability
ML	maximum likelihood
mod	modulus operation

n	signal length
N	length of DFT
N_1	watermark signal
N_2	attack noise
N_f	additive noise due to filter operation
PDF	probability density function
$p_{X,Y}(x,y)$	joint distribution
$p_{Y\|X}(y\|x)$	conditional probability of Y given X
P_e	probability of error
$P_{FA}^{(n)}$	probability of false alarm for dimension n
$P_M^{(n)}$	probability of miss for dimension n
$Pr(\cdot)$	probability of an event
$q(x)$	equivalent to $\int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$
$Q(\cdot)$	quantization
RDM	rational dither modulation
$R(\cdot)$	rate-distortion function
R_c, R_e	covering and equivalent radius of a lattice
R_N	residual term due to finite DFT
$SAWGN$	scale additive white Gaussian noise
SS	spread-spectrum
$t(k)$	k sample of the impulse response of multiband scale attack
T	threshold
$T(\omega)$	transfer function of multiband scale attack
U	output of a quantizer
V_0	Voronoi cell
W	index of a watermark message
WNR	watermark-to-noise ratio
$\hat{X}$	estimate of X
$\mathbb{R}^n$	n-dimensional vector space
$X(k)$	k sample of X
$\widetilde{X}$, $\widetilde{\boldsymbol{X}}$	host data random variable and vector
X, $\boldsymbol{X}$	watermarked data random variable and vector
X, $\boldsymbol{X}'$	multiband attacked random variable and vector
Y, $\boldsymbol{Y}$	attacked data random variable and vector
$\boldsymbol{Z}$	equivalent to $(1-\alpha)\boldsymbol{N}_1 + \alpha\boldsymbol{N}_2$, where $\alpha = \frac{\sigma_{N_1}^2}{\sigma_{N_1}^2+\sigma_{N_2}^2}$
β, $\boldsymbol{\beta}$	amplitude scale factor scalar and vector
Δ	scalar quantizer step size
$\mathcal{F}_\omega[X]$	Fourier operator applied on X
$\mu(s)$	moment generating function of χ^2
$\phi_X(\omega)$	Fourier transform of X
$\Phi_X(\omega)$	characteristic function of X
Λ	Lattice
$\mathcal{L}(m,\sigma^2)$	Laplacian distribution with mean m and variance σ^2
$\mathcal{N}(m,\sigma^2)$	Normal distribution with mean m and variance σ^2

$\sigma^2_{\widetilde{X}}$	variance of host signal
$\sigma^2_{N_1}$	variance of watermark
$\sigma^2_{N_2}$	variance of attack noise
σ^2_{Z}	variance of $\boldsymbol{Z}$
$\mathcal{U}(m, \sigma^2)$	uniform distribution with mean m and variance σ^2
ω_c	cutoff frequency
χ^2	chi-squared distribution

Chapter 1

Introduction

1.1 The Need for Watermarking

Digital multimedia devices are abundant in our lives. Such devices can easily produce identical copies of the multimedia data. The ease with which digital data can be duplicated has led to the need of its protection. With the advent of broadband internet connections, digital data can be easily spread around the world. Companies producing multimedia data (audio, images, video) suffer major losses due to piracy.

One method to protect digital data is by the use of cryptographic means. Cryptography [1, 2] protects the data during transmission over a hostile channel. At the transmitter and receiver though the data is not in encrypted form, and therefore no longer protected. As a result of this, the attacker's efforts were shifted more on the transmitter and the receiver and less on the communication channel. There was a need for a new technology which can protect[1] the data in its raw format [4].

Watermarking is a technology that can protect the data in its clear form by embedding a permanent message into the data [5, 6, 7, 8, 9, 10]. The message should satisfy certain constraints that in most cases are application dependent. General models of the watermark embedding and detection processes for image sources are shown in Fig. 1.1 and Fig. 1.2 respectively. The embedding process is a function, the input of which is the original image, the message that we want to embed, and a secret key that is known only to the embedder and the intended detector. The output of the embedding function is the watermarked data. Analogously, the detection process is a function with inputs the watermarked data (possibly modified by an attacker) and the secret key. The output of the detection function is a decision determining the presence or absence of the message.

The challenges that watermarking technology faces [11] are broad and sometimes contradictory, depending on the particular application, with goals far more restricted than those of cryptography. Cryptography is mainly concerned with the *secrecy* [12, 2] and *authenticity* of the data [13], although there are other services that cryptography provides, like data integrity, non-repudiation, etc [2]. However, there are no restrictions on preserving the quality of the data. In watermarking we are interested in the reliable, sometimes secret (secure [14, 15]) communication of the watermark message while at the same time preserving the quality of the data being protected.

[1] See [3] for examples of modern applications of watermarking.

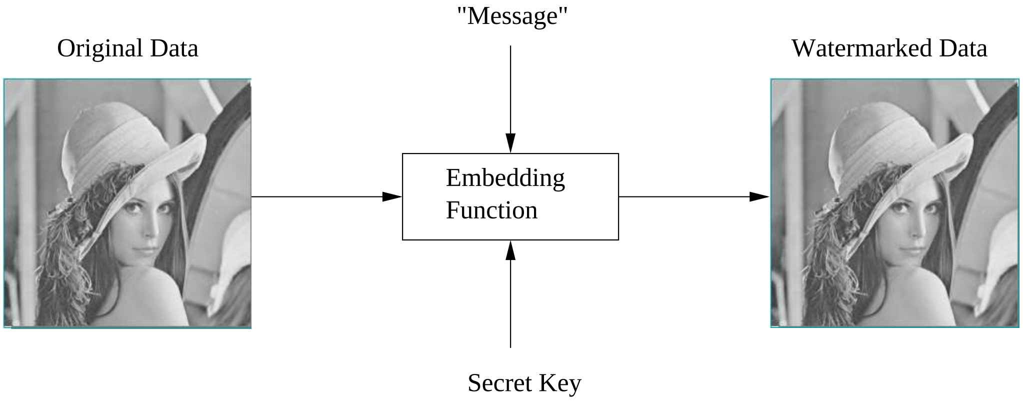

Figure 1.1: General model of watermark embedding.

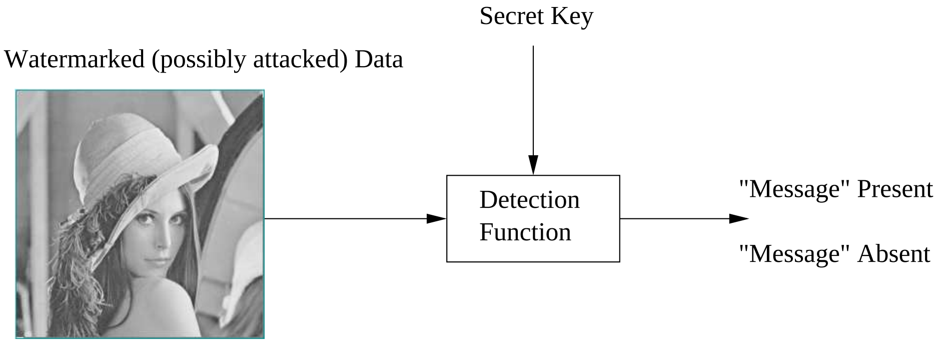

Figure 1.2: General model of watermark detection.

1.2 Requirements and Applications of Watermarking

Watermarking systems have to satisfy certain requirements, which in most cases are application-dependent. First, it is desirable that a watermarking system (algorithm) have a high watermark *payload*, or watermark *capacity*. In information theory, capacity is the maximum transmission rate at zero probability of error. In the watermarking literature, sometimes the term *capacity* is relaxed and usually means the number of watermark messages that can be embedded with a given algorithm irrespective of the probability of error of the system. We will use the term watermark *payload*. The term *capacity* will be used only when it refers to the information-theoretic meaning of capacity.

Robustness is another requirement that the watermarking system must satisfy, i.e. that the watermark should remain detectable by the intended receiver after attacks on the watermarked data.

The third requirement is that the embedding process should satisfy a *distortion* constraint, i.e. the watermarked data should be perceptually similar to the original data. The introduced distortion should be unnoticeable to the human perceptual system [16].

The fourth requirement is *security*. In this thesis, security is defined as the inability of an attacker to read the watermark message. This can be achieved by encrypting the watermark messages before embedding, or by incorporating a suitably chosen stochastic element directly into the embedding algorithm [17, 14]. Both methods require a secret key that has to be known to the intended decoder (detector). In most applications, it is more convenient to have the security independent of the other requirements.

Robustness and distortion are dependent parameters. Usually increasing the robustness leads to increased distortion. Although not generally independent of robustness, the payload can be chosen independent of distortion. An important and difficult problem in the design of watermarking systems is how to increase the robustness while at the same time keeping the distortion below a given level. This is most often achieved by a judicious choice and application of a perceptual model.

A block diagram with general requirements that a watermarking system should satisfy is shown in Fig. 1.3. The arrows illustrate the dependencies between the different requirements. Security is independent of the other requirements.

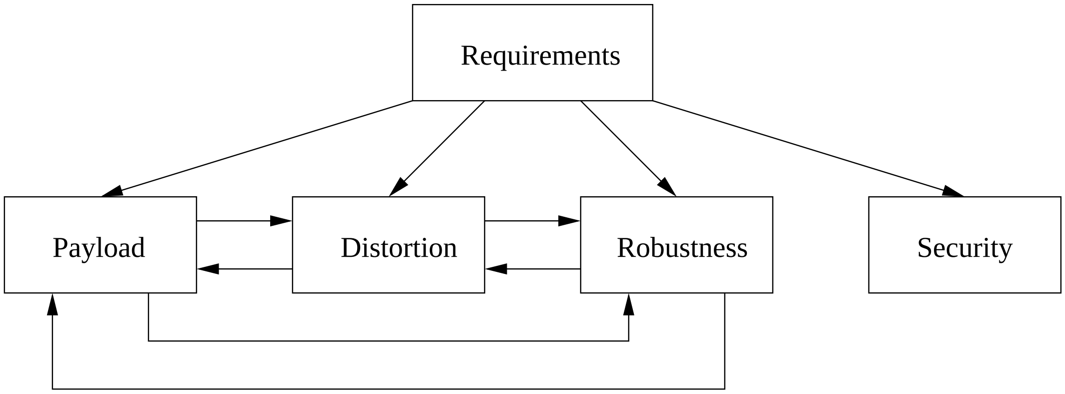

Figure 1.3: Requirements on watermarking systems. The arrows indicate the connection between the different requirements. Generally, *Distortion* is proportional to *Robustness*, *Payload* is inversely proportional to *Robustness*, and *Distortion* is proportional to *Payload*. *Security* is independent of the other parameters.

A classification of watermarking systems is shown in Fig. 1.4. In terms of the signal that has to be protected, watermarking techniques can be divided into the following categories:

- Audio watermarking. The host signal is an audio signal. This is an example of a typical one-dimensional signal processing. A model of the human auditory system is usually incorporated into the watermarking system to achieve low perceptual distortion.

- Image watermarking. This is a two-dimensional signal processing. Usually a good model of the human visual system is incorporated into the watermarking system.

- Video watermarking. The same as image watermarking, but now emphasizing on embedding in the temporal dimension.

- Watermarking of 3-D models.

- Watermarking of text files.

- etc.

Watermarking systems can also be categorized in terms of the purpose of the embedded message. We distinguish the following categories:

- Copyright protection. This is the case when we have a multimedia data and a player that can reproduce the data. The player checks for a watermark and based on it decides wether or not to play the multimedia data. The payload for this scenario is usually one bit for the whole data.

- Copy tracing. This is the case when a company wants to trace individual copies of its multimedia content. A different watermark is embedded in each copy, identifying the owner of the copy and possibly other information like when the copy was sold, etc. The purpose is to prevent illegal copies and sells of the multimedia data by the owner. The company can trace the illegal copy and connect it to the owner by identifying the watermark. Since the number of owners can be large, such applications require watermarking algorithms with high payload.

- Forensic watermarking. This is a scenario when someone wants to prove the authenticity of a multimedia data, for example in a court of law. Suppose we have a picture taken at a crime scene with a content that is likely to determine the outcome of the court case. Determining that the picture is not forged and is taken by the person who was on the crime scene can be done with a watermark. In such applications the robustness of the embedded watermark is at the highest priority.

- Steganography (secret communication). This application requires that the embedded message should be undetectable by an attacker.

- Content identification. This is an application where someone wants to identify an image or song. The purpose for example can be to measure the time during which the content is present in an advertisement.

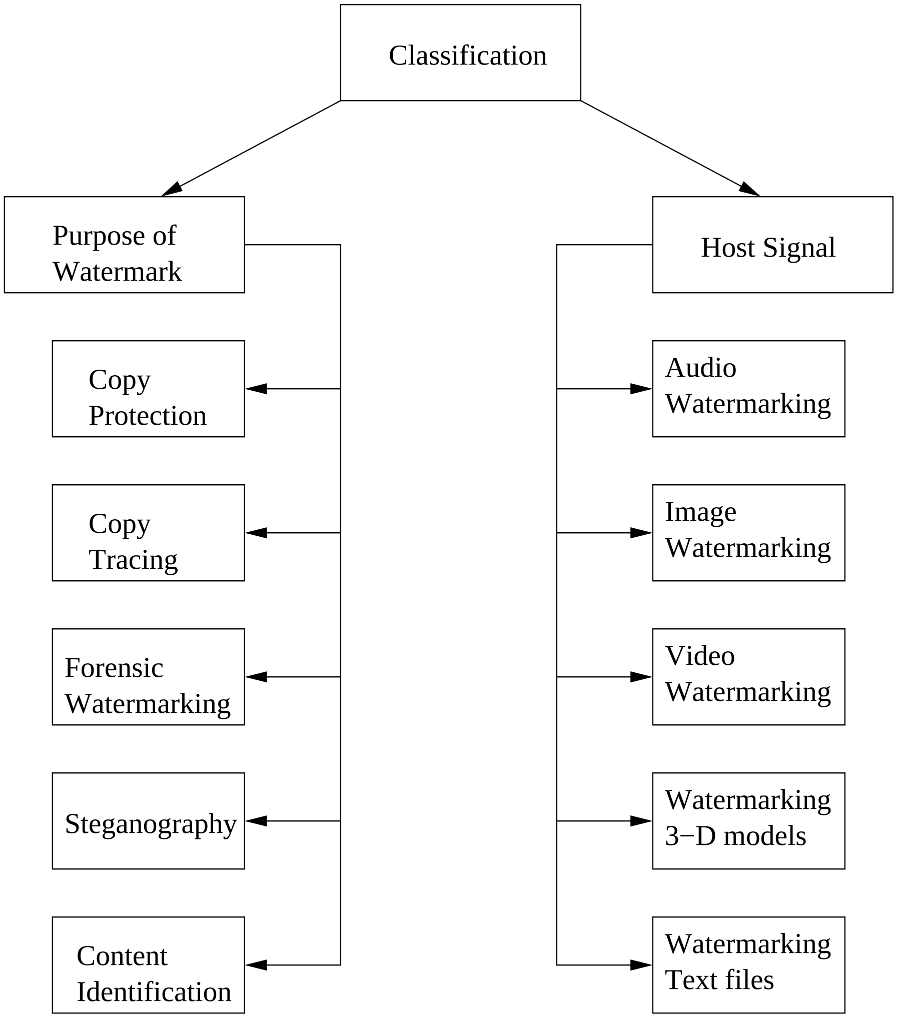

Figure 1.4: Classification of watermarking systems.

Watermarking systems are evaluated with respect to a given class of attacks. The simplest model of an attack operation consists of adding a source of noise to the watermarked data. There could be correlation between the noise samples and between the noise and watermark samples. This attack model is borrowed from communications theory due to its mathematical tractability and generality (especially the additive Gaussian noise channel).

Other models for attack channels are also of interest. Examples are deterministic operations like amplitude scaling, linear filtering. More complex operations include compression, linear and non-linear operations like gamma correction. These operations are quite common in many signal processing applications.

The attacks can be unintentional, i.e. they are part of a signal processing chain, or they can be malicious, i.e. due to an adversary. In both cases there should be a model of the attack channel. In the first case the attack channel parameters are fixed, while in the later case the attacker optimally chooses the parameters according to a given criterion (like capacity, probability of error, error exponent), under some constraints (such as distortion), after knowing the encoding and possibly decoding strategies. It is clear that malicious attacks are harder to combat than unintentional ones. However, as it is the case in this thesis, if the decoding operation is based on estimating the attacker's operation and inverting it prior to decoding the watermark, then clearly it doesn't matter if the operation was malicious or unintentional.

1.3 Objective of the Thesis

Depending on the principles of embedding and decoding, watermarking techniques can be grouped into two main categories - spread-spectrum and quantization-based watermarking. Spread-spectrum watermarking is efficient when the variance of the host signal is not too big with comparison to the variance of the watermark. In practical cases the watermark should preserve the quality of the original signal. In such cases the variance of the host signal must be much bigger than that of the watermark, in which case the capacity of the watermarking system is approximately zero. Quantization-based watermarking systems are able to completely cancel the influence of the host signal on the system performance. Theoretical analysis [18, 19, 20] showed that quantization-based watermarking achieves the highest capacity in terms of additive noise attacks among all other watermarking techniques. The theoretical results are valid for infinite dimensional quantizers. Later it was shown that schemes constructed with practical finite dimensional quantizers can perform very close to the theoretical limits.

Initially, it was not clear how quantization-based watermarking would perform against other than additive noise attacks. Moreover, it was difficult to obtain theoretical performance limits for more practical attacks like linear filtering, compression, non-linear signal processing operations, etc. In fact it was experimentally shown [21, 22, 14] that quantization-based techniques are vulnerable to a simple amplitude scale multiplication of the watermarked data, since this attack causes mismatch between the encoder and decoder.

This thesis is concerned with the study of quantization-based watermarking. We study this class of techniques in the context of unintentional, non-additive attacks, i.e. attacks that are standard operations in signal processing applications.

Due to the severity of the amplitude scale attack to quantization-based watermarking and the fact that it is one of the most common operations in signal processing applications, the amplitude scale attack in combination with additive noise is the primary attack channel model in this thesis. Furthermore, some practical and more complex operations like linear filtering can also be modeled by amplitude scaling, and additive noise.

In this thesis we propose statistical techniques for estimation of the amplitude scale attack. The effect of the amplitude scale is inverted by dividing the received data with the estimate before watermark decoding. Additionally we also study the secrecy (security) of the watermarking system by incorporating random dither into the system. Solutions are proposed in the presence and absence of the random dither. We extend those solutions to multi-band scaling attacks which are closely related to the linear filtering attacks. Finally, we study modifications for improving the robustness of quantization-based watermarking against linear filtering attacks.

1.4 Outline of the Thesis

In chapter 2, we present background on the information theoretic principles behind watermarking. These principles are based on spread-spectrum communications and channel coding with side information at the encoder. We discuss spread-spectrum, and quantization-based watermarking. We compare their performance in terms of capacity and error exponents. We discuss practical quantization-based watermarking techniques and evaluate their performance in terms of probability of error. We discuss the main types of attack channel models, which are the additive noise attack, the amplitude scale attack, and the linear time-invariant filtering attack.

In chapter 3, we derive probability density models in the absence of dither and describe our maximum likelihood, and Fourier based estimation procedures. The Fourier estimation technique is based on estimating the periodicity in the characteristic function of the watermarked data. The periodicity is due to the discontinuities in the density of the watermarked data, created by the encoding process. For certain watermark-to-noise ratios this periodicity is present in the attacked data. Experimental results with synthetic and real audio data are presented. We discuss the advantages and disadvantages of the proposed techniques.

In chapter 4, we derive probability density models in the presence of dither. We give approximations to these models for the case of small dither. We derive sufficient conditions for the dither to achieve a given level of security for the watermarking system. These conditions are not dependent on the approximations. We propose a maximum likelihood estimation procedure of the amplitude scale factor through the use of the dither. We propose a computationally efficient joint maximum likelihood estimation of the attacker's amplitude scale and noise variance. The reduction in computational complexity is due to a transformation of the attack channel into one that is equivalent but computationally less expensive for computing the likelihood function. Experimental results with synthetic and real data are presented.

In chapter 5, we extend our maximum likelihood estimation approach to the estimation of multi-band scaling. We derive probability density models of the filtered data. The probability density of the filtered signal is a convolution of the scaled with the filter coefficients individual probability densities derived in chapter 3. Experimental results are presented with synthetic and real audio signals. It is shown that only a few filter coefficients, namely the largest ones, are sufficient for constructing accurate density models.

In chapter 6 contains recent, still unpublished results. We discuss modifications for making quantization-based watermarking robust to linear filtering attacks. The main principle is based on frequency rational dither modulation, which is an application of rational dither modulation in the frequency domain. Since there is decoding error in the pass-band zone due to the finite length of the Fourier transform, we propose improvements by applying Hamming windows on each signal frame before taking the Fourier transform. While decreasing the error in the pass-band zone, the Hamming window causes additional distortion to the watermarked signal. To eliminate this additional distortion, instead of Hamming windows we apply overlapped cosine squared windows with overlap 50%. Experimental results of the proposed techniques are presented.

In chapter 7, we present discussion and directions for future research.

Chapters 3, 4, and 5 are parts of papers and for that reason there is inherently some overlap with chapter 2.

Chapter 3 is part of I. D. Shterev and R. L. Lagendijk "Amplitude Scale Estimation for Quantization-Based Watermarking", *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4146-4155, November 2006, and I. D. Shterev, R. L. Lagendijk, and R. Heusdens, "Statistical Amplitude Scale Estimation for Quantization-Based Watermarking", *SPIE Security, Steganography, and Watermarking of Multimedia Contents VI*, San Jose, CA, January 2004.

Chapter 4 is part of I. D. Shterev and R. L. Lagendijk "Amplitude Scale Estimation for Quantization-Based Watermarking, *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4146-4155, November 2006, I. D. Shterev and R. L. Lagendijk "Maximum Likelihood Amplitude Scale Estimation for Quantization-Based Watermarking in the Presence of Dither", *SPIE Security, Steganography, and Watermarking of Multimedia Contents VII*, San Jose, CA, January 2005, and R. L. Lagendijk and I. D. Shterev "Estimation of Attacker's Scale and Noise Variance for QIM-DC Watermark Embedding, *IEEE International Conference on Image Processing*, Singapore, October 2004.

Chapter 5 is published as J. Wang, I. D. Shterev, and R. L. Lagendijk "Scale Estimation in Two-Band Filter Attacks on QIM Watermarks, *SPIE Security, Steganography, and Watermarking of Multimedia Contents VIII*, San Jose, CA, January 2006, and part of J. Wang, I. D. Shterev, and R. L. Lagendijk "Two-Band Amplitude Scale Estimation for Quantization-Based Watermarking", *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, Hong Kong, December 2005.

Chapter 6 contains recent, still unpublished results.

Chapter 2

Information Theoretic Approaches to Watermarking[*]

Watermarking borrows many concepts from communications theory. In this chapter we describe the most important classes of watermarking schemes seen in the literature, namely spread spectrum watermarking and quantization-based watermarking [23, 20]. We discuss their advantages and disadvantages. We discuss some of the most important models of attack channels and their relevance to applications.

A general model of the watermarking problem is presented in Fig. 2.1. We want to embed a message W into a host signal vector $\widetilde{\boldsymbol{X}} \in \mathbb{R}^n$, at a rate of R_W bits per signal sample. Therefore $W \in \{1, 2, ..., 2^{nR_W}\}$.

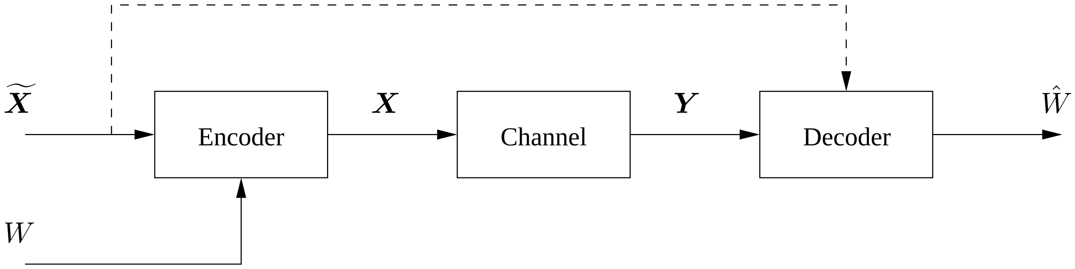

Figure 2.1: General model of the watermarking problem.

The encoder is a function that maps the host signal $\widetilde{\boldsymbol{X}} \in \mathbb{R}^n$ and the message W into a watermarked signal $\boldsymbol{X} \in \mathbb{R}^n$ subject to a distortion constraint.

The attack channel is a function that maps the watermarked signal $\boldsymbol{X}$ into an attacked signal $\boldsymbol{Y} \in \mathbb{R}^n$ subject to a distortion constraint.

When the host signal is available to the decoder (see Fig. 2.1), then we say that we have the non-blind watermarking scenario. It is logical that in this case, the watermarking scheme is expected to achieve better performance than in the blind case. However, such scenario has limited applications and we will be mostly interested in the blind case in which the decoder does not have access to the host signal.

[*] This chapter provides background information.

Various distortion measures are used in the literature depending on their relevance to the human perceptual system and their mathematical tractability [24, 25, 26]. The most common distortion measure is the squared-error distortion, which will be adopted here.

The distortion constraint on the encoder is given as

$$D(\boldsymbol{X}, \widetilde{\boldsymbol{X}}) = \frac{1}{n}\|\boldsymbol{X} - \widetilde{\boldsymbol{X}}\|^2 \tag{2.1}$$

The distortion constraint on the attack channel is given as

$$D(\boldsymbol{Y}, \boldsymbol{X}) = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\|^2 \tag{2.2}$$

2.1 Watermarking Techniques

Watermarking techniques can generally be divided into two categories: spread spectrum techniques and quantization-based watermarking.

2.2 Spread Spectrum Watermarking

Spread-spectrum watermarking schemes borrow principles from spread spectrum communications. The simplest form of embedding is given by

$$\boldsymbol{X} = \widetilde{\boldsymbol{X}} + \boldsymbol{N}_1(W), \tag{2.3}$$

where $\boldsymbol{N}_1(W)$ is a pseudorandomly generated sequence of length n based on the watermark message W.

The power of $\boldsymbol{N}_1(W)$ determines the strength of the watermark and therefore the distortion introduced to the host signal. For this scheme

$$D(\boldsymbol{X}, \widetilde{\boldsymbol{X}}) = \frac{1}{n}\|\boldsymbol{N}_1\|^2, \tag{2.4}$$

where we dropped the argument of $\boldsymbol{N}_1(\cdot)$ for notational conciseness.

All $\boldsymbol{N}_1(W)$ sequences are assumed to be known to the decoder. The decoder's task is to determine which $\boldsymbol{N}_1(W)$ is present in the received signal. The decoder performs correlation of $\boldsymbol{Y}$ with all sequences $\boldsymbol{N}_1(W)$ and makes an estimate of the embedded message $\hat{W}$. A block diagram of spread spectrum watermarking scheme is shown in Fig. 2.2.

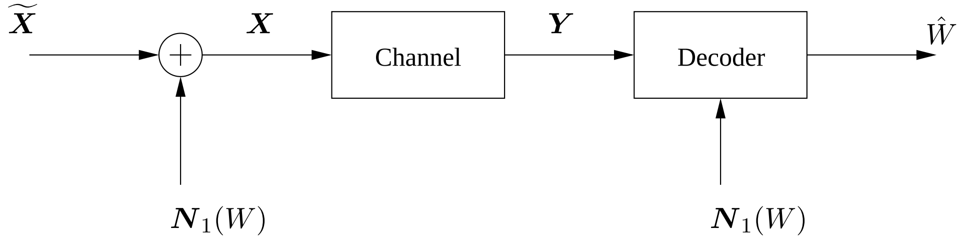

Figure 2.2: Spread Spectrum Watermarking Scheme.

For i.i.d. Gaussian host signals and additive Gaussian attack channels, the correlation detector coincides with the Maximum Likelihood decoder and is therefore the optimal detector [27].

As in communications [28, 29], watermarking schemes are often evaluated with respect to additive Gaussian noise channels, in this case representing the attack channel. Such a channel is presented in Fig. 2.3, where $\widetilde{X} \sim \mathcal{N}(0, \sigma^2_{\widetilde{X}} I)$, $N_2 \sim \mathcal{N}(0, \sigma^2_{N_2} I)$ is the attacker's noise, and I is an $n \times n$ identity matrix.

In most cases the performance criteria is the channel capacity which is formulated as the maximum amount of information that can be transmitted through the channel and decoded without any errors.

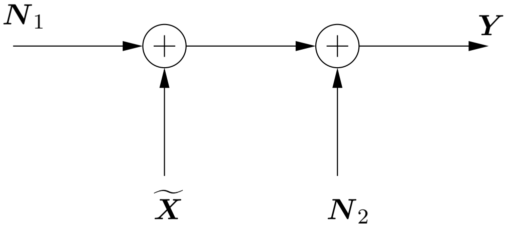

Figure 2.3: Gaussian Channel for Spread Spectrum Watermarking.

By definition the capacity of the channel in Fig. 2.3 is given as

$$C = \max_{f_{N_1}(n_1)} I(\boldsymbol{N}_1; \boldsymbol{Y}), \tag{2.5}$$

where the maximization is over all distributions $f_{N_1}(n_1)$. It can be shown [30] that

$$\max_{f_{N_1}(n_1)} I(\boldsymbol{N}_1; \boldsymbol{Y}) = \frac{1}{2} \log \left(1 + \frac{\sigma^2_{N_1}}{\sigma^2_{\widetilde{X}} + \sigma^2_{N_2}}\right), \text{ bits/sample} \tag{2.6}$$

and is achieved by $\boldsymbol{N}_1 \sim \mathcal{N}(0, \sigma^2_{N_1} I)$. In other words, Gaussian watermarks are optimal when the host signal and attack channel are Gaussian.

For the non-blind scenario the capacity is given as

$$\max_{f_{N_1}(n_1)} I(\boldsymbol{N}_1; \boldsymbol{Y}) = \frac{1}{2} \log \left(1 + \frac{\sigma^2_{N_1}}{\sigma^2_{N_2}}\right), \text{ bits/sample} \tag{2.7}$$

which is of course larger than in the blind case.

It can be seen that the host signal $\widetilde{X}$ introduces interference in the blind case and contributes to the attack channel. Therefore, spread spectrum watermarking schemes achieve good performance when the host variance is small in comparison to the watermark distortion and the power of the attack channel. However, such assumption in watermarking applications is unrealistic, since the watermark has to preserve the quality of the host signal. In real applications $\sigma^2_{\widetilde{X}} \gg \sigma^2_{N_1}, \sigma^2_{N_2}$, for which $C \approx 0$. This is the main limitation of spread spectrum watermarking in terms of additive noise attacks [20, 18].

2.3 One-Bit Spread-Spectrum Watermarking

In many watermarking applications like copy protection, it is enough to embed only one bit of information. Therefore, watermarking schemes that embed one bit of information in a signal vector are also of interest [31]. In such cases probability of error [32] or error exponents are the primary parameters for evaluating the performance of the scheme.

The scheme is a simplified version of the one discussed in the previous section. We add a pseudo-random sequence $\mathbf{N_1}$ if we wish to embed bit "1". The difference here is that we pass the host signal unchanged if we wish to embed bit "0". The detector tries to determine the presence or absence of $\mathbf{N_1}$ from the received signal.

The detection process is a binary hypothesis testing problem, from which the detector makes an estimate $\hat{W}$ of the embedded message W. The hypothesis testing can be written as

$$\hat{W} = \begin{cases} 0 & \text{if } \mathbf{Y} = \widetilde{\boldsymbol{X}} + \boldsymbol{N}_2 \\ 1 & \text{if } \mathbf{Y} = \boldsymbol{N}_1 + \widetilde{\boldsymbol{X}} + \boldsymbol{N}_2 \end{cases} \tag{2.8}$$

The likelihood ratio test is the correlation test (under Gaussian assumption).

$$\boldsymbol{N}_1 \cdot \boldsymbol{Y} \quad \overset{\hat{W}=1}{\underset{\hat{W}=0}{\gtrless}} \quad nT, \tag{2.9}$$

where $\cdot$ is the dot product and T is a threshold. The probabilities of false alarm and miss are given as

$$P_{FA}^{(n)} = q\Big(\sqrt{\frac{nT^2}{\sigma_{N_1}^2(\sigma_{\widetilde{X}}^2 + \sigma_{N_2}^2)}}\Big) \tag{2.10}$$

$$P_{M}^{(n)} = q\Big(\sqrt{\frac{n(\sigma_{N_1}^2 - T)^2}{\sigma_{N_1}^2(\sigma_{\widetilde{X}}^2 + \sigma_{N_2}^2)}}\Big), \tag{2.11}$$

where $q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ is the error function.

The exponents corresponding to $P_{FA}^{(n)}$ and $P_M^{(n)}$ are by definition

$$E_{FA} = \lim_{n\to\infty} -\frac{1}{n} \ln P_{FA}^{(n)} \tag{2.12}$$

$$E_M = \lim_{n\to\infty} -\frac{1}{n} \ln P_M^{(n)} \tag{2.13}$$

Using the property [33] $\lim_{n\to\infty} -\frac{1}{n}\ln\big(q(\sqrt{nx})\big) = \frac{1}{2}x$, for $x > 0$, (2.11), and (2.11), we have

$$\begin{aligned}
E_{FA} &= \lim_{n\to\infty} -\frac{1}{n}\ln P_{FA}^{(n)} \\
&= \lim_{n\to\infty} -\frac{1}{n}\ln q\Big(\sqrt{\frac{nT^2}{\sigma_{N_1}^2(\sigma_{\widetilde{X}}^2+\sigma_{N_2}^2)}}\Big) \\
&= \frac{T^2}{2\sigma_{N_1}^2(\sigma_{\widetilde{X}}^2+\sigma_{N_2}^2)} \qquad (2.14) \\
E_M &= \lim_{n\to\infty} -\frac{1}{n}\ln P_M^{(n)} \\
&= \lim_{n\to\infty} -\frac{1}{n}\ln q\Big(\sqrt{\frac{n(\sigma_{N_1}^2-T)^2}{\sigma_{N_1}^2(\sigma_{\widetilde{X}}^2+\sigma_{N_2}^2)}}\Big) \\
&= \frac{(\sigma_{N_1}^2-T)^2}{2\sigma_{N_1}^2(\sigma_{\widetilde{X}}^2+\sigma_{N_2}^2)} \qquad (2.15)
\end{aligned}$$

For non-blind watermarking, the detection test is

$$\boldsymbol{N}_1\cdot\boldsymbol{Y}-\boldsymbol{N}_1\widetilde{\boldsymbol{X}} \underset{\hat{W}=0}{\overset{\hat{W}=1}{\gtrless}} nT, \qquad (2.16)$$

with error exponents given as

$$\begin{aligned}
E_{FA} &= \frac{T^2}{2\sigma_{N_1}^2\sigma_{N_2}^2} \qquad (2.17) \\
E_M &= \frac{(\sigma_{N_1}^2-T)^2}{2\sigma_{N_1}^2\sigma_{N_2}^2}. \qquad (2.18)
\end{aligned}$$

The probability of error is defined as

$$P_E^{(n)} = Pr(W=0)P_{FA}^{(n)}+Pr(W=1)P_M^{(n)}, \qquad (2.19)$$

with corresponding error exponent

$$E_E = \lim_{n\to\infty} -\frac{1}{n}\ln P_E^{(n)}. \qquad (2.20)$$

Since $Pr(W=0)$ and $Pr(W=1)$ are independent of n and $\lim_{n\to\infty} P_{FA}^{(n)}=0$ for appropriately chosen T, it follows that $E_E=\min\{E_{FA},E_M\}$.

The threshold T is commonly chosen such that $E_{FA}=E_M$. Therefore, for $T=\frac{\sigma_{N_1}^2}{2}$, we have

$$\begin{aligned}
E_E &= \frac{\sigma_{N_1}^2}{8(\sigma_{\widetilde{X}}^2+\sigma_{N_2}^2)} \text{ , for blind detection} \qquad (2.21) \\
E_E &= \frac{\sigma_{N_1}^2}{8\sigma_{N_2}^2} \text{ , for non-blind detection} \qquad (2.22)
\end{aligned}$$

Higher error exponents means better performance. The expressions (2.21) and (2.22) are shown in Fig. 2.4. It can be seen that (2.22) is strictly higher than (2.21) as expected. The later depends on DWR. The difference between (2.21) and (2.22) is more pronounced at high DWR. The two expressions coincide at low DWR, when $\frac{\sigma^2_{\widetilde{X}}}{\sigma^2_{N_1}} \to 0$.

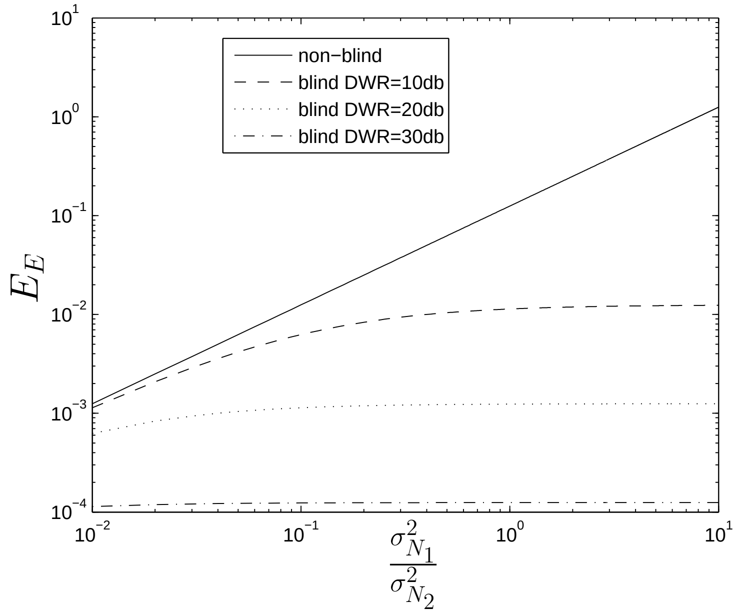

Figure 2.4: Error exponents for blind and non-blind spread-spectrum watermarking.

2.4 Quantization-Based Watermarking

As mentioned in section 2.2, the main limitation of spread-spectrum watermarking is the effect of the host signal on the capacity of the system. To overcome this limitation, after the development of spread spectrum watermarking, several scientists with information theory background [18, 19, 34, 35, 36, 37, 38] recognized the similarity between watermarking and coding for a channel with random parameters [39, 40, 41].

The channel with random parameters is shown in Fig. 2.5. The transition probability of the channel is $p_{Y|N_1\widetilde{X}}(y|n_1, \widetilde{x})$, where $\boldsymbol{Y}$ is the output of the channel, $\boldsymbol{N}_1$ is the input to the *channel*, and $\widetilde{\boldsymbol{X}}$ is the random parameter called the *state* of the channel. In watermarking applications, the state of the channel is the host signal. Since all random variables are i.i.d. Gaussian, we drop the bold notation where possible in the subsequent analysis of this section.

The capacity of this channel is given [39] as

$$C = \max_{p_{U\widetilde{X}N_1}(u,\widetilde{x},n_1)} \left\{ I(U;Y) - I(U;\widetilde{X}) \right\}, \tag{2.23}$$

where U is an auxiliary random variable, and the maximization is over the joint distribution $p_{U\widetilde{X}N_1}(u, \widetilde{x}, n_1)$.

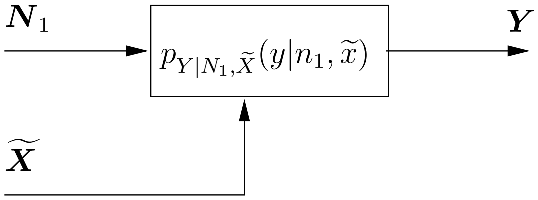
$\boldsymbol{N}_1$
$p_{Y|N_1,\widetilde{X}}(y|n_1,\widetilde{x})$
$\boldsymbol{Y}$
$\widetilde{\boldsymbol{X}}$

Figure 2.5: Channel with random parameters.

The capacity for the additive Gaussian noise channel and Gaussian state was found in [42]. For the case when the state is not known to both encoder and decoder, the capacity can be calculated as

$$C = \frac{1}{2}\log\Big(1+\frac{\sigma_{N_1}^2}{\sigma_{\widetilde{X}}^2+\sigma_{N_2}^2}\Big). \tag{2.24}$$

For the case when the state is known to the encoder but not to the decoder, the capacity is calculated as

$$C = \frac{1}{2}\log\Big(1+\frac{\sigma_{N_1}^2}{\sigma_{N_2}^2}\Big), \tag{2.25}$$

and is independent of the state $\widetilde{X}$. Therefore, it is possible to eliminate the interference of $\widetilde{X}$ without having to know it at the decoder side, by applying *channel coding with side information at the encoder*. It can be seen also that (2.24) and (2.25) are the capacities of blind (2.6) and non-blind (2.7) spread-spectrum watermarking respectively.

A simple block diagram of channel coding with side information at the encoder is shown in Fig. 2.6. With this type of coding, the encoder optimally assigns a codeword to each source vector, such that it is possible to cancel the interference of the powerful source with much less powerful codewords. The scheme works as follows. The encoder first generates $2^{n\big(I(\boldsymbol{U};\boldsymbol{Y})-\epsilon\big)}$ i.i.d. sequences $\boldsymbol{U}$, where $\epsilon > 0$ is arbitrarily chosen. These sequences are uniformly distributed in 2^{nR_W} bins. Each bin contains $2^{n\big(I(\boldsymbol{U};\widetilde{\boldsymbol{X}})+\delta\big)}$ sequences, where $\delta > 0$ is arbitrarily chosen. The union of these bins form the codebook which is given also to the decoder.

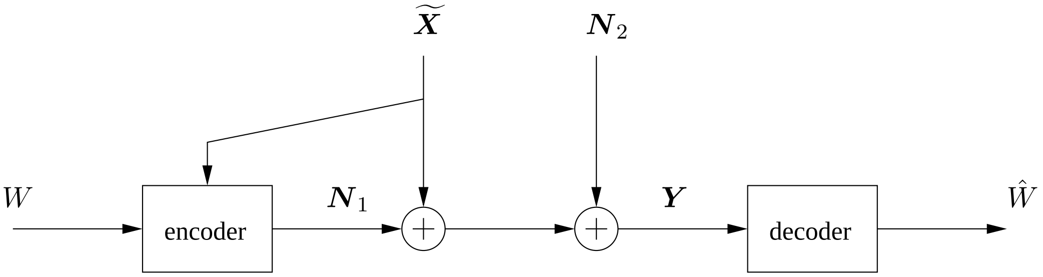

Figure 2.6: Channel with state known to the encoder.

Given the message W and the state $\widetilde{\boldsymbol{X}}$, the encoder looks in bin W for a sequence $\boldsymbol{U}$ such that $(\boldsymbol{U},\boldsymbol{S})$ are jointly typical. The encoder declares an error if no such sequence is found. The probability of such an error goes to zero as $n \to \infty$. Next the encoder chooses $\boldsymbol{N}_1$ such that $(\boldsymbol{N}_1, \boldsymbol{U}, \widetilde{\boldsymbol{X}})$ are jointly typical and sends it over the channel.

The decoder searches its codebook for a sequence $\boldsymbol{U}$ such that $(\boldsymbol{U}, \boldsymbol{Y})$ are jointly typical. The decoder makes an estimate $\hat{W}$ which is the index of the bin in which $\boldsymbol{U}$ is found. There is an error when no jointly typical pair can be found or when $\hat{W} \neq W$. The probability of not finding a jointly typical pair goes to zero as $n \to \infty$.

It has been shown [42] that capacity is achievable with the auxiliary random variable $U = N_1 + \alpha\widetilde{X}$, where α is a constant to be determined later. Substituting with $U = N_1 + \alpha\widetilde{X}$ in the capacity definition (2.23) we get

$$\begin{aligned} I(U;Y) - I(U;\widetilde{X}) &= H(U) - H(U|Y) - H(U) + H(U|\widetilde{X}) \\ &= H(U|\widetilde{X}) - H(U|Y) \\ &= H(U,\widetilde{X}) - H(\widetilde{X}) - H(U,Y) + H(Y) \\ &= \frac{1}{2}\log\frac{\sigma^2_{N_1} + \sigma^2_{\widetilde{X}} + \sigma^2_{N_2}}{\sigma^2_{\widetilde{X}}} \\ &+ \frac{1}{2}\log\frac{\big|K[N_1 + \alpha\widetilde{X}, \widetilde{X}]\big|}{\big|K[N_1 + \alpha\widetilde{X}, N_1 + \widetilde{X} + N_2]\big|}, \end{aligned} \tag{2.26}$$

where $\big|K[\cdot]\big|$ denotes the determinant of the correlation matrix $K[\cdot]$. We can write

$$\begin{aligned} \big|K[N_1 + \alpha\widetilde{X}, \widetilde{X}]\big| &= (\sigma^2_{N_1} + \alpha^2\sigma^2_{\widetilde{X}})\sigma^2_{\widetilde{X}} - \alpha^2\sigma^4_{\widetilde{X}} \\ \big|K[N_1 + \alpha\widetilde{X}, N_1 + \widetilde{X} + N_2]\big| &= (\sigma^2_{N_1} + \sigma^2_{\widetilde{X}} + \sigma^2_{N_2})(\sigma^2_{N_1} + \alpha^2\sigma^2_{\widetilde{X}}) - (\sigma^2_{N_1} + \alpha\sigma^2_{\widetilde{X}})^2 \end{aligned}$$

Substituting in (2.26) and simplifying we obtain

$$I(U;Y) - I(U;\widetilde{X}) = \frac{1}{2}\log\frac{(\sigma^2_{N_1} + \sigma^2_{\widetilde{X}} + \sigma^2_{N_2})\sigma^2_{N_1}}{(\alpha^2-1)^2\sigma^2_{N_1}\sigma^2_{\widetilde{X}} + (\sigma^2_{N_1} + \alpha^2\sigma^2_{\widetilde{X}})\sigma^2_{N_2}} \tag{2.27}$$

We want to find the value of α for which $I(U;Y) - I(U;\widetilde{X})$ is maximized. Noting that the nominator of (2.27) is independent of α, we need only minimize the denominator $(\alpha^2 - 1)^2\sigma^2_{N_1}\sigma^2_{\widetilde{X}} + (\sigma^2_{N_1} + \alpha^2\sigma^2_{\widetilde{X}})\sigma^2_{N_2}$. Taking the derivative with respect to α and making it equal to 0, we have

$$2\alpha\sigma^2_{N_1}\sigma^2_{\widetilde{X}} - 2\sigma^2_{N_1}\sigma^2_{\widetilde{X}} + 2\alpha\sigma^2_{\widetilde{X}}\sigma^2_{N_2} = 0 \tag{2.28}$$

Solving for α we get

$$\alpha = \frac{\sigma^2_{N_1}}{\sigma^2_{N_1} + \sigma^2_{N_2}} \tag{2.29}$$

Since $(\alpha^2 - 1)^2\sigma^2_{N_1}\sigma^2_{\widetilde{X}} + (\sigma^2_{N_1} + \alpha^2\sigma^2_{\widetilde{X}})\sigma^2_{N_2}$ is a convex function of α, (2.29) is the global minimum. Substituting with (2.29) in (2.27) and simplifying we get

$$\max_{\alpha}\big\{I(U;Y) - I(U;\widetilde{X})\big\} = \frac{1}{2}\log\Big(1 + \frac{\sigma^2_{N_1}}{\sigma^2_{N_2}}\Big) \tag{2.30}$$

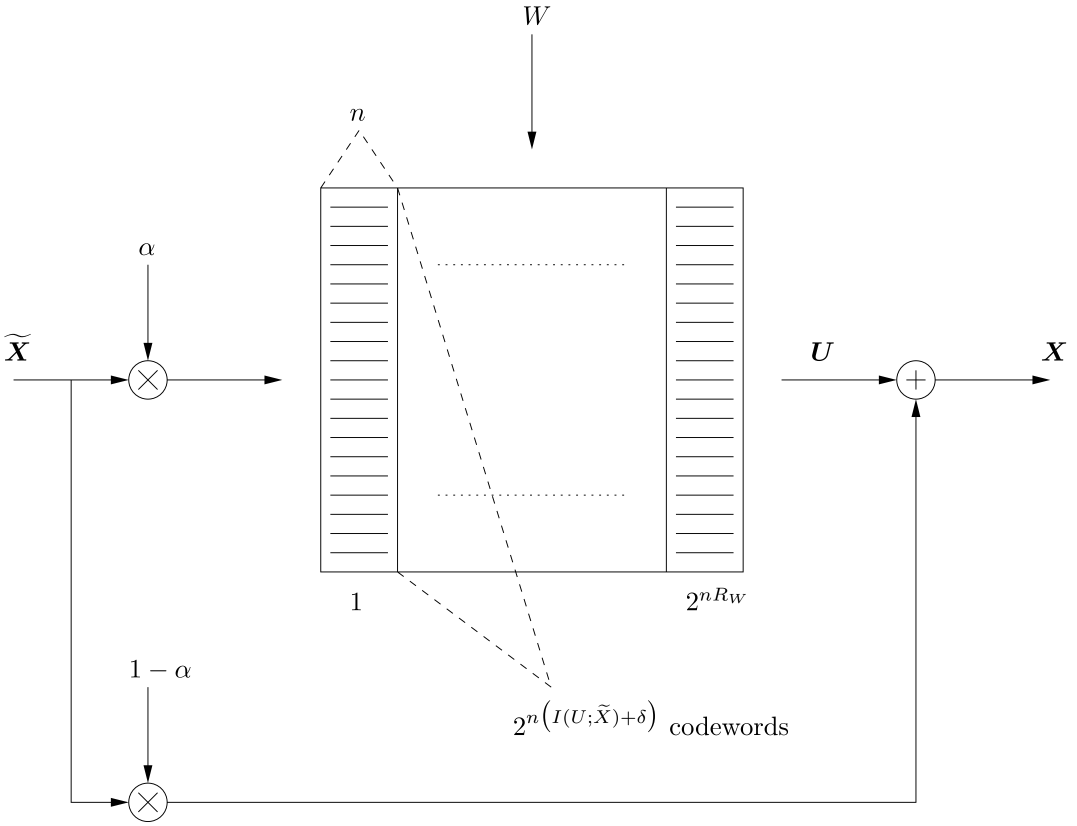
W
n
α
$\widetilde{\boldsymbol{X}}$
$\boldsymbol{U}$
$\boldsymbol{X}$
1
2^{nR_W}
$1-\alpha$
$2^{n\left(I(U;\widetilde{X})+\delta\right)}$ codewords

Figure 2.7: Detailed encoder scheme. The total number of codewords is $2^{n\left(I(U;Y)-\epsilon\right)}$.

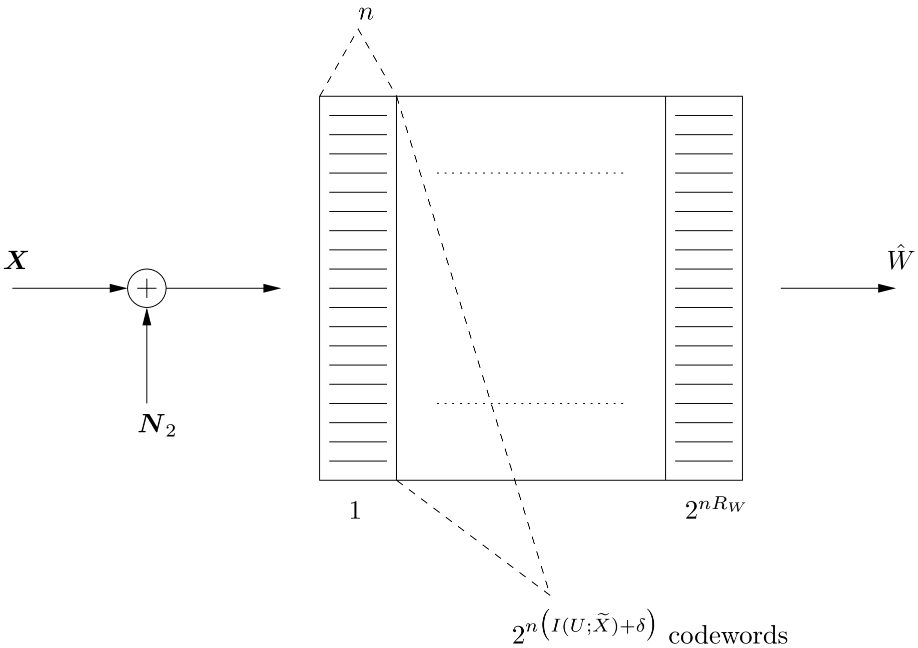

Figure 2.8: Detailed decoder scheme. The total number of codewords is $2^{n\left(I(U;Y)-\epsilon\right)}$.

Detailed descriptions of the encoder and decoder are shown in Fig. 2.7 and Fig. 2.8 respectively.

The encoder and decoder of Fig. 2.7 and Fig. 2.8 respectively are idealistic and only show the existence of good deterministic codes[1]. They do not give a clue as to how to construct good practical codes that can achieve the capacity (2.25). Lattice codes [44] are attractive to physically construct optimal encoders and decoders, because of their optimality in high dimensions, and the existence of computationally efficient decoding algorithms for lattices [45, 46].

2.5 One-Bit Quantization-Based Watermarking

In this section we describe and evaluate one-bit quantization-based watermarking and compare it with one-bit spread-spectrum watermarking. The scheme is developed in [47]. As mentioned in the previous section, lattices are among the best candidates for constructing practical, capacity achieving codes. The encoder employs an n-dimensional lattice quantizer $Q(\cdot)$ [44]. The message set is $W \in \{0,1\}$. Analogously to one-bit spread-spectrum watermarking, for $W = 0$ the host signal is left unchanged. The watermark encoder for $W = 1$ is shown in Fig. 2.9. The watermarked signal is given as

$$\mathbf{X} = \begin{cases} \widetilde{\mathbf{X}} + \mathbf{N}_1 & \text{if } W = 1 \\ \widetilde{\mathbf{X}} & \text{if } W = 0 \end{cases} \tag{2.31}$$

where $\mathbf{N}_1$ is the quantization noise. The distortion introduced by the encoder is

$$\sigma_{N_1}^2 = \frac{1}{n} E\|\mathbf{N}_1\|^2 \tag{2.32}$$

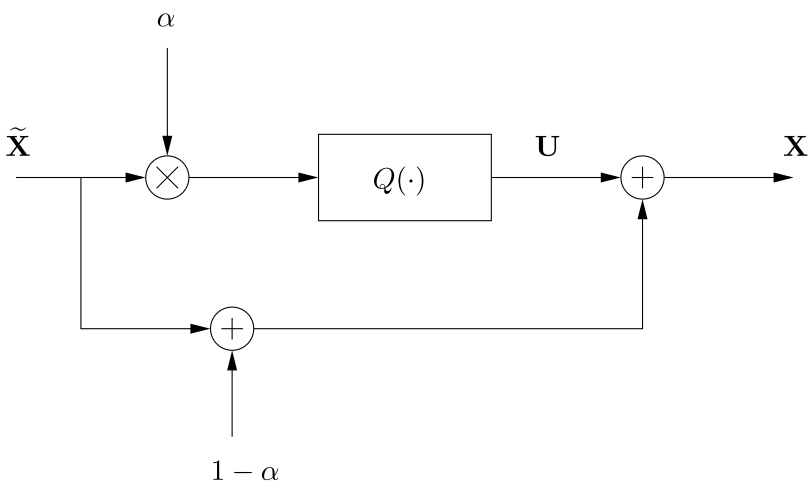

Figure 2.9: One-bit lattice encoder for $W = 1$.

[1] In Information Theory, the term *code* refers to a pair of encoding and decoding functions (mappings) [43].

The attack channel and the decoder are shown in Fig. 2.10. The received data is scaled by α and then quantized using the lattice quantizer $Q(\cdot)$. Then the decoder computes the squared L_2 norm of the quantization noise, compares it to a threshold nT, and makes an estimate $\hat{W}$.

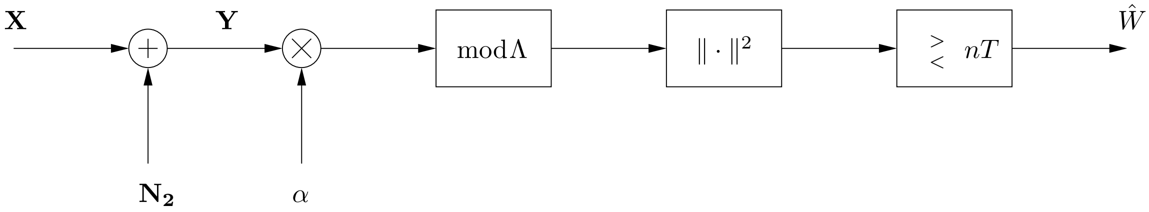

Figure 2.10: Attack channel and one-bit lattice decoder.

The quantization noise at the decoder is given as

$$\alpha\mathbf{Y} \bmod \Lambda \;=\; \alpha\mathbf{Y} - Q(\alpha\mathbf{Y}) \tag{2.33}$$

The decoding process is based on computing the square of the norm of the vector $\alpha\mathbf{Y} \bmod \Lambda$ and comparing the result to a threshold. The decoding is written as

$$\hat{W} \;=\; \begin{cases} 0 & \text{if } \|\alpha\mathbf{Y} \bmod \Lambda\|^2 > nT \\ 1 & \text{if } \|\alpha\mathbf{Y} \bmod \Lambda\|^2 < nT \end{cases}, \tag{2.34}$$

where T is a threshold, the optimal value of which will be derived later. The probability of false alarm can be bounded as follows

$$\begin{aligned} P_{FA}^{(n)} &= Pr\big(\|\alpha\mathbf{Y} \bmod \Lambda\|^2 < nT | W = 0\big) \\ &= \frac{Vol\big(B^n(\sqrt{nT}) \cap V_0\big)}{Vol(V_0)} \\ &\leq \frac{Vol\big(B^n(\sqrt{nT})\big)}{Vol(V_0)} \\ &= \Big(\frac{n}{n+2}\frac{G_n(\Lambda)}{G_n^*}\frac{T}{\sigma_{N_1}^2}\Big)^{n/2}, \end{aligned} \tag{2.35}$$

where V_0 is the base Voronoi cell of Λ and $G_n(\Lambda)$, and G_n^* are the normalized second moments of the lattice Λ, and the n-dimensional ball $B^n(\sqrt{nT})$ with radius $\sqrt{nT}$ respectively.

The probability of miss is written as

$$P_M^{(n)} \;=\; Pr\big(\|\alpha\mathbf{Y} \bmod \Lambda\|^2 > nT | W = 1\big) \tag{2.36}$$

The quantization noise at the decoder can be written as

$$\alpha\mathbf{Y} \bmod \Lambda \;=\; (1-\alpha)\mathbf{N_1} + \alpha\mathbf{N_2} \tag{2.37}$$

Therefore $P_M^{(n)}$ can be written as

$$P_M^{(n)} \;=\; Pr\big(\|(1-\alpha)\mathbf{N_1} + \alpha\mathbf{N_2}\|^2 > nT | W = 1\big). \tag{2.38}$$

For notational convenience we make the following substitution

$$\mathbf{Z} \quad = \quad (1-\alpha)\mathbf{N_1} + \alpha\mathbf{N_2} \tag{2.39}$$

In [48] it was proved that there exists a Gaussian vector $\mathbf{Z}^* \sim \mathcal{N}\Big(0, \sigma^2_{Z^*}I\Big)$ with variance bounded as

$$\frac{n}{n+2}\frac{\sigma^2_{N_1}\sigma^2_{N_2}}{\sigma^2_{N_1}+\sigma^2_{N_2}} \le \sigma^2_{Z^*} < \left(\frac{R_c}{R_e}\right)^n \frac{\sigma^2_{N_1}\sigma^2_{N_2}}{\sigma^2_{N_1}+\sigma^2_{N_2}} \tag{2.40}$$

such that

$$f_{\mathbf{Z}}(\mathbf{z}) \quad \le \quad e^{\epsilon_1(\Lambda)n} f_{\mathbf{Z}^*}(\mathbf{z}^*), \tag{2.41}$$

where

$$\epsilon_1(\Lambda) \quad = \quad \log\frac{R_c}{R_e} + \frac{1}{2}\log 2\pi e G(\Lambda), \tag{2.42}$$

and R_c, and R_e are the covering and equivalent radius respectively of Λ.

We can use (2.41) to upper-bound $P_M^{(n)}$, i.e.

$$\begin{aligned} P_M^{(n)} \quad &= \quad Pr\big(\|(1-\alpha)\mathbf{N_1} + \alpha\mathbf{N_2}\|^2 > nT|W=1\big) \\ &\le \quad e^{\epsilon_1(\Lambda)n} Pr\big(\|\mathbf{Z}^*\|^2 > nT|W=1\big). \end{aligned} \tag{2.43}$$

The distribution of $\|\mathbf{Z}^*\|^2$ is the $\mathcal{X}^2$ (chi-square) distribution [49], i.e.

$$f_{\|\mathbf{Z}^*\|^2}(z) \quad = \quad \begin{cases} \frac{z^{n/2-1}e^{-z/2\sigma_Z^2}}{2^{n/2}\sigma_Z^n\Gamma(\frac{n}{2})} & \text{if } z \ge 0 \\ 0 & \text{if } z < 0 \end{cases} \tag{2.44}$$

Applying the Chernoff bound on $\mathcal{X}^2$, we get

$$P_M^{(n)} \le e^{\epsilon_1(\Lambda)n} Pr\Big(\|\mathbf{Z}^*\|^2 > nT|W=1\Big) \le e^{\epsilon_1(\Lambda)n} e^{-snT}\mu(s), \text{ for } s > 0 \tag{2.45}$$

and

$$\mu(s) = e^{\ln \int\limits_0^\infty e^{sz}\frac{z^{n/2-1}e^{-z/2\sigma_Z^2}}{2^{n/2}\sigma_Z^n\Gamma(\frac{n}{2})}dz} = e^{\frac{n}{2}\ln\frac{1}{1-s\sigma^2_{Z^*}}} \tag{2.46}$$

is the moment generating function of $\mathcal{X}^2$. The expression $e^{-snT}\mu(s)$ is always convex in $s > 0$ [25]. Therefore, to find s which minimizes (2.45) we write

$$nT \quad = \quad \frac{\partial\mu(s)}{\partial s}\Big/\mu(s). \tag{2.47}$$

Solving for s we get

$$s \quad = \quad \frac{1}{2\sigma^2_{Z^*}} - \frac{1}{2T}. \tag{2.48}$$

Substituting in (2.46) and (2.45) we get

$$P_M^{(n)} \leq e^{\epsilon_1(\Lambda)n} e^{-\frac{n}{2}\left(\frac{T}{\sigma^2_{Z^*}} - \ln\frac{T}{\sigma^2_{Z^*}} - 1\right)} \tag{2.49}$$

Taking the limit $n \to \infty$ we have

$$\begin{aligned} G_n(\Lambda) &\to G^* \text{ , as } n \to \infty && (2.50)\\ \epsilon_1(\Lambda) &\to 0 \text{ , as } n \to \infty && (2.51)\\ \sigma^2_{Z^*} &\to \frac{\sigma^2_{N_1}\sigma^2_{N_2}}{\sigma^2_{N_1}+\sigma^2_{N_2}} \text{ , as } n \to \infty && (2.52)\end{aligned}$$

Taking the above limits into account, the error exponents become

$$\begin{aligned} E_{FA} &= \lim_{n\to\infty} -\frac{1}{n}\ln P_{FA}^{(n)}\\ &\geq \lim_{n\to\infty} -\frac{1}{n}\ln\left(\frac{n}{n+2}\frac{G_n(\Lambda)}{G_n^*}\frac{T}{\sigma^2_{N_1}}\right)^{n/2}\\ &= \frac{1}{2}\ln\frac{\sigma^2_{N_1}}{T} && (2.53)\end{aligned}$$

$$\begin{aligned} E_M &= \lim_{n\to\infty} -\frac{1}{n}\ln P_M^{(n)}\\ &\geq \lim_{n\to\infty} -\frac{1}{n}\ln e^{\epsilon_1(\Lambda)n} e^{-\frac{n}{2}\left(\frac{T}{\sigma^2_{Z^*}} - \ln\frac{T}{\sigma^2_{Z^*}} - 1\right)}\\ &= \frac{1}{2}\Big(T\frac{\sigma^2_{N_1}+\sigma^2_{N_2}}{\sigma^2_{N_1}\sigma^2_{N_2}} - \ln T\frac{\sigma^2_{N_1}+\sigma^2_{N_2}}{\sigma^2_{N_1}\sigma^2_{N_2}} - 1\Big) && (2.54)\end{aligned}$$

To have $E_{FA}, E_M \geq 0$ we need the condition $\frac{\sigma^2_{N_1}\sigma^2_{N_2}}{\sigma^2_{N_1}+\sigma^2_{N_2}} \leq T \leq \sigma^2_{N_1}$.

The probability of error is

$$P_E^{(n)} = Pr(W=0)P_{FA}^{(n)} + Pr(W=1)P_M^{(n)} \tag{2.55}$$

Since probabilities can be expressed as exponentials, we have

$$E_E = \min\{E_{FA}, E_M\}. \tag{2.56}$$

Therefore, to maximize the error exponent bound we should choose the threshold nT such that the lower bounds to E_{FA} and E_M are equalized, i.e. $T = \frac{\sigma^2_{N_1}\sigma^2_{N_2}}{\sigma^2_{N_1}+\sigma^2_{N_2}}\big(1 + \ln(1 + \frac{\sigma^2_{N_1}}{\sigma^2_{N_2}})\big)$. For this case we have

$$E_E \geq \ln(1 + \frac{\sigma^2_{N_1}}{\sigma^2_{N_2}}) - \ln\big(1 + \ln(1 + \frac{\sigma^2_{N_1}}{\sigma^2_{N_2}})\big) \tag{2.57}$$

The bound (2.57) is shown in Fig. 2.11, together with (2.21) and (2.22). We can see that Quantization-Based watermarking clearly outperforms blind spread-spectrum watermarking for $DWR > 35db$. We also observe that for some ratios $0.5 < \frac{\sigma^2_{N_1}}{\sigma^2_{N_2}} < 10$, QIM clearly outperforms even non-blind spread-spectrum.

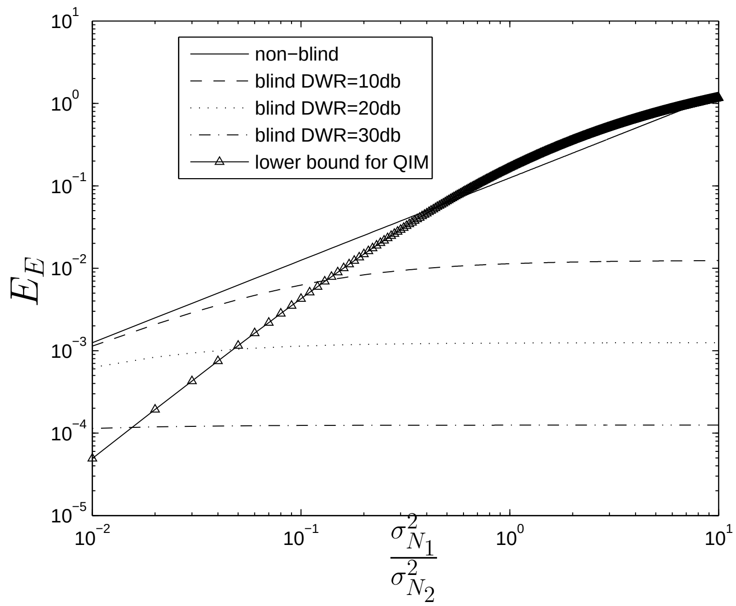

Figure 2.11: Error-exponents for one-bit watermarking.

2.6 Scalar Quantization-Based Watermarking

The simplest implementation of quantization-based watermarking employs scalar quantizers [50, 51]. An example of scalar QIM encoder is shown in Fig. 2.12, where $Q(\cdot)$ denotes uniform quantization with step size Δ. For $W = 1$ the quantizer's input and output are shifted with $\frac{\Delta}{2}$ and $-\frac{\Delta}{2}$ respectively. The quantizer's input-output characteristics are shown in Fig. 2.13. One of the two input-output characteristics is chosen, depending on the message we want to embed. The quantization noise is defined as

$$\begin{aligned} N_1 &= \alpha\widetilde{X} - Q(\alpha\widetilde{X}) \\ &= \widetilde{X} - X \end{aligned} \tag{2.58}$$

The quantizer output is given as

$$U = \begin{cases} k\Delta & \text{if } W = 0 \\ (2k+1)\frac{\Delta}{2} & \text{if } W = 1 \end{cases} \tag{2.59}$$

The attack channel and the decoder are shown in Fig. 2.14. The received signal is written as

$$\begin{aligned} Y &= X + N_2 \\ &= U + (1-\alpha)\widetilde{X} + N_2 \\ &= \frac{1}{\alpha}\big(U + (1-\alpha)N_1\big) + N_2, \end{aligned} \tag{2.60}$$

where the last line follows from using the relation $\alpha\widetilde{X} = U + N_1$.

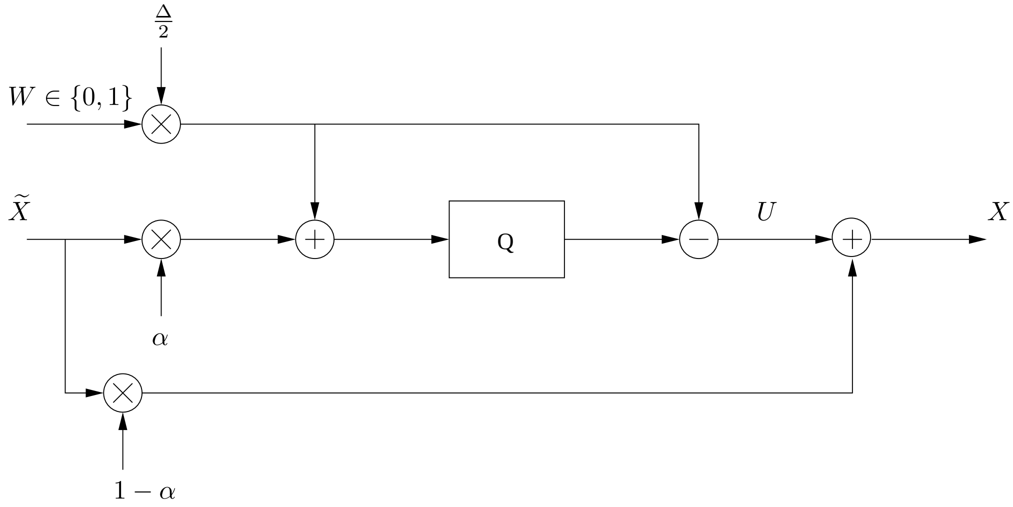

Figure 2.12: Watermark encoder.

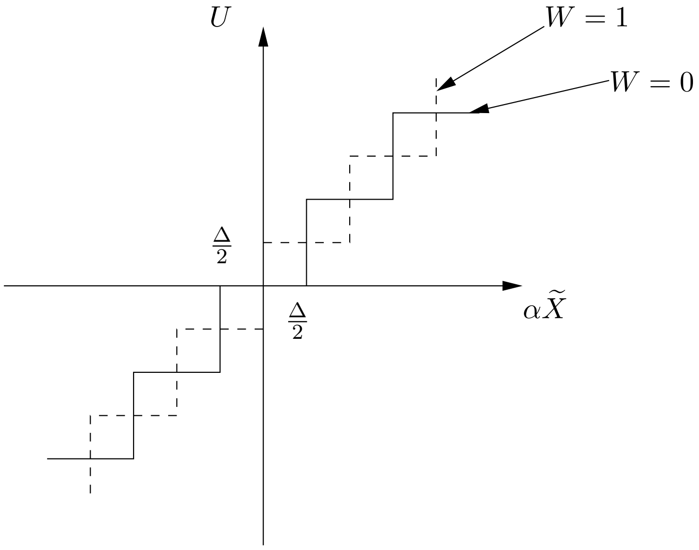

Figure 2.13: Quantizer input-output characteristics

The decoder multiplies Y by α, thus obtaining

$$\begin{aligned}\hat{U} &= \alpha Y \\ &= U + (1-\alpha)N_1 + \alpha N_2, \end{aligned} \tag{2.61}$$

where we used (2.60) in the last line. The decoder computes $|\hat{U} - Q(\hat{U})|$ and estimates the watermark as follows:

$$\hat{W} = \begin{cases} 0 & \text{if } |\hat{U} - Q(\hat{U})| \leq \frac{\Delta}{4} \\ 1 & \text{if } |\hat{U} - Q(\hat{U})| > \frac{\Delta}{4} \end{cases} \tag{2.62}$$

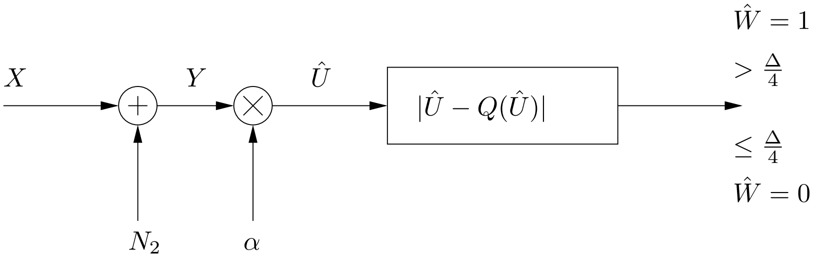

Figure 2.14: Watermark decoder.

When $\alpha = 1$, irrespective of the attack channel, then obviously the encoder and decoder do not know the variance of the attack channel. The performance for the case $\alpha = 1$ is therefore expected to be reduced. Experimental results of the probability of error for $\alpha = \frac{\sigma_{N_1}^2}{\sigma_{N_1}^2 + \sigma_{N_2}^2}$ and $\alpha = 1$ are shown in Fig. 2.15. It can be seen that for the case $\alpha = 1$, P_e approaches 0.5 much faster than in the case when the encoder and decoder know the variance of the attack channel.

It is possible to evaluate the performance of QIM watermarking in terms of capacity, using the formula

$$\begin{aligned} C &= 1 + H(P_e) \\ &= 1 + P_e \log P_e + (1-P_e)\log(1-P_e) \end{aligned} \tag{2.63}$$

The formula (2.63) is the capacity of the binary symmetric channel (BSC) [30] with crossover probability P_e. It can be straightforwardly shown that the QIM watermarking system together with the attack channel can be modeled as a BSC with crossover probability equal to the probability of error of the watermarking system, see Fig. 2.16. Comparison of the experimental capacity with theoretical limits is shown in Fig. 2.17. We can see that there is a gap to the theoretical limits (2.25) especially at high WNR, because the capacity according to (2.63) can be at most 1 bit.

Improvements to the basic scalar QIM are possible by incorporating non-uniform scalar quantizers. In [52] it was shown that the fidelity of the embedding process can be improved by selectively changing the quantization step size based on the host signal statistics.

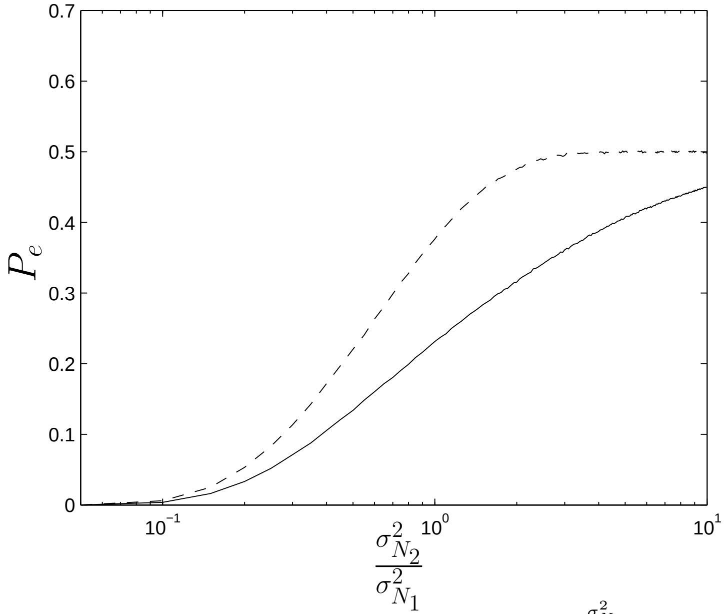

Figure 2.15: Probability of error for additive Gaussian attacks, $\alpha = \frac{\sigma^2_{N_1}}{\sigma^2_{N_1}+\sigma^2_{N_2}}$ (solid) and $\alpha = 1$ (dashed). $DWR = 20db$.

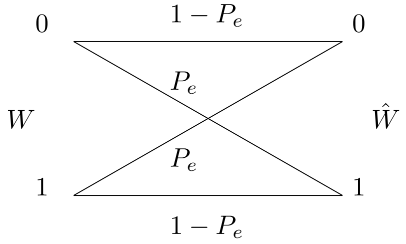

Figure 2.16: Watermarking scheme as a BSC.

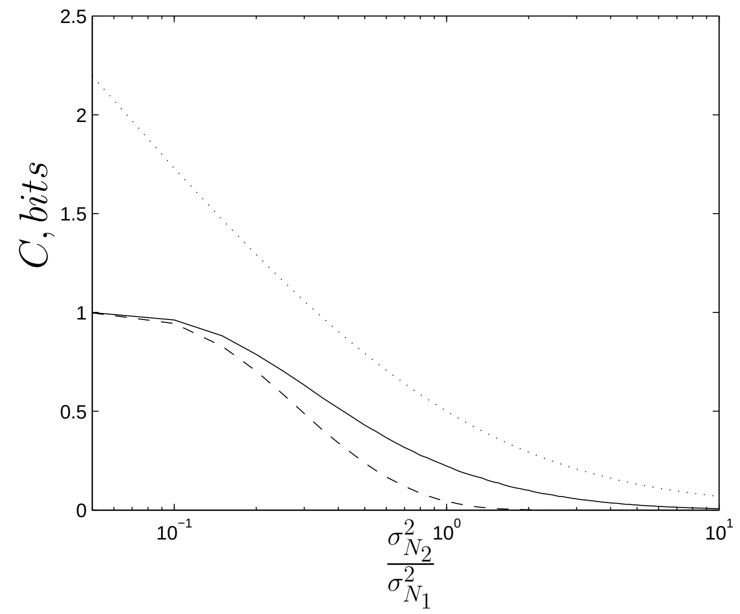

Figure 2.17: Capacity of QIM watermarking for additive Gaussian attacks, $\alpha = \frac{\sigma^2_{N_1}}{\sigma^2_{N_1}+\sigma^2_{N_2}}$ (solid), $\alpha = 1$ (dashed), and theoretical curve according to (2.25) (dotted). $DWR = 20db$.

2.7 Attacks

In this section we consider attack channels that model signal processing operations commonly seen in watermarking applications. These attack channels are of particular interest to quantization-based watermarking.

2.7.1 Additive Attacks

This type of attacks consists of adding a source of noise to the watermarked data subject to distortion constraint. An example is shown in Fig. 2.18. The noise $\boldsymbol{N_2}$ has covariance function $K_{N_2}(t,u)$. When $K_{N_2}(t,u) = \frac{N_0}{2}\delta(u-t)$ then the noise is white, otherwise it is said to be colored. Obviously, the colored noise can better adapt to the watermarked data statistics, and therefore be more devastating than the white noise. Surprisingly in [19] it was shown, under general conditions, that memoryless attacks are optimal, in particular the i.i.d. Gaussian noise is the most malevolent power limited noise in terms of watermark capacity.

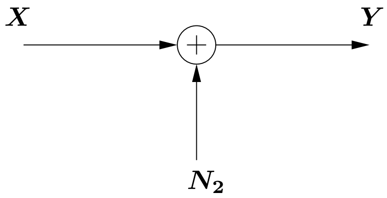

Figure 2.18: Additive Gausian noise attack.

As mentioned above, the results obtained in [19] concern the capacity of ideal (theoretical) schemes. For some practical schemes, or other performance evaluation parameters (like probability of error), the Gaussian attack may not be the most malevolent additive attack. Here is an example. If we take (4.13) with $D = 0$, $\alpha = 1$, and $\beta = 1$, we get the probability of error for QIM without distortion compensation.

$$P_e \quad = \quad \sum_m \Pr[m\Delta - \frac{3\Delta}{4} \leq N_2 \leq m\Delta - \frac{\Delta}{4}] \tag{2.64}$$

It can be shown that $P_e = 0.5$ when $N_2 \sim \mathcal{U}(0, \sigma^2_{N_1})$. Theoretically, to achieve $P_e = 0.5$ for Gaussian attacks, we must have $N_2 \sim \mathcal{N}(0, \infty)$, i.e. the power of the Gaussian attack must be infinite. In practice (see Fig.2.15) $P_e \approx 0.5$ when $N_2 \sim \mathcal{N}(0, 2\sigma^2_{N_1})$. Therefore, for scalar QIM watermarking, the uniform noise is more malevolent than the Gaussian noise.

The Gaussian attack channel is often used in theoretical analysis for establishing bounds and due to its mathematical tractability. Detailed analysis of the impact of additive noise channels with different distributions on the performance of practical watermarking schemes is given in [14].

2.7.2 Nonadditive Attacks

In [19] it was shown that, under squared-error constraints, the capacity when the attacker is restricted to additive attacks is strictly larger than the capacity for general attacks (attack channel specified by a conditional distribution). This demonstrates that additive attacks are suboptimal. The result actually is not surprising, because the set of non-additive attacks is larger than the set of additive ones.

Many signal processing operations can hardly be modeled as additive noise. In this thesis we are mostly concerned with attacks that are common operations in signal processing applications. An example of such an application is digital audio broadcasting (DAB). A DAB encoder is shown in Fig. 2.19.

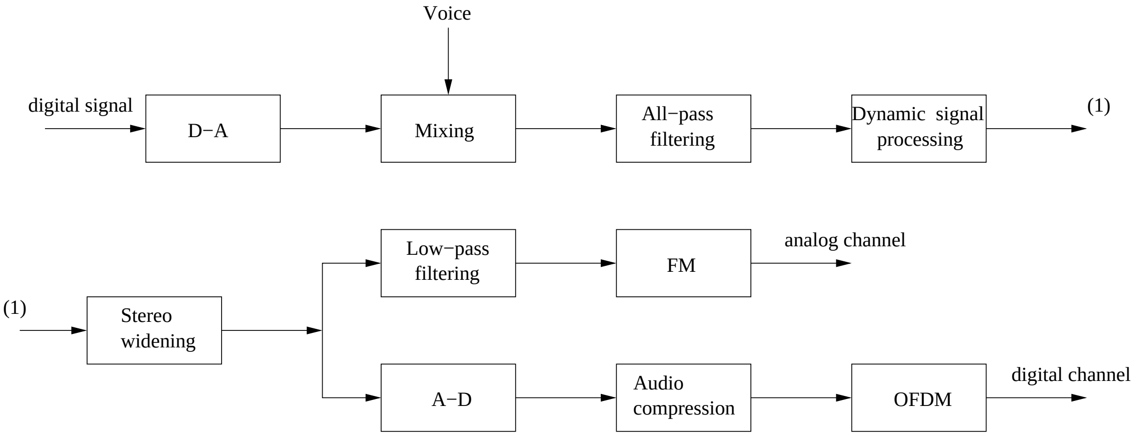

Figure 2.19: Block diagram of DAB encoder.

Clearly, the operations within the DAB encoder cannot be modeled as pure additive noise. They can be classified into four types of operations that we show in Fig. 2.20. The amplitude scaling operation is shown as a separate group, due to its central role as an attack channel in this thesis.

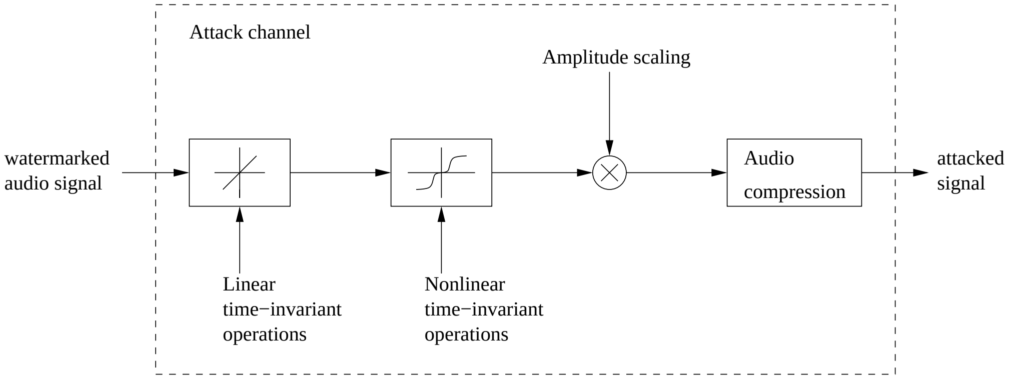

Figure 2.20: Classification of signal processing operations within the DAB encoder.

The operations shown in Fig. 2.19 are of particular interest to watermarking, because they do not induce big perceptual distortions to the signal, thus also making them attractive for use by a malicious attacker. Linear time-invariant attacks are tackled in chapter 5 and chapter 6. The class of non-linear time-invariant and compression attacks are beyond the scope of this thesis and the reader is referred to [53, 54, 55] for theoretical analysis of the compression attack.

Another operation (attack) that is important both to watermarking and communications systems is the de-synchronization attack [56, 57] (time warping in audio), which can be produced by a malicious attacker, or a byproduct of standard signal processing operations such as sampling [58], analog-to-digital and digital-to-analog conversions [59]. The treatment of de-synchronization attacks is outside the scope this thesis and the reader is referred to [60, 61, 62, 63, 64] for analysis and countermeasures.

2.7.3 Amplitude Scale Attacks

As mentioned in the previous subsection, amplitude scale attacks constitute a central point in this thesis. Amplitude scaling operations happen in every signal processing application. Moreover, quantization-based watermarking schemes are extremely vulnerable to amplitude scaling, because such operations introduce mismatch between the codebooks of the encoder and decoder. Another important aspect is that although amplitude scaling introduces large mean squared-error, the perceptual distortion of this operation is very low, i.e. only the brightness (in images) or the loudness (in audio) is changed. The effect of unknown scaling on probability of decoding error is shown in Fig. 2.21. It can be seen that probability of error approaches 0.5 very fast as the scaling factor deviates from unity. It has to be noticed that the effect of the amplitude scale attack is more pronounced at high DWR, because then the mismatch between the codebooks is greater. In other words, the residual noise after lattice decoding has bigger variance than in the case of low DWR.

The amplitude scale attack has also theoretical significance. For Gaussian sources, under squared-error constraints, amplitude scale and additive noise is the optimal attack channel from rate-distortion point of view. To see this, suppose that we want to maximally compress

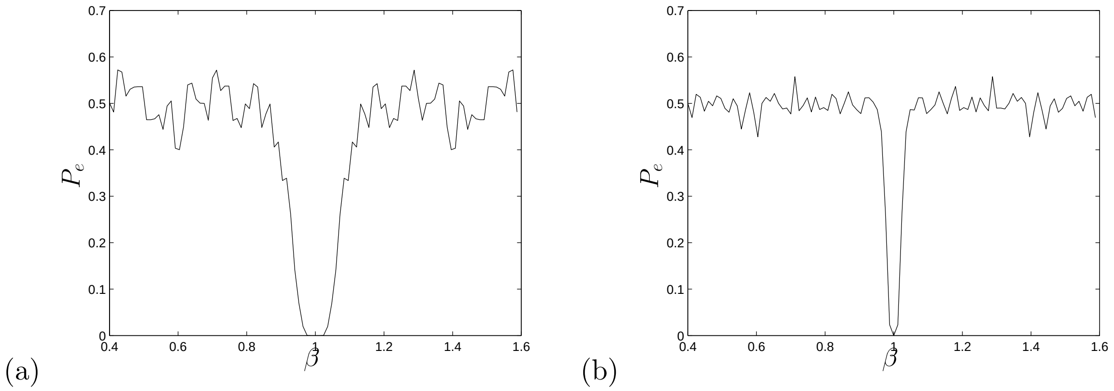

Figure 2.21: The effect of scaling β on probability of error. Experiments are performed for $\widetilde{X} \sim \mathcal{N}(0,1)$, (a) $DWR = 20db$, (b) $DWR = 30db$.

X into Y (see Fig. 2.22) at rate R subject to distortion $\sigma^2_{N_2}$, it is well known that this is given by the rate-distortion function

$$R(\sigma^2_{N_2}) \quad = \quad \min_{p(y;x)} I(Y;X), \tag{2.65}$$

where $I(Y;X)$ is the mutual information between X and Y, and the minimization is over all joint distributions $p(x;y)$ that satisfy the distortion constraint $\sigma^2_{N_2}$.

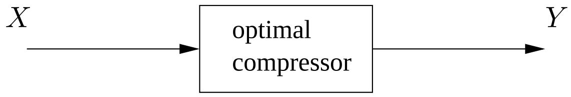

Figure 2.22: Optimal compressor.

The rate-distortion function of a memoryless Gaussian source with respect to the squared-error criterion is [24]

$$\begin{aligned} R(\sigma^2_{N_2}) \quad &= \quad \frac{1}{2}\max\left(0, \log\frac{\sigma^2_{\widetilde{X}} + \sigma^2_{N_1}}{\sigma^2_{N_2}}\right) \\ &= \quad \begin{cases} \frac{1}{2}\log\frac{\sigma^2_{\widetilde{X}}+\sigma^2_{N_1}}{\sigma^2_{N_2}} & , \; 0 \le \sigma^2_{N_2} \le \sigma^2_{\widetilde{X}} + \sigma^2_{N_1} \\ 0 & , \; \sigma^2_{N_2} \ge \sigma^2_{\widetilde{X}} + \sigma^2_{N_1} \end{cases} \end{aligned} \tag{2.66}$$

where $X \sim \mathcal{N}(m, \sigma^2_{\widetilde{X}} + \sigma^2_{N_1})$.

From [24] the optimal test channel is

$$p(y|x) \quad = \quad \frac{1}{\sqrt{2\pi\beta\sigma^2_{N_2}}}\exp\frac{-(y-\beta x)^2}{2\beta\sigma^2_{N_2}}, \tag{2.67}$$

where $\beta = 1 - \frac{\sigma^2_{N_2}}{\sigma^2_{\widetilde{X}}+\sigma^2_{N_1}}$. From (2.67) the test channel is constructed and shown in Fig. 2.23. We can see that Fig. 2.23 is nothing else but a scaling operation followed by additive

Gaussian noise. Since the test channel minimizes $I(X;Y)$, we can also say that it is the optimal attack channel that minimizes the information about the watermarked signal carried by the attacked signal.

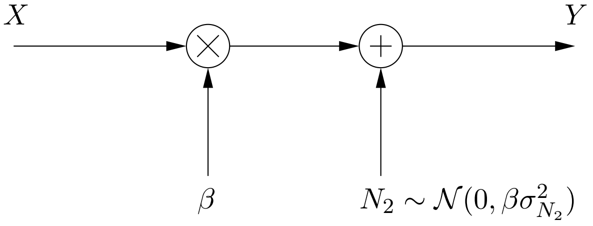

Figure 2.23: Test channel for Gaussian sources and squared-error criterion.

2.7.4 Linear Time-Invariant Filtering Attacks

Linear time-invariant (LTI) filtering operations are abundant in virtually all signal processing applications. Moreover, signal processing operations that are not explicitly implemented with filters can be modeled by a convolution with kernel. A malicious attacker can convolve the watermarked signal with such an optimally designed kernel to prevent the communication of the watermark message.

An example of linear filtering is shown in Fig. 2.24, where $h(\tau)$ is the filter impulse response. The output of the filter can be written as

$$\begin{aligned} Y(k) &= \sum_{\tau=0}^{L_f} h(\tau)X(k-\tau) \\ &= h(0)X(k) + h(1)X(k-1) + \cdots , \end{aligned} \tag{2.68}$$

where L_f is the filter length.

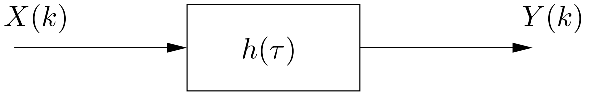

Figure 2.24: LTI filtering attack.

Each sample of the attacked data $Y(k)$ is a linear combination of the watermarked data X weighted by the filter coefficients. It is well known that the linear filtering operation introduces inter-symbol interference. Therefore, the noise that hampers the decoding of the watermark bits is the watermarked data itself. The larger the DWR the larger the variance of the host (and therefore watermarked) signal with respect to the watermark variance and the larger the probability of decoding error. If we assume that $h(0) = 1$, then from (2.64), the probability of error under linear filtering attacks can be written as

$$P_e = \sum_m \Pr[m\Delta - \frac{3\Delta}{4} \le N_f \le m\Delta - \frac{\Delta}{4}]. \tag{2.69}$$

If the watermarked signal is i.i.d. $X \sim \mathcal{N}(0, \sigma_X^2)$, then using (2.68) we have $N_f \sim \mathcal{N}\big(0, \sigma_X^2 \sum_{\tau=1}^{L_f} h^2(\tau)\big)$. If DWR is fixed, P_e depends mostly on the value of the expression $\sum_{\tau=1}^{L_f} h^2(\tau)$. For high DWR and $\sum_{\tau=1}^{L_f} h^2(\tau) \geq 1$, P_e is expected to be high.

If the watermarked signal is correlated, the variance of N_f is given as

$$\sigma_{N_f}^2 \;=\; \sum_{k=1}^{L_f}\sum_{l=1}^{L_f} h(\tau - k) K_X(k,l) h(\tau - l), \tag{2.70}$$

where $K_X(k,l)$ is the correlation function of the watermarked signal X.

If $h(0) \neq 1$, then we have the more devastating amplitude scale and noise attack[2] shown in Fig. 2.25.

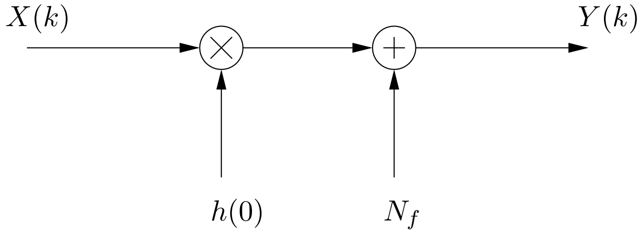

Figure 2.25: Amplitude scale and noise channel model for an LTI filtering attack.

2.8 Rational Dither Modulation

Quantization-based watermarking schemes are vulnerable to amplitude scaling operations on the watermarked data, because such operations cause a mismatch between the encoder and decoder code books. Amplitude scaling operations can be either unintentional (due to signal processing applications) or malicious (due to adversary).

To cope with amplitude scale attacks, a quantization-based scheme was proposed in [22], named Rational Dither Modulation (RDM). The advantage of the scheme is its invariance to amplitude scale attacks. The simplest version of a first order RDM scheme is shown in Fig. 2.26, together with the attack channel. The embedding rule is given as

$$X(k) \;=\; \big|X(k-1)\big| Q_W\Big(\frac{\widetilde{X}(k)}{\big|X(k-1)\big|}\Big), \tag{2.71}$$

where $Q_W(\cdot)$ denotes uniform scalar quantizer associated with watermark message W, see Fig. 2.13. The quantization stepsize is Δ. The quantizer operates not on the host signal sample $\widetilde{X}(k)$, but on the ratio $\frac{\widetilde{X}(k)}{|X(k-1)|}$ of the current host sample and the previous watermarked sample $X(k-1)$.

[2] Note that, since the attacked samples are linear combinations of the watermarked samples, the *noise* is not additive.

The attack channel consists of the additive Gaussian noise $N_2(k)$ and the constant amplitude scaling factor β.

The decoding rule is given as

$$\hat{W} \quad = \quad \arg\min_{W} \Big| \frac{Y(k)}{\big|Y(k-1)\big|} - Q_W\Big(\frac{Y(k)}{\big|Y(k-1)\big|}\Big)\Big| \tag{2.72}$$

First, the decoder computes the ratio $\frac{Y(k)}{\left|Y(k-1)\right|}$. Then the decoder determines the quantizer whose representation level is closest to the ratio $\frac{Y(k)}{\left|Y(k-1)\right|}$ and makes an estimate $\hat{W}$ of the embedded message. From (2.72) we can see that if the watermarked data $X(k)$ is scaled by β, β will be present in $Y(k)$ and in $Y(k-1)$. Therefore the ratio $\frac{Y(k)}{\left|Y(k-1)\right|}$ is independent of β, which makes the performance of the scheme invariant to amplitude scaling attacks.

There are some disadvantages of the first order RDM, which are as follows. Since the quantizer operate on the ratios, this operation can also be seen as quantizing the current samples with a variable step size quantizer, whose step size depends on the previous sample. Since in the attack channel there is also additive noise, the variable step sizes at the encoder and decoder will be different. This mismatch increases the probability of error with respect to additive noise attacks. Another disadvantage is that the distribution of the watermarked data becomes non-stationary, which makes the system difficult to analyze theoretically.

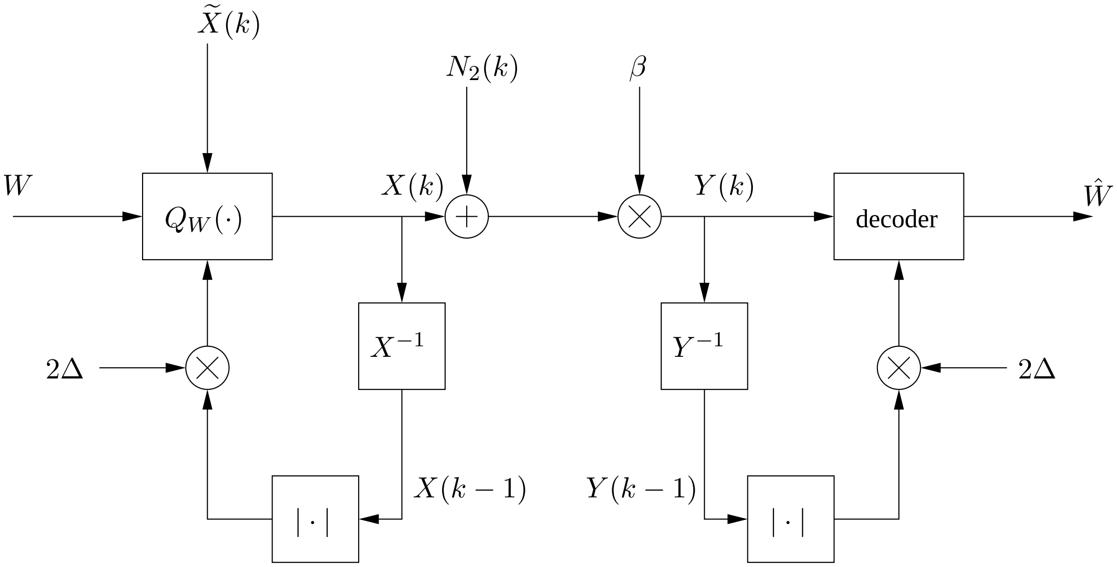

Figure 2.26: First order RDM scheme.

The quantization step can be made asymptotically constant by increasing the memory of the system. This can be done by introducing the function

$$g(X,k,L,p) \quad = \quad \Big(\frac{1}{L}\sum_{m=k-L}^{k-1} |X(m)|^p\Big)^{\frac{1}{p}}, \tag{2.73}$$

where L is the memory of the system and $p \geq 1$. See [22] for optimal choices of the parameter p. The Lth order RDM scheme is shown in Fig. 2.27. The encoding rule is given as

$$X(k) \quad = \quad g\big(X(k-1)\big)Q_W\Big(\frac{\widetilde{X}(k)}{g(X,k,L,p)}\Big) \tag{2.74}$$

Instead of $X(k-1)$, the encoder uses $g(X,k,L,p)$. The decoding rule is given as

$$\hat{W} \quad = \quad \arg\min_{W}\Big|\frac{Y(k)}{g(X,k,L,p)} - Q_W\Big(\frac{Y(k)}{g(X,k,L,p)}\Big)\Big| \tag{2.75}$$

As $L \to \infty$, $g(\cdot)$ approaches a constant, and the performance of RDM approaches that of QIM [22].

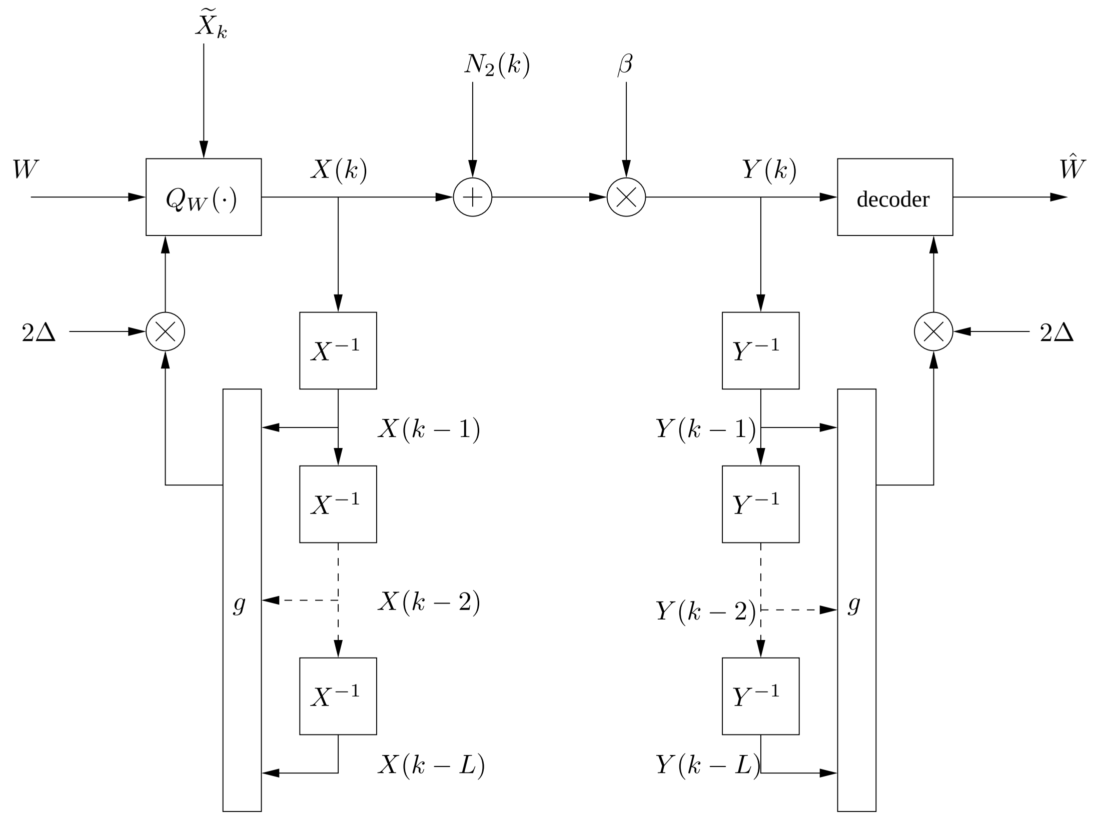

Figure 2.27: General order RDM scheme.

Experimental results of the probability of error are shown in Fig. 2.28. It can be seen that increasing the memory of the system reduces P_e mostly at high WNR, which is due to the fact that the quantization step sizes at the encoder and decoder become almost equivalent.

RDM can be further improved by incorporating perceptual models [65], or error correcting codes [66].

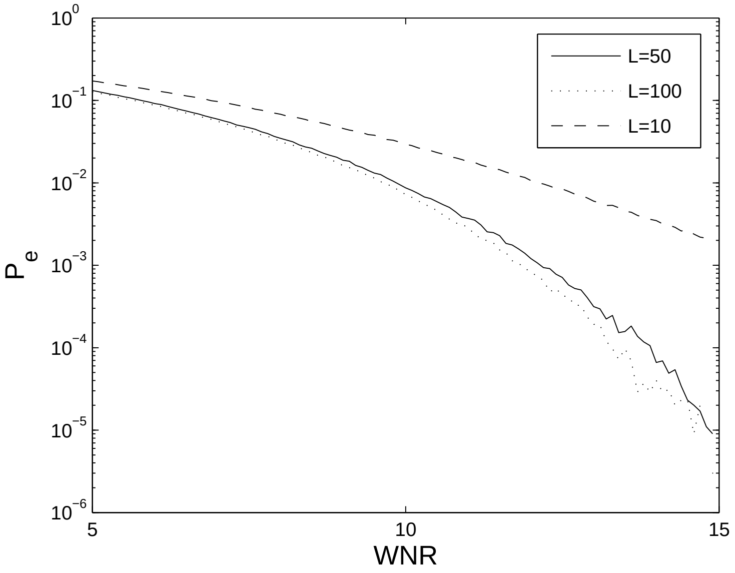

Figure 2.28: Probability of error for RDM. Chosen settings are $DWR = 25db$, $p = 2$, $\beta = 1.05$.

2.9 Discussion

In this section we have reviewed the information theoretic principles that form the foundation for studying the watermarking problem. We analyzed spread spectrum and quantization-based watermarking, and compared these two classes in terms of capacity and error exponents, subject to additive Gaussian noise attacks. Quantization-based watermarking outperforms blind spread spectrum watermarking in terms of capacity and error exponents. In terms of capacity the performance of quantization-based watermarking is equal to that of non-blind spread spectrum watermarking, and in terms of error exponents the performance of both classes is more or less the same.

We made a simple classification of the types of attacks in signal processing applications, and gave models for each of the most important attack operations, with a particular emphasis on the amplitude scale attack. Although quantization-based watermarking outperforms spread-spectrum watermarking in terms of additive noise attacks, the former remains vulnerable to amplitude scale and linear filtering attacks.

We discussed a novel technique called rational dither modulation, that is invariant to amplitude scale attacks. However, this technique has a lower capacity than distortion compensated QIM in terms of additive noise attacks. Another drawback is its insecurity. An attacker knowing the step size of the quantizer can easily decode the watermark. In the next chapters we present techniques overcoming these two drawbacks.

The performance of the described watermarking schemes can be further improved, irrespective of the attack channel, by encoding the watermark bits with error correcting codes [67, 68, 69] before embedding. The error correcting code will introduce redundancy, thus reducing the watermark payload, and of course reducing probability of error (if the errors are sufficiently low). The more powerful the error correcting code (Reed-Solomon, turbo codes) the bigger the improvement of the whole system.

Chapter 3

Amplitude Scale Estimation [*]

Watermarking schemes based on Quantization Theory have recently emerged as a result of Information Theoretic analysis [18, 19, 70, 71]. These schemes prove to perform better than the well known spread spectrum watermarking in the context of additive attacks [72, 73]. However, the resulting watermarking schemes fail to perform well for a number of important non-additive attacks (operations) [74, 75]. One such operation is amplitude scaling which is a common operation in many applications, such as audio play out and recording. Another application is Digital Audio Broadcasting (DAB), where amplitude scaling is even more complex, because different frequency bands are scaled (filtered) with different factors. Nonlinear scaling such as gamma correction can be seen in image processing applications. Quantization-based watermarking schemes are vulnerable against amplitude scaling. The reason for this is the fact that in order to assist the structured decoder, a maximum a posteori (MAP) estimation of the codeword used in the embedding stage is needed. Therefore, the amplitude scaling factor has to be known at the detection side for reliable codeword estimation.

Two approaches have been proposed in the literature to combat the scaling attack. One of them is based on estimating the scaling factors using the histogram of the received data. Once a good estimate is obtained, the scaling factors can be accounted for by dividing the received data by the estimated scaling factors, or by an appropriate modification of the watermark detector. We distinguish estimation based on pilot signal [21, 76] and blind estimation [77]. Another approach is based on optimized for the scaling attack codes [78] , such as modified trellis codes [79]. Invariance to amplitude scaling factors can also be achieved by incorporating into the QIM scheme a suitably modified perceptual model [80], such as the Watson model [81].

In this chapter we derive the probability density model of the received watermarked and attacked data when the encoder is Quantization Index Modulation (QIM) with distortion compensation (DC). Based on this model we derive two approaches for estimation of amplitude scaling modifications. In section 3.1, a mathematical model of the problem is introduced. In section 3.2, the model of the probability density function (PDF) of the

[*] This chapter is part of I. D. Shterev and R. L. Lagendijk "Amplitude Scale Estimation for Quantization-Based Watermarking", *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4146-4155, November 2006, and I. D. Shterev, R. L. Lagendijk, and R. Heusdens, "Statistical Amplitude Scale Estimation for Quantization-Based Watermarking", *SPIE Security, Steganography, and Watermarking of Multimedia Contents VI*, San Jose, CA, January 2004.

received data is derived. In section 3.3, a procedure based on Fourier Analysis is examined. In section 3.4, the maximum likelihood estimator is described. In section 3.5 we compare the two proposed estimation techniques, and in section 3.6, we describe the case when different messages are embedded. Finally, conclusions and discussion are presented in section 3.7.

3.1 Mathematical Formulation of QIM with DC

In this section we focus on the most popular quantization-based watermarking scheme, namely scalar Quantization Index Modulation (QIM). Random variables are denoted by capital letters and their realizations by the respective small letters. The notation $X \sim f_X(x)$ indicates that the random variable X has a PDF $f_X(x)$.

Fig. 4.1 shows the watermark encoder, where $W \in \{0, 1\}$ denotes the message bits that are embedded in the host data, $\widetilde{X}$ is the host signal itself with a variance $\sigma^2_{\widetilde{X}}$, X is the watermarked signal. The variable U is the output of the quantizer. $Q(\cdot)$ denotes uniform quantization with step size Δ. The quantization noise, which is the difference between the quantizer input and output, is defined as

$$\begin{aligned} N_1 &= \alpha\widetilde{X} - Q(\alpha\widetilde{X}) \\ &= \alpha\widetilde{X} - (X - (1-\alpha)\widetilde{X}) \\ &= \widetilde{X} - X, \end{aligned} \tag{3.1}$$

where α is a coefficient to be defined later.

From (4.1) we see that the watermark $\widetilde{X} - X$ and the quantization noise are equal. The quantizer input-output characteristic is shown in Fig. 4.2 for the watermark message $W \in \{0, 1\}$. The output of the quantizer can be written as:

$$U = \begin{cases} k\Delta & \text{if } W = 0 \\ (2k+1)\frac{\Delta}{2} & \text{if } W = 1 \end{cases} \tag{3.2}$$

where $k \in (-\infty, \infty)$ is an integer.

The attack channel is shown in Fig. 4.3. It consists of the constant amplitude scale factor β and the noise $N_2 \sim \mathcal{N}(0, \sigma^2_{N_2})$. The noise N_2 is independent of $\widetilde{X}$ and N_1. We choose the coefficient $\alpha = \frac{\sigma^2_{N_1}}{\sigma^2_{N_1}+\sigma^2_{N_2}}$ as in [42], where $\sigma^2_{N_1}$ is the variance of N_1. Other choices for α are also possible [82].

The attacked (received) signal Y can be written in the following way:

$$\begin{aligned} Y &= \beta X + N_2 \\ &= \beta(U + (1-\alpha)\widetilde{X}) + N_2. \end{aligned} \tag{3.3}$$

Using the relation $\alpha\widetilde{X} = U + N_1$, we obtain the received data Y in terms of N_1, N_2, and the watermark-bearing signal U:

$$Y = \frac{\beta}{\alpha}(U + (1-\alpha)N_1) + N_2. \tag{3.4}$$

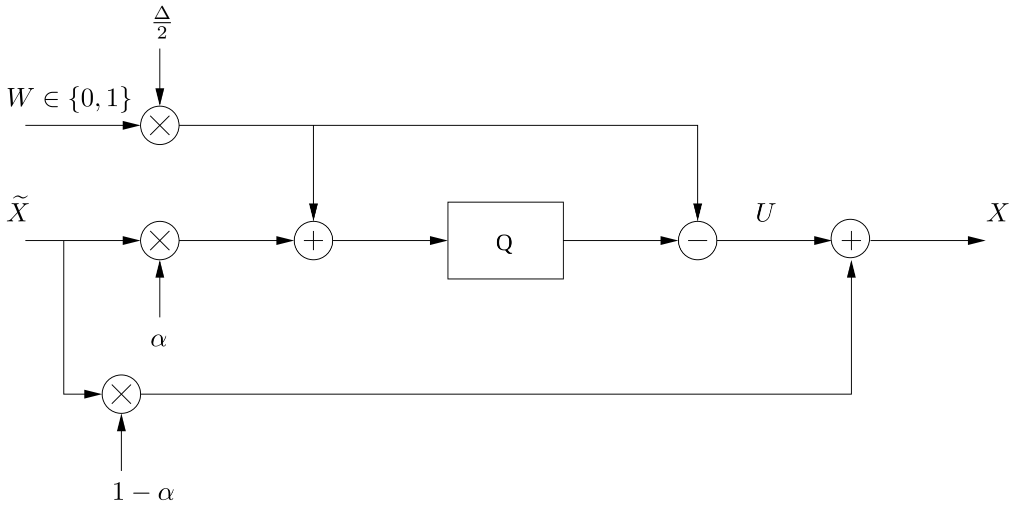

Figure 3.1: Watermark encoder.

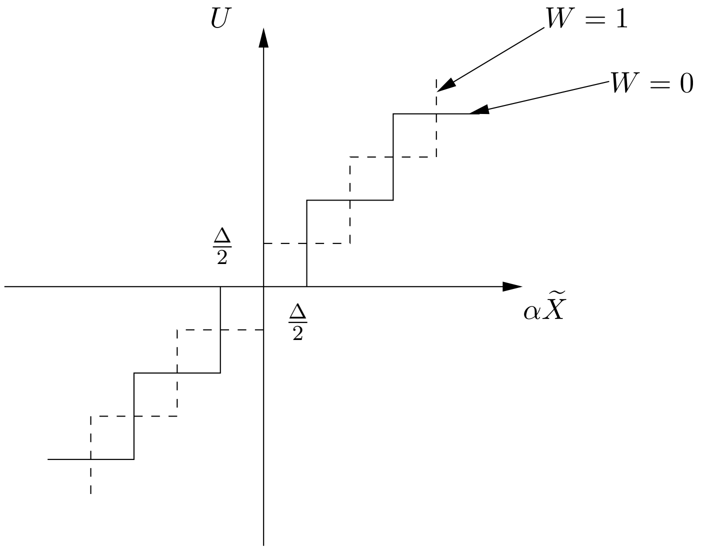

Figure 3.2: Quantizer input-output characteristics

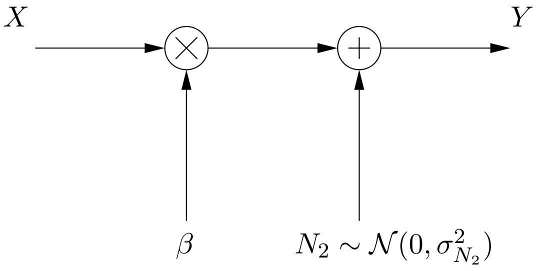

Figure 3.3: Attack Channel.

The watermark decoder is shown in Fig. 4.4. From the received signal Y, the decoder first performs Maximum a Posteriori Probability (MAP) estimation of the signal U, which under mild assumptions [18] is equivalent to multiplication by $\frac{\alpha}{\beta}$[1]. The decoder obtains:

$$\begin{aligned}\hat{U} &= \frac{\alpha}{\beta}Y \\ &= U + (1-\alpha)N_1 + \frac{\alpha}{\beta}N_2. \qquad (3.5)\end{aligned}$$

The decoder then computes the absolute value of the quantization noise $|\hat{U} - Q(\hat{U})|$ and makes an estimate of the embedded watermark in the following way:

$$\hat{W} = \begin{cases} 0 & \text{if } |\hat{U} - Q(\hat{U})| \leq \frac{\Delta}{4} \\ 1 & \text{if } |\hat{U} - Q(\hat{U})| > \frac{\Delta}{4} \end{cases} \qquad (3.6)$$

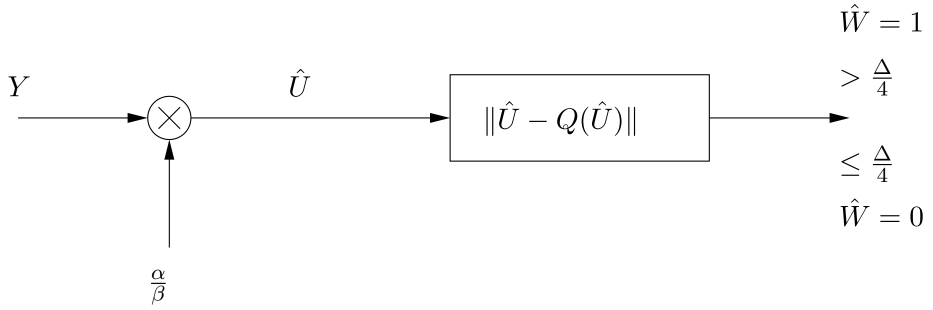

Figure 3.4: Watermark decoder.

Throughout we denote $WNR = 10\log\frac{\sigma^2_{N_1}}{\sigma^2_{N_2}}$, and the document-to-watermark ratio $DWR = 10\log\frac{\sigma^2_{\tilde{X}}}{\sigma^2_{N_1}}$.

Experimental results of the effect of unknown β on probability of error are shown in Fig. 3.5. We can see that the amplitude scale attack is more devastating at high WNR. At low WNR, the effect of the attack is less pronounced, because P_e is already quite large for $\beta = 1$.

3.2 PDF Models

In this section we first derive exact PDF models of the watermarked and attacked signals.

From Fig. 4.1, the PDF of the watermarked data X is given as:

$$f_X(x) = f_{X|W=0}(x)Pr[W=0] + f_{X|W=1}(x)Pr[W=1], \qquad (3.7)$$

where $Pr[W=0]$ and $Pr[W=1]$ are the probabilities of occurrence of bit 0 and 1, respectively, and $f_{X|W=0}(x)$ and $f_{X|W=1}(x)$ are the conditional PDFs of the watermarked data corresponding to $W=0$ and $W=1$, respectively.

[1] Here we assume that we are able to perfectly estimate β

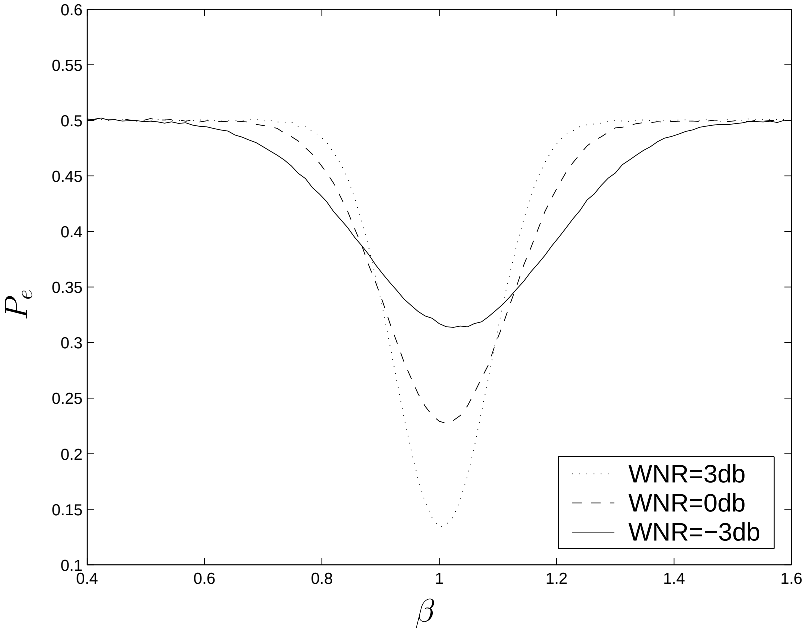

Figure 3.5: The effect of β on probability of error. Experiments are performed for $\widetilde{X} \sim \mathcal{N}(0,1)$, $N_2 \sim \mathcal{N}(0,\sigma_{N_2}^2)$, $DWR = 20db$.

Taking β and N_2 into account and using the fact that for any $\beta > 0$ we have $f_{\beta X|W}(x) = \frac{1}{\beta} f_{X|W}(\frac{x}{\beta})$, we obtain the PDF of the received data Y as:

$$\begin{aligned} f_Y(y) &= f_{N_2}(n_2) * f_{\beta X|W=0}(x) Pr[W=0] + f_{N_2}(n_2) * f_{\beta X|W=1}(x) Pr[W=1] \\ &= f_{N_2}(n_2) * \Big[\frac{1}{\beta} f_{X|W=0}\Big(\frac{x}{\beta}\Big) Pr[W=0]\Big] \\ &+ f_{N_2}(n_2) * \Big[\frac{1}{\beta} f_{X|W=1}\Big(\frac{x}{\beta}\Big) Pr[W=1]\Big], \end{aligned} \tag{3.8}$$

where the convolution $*$ follows from the independence between βX and N_2.

We derive the expression for $f_{X|W=0}(x)$. The derivation of $f_{X|W=1}(x)$ follows using similar reasoning.

Let us consider the case where the input to the quantizer is in the kth quantization cell, i.e. the output of the quantizer is $U = k\Delta$. We have

$$\Delta\Big(k - \frac{1}{2}\Big) < \alpha\widetilde{X} < \Delta\Big(k + \frac{1}{2}\Big). \tag{3.9}$$

Multiplying all sides by the positive term $\frac{1-\alpha}{\alpha}$, we get

$$\frac{(1-\alpha)\Delta}{\alpha}\Big(k - \frac{1}{2}\Big) < (1-\alpha)\widetilde{X} < \frac{(1-\alpha)\Delta}{\alpha}\Big(k + \frac{1}{2}\Big). \tag{3.10}$$

Adding $k\Delta$ to all sides and reorganizing, we obtain

$$\frac{\Delta}{\alpha}\Big(k - \frac{1-\alpha}{2}\Big) < (1-\alpha)\widetilde{X} + k\Delta < \frac{\Delta}{\alpha}\Big(k + \frac{1-\alpha}{2}\Big). \tag{3.11}$$

We define the indicator function

$$I_{A_{k|W=0}}(x) = \begin{cases} 1 & \text{if } x \in A_{k|W=0} \\ 0 & \text{if } x \notin A_{k|W=0} \end{cases} \tag{3.12}$$

where

$$A_{k|W=0} = \Big[\frac{\Delta}{\alpha}(k - \frac{1-\alpha}{2}), \frac{\Delta}{\alpha}(k + \frac{1-\alpha}{2})\Big]. \tag{3.13}$$

Therefore, the PDF of $(1-\alpha)\widetilde{X} + k\Delta$ over the support set $A_{k|W=0}$ is $\frac{1}{1-\alpha} f_{\widetilde{X}}\Big(\frac{x-k\Delta}{1-\alpha}\Big) I_{A_{k|W=0}}(x)$. Recognizing that $(1-\alpha)\widetilde{X} + k\Delta$ is the watermarked data X for a particular k, we can find the PDF of X by summing over k. Thus we have:

$$f_{X|W=0}(x) = \sum_{k=-\infty}^{\infty} \frac{1}{1-\alpha} f_{\widetilde{X}}\Big(\frac{x-k\Delta}{1-\alpha}\Big) I_{A_{k|W=0}}(x). \tag{3.14}$$

In the same fashion we can express the PDF of the watermarked data for $W = 1$ as

$$f_{X|W=1}(x) = \sum_{k=-\infty}^{\infty} \frac{1}{1-\alpha} f_{\widetilde{X}}\Big(\frac{x - \frac{2k+1}{2}\Delta}{1-\alpha}\Big) I_{A_{k|W=1}}(x) \tag{3.15}$$

where

$$A_{k|W=1} = \Big[\frac{\Delta}{\alpha}(k + \frac{\alpha}{2}), \frac{\Delta}{\alpha}(k + \frac{2-\alpha}{2})\Big]. \tag{3.16}$$

An illustration of (3.14) is shown in Fig. 3.6.

Referring to the above equations, we can now take the scaling factor β into account:

$$f_{\beta X|W=0}(x) = \frac{1}{\beta(1-\alpha)} \sum_{k=-\infty}^{\infty} f_{\widetilde{X}}\Big(\frac{x - k\beta\Delta}{\beta(1-\alpha)}\Big) I_{A_{k|\beta,W=0}}(x) \tag{3.17}$$

$$f_{\beta X|W=1}(x) = \frac{1}{\beta(1-\alpha)} \sum_{k=-\infty}^{\infty} f_{\widetilde{X}}\Big(\frac{x - \frac{2k+1}{2}\beta\Delta}{\beta(1-\alpha)}\Big) I_{A_{k|\beta,W=1}}(x) \tag{3.18}$$

where the indicator sets are given as

$$A_{k|\beta,W=0} = \Big[\frac{\beta\Delta}{\alpha}(k - \frac{1-\alpha}{2}), \frac{\beta\Delta}{\alpha}(k + \frac{1-\alpha}{2})\Big] \tag{3.19}$$

$$A_{k|\beta,W=1} = \Big[\frac{\beta\Delta}{\alpha}(k + \frac{\alpha}{2}), \frac{\beta\Delta}{\alpha}(k + \frac{2-\alpha}{2})\Big]. \tag{3.20}$$

An illustration of (3.17) is shown in Fig. 3.7. The regular pattern that carries information about the quantity $\frac{\beta\Delta}{\alpha}$ in the PDF of the watermarked data can clearly be seen. The work [21] exploits similar modeling.

Finally an illustration of (3.8) with $\beta = 1$ and $Pr[W = 0] = 1$ is given in Fig. 3.8.

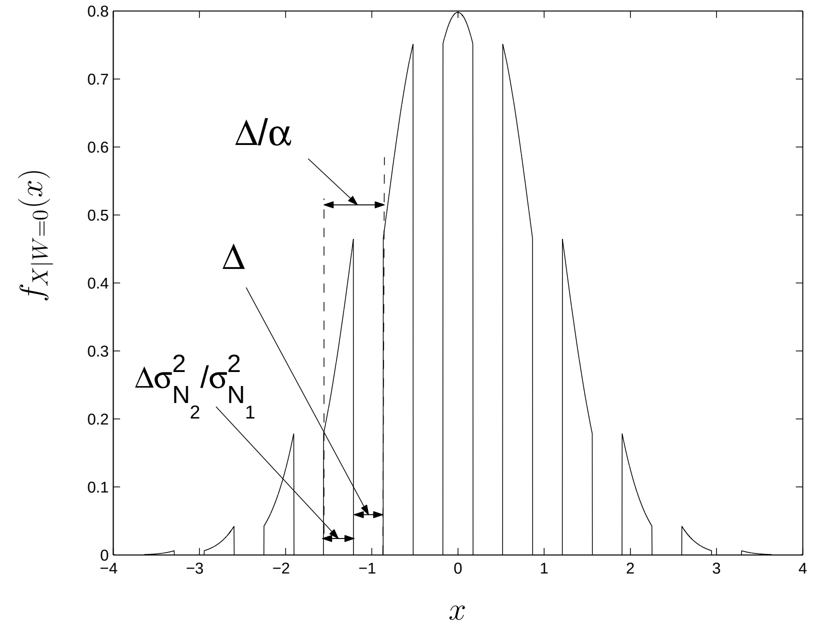

Figure 3.6: Graph of $f_{X|W=0}(x)$. Chosen settings are $\widetilde{X} \sim \mathcal{N}(0,1)$, $N_2 \sim \mathcal{N}(0,0.01)$, $\sigma^2_{N_1} = 0.01$, $\Delta \approx \sqrt{0.12}$.

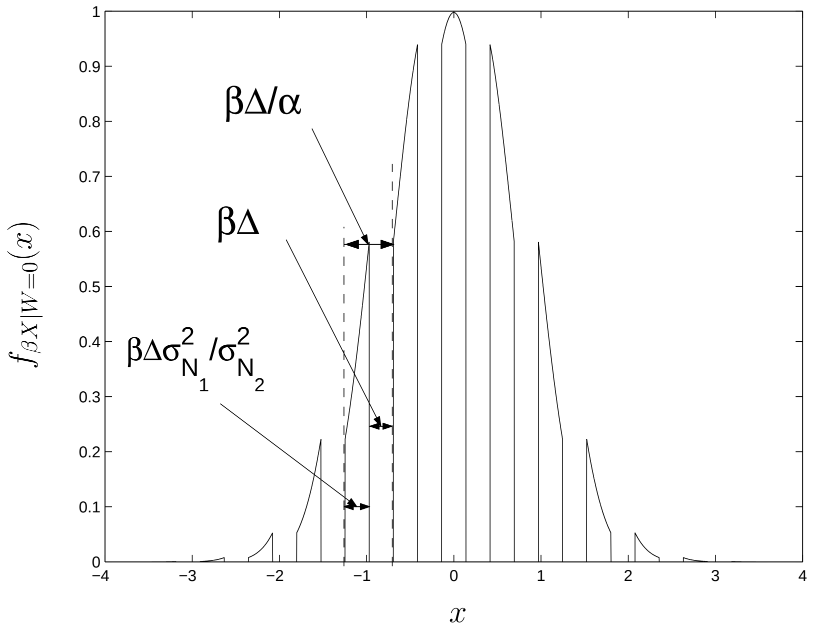

Figure 3.7: Graph of $f_{\beta X|W=0}(x)$. Chosen settings are $\widetilde{X} \sim \mathcal{N}(0,1)$, $N_2 \sim \mathcal{N}(0,0.01)$, $\sigma^2_{N_1} = 0.01$, $\Delta \approx \sqrt{0.12}$, $\beta = 0.8$.

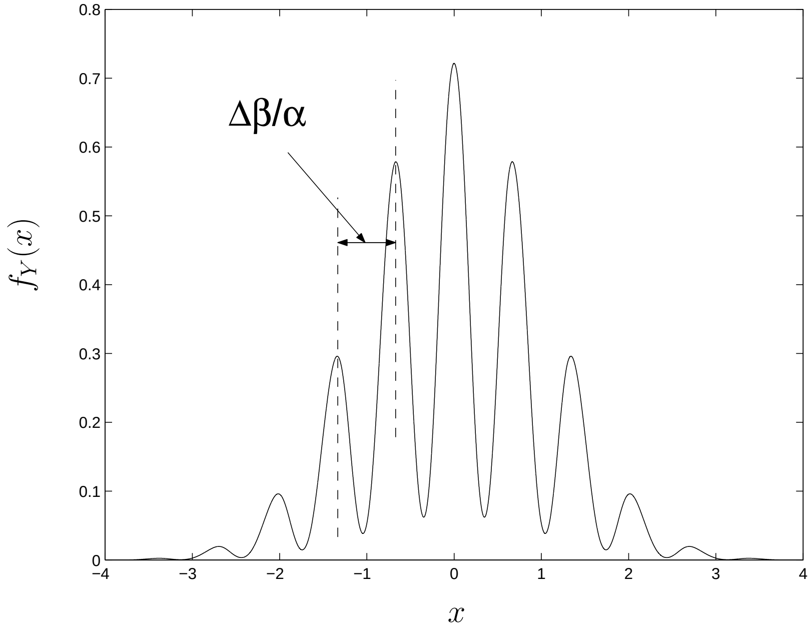

Figure 3.8: Graph of $f_Y(x)$ with $Pr(W=0)=1$. Chosen settings are $\widetilde{X}\sim\mathcal{N}(0,1)$, $N_2\sim\mathcal{N}(0,0.01)$, $\sigma^2_{N_1}=0.01$, $\beta=1$.

3.3 Estimation Based on Fourier Analysis

In this section we derive a procedure for scale estimation based on Fourier Analysis of the expression (10). A similar procedure was derived by Eggers [21] for the watermark encoding function $X=(1-\alpha^*)\widetilde{X}+\alpha^*Q(\widetilde{X})$ in the presence of dither. In [21] the authors choose the optimal values for the quantizer step size and the coefficient α^* numerically [83]. We noticed, though that there is no significant difference in performance between our procedure based on Fourier analysis and that described in [21].

We will need to define the characteristic function (c.f.) of a random variable X with pdf $f_X(x)$ as:

$$\Phi_X(\omega) \quad = \quad \int_{-\infty}^{+\infty} f_X(x)e^{i\omega x}dx \tag{3.21}$$

From (3.19) and also from Fig. 3.7 we can see that $f_{X|W}(x|w=0)$ has a regular structure of discontinuity and continuity regions with width of $\Delta\beta$ and $\Delta\beta\frac{\sigma^2_{N_2}}{\sigma^2_{N_1}}$ respectively. The total distance between the discontinuities is $\Delta\beta+\Delta\beta\frac{\sigma^2_{N_2}}{\sigma^2_{N_1}}=\frac{\Delta\beta}{\alpha}$. Therefore $\Phi_X(\omega)$ will have a periodic-like structure with a period $2\pi\frac{\alpha}{\Delta\beta}$. Observing (11), we can say that the periodicity of $\Phi_X(\omega)$ will not change if we embed only ones, i.e. P(W=1)=1, showing the advantage of working in the Fourier domain. From (10) it follows that the c.f. of the received data can be written as

$$\Phi_Y(\omega) \quad = \quad \Phi_X(\omega)\Phi_{N_2}(\omega) \tag{3.22}$$

where $\Phi_{N_2}(\omega)$ is the c.f. of N_2. In the estimation procedure we will need to estimate the periodicity $2\pi\frac{\alpha}{\Delta\beta}$ from $\Phi_Y(\omega)$, which due to the additive part in the attack channel will be disturbed in a degree depending on the strength of N_2 (see Fig. 4.16).

An illustration of characteristic functions for host, watermarked, and attacked data, for Gaussian sources and different ratios of WNR is shown in Fig. 4.16. The first dominating peak away from zero frequency is always at $\omega = 2\pi\frac{\alpha}{\Delta\beta}$.

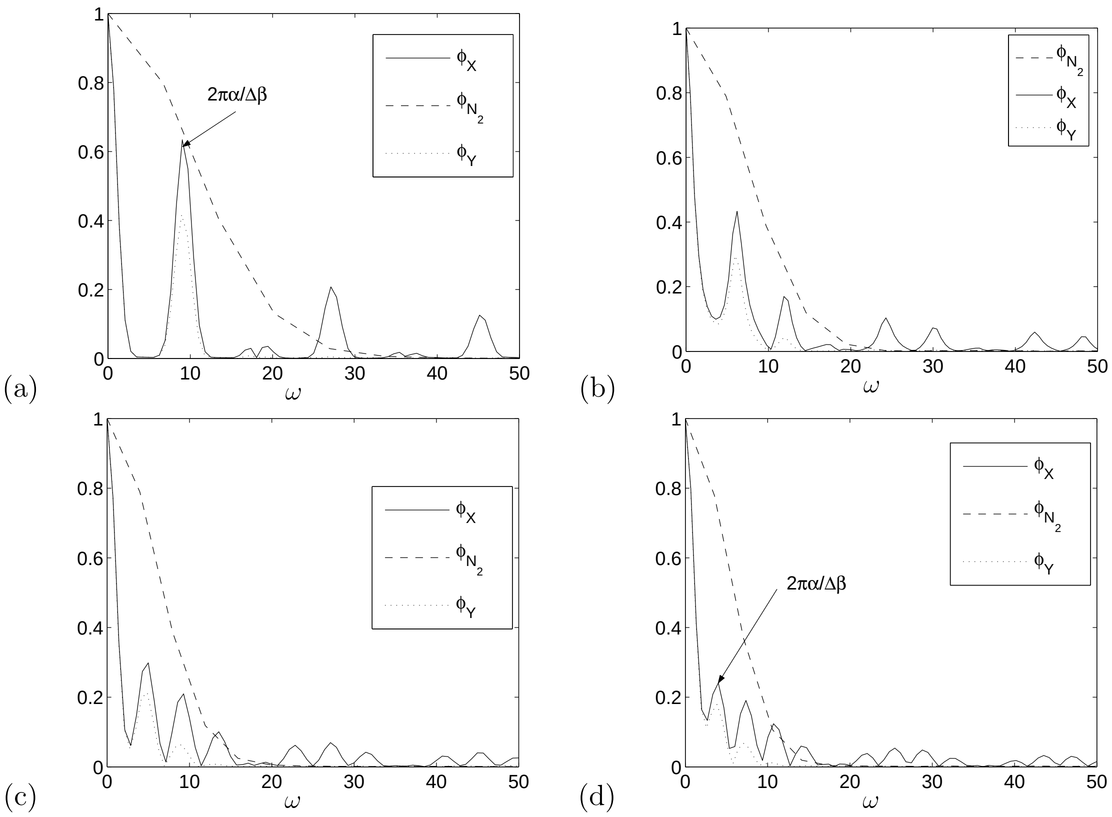

Figure 3.9: *Plot of characteristic functions for Gaussian sources. (a) $WNR = 0db$, (b) $WNR = -3db$, (c) $WNR = -4.8db$, (d) $WNR = -6db$.*

There are two interesting features of the encoder that are due to the presence of α in the periodicity of $\Phi_X(\omega)$. The first one improves the estimation robustness, while the second one hampers it. Increasing $\sigma^2_{N_2}$, the slope of $\Phi_{N_2}(\omega)$ becomes steeper, and since $\Phi_{N_2}(\omega) \leq \Phi_{N_2}(0) = 1$ for every ω (which is true for every valid c.f., see [84]), we can say that for low WNR, only those peaks of $\Phi_X(\omega)$ that are nearer to $\omega = 0$ will survive. Fortunately decreasing WNR will also decrease α and the peaks of $\Phi_X(w)$ will be shifted towards lower frequencies, countering the effect of increasing $\sigma^2_{N_2}$. The negative impact of α consists of the fact that with decreasing WNR, a bigger part of the host signal will pass through the $(1-\alpha)\widetilde{X}$ branch, therefore reducing the part that passes through the quantizer. As a result of that the zero regions in the PDF of the watermarked data will tend to disappear, the peaks in $\Phi_X(\omega)$ will become flatter (even before multiplying with $\Phi_{N_2}(\omega)$) as illustrated in Fig. 4.16, from which it would be more difficult to estimate the scaling factor. However, experiments showed that the positive feature prevails and knowing the statistics at the encoder side gives better results than the case of QIM.

3.4 Maximum Likelihood Estimation

In this section we will derive the Maximum Likelihood (ML) functional of β and study its properties. A derivation of an analytical expression for this method is quite tedious and in most cases is not possible. That is why we have to constrain ourselves to working with convolution of PDFs.

The ML estimator of β can be written as:

$$\hat{\beta} \quad = \quad \arg\max_{\beta} f_Y(y|\beta) \tag{3.23}$$

We will assume that the samples of the received data are independent, for which we can write the joint PDF of the received data as a product of the individual densities, i.e. $f_Y(y) = f_Y(y_1)f_Y(y_2)...f_Y(y_n)$. We note however that such an assumption may result in a source of substantial loss for real audio signals, exhibiting high correlation between the samples. Expanding (3.23), we get:

$$\begin{aligned} \hat{\beta} &= \arg\max_{\beta} f_Y(y_1, y_2, ..., y_n|\beta) \\ &= \arg\max_{\beta} f_Y(y_1|\beta)f_Y(y_2|\beta)...f_Y(y_n|\beta) \\ &= \arg\max_{\beta} \sum_{i}^{n} \ln f_Y(y_i|\beta) \qquad (3.24) \end{aligned}$$

where the last line follows from the monotonicity of the logarithm.

Since it is difficult to further simplify the expression of $f_Y(y)$ for general (even for Gaussian[2]) sources, we perform experiments with Gaussian sources to see the behavior of the likelihood function (LF) $\sum_i^n \ln f_Y(y_i|\beta)$ as a function of β. In Fig. 3.10(a), curves are shown for different β. In Fig. 3.10(b) we plot the likelihood function for different WNR. The maximum in the likelihood function curves indicating the right scaling factor β used in the attack channel is clearly visible in all cases. We can see that around the maximum, the likelihood function exhibits almost concave behavior.

3.5 Experimental Results

In this section we compare the performance of the proposed estimation techniques in terms of the ratio $\frac{\sigma_{N_2}^2}{\sigma_{N_1}^2}$, and the number of available signal samples, for different audio signals. Experimental results for the estimation procedure based on Fourier analysis with real audio host signals are shown in Fig. 3.11. It can be seen that reliable estimation of β is possible in the presence of additive noise with ratios up to $\frac{\sigma_{N_2}^2}{\sigma_{N_1}^2} = 2$. Fig. 3.11 shows that reliable estimation is also possible from around 1000 signal samples.

[2] Because of the discontinuity in $f_X(x)$ it is difficult to obtain a convenient analytical expression for $f_Y(y)$ trough characteristic functions.

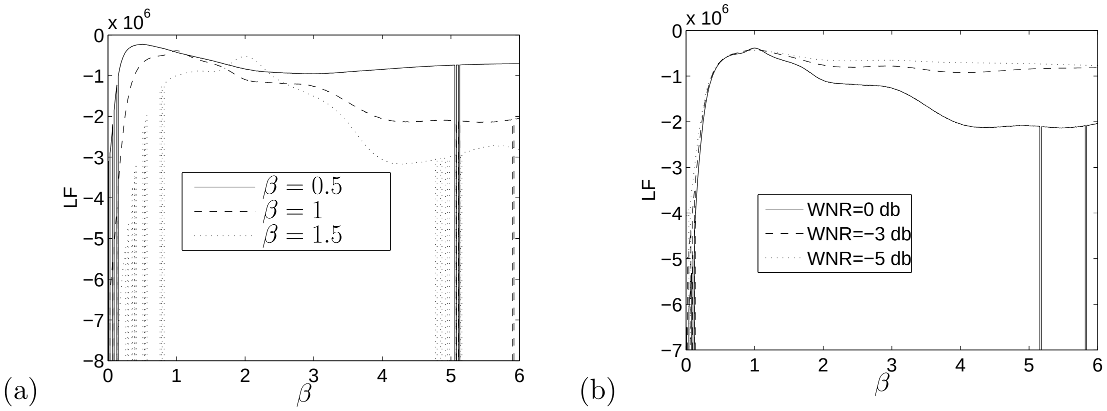

Figure 3.10: *Plot of experimental LF with Gaussian sources, $DWR = 20db$. (a) Different values of β, and fixed $WNR = 0db$. (b) Different values of WNR, and $\beta = 1$.*

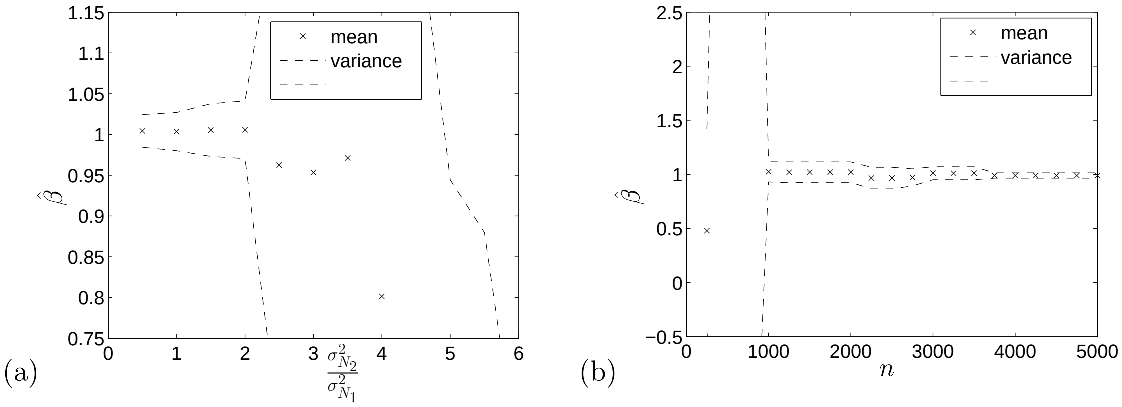

Figure 3.11: *Plot of $\hat{\beta}$ for Fourier-based estimation averaged over different audio signals with $DWR = 20db$, $\beta = 1$. The crosses represent the estimation mean. The dashed curves represent the variance of the estimation in both directions.*

In Fig. 3.12, the ML approach is evaluated with real audio signals (model mismatch) and $\beta = 1$. In the experiments the decoder assumes a Laplacian host signal with variance $\sigma^2_{\widetilde{X}} + \sigma^2_{N_1} + \sigma^2_{N_2}$. For the small distortion case $\sigma^2_{\widetilde{X}} \gg \sigma^2_{N_1}, \sigma^2_{N_2}$, $\sigma^2_{\widetilde{X}} + \sigma^2_{N_1} + \sigma^2_{N_2} \approx \sigma^2_{\widetilde{X}}$. Therefore, in practical applications, guessing the host signal variance at the detection side is not a big issue. In terms of the ratio $\frac{\sigma^2_{N_2}}{\sigma^2_{N_1}}$, the ML approach outperforms the Fourier-based approach, especially at high ratios $\frac{\sigma^2_{N_2}}{\sigma^2_{N_1}}$. In terms of the number of available signal samples, it can be seen that reliable estimation of the amplitude scaling factor with the ML approach is possible from around 2500 samples, which is higher than the minimum number of signal samples needed for estimation with the Fourier-based method.

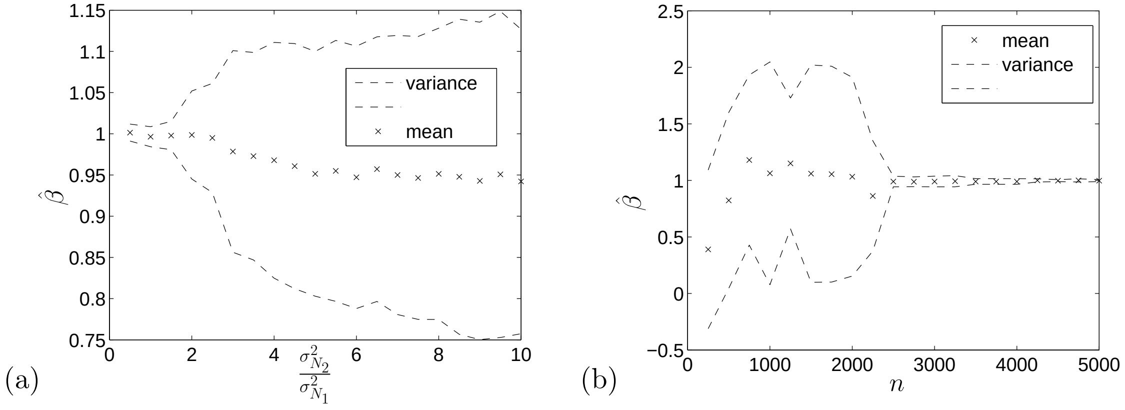

Figure 3.12: *Plot of $\hat{\beta}$ for ML estimation averaged over different audio signals, for fixed $\beta = 1$. The crosses represent the estimation mean. The dashed curves represent the variance of the estimation in both directions.*

3.6 A Note on Different Messages

In this section we discuss the case of imbedding different messages with specified probabilities $Pr[W = 0], Pr[W = 1] \neq 0$ and its influence on the proposed estimation procedures. Since this case will mostly affect the discontinuity of $f_X(x)$, we will concentrate on the Fourier based estimation method.

Lets assume that $f(W = 0) \approx f(W = 1) \approx 0.5$, or in other words there is a large enough number of zeros and ones in the watermark bitstream. From (3.19) and (3.20) we can see that when $\alpha = 0.5$, the union of $A_{k,W=0}(x)$ and $A_{k,W=1}(x)$ completely covers the real line, $f_X(x)$ will be absolutely continuous, and there will be no periodicity in $\Phi_X(\omega)$. Further decreasing α will cause $A_{k,W=0}(x)$ and $A_{k,W=1}(x)$ to overlap. An illustration of these cases is shown in Fig. 3.13.

We can conclude that in the case of large enough number of zeros and ones in the watermark bitstream, the Fourier based approach will work only within the restriction $\sigma^2_{N_2} < \sigma^2_{N_1}$. The ML approach does not rely on the discontinuity in the PDF of the watermarked data and therefore is invariant to this restriction.

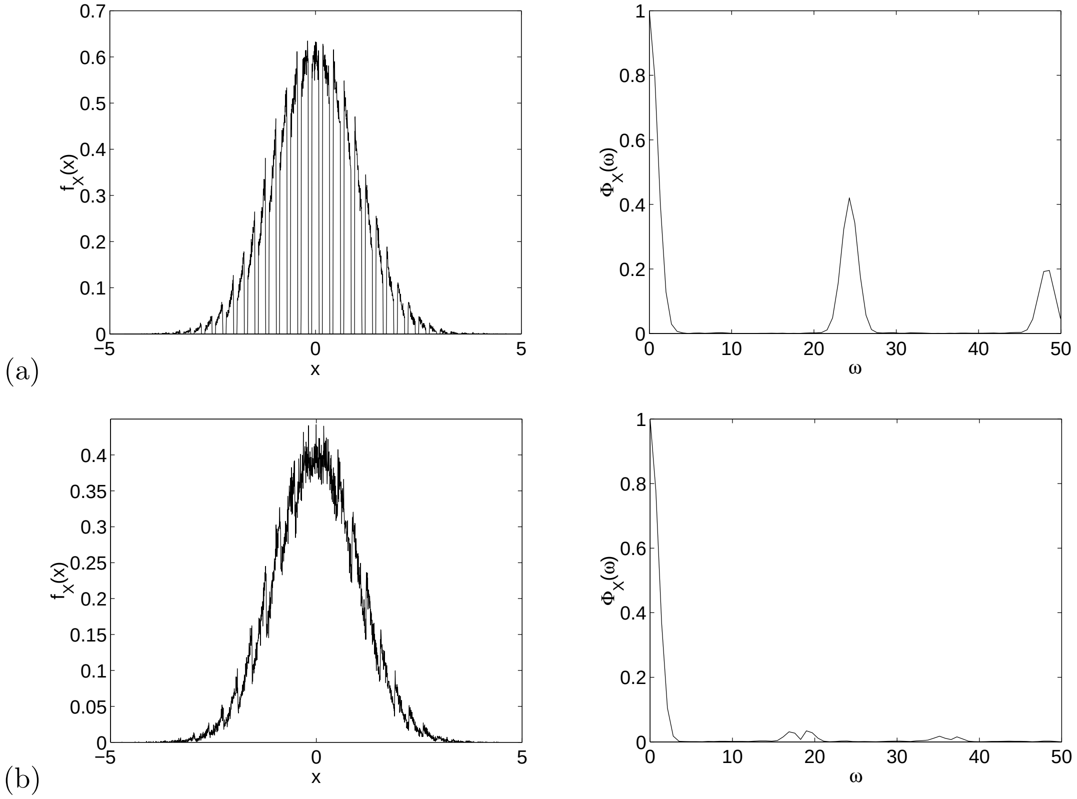

Figure 3.13: Graphs of $f_X(x)$ and their corresponding $\Phi_X(\omega)$ in the case of $Pr[W = 0] = Pr[W = 1] = 0.5$. (a) $WNR = 3db$ and (b) $WNR = 0db$.

3.7 Discussion

We presented two statistical procedures for estimation of scaling factors in attack channels consisting of amplitude scaling followed by additive noise. The advantage of the procedure based on characteristic functions is that the method relies on the discontinuity of the PDF of the watermarked data, and is not "generally" dependent on the host signal. However, for too strong noise in the attack channel, the method fails. Another disadvantage is the insecurity. An attacker can easily determine the quantity $\frac{\alpha}{\Delta\beta}$ from the characteristic function of the received data and decode the watermark by directly applying lattice decoding with step size $\frac{\beta\Delta}{\alpha}$. The method is computationally cheap and suitable for real-time applications. The second method based on ML estimation is computationally more expensive than the method based on characteristic functions. In our implementation, though, we managed to estimate β in around 50 sec. from 10000 signal samples. Another disadvantage of the method is that it is theoretically dependent on the host signal statistics (although the experimental results indicate good performance in case of model mismatch for a variety of audio host signals). The advantage of the method is the high estimation accuracy even in the presence of very strong noise in the attack channel. In the case of model mismatch, in terms of the ratio $\frac{\sigma^2_{N_2}}{\sigma^2_{N_1}}$, the ML approach gives better results than the Fourier based approach, while in terms of available signal samples, the Fourier based method gives superior estimation.

The quantization-based watermarking scheme on which the estimation procedures are based is not secure. An attacker knowing the distortion of the encoder can easily decode the watermark. In the next section we propose a technique for making the watermarking system secure, and develop amplitude scale estimation procedure for it.

Chapter 4

Amplitude Scale Estimation in the Presence of Dither*

Watermarking is the process of imperceptibly embedding a message (watermark) into a host signal (audio, video). The resulting signal is called a watermarked signal. The message should introduce only tolerable distortion to the host signal and it should be recoverable by the intended receiver after signal processing operations on the watermarked data.

Watermarking schemes based on quantization theory have recently emerged as a result of information theoretic analysis [18, 20, 39, 42]. In terms of additive noise attacks, these schemes have proven to perform better than traditional spread spectrum watermarking because they can completely cancel the host signal interference, which makes them invariant to the host signal [85]. The existence of good lattices in high dimensions [24, 86, 87] that can be directly and efficiently implemented has made quantization-based schemes of practical interest.

Lattice-based schemes are vulnerable to amplitude scale attacks, because these attacks introduce mismatch between the encoder and the decoder lattice volumes. Furthermore, amplitude scaling induces a large amount of distortion with respect to the mean squared error, but does not cause significant perceptual degradations. Such operation on watermarked signals is quite common in many applications. One example is audio play-out and capturing, where the watermarked signal is passed through a D-A converter, transmitted through an analog noisy channel, captured by a microphone, and converted back to a digital representation. Clearly the microphone will capture a less powerful and degraded watermarked signal, which has led us to model the noisy channel as an amplitude scaling operation followed by additive noise. In this paper we concentrate on operations consisting of amplitude scaling followed by additive white Gaussian noise (AWGN), often called scale additive white Gaussian noise (SAWGN) channel.

Several techniques are known in the literature for combating amplitude scale attacks.

[*] This chapter is part of I. D. Shterev and R. L. Lagendijk "Amplitude Scale Estimation for Quantization-Based Watermarking, *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4146-4155, November 2006, I. D. Shterev and R. L. Lagendijk "Maximum Likelihood Amplitude Scale Estimation for Quantization-Based Watermarking in the Presence of Dither", *SPIE Security, Steganography, and Watermarking of Multimedia Contents VII*, San Jose, CA, January 2005, and R. L. Lagendijk and I. D. Shterev "Estimation of Attacker's Scale and Noise Variance for QIM-DC Watermark Embedding, *IEEE International Conference on Image Processing*, Singapore, October 2004.

One of the approaches is based on designing watermarking codes that are invariant to amplitude scale operations, such as modified trellis codes [79], order preserving lattice codes [88], and rational dither modulation [22]. Another approach is based on estimating the non-additive operations and inverting them prior to watermark decoding, using pilot signals [21], or blind estimation [89, 83, 78]. More recently an iterative estimation procedure in combination with error correcting codes (equalization [90]) was proposed [91, 92, 93], which proved to perform well even for low watermark-to-noise ratios (WNR). The advantage of the approach in [21] is the ability to estimate the scaling factor from a small number of signal samples, which makes the estimation procedure applicable in situations where the scaling factor slowly varies. The disadvantage of the method is that the pilot signals consume part of the capacity of the watermarking system. The method proposed in [83] performs well for low WNR, but lacks security, in the sense that an attacker knowing the distortion of the embedder is able to estimate the scaling factors and decode the watermark. The methods based on invariant codes give small probability of error with respect to amplitude scale attacks at the expense of increased probability of error [22, 88] with respect to additive noise attacks and reduced payload [79].

In this chapter we propose a Maximum Likelihood approach for estimating amplitude scaling factors. Our estimation technique is blind and only assumes knowledge of the watermark message priors. No knowledge of the position of the message bits in the watermark bitstream is required. We also introduce subtractive dither [94] in the encoder. The realization of the dither is assumed to be known to the decoder. An application of subtractive dither to watermarking appeared first in [17], but with no theoretical analysis of the system security. In this chapter we design the dither statistics such that an attacker without knowing the dither realization is not able to decode the watermark. Thus the dither serves as the *key* ensuring security of the system.

The chapter is organized as follows. In Section 4.1 we formulate the attack channel, the watermark encoder and decoder. In Section 4.2 we derive the probability density function (PDF) of the received data in the presence of dither [14]. In Section 4.4 we give conditions for the dither sequence statistics, such that a given level of security is achieved and at the same time the dither variance is as small as possible, using the probability of error of the watermarking system as an objective function. A description of the ML estimation procedure is given in Section 4.5. Section 4.6 contains experimental results with synthetic and real audio host signals, and Section 4.8 concludes the chapter.

4.1 Mathematical Formulation

In this section we again focus on the most popular quantization-based watermarking scheme, namely scalar Quantization Index Modulation (QIM).

Fig. 4.1 shows the watermark encoder, where $D \sim f_D(d)$ is the dither sequence with a variance σ_D^2. The statistics of the dither sequence will be derived in Section 4.4. The quantization noise, which is the difference between the quantizer input and output, is defined as

$$\begin{aligned} N_1 &= \alpha\widetilde{X} + D - Q(\alpha\widetilde{X} + D) \\ &= \alpha\widetilde{X} - \big(X - (1-\alpha)\widetilde{X}\big) \\ &= \widetilde{X} - X, \end{aligned} \tag{4.1}$$

The output of the quantizer can be written as:

$$U = \begin{cases} k\Delta & \text{if } W = 0 \\ (2k+1)\frac{\Delta}{2} & \text{if } W = 1 \end{cases} \tag{4.2}$$

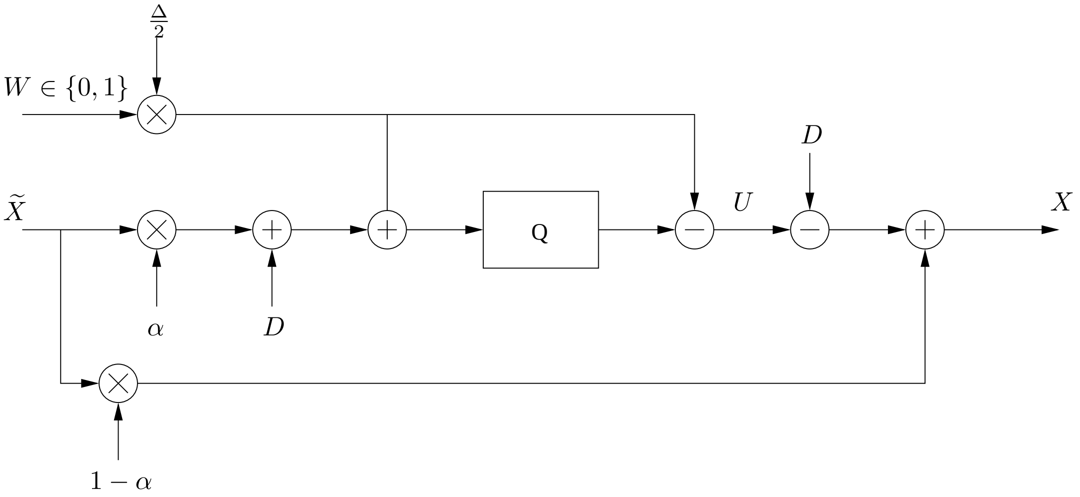

Figure 4.1: Watermark encoder.

The attack channel is shown in Fig. 4.3. The attacked (received) signal Y can be written in the following way:

$$\begin{aligned} Y &= \beta X + N_2 \\ &= \beta\big(U - D + (1-\alpha)\widetilde{X}\big) + N_2. \end{aligned} \tag{4.3}$$

Using the relation $\alpha\widetilde{X} = U - D + N_1$, we obtain the received data Y in terms of N_1, N_2, and the watermark-bearing signal U:

$$Y = \frac{\beta}{\alpha}\big(U - D + (1-\alpha)N_1\big) + N_2. \tag{4.4}$$

The watermark decoder is shown in Fig. 4.4. From the received signal Y, the decoder first performs Maximum a Posteriori Probability (MAP) estimation of the signal $U - D$, which under mild assumptions [18] is equivalent to multiplication by $\frac{\alpha}{\beta}$[1]. Then the decoder adds the dither D, obtaining:

$$\begin{aligned} \hat{U} &= \frac{\alpha}{\beta}Y + D \\ &= U + (1-\alpha)N_1 + \frac{\alpha}{\beta}N_2. \end{aligned} \tag{4.5}$$

[1] Here we assume that we are able to perfectly estimate β

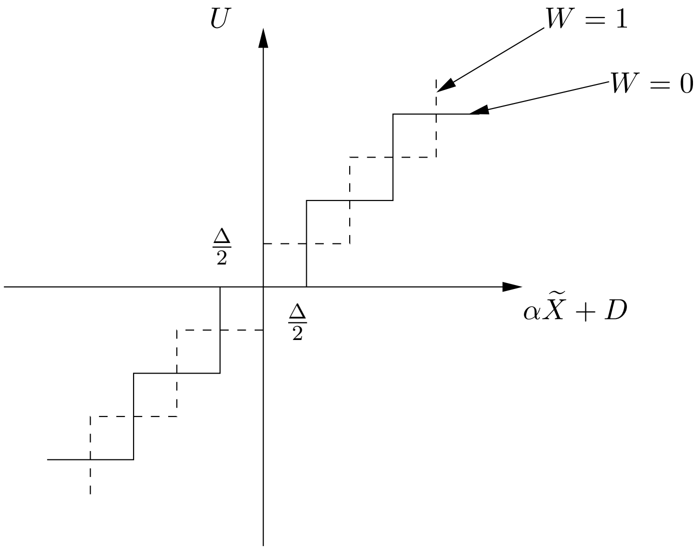

Figure 4.2: Quantizer input-output characteristics

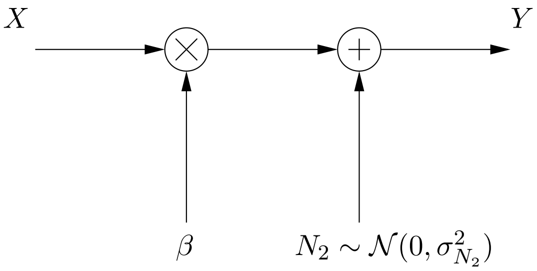

Figure 4.3: Attack Channel.

The decoder then computes the absolute value of the quantization noise $|\hat{U} - Q(\hat{U})|$ and makes an estimate of the embedded watermark in the following way:

$$\hat{W} = \begin{cases} 0 & \text{if } |\hat{U} - Q(\hat{U})| \leq \frac{\Delta}{4} \\ 1 & \text{if } |\hat{U} - Q(\hat{U})| > \frac{\Delta}{4} \end{cases} \tag{4.6}$$

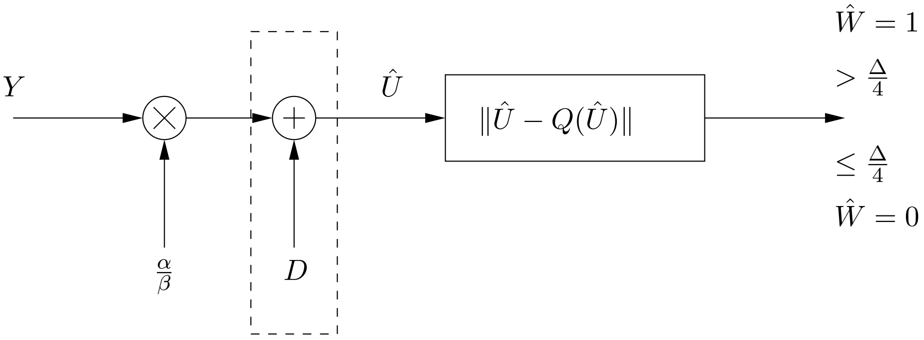

Figure 4.4: Watermark decoder.

4.2 PDF Models

Since in the presence of subtractive dither $f_X(x)$ will be perturbed by D, it is difficult to derive a useful exact mathematical expression for it. That is why we choose to manipulate X in a convenient way, having knowledge of D, so that we are able to mathematically describe the structure of the PDF of the resulting random variable.

For simplicity, we will assume that only message $W = 0$ is embedded, therefore working only with the first part of (3.8). Extension to the more general case of embedding zeros and ones is straightforward: use the whole expression (3.8).

Referring to Fig. 4.1, let us assume that $Pr[W = 0] = 1$ and $\alpha\widetilde{X} + D$ belongs to the k-th quantization cell, i.e.:

$$\Delta\Big(k - \frac{1}{2}\Big) < \alpha\widetilde{X} + D < \Delta\Big(k + \frac{1}{2}\Big). \tag{4.7}$$

Multiplying by $\frac{1-\alpha}{\alpha}$ and adding $k\Delta$, we obtain:

$$\frac{\Delta}{\alpha}\Big(k - \frac{1-\alpha}{2}\Big) < (1-\alpha)\widetilde{X} + k\Delta + \frac{1-\alpha}{\alpha}D < \frac{\Delta}{\alpha}\Big(k + \frac{1-\alpha}{2}\Big). \tag{4.8}$$

Recognizing that the leftmost and rightmost parts of (4.8) are the indicator set $A_{k|W=0}$ as given by (3.13) and taking into account the fact that $(1-\alpha)\widetilde{X} + k\Delta + \frac{1-\alpha}{\alpha}D = X + \frac{1}{\alpha}D$ (see Fig. 4.1), we can write the PDF of $X + \frac{1}{\alpha}D$ for a particular k as

$$f'_{X+\frac{1}{\alpha}D}(x) \quad = \quad f_{(1-\alpha)\widetilde{X}+k\Delta+\frac{1-\alpha}{\alpha}D}(x) I_{A_{k|W=0}}(x). \tag{4.9}$$

Generalizing for all k, we have

$$f_{X+\frac{1}{\alpha}D}(x) \quad = \quad \sum_{k=-\infty}^{+\infty} f_{(1-\alpha)\widetilde{X}+k\Delta+\frac{1-\alpha}{\alpha}D}(x) I_{A_{k|W=0}}(x). \tag{4.10}$$

The key expression for the estimation procedure in the presence of subtractive dither is (4.10). We can see that although X is perturbed by the dither, if we add the term $\frac{1}{\alpha}D$ to the watermarked signal, we are able to obtain a signal that has a PDF with an indicator function equal to that when no dither is used. In other words, we are able to *recover* the structure of the watermarked signal PDF by the use of the dither.

Taking into account β and the additive noise N_2, we now have:

$$f_{Y+\frac{\beta}{\alpha}D}(y+\frac{\beta}{\alpha}d) \quad = \quad f_{N_2}(n_2) * f_{\beta X+\frac{\beta}{\alpha}D}(x+\frac{\beta}{\alpha}d), \tag{4.11}$$

where the convolution $*$ follows from the independence between N_2 and $\beta X + \frac{\beta}{\alpha}D$.

4.3 Approximation to the PDF Models

Since expression (4.10) is very complex to implement, we make approximations to it. We can see that there are only two random variables involved in (4.10), namely $\widetilde{X}$ and D. Assuming that $\sigma^2_{\widetilde{X}} \gg \sigma^2_D$, we can approximate $f_{X+\frac{1}{\alpha}D}(x)$ in the following way:

$$\begin{aligned} f_{X+\frac{1}{\alpha}D}(x) &= \sum_{k=-\infty}^{+\infty} f_{(1-\alpha)\widetilde{X}+k\Delta+\frac{1-\alpha}{\alpha}D}(x) I_{A_{k|W=0}}(x) \\ &\approx \sum_{k=-\infty}^{+\infty} f_{(1-\alpha)\widetilde{X}+k\Delta}(x) I_{A_{k|W=0}}(x) \end{aligned} \tag{4.12}$$

Note that the output of the quantizer depends both on $\widetilde{X}$ and D, but since the variance of the first is assumed to be much larger, the term $k\Delta$ is present in the approximation together with $\widetilde{X}$. An illustration of $f_{X+\frac{1}{\alpha}D}(x)$, its approximation as given by (4.12), and $f_X(x)$ is shown in Fig. 4.5. The difference between $f_{X+\frac{1}{\alpha}D}(x)$ and its approximation can hardly be recognized. We can also see the huge difference between $f_{X+\frac{1}{\alpha}D}(x)$ and $f_X(x)$.

4.4 Design of the Dither Sequence

In the previous section we saw that $\sigma^2_{\widetilde{X}} \gg \sigma^2_D$ in order for (4.12) to be an accurate approximation. Approximation is perfect if $\sigma^2_D = 0$, but this is unacceptable from security point of view. In this section we find sufficient conditions for the dither sequence statistics such that, for σ^2_D as small as possible, an attacker is not able to decode the watermark with an error probability[2] different than 0.5.

To derive the conditions, we first need to derive the error probability, which is given by the following theorem.

Theorem 1: When the dither sequence D is not known to the decoder, the error probability P_e is given by the expression:

$$P_e \quad = \quad \sum_m \Pr[m\Delta - \frac{3\Delta}{4} \leq (1-\alpha)N_1 + \frac{\alpha}{\beta}N_2 - D \leq m\Delta - \frac{\Delta}{4}], \tag{4.13}$$

[2] Since we have one-dimensional, one-bit watermarking, the error probability and bit error probability are equal.

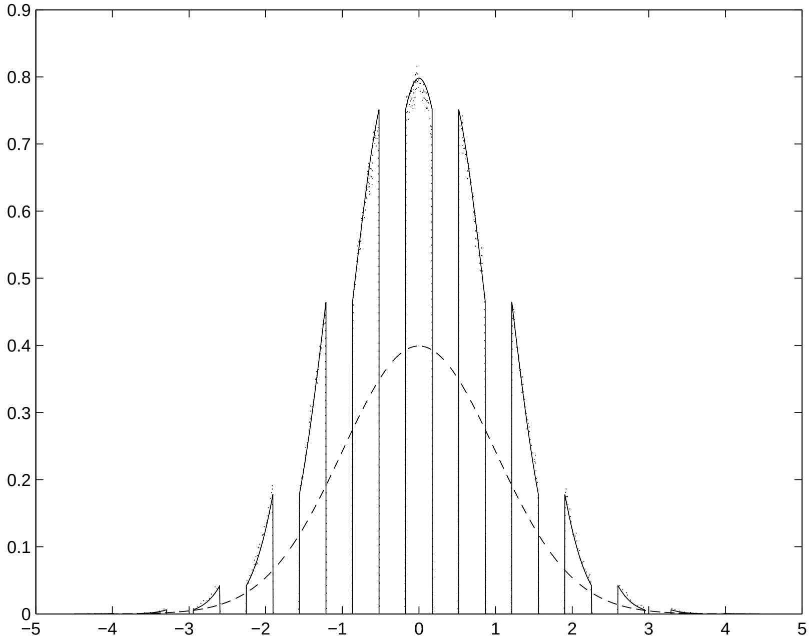

Figure 4.5: An illustration of $f_X(x)$ (dashed line), $f_{X+\frac{1}{\alpha}D}(x)$ (dotted line), and its approximation $\sum_{k=-\infty}^{+\infty} f_{(1-\alpha)\widetilde{X}+k\Delta}(x) I_{A_{k|W=0}}(x)$ (solid line). Chosen settings are $\widetilde{X} \sim \mathcal{N}(0,1)$, $\sigma^2_{N_1} = \sigma^2_{N_2} = \sigma^2_D = 0.01$, $\beta = 1$.

where $m \in (-\infty, \infty)$ is an integer.

Proof: The error probability P_e can be expressed as

$$\begin{aligned} P_e &= Pr[\hat{W} = 1|W = 0]Pr[W = 0] + Pr[\hat{W} = 0|W = 1]Pr[W = 1] \\ &= Pr[\hat{W} = 1|W = 0], \end{aligned} \tag{4.14}$$

where the last line follows from the fact that the encoder is a symmetric scheme of two quantizers, that the channel strategy is independent of the embedded message, i.e., $f_{N_2|W}(n_2|w) = f_{N_2}(n_2)$, and that $Pr[W = 0] + Pr[W = 1] = 1$. Therefore we can model the whole watermarking system, together with the attack channel, as a Binary Symmetric Channel with crossover probability P_e.

From (4.14) and (4.6), it is straightforward to show that the probability of error when D is *not added* at the decoder can be written as

$$\begin{aligned} P_e &= Pr[\hat{W} = 1|W = 0] \\ &= Pr\Big[\big|Q(\hat{U} - D) - (\hat{U} - D)\big| \geq \frac{\Delta}{4}\Big]. \end{aligned} \tag{4.15}$$

Observe that for any X and scalar quantizer $Q(\cdot)$ with step size Δ, we can write the relation

$$\big|Q(X) - X\big| \;=\; \big|(X + \frac{\Delta}{2}) \bmod \Delta - \frac{\Delta}{2}\big|. \tag{4.16}$$

Using (4.16) in (4.15) we have

$$\begin{aligned} P_e &= Pr\Big[\big|(\hat{U} - D + \frac{\Delta}{2}) \bmod \Delta - \frac{\Delta}{2}\big| \geq \frac{\Delta}{4}\Big] && (4.17)\\ &= Pr\Big[(\hat{U} - D + \frac{\Delta}{2}) \bmod \Delta \leq \frac{\Delta}{4} \bigcup (\hat{U} - D + \frac{\Delta}{2}) \bmod \Delta \geq \frac{3\Delta}{4}\Big] && (4.18)\end{aligned}$$

where $\bigcup$ denotes the union of two events.

Using (4.5) and taking into account that $U \in \Lambda$, the quantizer lattice, we can write

$$\begin{aligned} P_e &= Pr[((1-\alpha)N_1 + \frac{\alpha}{\beta}N_2 - D + \frac{\Delta}{2}) \bmod \Delta \leq \frac{\Delta}{4} \\ \bigcup & ((1-\alpha)N_1 + \frac{\alpha}{\beta}N_2 - D + \frac{\Delta}{2}) \bmod \Delta \geq \frac{3\Delta}{4}].\end{aligned}$$

Using Number theory [95], we can write that for any b, and any c such that $b > c > 0$, and any $a \neq mb$, where $m \in (-\infty, +\infty)$ is an integer, the solution to the inequalities

$$\begin{aligned} a \bmod b &\geq c && (4.19)\\ a \bmod b &\leq c && (4.20)\end{aligned}$$

is

$$\begin{aligned} mb + c \leq a \leq (m+1)b \text{ and} && (4.21)\\ mb \leq a \leq mb + c \text{ , respectively} && (4.22)\end{aligned}$$

Therefore, after simple arithmetics, we arrive at (4.13).

We would like to choose the dither sequence statistics such that the error probability $Pe = 0.5$ for all choices of the attacker noise N_2. We state the following theorem.

Theorem 2: For the probability of error $P_e = 0.5$ it is sufficient to choose the dither uniformly distributed over the base quantization cell[3], i.e. $D \sim \mathcal{U}(0, \sigma^2_{N_1})$.

Proof: For notational simplicity we make the following substitution

$$Z = (1-\alpha)N_1 + \frac{\alpha}{\beta}N_2 - D \qquad (4.23)$$

We can express (4.13) in the following way:

$$P_e = \sum_m \int_{m\Delta - \frac{3\Delta}{4}}^{m\Delta - \frac{\Delta}{4}} f_Z(z)dz \qquad (4.24)$$

By definition, we can write

$$f_Z(z) = \int\int f_{Z|N_1,N_2}(z|n_1,n_2) f_{N_1,N_2}(n_1,n_2) dn_1 dn_2 \qquad (4.25)$$

[3] In [94], it was shown that this is sufficient for making the quantization noise independent of the input signal.

Substituting with (4.25) in (4.24) we get

$$\begin{aligned} P_e &= \sum_m \int\limits_{m\Delta-\frac{3\Delta}{4}}^{m\Delta-\frac{\Delta}{4}} \int\int f_{Z|N_1,N_2}(z|n_1,n_2) f_{N_1,N_2}(n_1,n_2) dn_1 dn_2 dz \\ &= \int\int \sum_m \int\limits_{m\Delta-\frac{3\Delta}{4}}^{m\Delta-\frac{\Delta}{4}} f_{Z|N_1,N_2}(z|n_1,n_2) dz f_{N_1,N_2}(n_1,n_2) dn_1 dn_2, \end{aligned} \tag{4.26}$$

where in the second equality we interchanged the order of integration and summation.

From (4.23) we can write

$$f_{Z|N_1,N_2}(z|n_1,n_2) \;=\; f_D\big((1-\alpha)n_1 + \frac{\alpha}{\beta}n_2 - z\big) \tag{4.27}$$

Therefore, P_e can be written as

$$P_e \;=\; \int\int \sum_m \int\limits_{m\Delta-\frac{3\Delta}{4}}^{m\Delta-\frac{\Delta}{4}} f_D\big((1-\alpha)n_1 + \frac{\alpha}{\beta}n_2 - z\big) dz f_{N_1,N_2}(n_1,n_2) dn_1 dn_2 \tag{4.28}$$

From (4.27) we see that the term $(1-\alpha)n_1 + \frac{\alpha}{\beta}n_2$ affects only the mean of $f_D(-z)$. If we choose the dither to be uniform over the base quantization cell, i.e., $D \sim \mathcal{U}(0, \sigma^2_{N_1})$, then we can show that (see Fig.4.6)

$$\sum_m \int\limits_{m\Delta-\frac{3\Delta}{4}}^{m\Delta-\frac{\Delta}{4}} f_D\big((1-\alpha)n_1 + \frac{\alpha}{\beta}n_2 - z\big) dz \;=\; 0.5 \tag{4.29}$$

Therefore,

$$P_e \;=\; \int\int 0.5 f_{N_1,N_2}(n_1,n_2) dn_1 dn_2 = 0.5 \tag{4.30}$$

Note that in the proof of the theorems, we don't need the assumption $\sigma^2_{\tilde{X}} \gg \sigma^2_D$, and therefore the high resolution quantization assumption [89] is not necessary for the system security. However, we will assume the low distortion case $\sigma^2_{\tilde{X}} \gg \sigma^2_{N_1} = \sigma^2_D$ because of the approximation assumptions in the previous section.

Using Fig. 4.6 it can also be shown that $P_e \leq \frac{2}{3}$, with equality when $\sigma^2_D = \frac{9}{4}\sigma^2_{N_1}$ and $\sigma^2_{N_2} = 0$.

Experimental curves for the probability of error P_e as a function of σ^2_D for different values of $\sigma^2_{N_2}$ are shown in Fig.4.7. It can be seen that $P_e = 0.5$ independently of $\sigma^2_{N_2}$, as long as $D \sim \mathcal{U}(0, \sigma^2_{N_1})$.

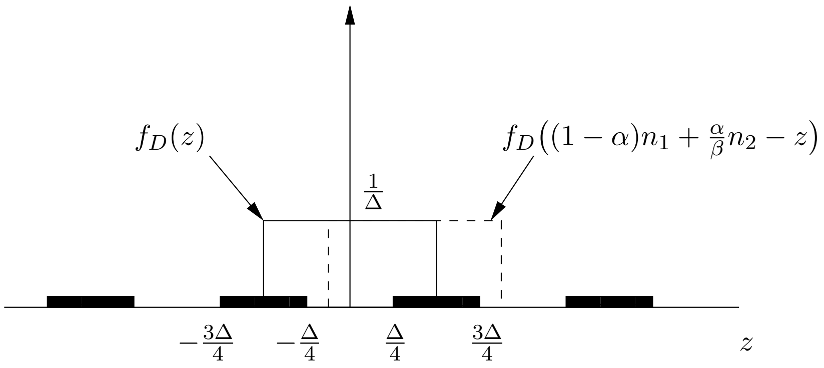

Figure 4.6: An illustration of $\sum_m \int_{m\Delta-\frac{3\Delta}{4}}^{m\Delta-\frac{\Delta}{4}} f_{Z|N_1,N_2}(z|n_1,n_2)dz$ for $D\sim\mathcal{U}(0,\sigma^2_{N_1})$.

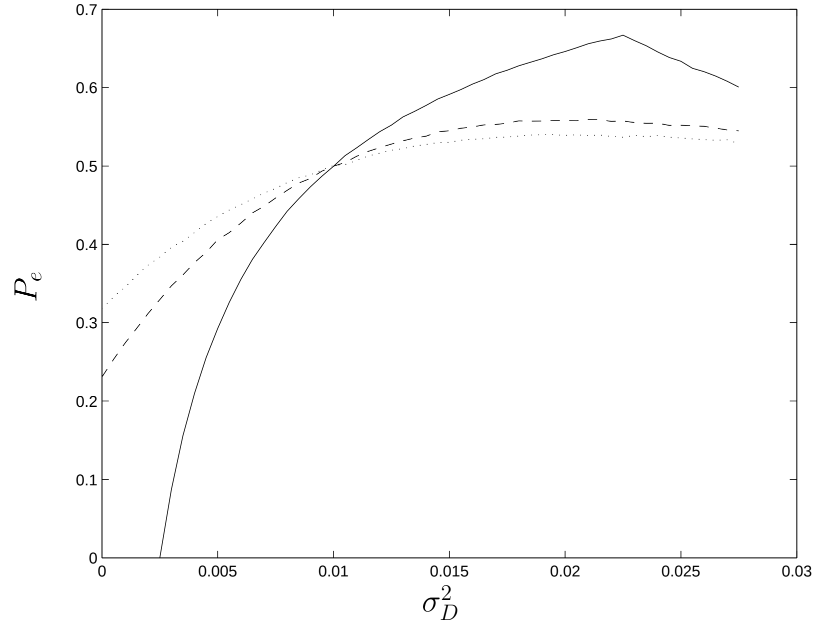

Figure 4.7: Experimental curves for P_e as a function of σ_D^2 for different values of $\sigma_{N_2}^2$. The solid curve is for $\sigma_{N_2}^2 = 0$, the dashed curve is for $\sigma_{N_2}^2 = 0.01$, and the dotted curve is for $\sigma_{N_2}^2 = 0.02$. Chosen settings are $\widetilde{X}\sim\mathcal{N}(0,1)$, $D\sim\mathcal{U}(0,\sigma_D^2)$, $N_2\sim\mathcal{N}(0,\sigma_{N_2}^2)$, $\sigma_{N_1}^2 = 0.01$, and $\beta = 1$.

4.5 Maximum Likelihood Estimation

The PDF models of the watermarked and attacked data have been derived as a function of β in the previous sections. We are now able to use these models to estimate β from the observed signal Y.

We assume that the host signal and attack channel noise are i.i.d. vector sources, i.e., we consider all signals to be n dimensional vectors with i.i.d. components. The ML estimation of β is done based on the following relation:

$$f_{Y+\frac{\beta}{\alpha}D}(y+\frac{\beta}{\alpha}d) = f_{\beta X+\frac{\beta}{\alpha}D}(\beta x+\frac{\beta}{\alpha}d) * f_{N_2}(n_2) \tag{4.31}$$

Representing $f_{Y+\frac{\beta}{\alpha}D}(y+\frac{\beta}{\alpha}d)$ as a joint distribution, the ML estimation $\hat{\beta}$ of the parameter β [27] is given as:

$$\begin{aligned}\hat{\beta} &= \arg\max_{\beta} f_{Y_1+\frac{\beta}{\alpha}D_1,...,Y_n+\frac{\beta}{\alpha}D_n}(y_1+\frac{\beta}{\alpha}d_1,...,y_n+\frac{\beta}{\alpha}d_n)\\ &= \arg\max_{\beta} f_{Y_1+\frac{\beta}{\alpha}D_1}(y_1+\frac{\beta}{\alpha}d_1)...f_{Y_n+\frac{\beta}{\alpha}D_n}(y_n+\frac{\beta}{\alpha}d_n)\\ &= \arg\max_{\beta} \sum_i \log f_{Y_i+\frac{\beta}{\alpha}D_i}(y_i+\frac{\beta}{\alpha}d_i). \end{aligned} \tag{4.32}$$

Here the second line follows from the assumption that the received data consists of n i.i.d. samples, and therefore the joint PDF can be written as a product of the "n" marginal PDFs. The last line follows from the monotonicity of the logarithm.

Experimental curves for the maximum likelihood function (LF), which is the expression $\sum_i \log f_{Y_i+\frac{\beta}{\alpha}D_i}(y_i+\frac{\beta}{\alpha}d_i)$, are shown in Fig. 4.8 and Fig. 4.9.

4.6 Experimental Results

In this section we describe experiments carried out to test the estimation accuracy of the proposed technique in terms of WNR and the number of available signal samples n. In principle one aims at developing estimation techniques that require small amount of data, so that they can be applied in situations where the estimating parameter slowly varies. Since it is difficult to further manipulate (4.11) (even for Gaussian sources) because of the indicator function in $f_{\beta X+\frac{\beta}{\alpha}D}(\beta x+\frac{\beta}{\alpha}d)$, we do brute force search for the optimal β.

4.6.1 Synthetic Host Signals

Here we perform experiments with synthetic host signals. We assume that the estimator has perfect knowledge of the host signal variance. In Fig. 4.10 we present results for $\hat{\beta}$ as a function of WNR. It can be seen that for $WNR > -7db$, the mean of $\hat{\beta}$ is very close to the true value of β, and the standard deviation of $\hat{\beta}$ is always smaller than 1%. In Fig. 4.11 we present results for $\hat{\beta}$ as a function of number of signal samples. It can be seen that around 100 signal samples are needed for reliable estimation of β. Results of $\beta - \hat{\beta}$ as a function of β are presented in Fig. 4.12. We can see that the standard deviation of $\beta - \hat{\beta}$ is smaller than 1% for $\beta > 0.75$.

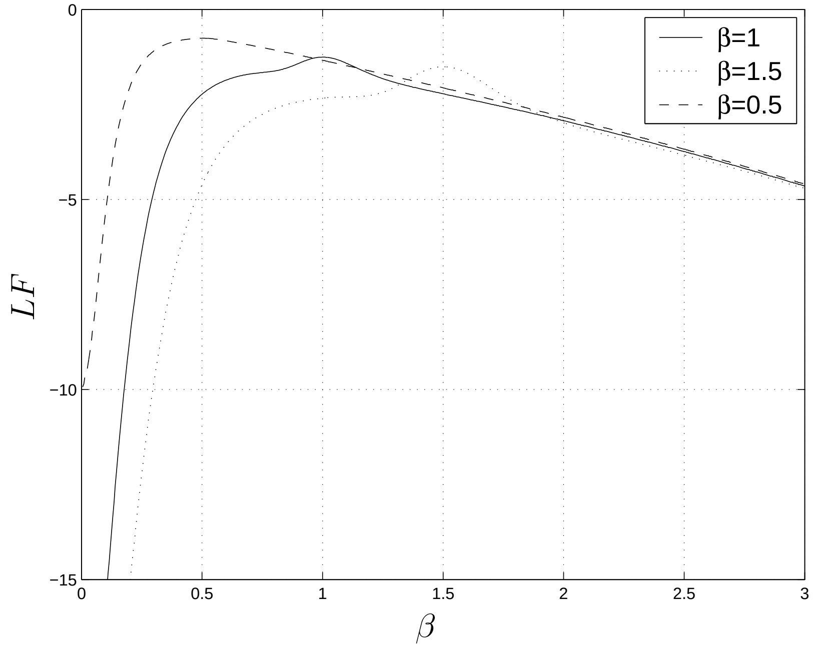

Figure 4.8: Graph of LF for different values of β. Chosen settings are $\widetilde{X} \sim \mathcal{N}(0,1)$, $D \sim \mathcal{U}(0,0.01)$, $N_2 \sim \mathcal{N}(0,0.01)$, and $\sigma^2_{N_1} = 0.01$.

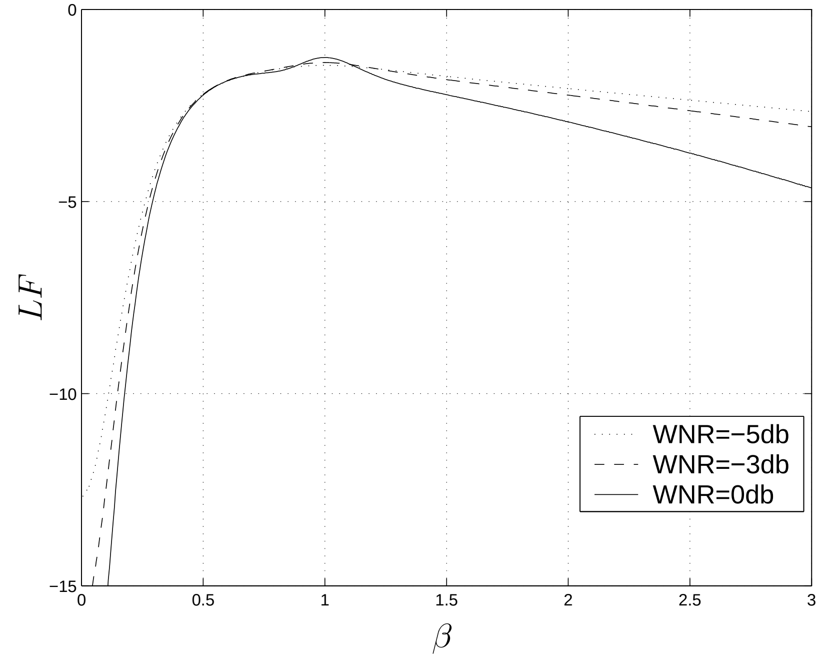

Figure 4.9: Graph of LF for different values of WNR. Chosen settings are $\widetilde{X} \sim \mathcal{N}(0,1)$, $D \sim \mathcal{U}(0,0.01)$, $\sigma^2_{N_1} = 0.01$, and $\beta = 1$.

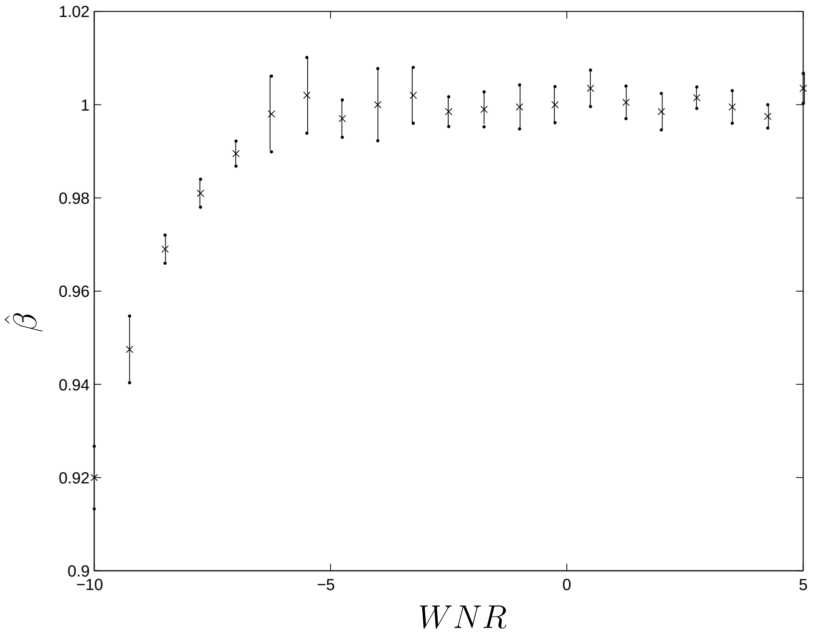

Figure 4.10: Graphs of $\hat{\beta}$ as a function of WNR. The crosses represent the mean, and the lines the standard deviation in both directions. The chosen settings are $\widetilde{X} \sim \mathcal{N}(0,1)$, $DWR = 20db$, $\beta = 1$, $n = 350000$.

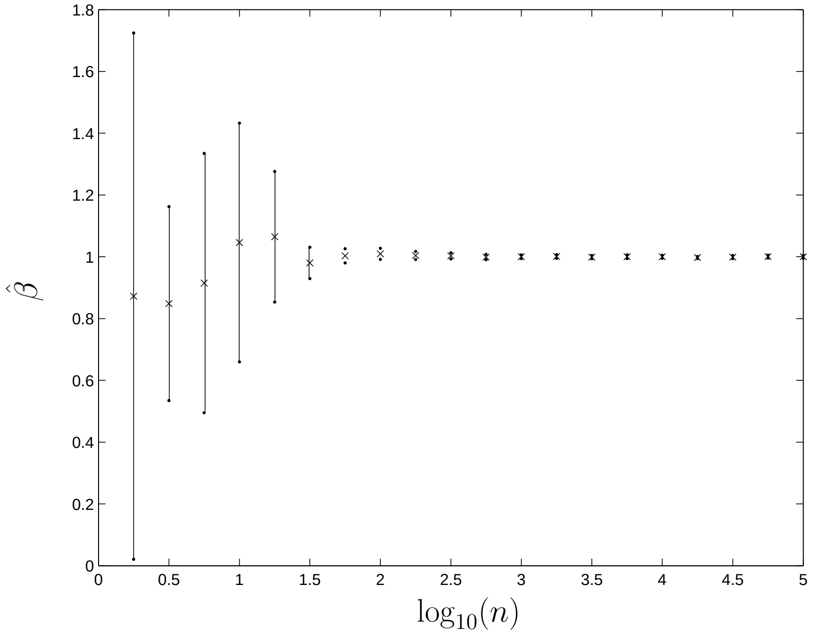

Figure 4.11: Graphs of $\hat{\beta}$ for synthetic host signals as a function of the number of signal samples n. The crosses represent the mean, and the lines the standard deviation in both directions. The chosen settings are $\widetilde{X} \sim \mathcal{N}(0,1)$, $DWR = 20db$, $WNR = 0db$, $\beta = 1$.

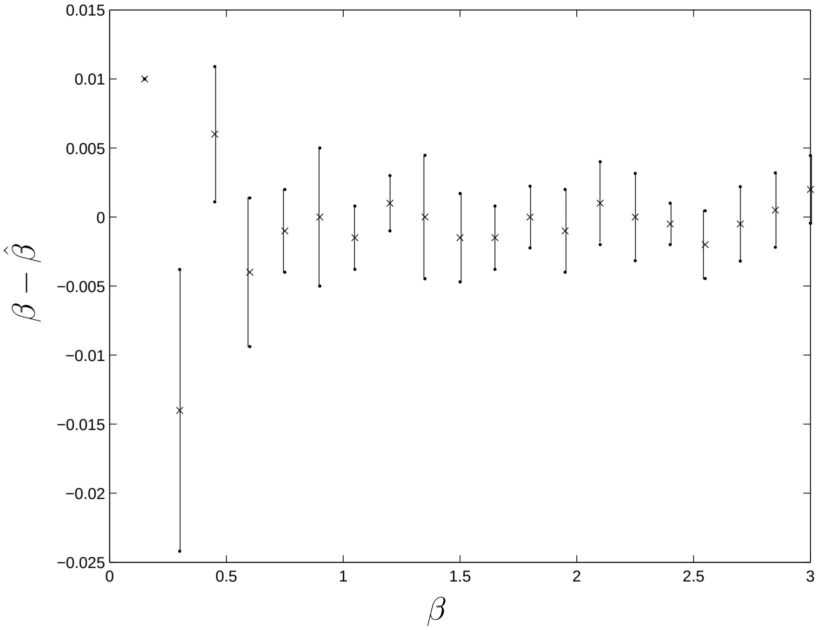

Figure 4.12: Graphs of $\beta - \hat{\beta}$ as a function of β. The crosses represent the mean, and the lines the standard deviation in both directions. The chosen settings are $\widetilde{X} \sim \mathcal{N}(0,1)$, $DWR = 20db$, $WNR = 0db$, $n = 350000$.

4.6.2 Real Host Signals

In this subsection we describe experiments with real audio signals (audio and speech with sampling frequency 48kHz). We choose more realistic settings than in the case of synthetic hosts, in which the estimator does not have a perfect knowledge of the host signal variance. The assumed PDF model of the host signal at the detection side is a zero-mean Laplacian PDF with variance equal to the variance of the received signal, i.e. $\widetilde{X} \sim \mathcal{L}\big(0, \beta^2(\sigma^2_{\widetilde{X}} + \sigma^2_{N_1}) + \sigma^2_{N_2}\big)$. This is a realistic assumption, because the decoder has access to the received data and can estimate its variance. Furthermore, in practice most audio signals have a marginal PDF that resembles the Laplacian PDF [96]. Experimental results in terms of WNR are shown in Fig. 4.13. It can be seen that the standard deviation of $\hat{\beta}$ is smaller than 1% for $WNR > -5db$. Experimental results of $\hat{\beta}$ as a function of number of signal samples are shown in Fig.4.14. It can be seen that reliable estimation of β is possible for $n > 31000$ samples. In Fig. 4.15 we plot experimental results of $\beta - \hat{\beta}$ as a function of β for different audio signals. It can be seen that the standard deviation of $\beta - \hat{\beta}$ is smaller than 1% for $\beta > 0.75$.

The experimental results with real signals are generally worse than in the case of synthetic signals. There are several reasons for that. First, the experimental settings are different. For real signals the estimator has access only to the received signal. The variance of the received signal differs from the variance of the host signal and the difference is especially pronounced when β deviates from 1. This causes a difference between the PDF of the watermarked data and the PDF assumed by the estimator. Secondly, real signals are non-stationary and exhibit correlation between the samples, which is not captured by our PDF models.

The ML estimation procedure is computationally very expensive, because of the brute force searching for the optimal β. The paper [97] treats the problem of jointly estimating β and $\sigma^2_{N_2}$ by transforming the attack channel into one that is equivalent but computationally less expensive for the ML approach processing chain. However, this transform does not improve the estimation.

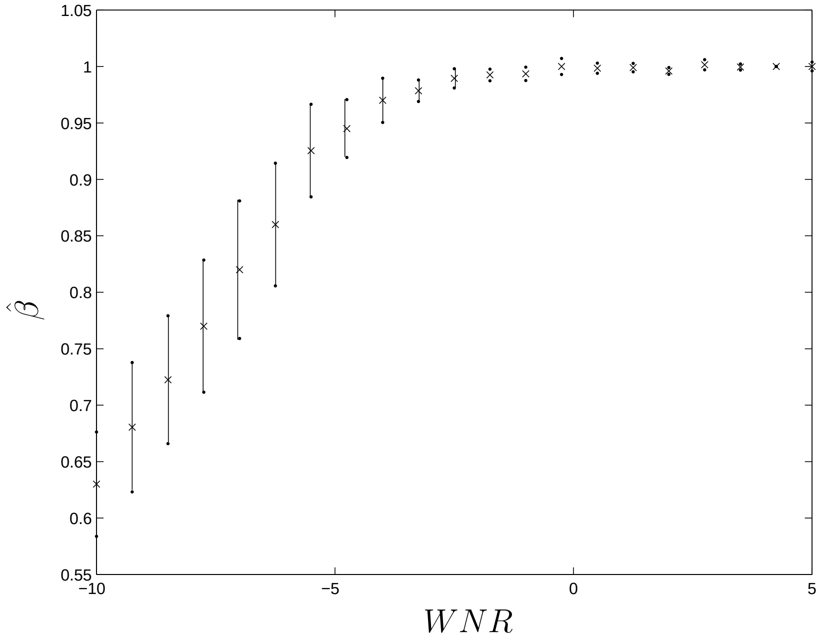

Figure 4.13: Graphs of $\hat{\beta}$ for real audio signals as a function of WNR. The crosses represent the estimation mean, and the lines the standard deviation in both directions. The chosen settings are $DWR = 20db$, $\beta = 1$, $n = 350000 \ldots 500000$.

4.7 Joint Estimation of Attack Parameters

In this section we propose a maximum likelihood (ML) approach to the estimation of a linear scale factor and the variance of the attacker's additive noise. The approach is based on the ML scale estimation procedure developed in the previous section. We extend this procedure to include estimation of the attacker's noise variance by transforming the attack channel into one that is less computationally expensive for the ML procedure. We apply the resulting estimation procedure on attacked watermarked images.

The model we use for the attack on the watermarked data is given by:

$$\begin{aligned} Y(i,j) &= \beta X(i,j) + N_2'(i,j) & (4.33) \\ &= \beta(X(i,j) + N_2(i,j)), & (4.34) \end{aligned}$$

where $Y(i,j)$ is the received watermarked and attacked image, $N_2(i,j)$ is the attackers i.i.d. (zero-mean Gaussian) noise with variance $\sigma^2_{N_2}$, and β is the amplitude scaling factor. The model (4.34) assumes that scaling is applied after adding the attacker's noise $N_2(i,j)$. This is slightly different from the model commonly used (4.33), where scaling is applied *before* the attacker's noise is added [83]. Since, in our estimation procedure we estimate *both* the

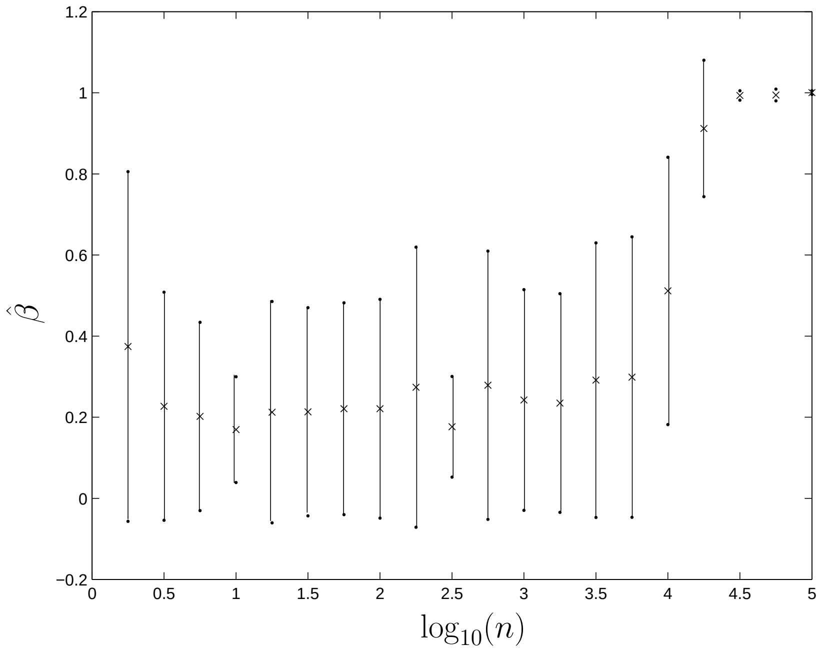

Figure 4.14: Graphs of $\hat{\beta}$ for real audio signals as a function of the number of signal samples n. The crosses represent the mean, and the lines the standard deviation in both directions. The chosen settings are $DWR = 20db$, $WNR = 0db$, $\beta = 1$.

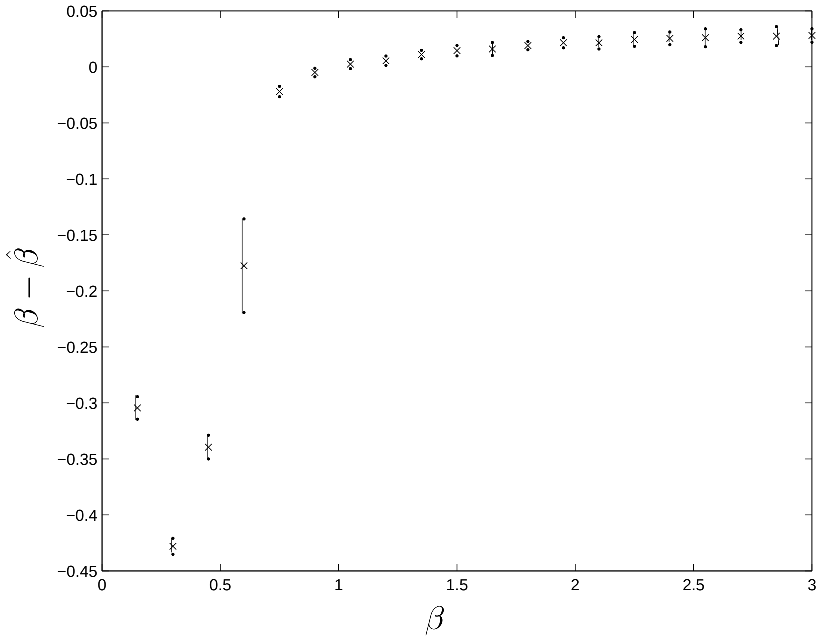

Figure 4.15: Graphs of $\beta - \hat{\beta}$ for real audio signals as a function of β. The crosses represent the mean, and the lines the standard deviation in both directions. The chosen settings are $DWR = 20db$, $WNR = 0db$, $n = 350000 \ldots 500000$.

scaling factor β and the noise variance $\sigma^2_{N_2}$, we can interchange noise addition and scaling without loss of generality. As we will see later on in this section, such model has significant modeling and computational advantages.

The watermarked data is corrupted by the attacker's noise. The resulting PDF of $\widetilde{Y}(i,j) = X(i,j) + N_2(i,j)$ is given by

$$f_{\widetilde{Y}}(x;\sigma^2_{N_2}) \;=\; f_X(x) * f_{N_2}(x;\sigma^2_{N_2}), \tag{4.35}$$

where $f_{N_2}(x;\sigma^2_{N_2})$ is the PDF of the attackers noise.

Finally, the PDF of the scaled version $Y(i,j) = \beta\widetilde{Y}(i,j)$ is given by

$$f_Y(x;\beta;\sigma^2_{N_2}) \;=\; \frac{1}{\beta} f_{\widetilde{Y}}\Big(\frac{x}{\beta};\sigma^2_{N_2}\Big). \tag{4.36}$$

Note that in the above PDFs, we explicitly indicate the dependency on the attacker's parameters, namely the amount of additive noise $\sigma^2_{N_2}$ and the amplitude scaling β. In this section we again assume that the host data $\widetilde{X}(i,j)$ and attacker noise $N_2(i,j)$ can be regarded as i.i.d. processes. Hence the joint PDF of the image $\mathbf{Y} = \{Y(i,j), 0 \le i \le M_1 - 1, 0 \le j \le M_2 - 1\}$ is equal to the product of the marginal PDFs:

$$f_{\mathbf{Y}}(\mathbf{x};\beta,\sigma^2_{N_2}) \;=\; \prod_{i,j} \frac{1}{\beta} f_{\widetilde{Y}(i,j)}\Big(\frac{x}{\beta};\sigma^2_{N_2}\Big). \tag{4.37}$$

Using the PDF $f_{\mathbf{Y}}(\mathbf{x};\beta,\sigma^2_{N_2})$ in (4.37) we can formulate the maximum likelihood estimator of the unknown parameters $\sigma^2_{N_2}$ and β as:

$$\begin{aligned}
(\hat{\beta},\hat{\sigma}^2_{N_2}) &= \arg\max_{\beta,\sigma^2_{N_2}} LF(\beta,\sigma^2_{N_2}) \\
&= \arg\max_{\beta,\sigma^2_{N_2}} \log f_{\mathbf{Y}}(\mathbf{x};\beta,\sigma^2_{N_2}) \\
&= \arg\max_{\beta,\sigma^2_{N_2}} \left\{ M_1 M_2 \log\Big(\frac{1}{\beta}\Big) + \sum_{i,j} \log f_{\widetilde{Y}(i,j)}\Big(\frac{x}{\beta};\sigma^2_{N_2}\Big) \right\}
\end{aligned} \tag{4.38}$$

The likelihood function $LF(\beta,\sigma^2_{N_2})$ can be evaluated for a given combination $(\beta,\sigma^2_{N_2})$. We remark that the actual evaluation of $\log f_{\widetilde{Y}}(\frac{x}{\beta};\sigma^2_{N_2})$ requires the PDF $f_{\widetilde{Y}}(x;\sigma^2_{N_2})$, which does *not* depend on β. In fact, into this PDF we substitute the *inversely scaled* (using the current estimate $\hat{\beta}$) amplitudes of the attacked watermarked image $Y(i,j)$. The efficient evaluation of the (rather complex expression of the) likelihood function is possible thanks to the model (4.34). In case the model (4.33) had been used, the expression for the likelihood function would be dependent on β in a more elaborate way, making efficient evaluation of $f_{\widetilde{Y}}(x;\sigma^2_{N_2})$ far more difficult.

Clearly, the sophistication of the optimization method used to maximize the likelihood function depends greatly on the behavior $LF(\beta,\sigma^2_{N_2})$. Fig. 4.16 illustrates the behavior of the likelihood function for $\beta \in [0.5, 1.5]$ and $\sigma^2_{N_2} \in [0.1, 10.0]$. In this case we have assumed that the host image $\widetilde{X}(i,j)$ is Gaussian distributed, $DWR = 30db$ (yielding $\Delta = 3.29$), $WNR = 3db$, $\alpha = 0.67$, and $\beta = 1.21$. The optimum of the likelihood value can be found close to the actual attacker's parameters, but we also observe that $LF(\beta, \sigma_{N_2}^2)$ consists of ridges with deep valleys in between. In fact, in case $f_{\widetilde{Y}}(x; \sigma_{N_2}^2) = 0$ for certain amplitudes, the likelihood function may become equal to negative infinity if the (inversely scaled) amplitudes of $Y(i,j)$ fall in these zero regions of the PDF of $\widetilde{Y}$. Such zero regions are more likely to occur for larger WNR, for which the behavior of $LF(\beta, \sigma_{N_2}^2)$ becomes more irregular and efficient numerical optimization procedures for (4.38) (such as gradient-based optimization) become less likely to be successful.

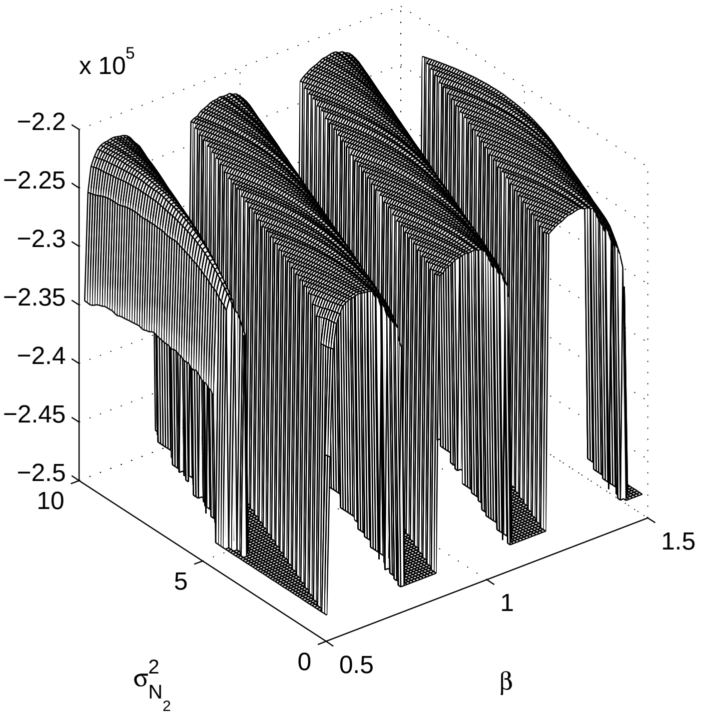

Figure 4.16: *Illustration of $LF(\beta, \sigma_{N_2}^2)$ without dithered quantization.*

In the above approach we have excluded dithering of the quantization process. To include the dither sequence $D(i,j)$ in the estimation of β and $\sigma_{N_2}^2$, we observe that:

- the input to the quantizer has a PDF that is different from the case when dither is not used. However, if the variance of the dither sequence is small compared to the variance of the host data, we can approximately ignore the effect of the dither on the PDF of $X(i,j)$. In our QIM-DC scheme, the dither is uniformly distributed in $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$,

- the subtracted dither $D(i,j)$ after the quantizer in the watermark embedding scheme can be compensated for in the parameter estimation process by simply re-adding $D(i,j)$ to the inversely scaled (using the current estimate of β) attacked watermarked image $Y(i,j)$. Hence, in (4.38) we simply replace the argument $\frac{x}{\beta} = \frac{x(i,j)}{\beta}$ by $\left(\frac{x(i,j)}{\beta} - D(i,j)\right)$. Again, this makes possible an efficient evaluation of the likelihood function.

Fig. 4.17 illustrates the behavior of the likelihood function under the same conditions as those in Fig. 4.16, but now taking into account dithered quantization. The maximum of the likelihood function can still be found in approximately the same location, and the behavior of the likelihood function itself has changed marginally. This confirms the validity of the assumption that we can safely ignore the effect of the dither on the PDF of $X(i,j)$ in the maximum likelihood parameter estimation.

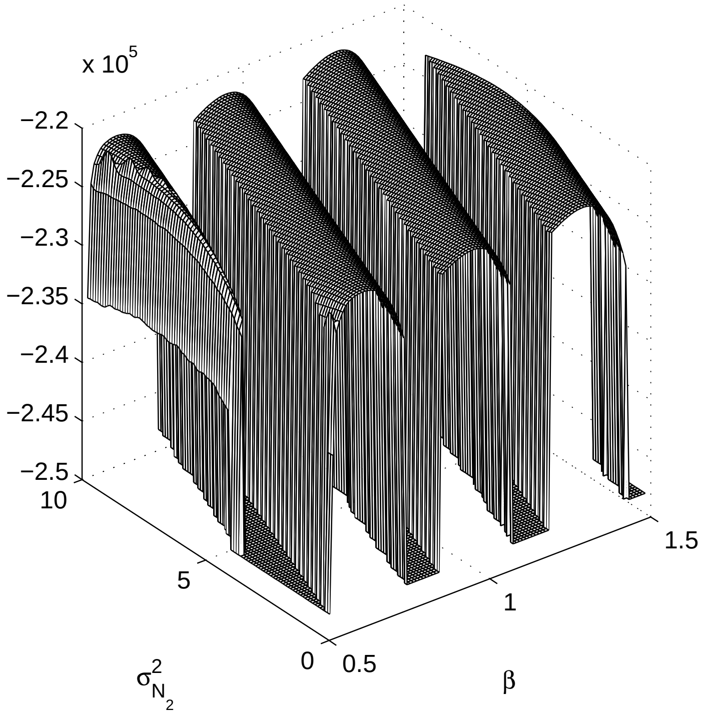

Figure 4.17: *Illustration of* $LF(\beta, \sigma^2_{N_2})$ *including dithered quantization.*

An effect that we see in both Fig. 4.16 and Fig. 4.17 is that the (correct) optimum of the likelihood function is relatively insensitive to the variance of the attacker's noise. This suggests that in a practical context we can limit the search of a proper value of $\sigma^2_{N_2}$ to a limited set of values.

We have applied the proposed scale and variance estimation procedure on synthetic Gaussian distributed images of size 256 × 256. The numbers obtained in this way give a performance ceiling, since the PDF of real images can obviously be modeled less accurately. Various embedding settings have been used to benchmark the estimation procedure. Table 4.1 lists our experimental results based on synthetic (Gaussian) data, with $\beta = 0.91$, $DWR = 30db$, and $WNR = 0, -10, -20db$.

Similar results are obtained for other values of β and DWR. Our results show that the value of β can be estimated much more reliably than $\sigma^2_{N_2}$. However, as we already remarked, the value of $\sigma^2_{N_2}$ seems not be important in finding the correct scaling factor β. Estimating β below a WNR of −10 to −20 db is useless, as the attacker's noise will effectively make

WNR	0db	-3db	-10db
β	0.91	0.91	0.91
$\sigma^2_{N_2}$	0.9	1.8	9.0
$\hat{\beta}$	0.91	0.91	0.90
β search resolution	0.01	0.01	0.01
variance $\hat{\beta}$	0.00	0.01	0.05
$\hat{\sigma}^2_{N_2}$	1.4	1.4	2.1
$\sigma^2_{N_2}$ search resolution	0.1	0.1	0.1
variance $\hat{\sigma}^2_{N_2}$	0.3	0.7	5.0

Table 4.1: Experimental results using synthetic data

the extraction of message bit very difficult (probability of error approaching 0.5).

4.8 Discussion

In this chapter we presented an ML amplitude scale estimation technique for quantization-based watermarking. We also incorporated subtractive dither into the watermarking system and gave conditions for the dither sequence to achieve a given level of security. The estimation approach needs small amount of signal samples for estimating reliably β in the case of synthetic host signals, but relatively large amount of signal samples in the case of real audio host signals. Experiments showed that the proposed approach performs well under realistic conditions.

We developed a computationally efficient estimation of the amplitude scaling factor and variance of the noise of the attack channel. The estimation procedure is not affected by the presence or absence of dither. The optimum of the likelihood function is found around the correct values of the parameters β and $\sigma^2_{N_2}$ for a wide range of watermark-to-noise ratios. A major disadvantage of the current ML approach is that the likelihood function shows a very irregular behavior for varying β and $\sigma^2_{N_2}$. For that reason, we have optimized $LF(\beta, \sigma^2_{N_2})$ using a full search of the parameter space.

Although the amplitude scale attack channel can model linear filtering attacks (see section 2.7.4), it is difficult to employ the methods developed in this and the previous chapter to combat the linear filtering attack. The difficulty is due to the large amount of noise introduced by the filtering operation that hampers the watermark. In order to develop techniques for combating the linear filtering attack, we first try to study intermediate cases and develop countermeasures based on the already developed estimation techniques. In the next chapter we study an intermediate scenario in which the amplitude scale operation is applied in the frequency domain and develop a procedure for estimating the scaling factors.

Chapter 5

Extension to Two-Band Amplitude Scale Attacks *

Watermarking schemes based on quantization theory have recently emerged as a result of information theoretic analysis [18, 20]. In terms of additive noise attacks, these schemes have proven to perform better than traditional spread spectrum watermarking because the used lattice codes achieve capacity for the AWGN channel. Another important feature of quantization-based watermarking schemes is that they can completely cancel the host signal interference, which makes them invariant to the host signal. A similar phenomenon exists in channel coding with side information at the encoder [42]. Unfortunately, quantization-based watermarking schemes such as Quantization Index Modulation watermarking with Distortion Compensation (QIM with DC) [20] are not robust against LTI filtering attacks. Considering the implementation of a quantization-based scheme in a LTI filtering setting, it is likely that the scheme will fail. Weakness against LTI filtering is a serious drawback, since many normal operations on images and audio are explicitly implemented with linear filters. The bass and treble adjustments in a stereo system apply simple filtering operations. In addition, many other operations, although not explicitly implemented with filters, can be modeled by them. For example, playback of audio over loudspeakers can also be approximated as a filtering operation.

In this chapter, we focus on multi-band amplitude scaling problem in combination with additive noise attack. One of its applications of which is a multi-band equalizer that modifies the spectrum of the signal using the filter bank. The signal frequency range is divided into a number of frequency bands and the signal may be amplified or attenuated in each of these bands independently. To see how serious the problem can be, Fig. 5.1 shows the behavior of QIM with DC for a variety of Document to Watermark ratio (DWR), when the watermarked signal is attacked by a two-band filter bank with a scaling in the high frequency band depicted in Fig. 5.2.

The solutions proposed so far to deal with one channel amplitude scaling attack, in the

[*] This chapter is published as J. Wang, I. D. Shterev, and R. L. Lagendijk "Scale Estimation in Two-Band Filter Attacks on QIM Watermarks, *SPIE Security, Steganography, and Watermarking of Multimedia Contents VIII*, San Jose, CA, January 2006, and part of J. Wang, I. D. Shterev, and R. L. Lagendijk "Two-Band Amplitude Scale Estimation for Quantization-Based Watermarking", *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, Hong Kong, December 2005.

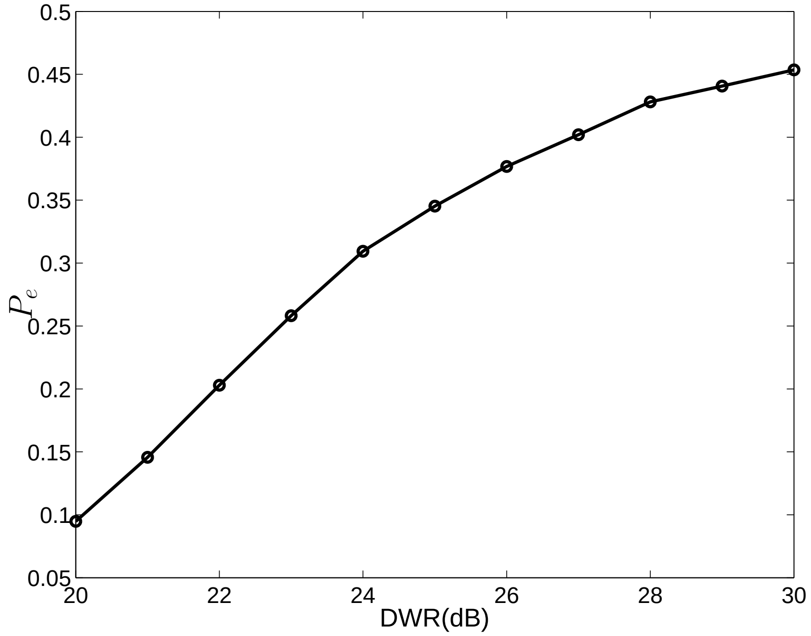

Figure 5.1: Probability of error for different values of DWR. $\beta = 0.95$, no noise.

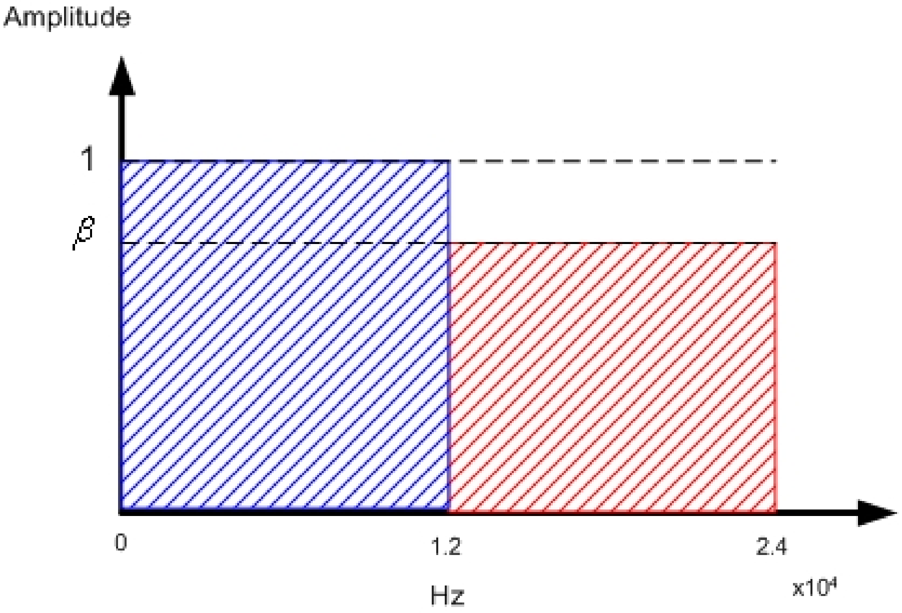

Figure 5.2: Filter transfer function.

framework of QIM watermarking, can be grouped into two main categories: One of the approaches is based on designing watermarking codes that are resilient to amplitude scaling operation, such as trellis codes [79, 98]. Another approach is based on estimation the amplitude scaling operation and inverting them prior to watermark decoding [89]. However, to the best of our knowledge, no earlier work with regard to multi-band amplitude scaling has been proposed before. The chapter is organized as follows: in Section 5.1 we formulate the multi-band amplitude scale attack and introduce some important notation. In Section 5.2 we derive the PDF models for frequency band amplitude scaled signal and attacked signal respectively. A description of the estimation procedure is given in Section 5.3. Section 5.4 contains experimental results from synthetic and real audio host signals, and Section 5.6 concludes the chapter.

5.1 Mathematical Formulation

In this section, we define some notational conventions. We assume that the host signal is arranged in an n-dimensional vector $\widetilde{\boldsymbol{X}}$, i.e., $\widetilde{\boldsymbol{x}} = \big(x(1), x(2), \ldots, x(n)\big)$, where $X(k)$ ($k \in 1, \ldots, n$) refers to the k-th element. Throughout the chapter, random variables are denoted by capital letters and their realizations by the respective small letters. Vectors will be denoted by bold letters. Fig. 5.3 and Fig. 5.4 show the watermark encoder together with the attack channel and the scale corrector together with the watermark decoder, respectively. The basic embedding and decoding procedures are based on QIM with DC, proposed by Chen and Wornell [20]. In the watermark encoder, where $W \in \{0, 1\}$ denotes the message bits that are embedded in the host data, $\widetilde{\boldsymbol{X}}$ is the host signal itself with a variance $\sigma^2_{\widetilde{X}}$, $\boldsymbol{X}$ is the watermarked signal.

The multi-band amplitude scaling attack consists of an analysis/synthesis filter bank and a constant scaling of the amplitude of the watermarked signal in each band. Furthermore, we will assume that zero-mean additive white Gaussian noise $\boldsymbol{N}_2$ with variance $\sigma^2_{N_2}$ and independent of the output of the filter attack $\boldsymbol{X}'$ is also added by the attacker. Let $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_M]$, where $\beta_i > 0$, for all i, denotes the Multi-band amplitude scaling factor vector, and M is the number of the frequency channel. Following our model, the Fourier transform of $\boldsymbol{X}'$ can be written as

$$\begin{aligned} X'(\omega) &= T(\omega)X(\omega) \\ &= \big(\beta_0 G_0(\omega)H_0(\omega) + \beta_1 G_1(\omega)H_1(\omega) + \ldots + \beta_M G_M(\omega)H_M(\omega)\big)X(\omega), \quad (5.1) \end{aligned}$$

where $G(\omega)$ and $H(\omega)$ are the transfer functions of the analysis and synthesis filters respectively.

Then the attacked vector $\boldsymbol{Y}$ is given as

$$\boldsymbol{Y} = \boldsymbol{X}' + \boldsymbol{N}_2. \quad (5.2)$$

Finally, it is useful to define some quantities that relate the powers of the host, the watermark and noise. The *Document-to-Watermark Ratio* (DWR) is given by $10 \log_{10} \frac{\sigma^2_{\widetilde{X}}}{\sigma^2_{N_1}}$; the *Watermark-to-Noise Ratio* (WNR) is $10 \log_{10} \frac{\sigma^2_{N_1}}{\sigma^2_{N_2}}$, where $\sigma^2_{N_1}$ is the variance of the watermark. These quantities are expressed in decibels.

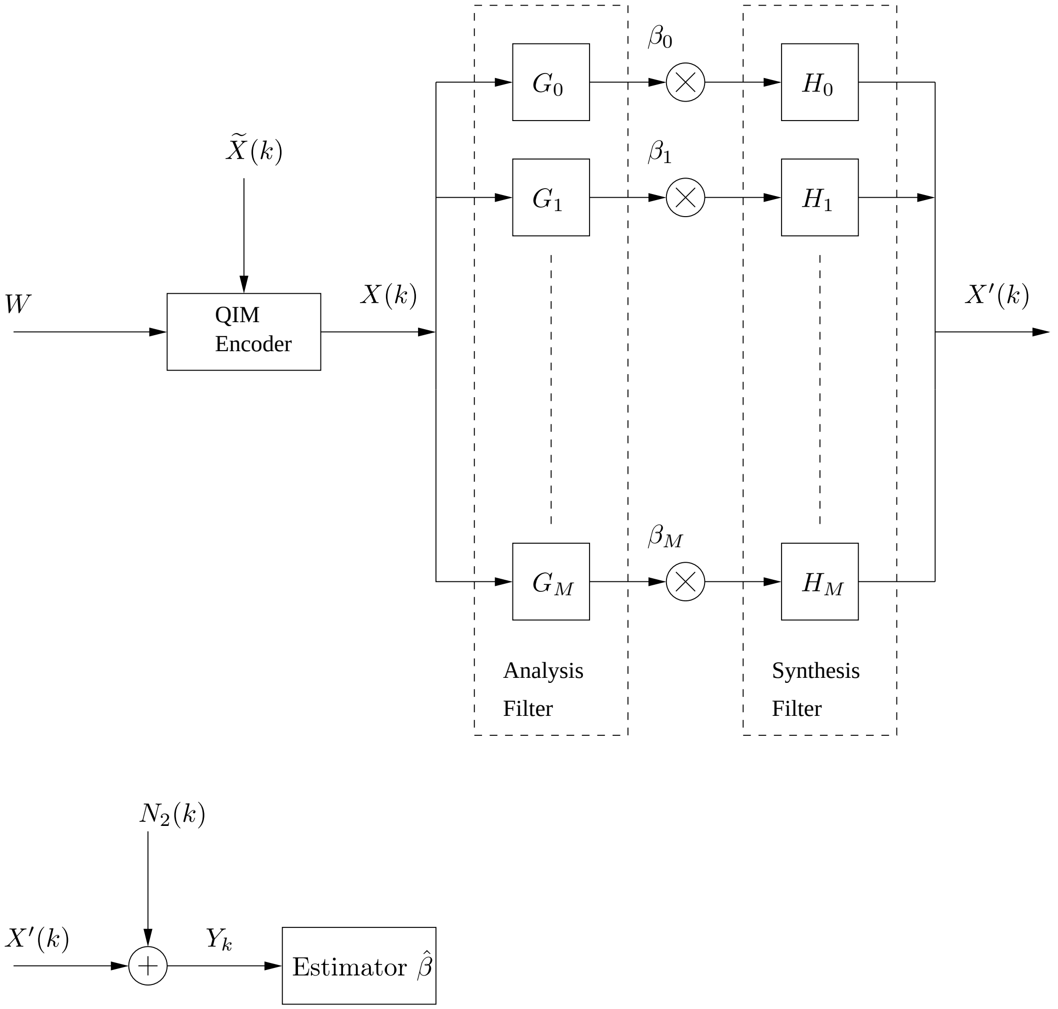

Figure 5.3: Block diagram of watermark encoder and attack channel.

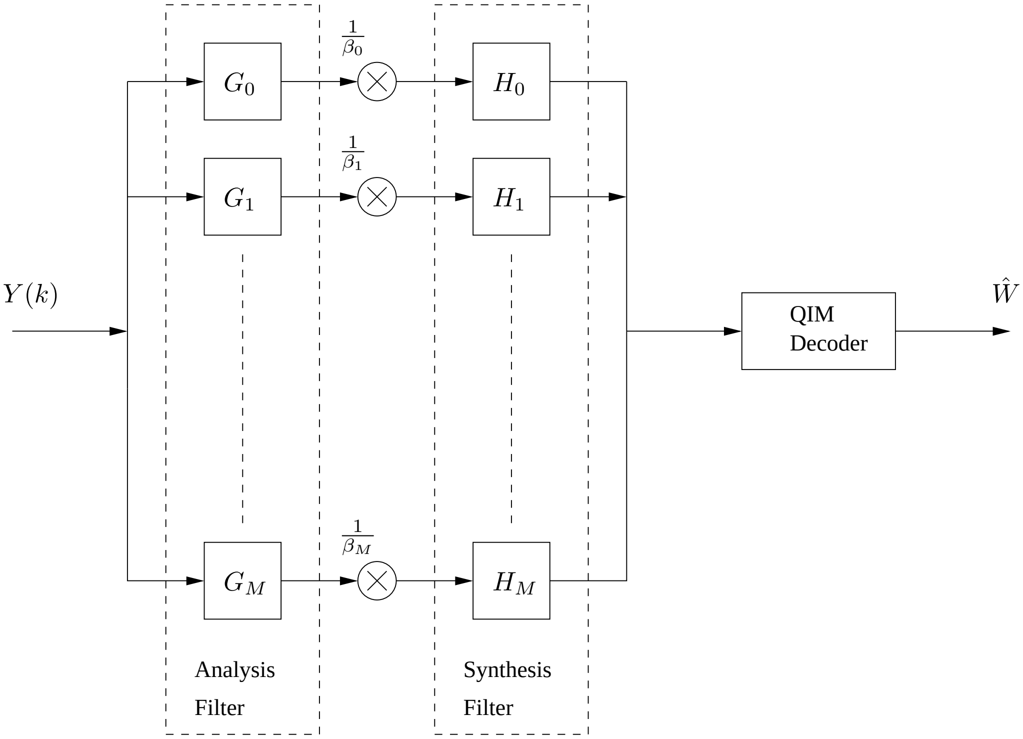

Figure 5.4: Block-diagram of scale corrector and watermark decoder.

5.2 PDF Models

In this section we derive the PDF models for frequency band amplitude scaled vector $\boldsymbol{X'}$ and attacked vector $\boldsymbol{Y}$ as a function of $\boldsymbol{\beta}$. These PDF models are the basis for the ML estimation procedures for estimating β developed in the next section.

Referring to Fig. 5.3 and Fig. 5.4, multi-band amplitude scaling attack in each frequency band consists of a twin LTI filters and a scaling factor β_i. Assume that the filter bank holds *Perfect Reconstruction* (PR) property, if the scaling factor $\boldsymbol{\beta} = \mathbf{1}$, and $\boldsymbol{N}_2 = 0$, we obtain:

$$X(k) \quad = \quad X'(k). \tag{5.3}$$

For $\boldsymbol{\beta} \neq \mathbf{1}$, (5.3) no longer holds; hense it leads to watermark detection error because the watermarked signal is moved away from the correct centroids. From (5.1), we can see that the transfer function $T(\omega)$ carries information about $\boldsymbol{\beta}$. Since our goal is to derive PDF of frequency band amplitude scaled vector $\boldsymbol{X'}$, it would be reasonable to use time domain representation of (5.1). Then $X'(k)$ can be written as

$$\begin{aligned} X'(k) &= t(k) * X(k) \\ &= t(0)X(k) + t(1)X(k-1) + t(2)X(k-2) + \ldots + t(k)X(0), \end{aligned} \tag{5.4}$$

where $*$ denotes convolution and $t(k)$ is the impulse response of $T(\omega)$. Note that the impulse response of the filters are known to the estimator.

We see that the overall filter operates by summing weighted delayed versions of the watermarked vector $\boldsymbol{X}$. In order to derive PDF of frequency band amplitude scaled vector

$\boldsymbol{X}'$, we assume that the host signal is an i.i.d. vector source. We note that this assumption is only an approximation for the real world case. Thus, the frequency band amplitude scaled vector sample $X'(k)$ is a weighted sum of i.i.d. random variables $X(k)$. In chapter 3, we have derived the PDF model for the watermarked data $\boldsymbol{X}$, i.e. $f_X(x)$. Then the PDF of $\boldsymbol{X}'$ is given as

$$f_{X'}(x') \quad = \quad \frac{1}{|t(0)|} f_X\Big(\frac{x}{t(0)}\Big) * \frac{1}{|t(1)|} f_X\Big(\frac{x}{t(1)}\Big) * \ldots * \frac{1}{|t(k)|} f_X\Big(\frac{x}{t(k)}\Big). \tag{5.5}$$

To simplify the multi-band amplitude scaling problem, we confine ourselves to use a simplified model, namely, a two-band filter bank, and the scaling factor only exists in the high frequency band, in other words, the scaling factor vector is $\boldsymbol{\beta} = [1, \beta]$.

Fig. 5.5 illustrates the statistical distribution of the output of the filter attack $\boldsymbol{X}'$, showing the sufficient accuracy in the predicted PDF model. For $\beta = 1$ the analytical PDF is that of the typical QIM watermarked signal.

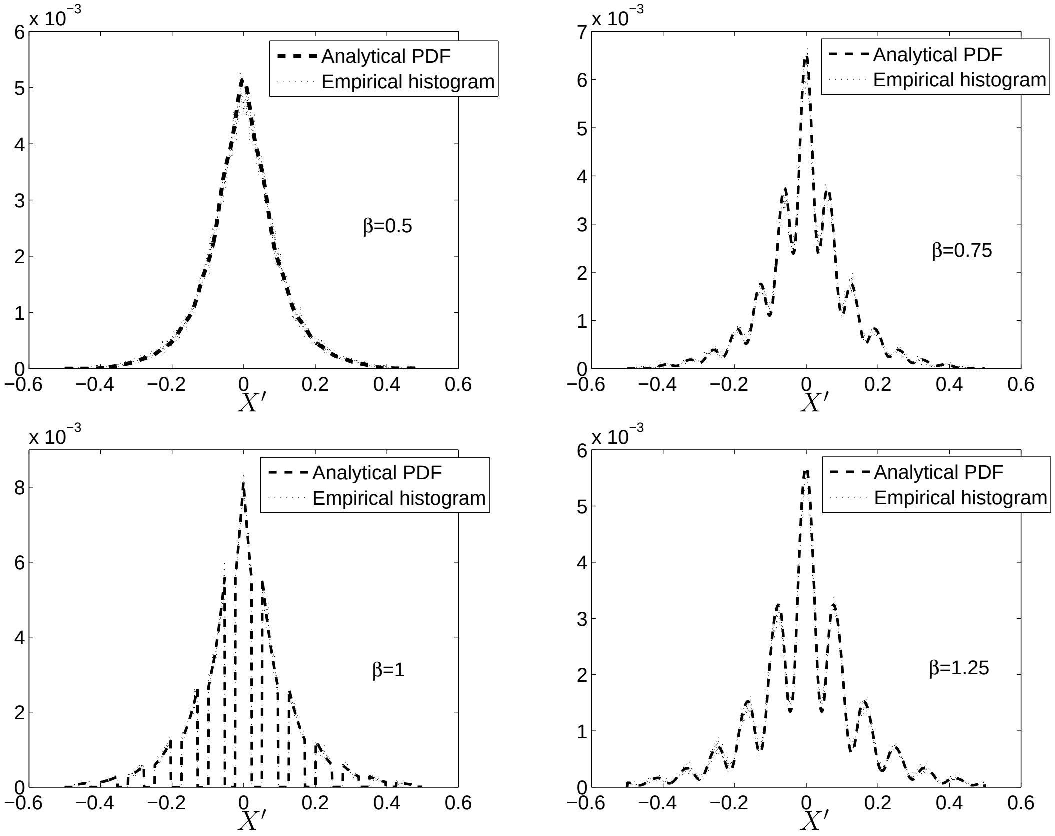

Figure 5.5: Analytical PDF for different β vs. empirical histogram for a Laplacian host, $DWR = 15db$. The filter transfer function is shown in Fig. 5.2.

In addition there are only several coefficients $t(k)$ which have relatively large magnitude. So it is reasonable to consider that these filter coefficients with larger magnitude play important role in (5.5). Therefore, $f_{X'}(x')$ can be simplified by substituting only a few filter coefficients with larger magnitude in (5.5), instead of using all filter coefficients. Let F

denote the necessary number of filter coefficients. Fig. 5.6 illustrates $f_{X'}(x')$ for different F.

From Fig. 5.6, we can see that in this case, $F = 3$ is sufficient for (5.5). For larger F, there is no evident improvement in the accuracy of the analytical PDF model, which verifies that (5.5) can be simplified by substituting only a few filter coefficients with larger magnitude.

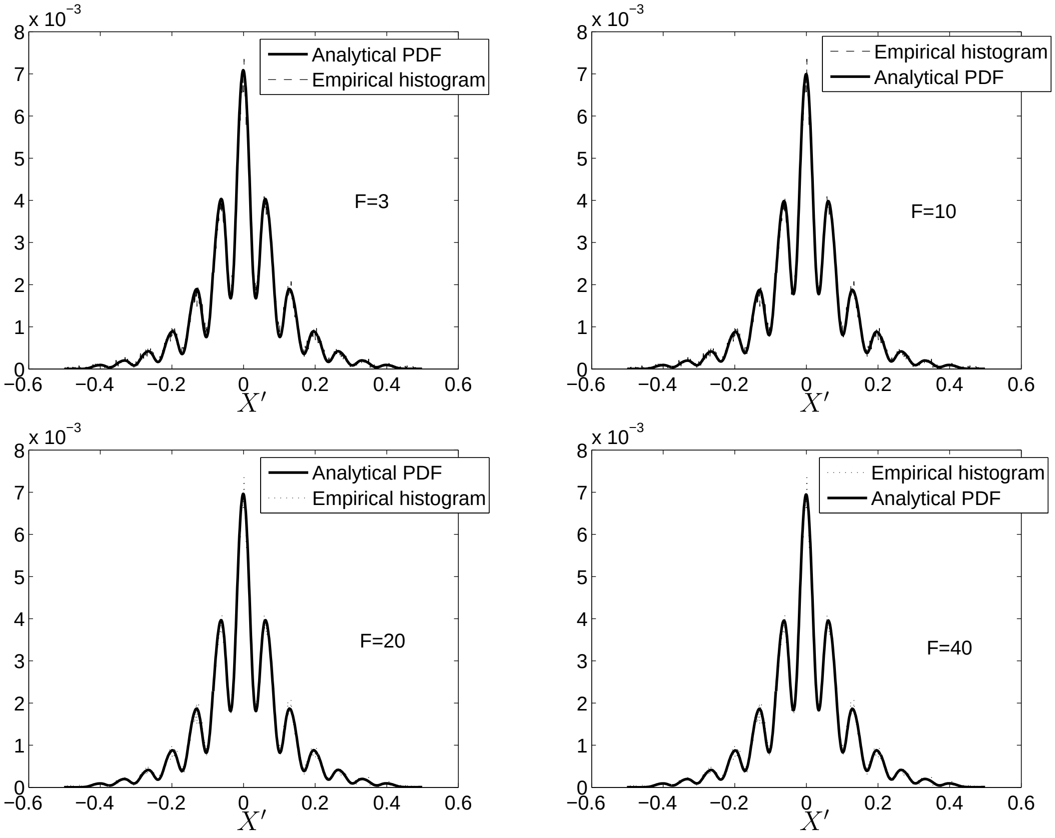

Figure 5.6: Analytical PDF for different F vs. empirical histogram for a Laplacian host, $\beta = 0.8$, $DWR = 15db$. The filter transfer function is shown in Fig. 5.2.

Taking into account the additive noise $\boldsymbol{N}_2$, we obtain the PDF of the attacked vector $\boldsymbol{Y}$:

$$f_Y(y) \quad = \quad f_{N_2}(n_2) * f_{X'}(x') \qquad (5.6)$$

where the convolution follows from the independence between $\boldsymbol{N}_2$ and $\boldsymbol{X'}$. $f_Y(y)$ is shown in Fig. 5.7. We see that the PDF model of the attacked vector matches the histogram quite well for additive noise case.

5.3 Maximum Likelihood Estimation

The PDF model of the attacked vector has been derived as a function of β in the previous section. We are now able to use the model to estimate β from the attacked vector $\boldsymbol{Y}$.

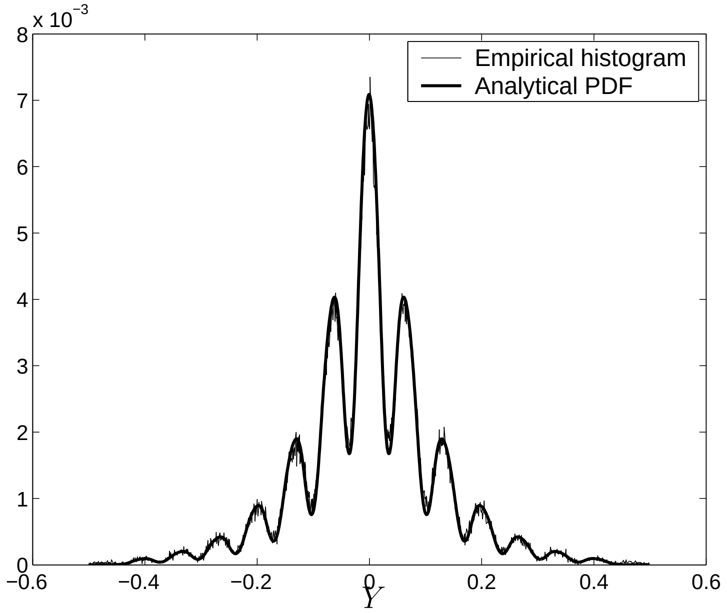

Figure 5.7: PDF of attacked vector Y vs. empirical histogram for Laplacian host, $\beta = 0.8$, $WNR = 3db$, $DWR = 15db$. The filter transfer function is shown in Fig. 5.2.

Maximum Likelihood (ML) estimation can be used to solve this problem. The ML estimation of β is done based on (5.6). By definition [27], the ML estimate $\hat{\beta}$ of the scaling factor β is given as:

$$\hat{\beta} \;=\; \arg\max_{\beta} f_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n|\beta). \tag{5.7}$$

However, it is difficult to derive the above joint PDF. Recall that for deriving (5.5), we have made an assumption that the frequency band amplitude scaled vector $\boldsymbol{X'}$ has i.i.d. components, so it is reasonable to consider that the vector $\boldsymbol{Y}$ will also have approximately i.i.d. components.

Therefore, the joint PDF can be approximately written as a product of the marginal PDFs, that is:

$$\hat{\beta} \;=\; \arg\max_{\beta} \prod_{i=1}^{n} f_{Y_i}(y_i|\beta) = \arg\max_{\beta} \sum_{i=1}^{n} \log f_{Y_i}(y_i|\beta) \tag{5.8}$$

The likelihood function is $\sum_i \log f_{Z_i}(z_i|\beta)$. Experimental curves of the LF for different values of β and WNR are shown in Fig. 5.8. Since it is difficult to find an analytical expression for $\hat{\beta}$, we do a brute force search for the optimal value of β based on (5.8).

5.4 Experiments

In this section we describe experiments with synthetic and real audio signals (with sampling frequency $48kH_z$) carried out to test the estimation accuracy of the proposed techniques in terms of WNR, the parameter β, and the number of available signal samples n. Furthermore,

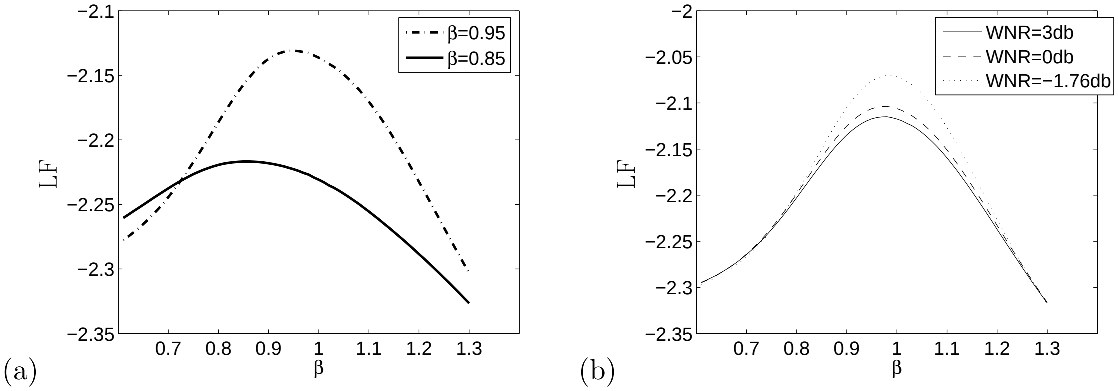

Figure 5.8: Graph of LF for different values of β (a) and different values of WNR (b). Chosen settings are $\widetilde{X} \sim \mathcal{L}(0, 0.02)$, $N_2 \sim \mathcal{N}(0, 0.01)$, and $\sigma^2_{N_1} = 0.01$. The filter transfer function is shown in Fig. 5.2.

we experimentally show how inverting the effect of the attack can significantly help to reduce the bit error rate.

Experimental results in terms of WNR and n are shown in Fig.5.9. The assumed PDF model of the host signal at the estimator side is a zero-mean Laplacian PDF with variance equal to the sum of the variances of the host signal, watermark, and the noise in the attack channel, i.e. $\mathcal{L}(0, \sigma^2_{\widetilde{X}} + \sigma^2_{N_1} + \sigma^2_{N_2})$. This is a realistic assumption, because the decoder has access to the received data and can estimate its variance. Furthermore, in practice most audio signals have a PDF that resembles the Laplacian PDF. The loss in performance of the ML approach is due to the approximation in $f_Y(y)$ and the fact that generally, ML estimation requires a large sample size [27]. In Fig. 5.10 we plot experimental results of $\beta - \hat{\beta}$ as a function of β for different audio signals.

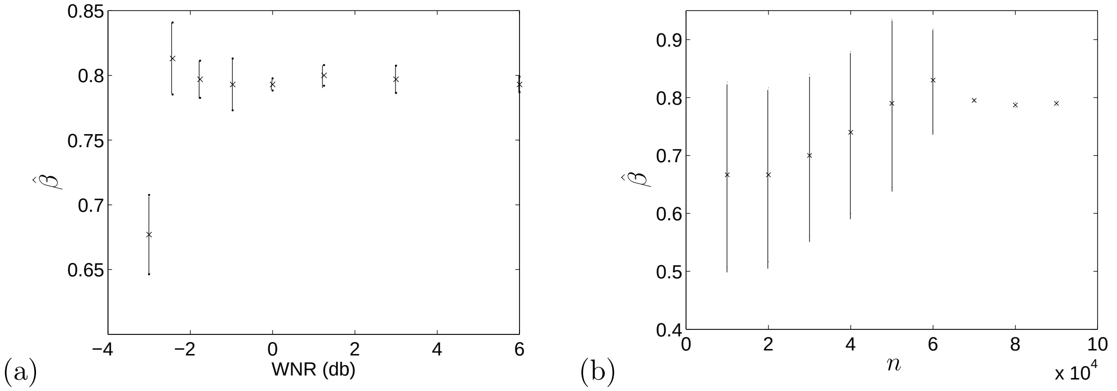

Figure 5.9: Graphs of $\hat{\beta}$ for real audio signals as a function of WNR (a) and as a function of available signal samples n (b). The crosses represent the estimation mean, and the lines the estimation standard deviation in both directions. $DWR = 15db$. The assumption for the estimator is $\widetilde{X} \sim \mathcal{L}(0, \sigma^2_{\widetilde{X}} + \sigma^2_{N_1} + \sigma^2_{N_2})$. The filter transfer function is shown in Fig. 5.2.

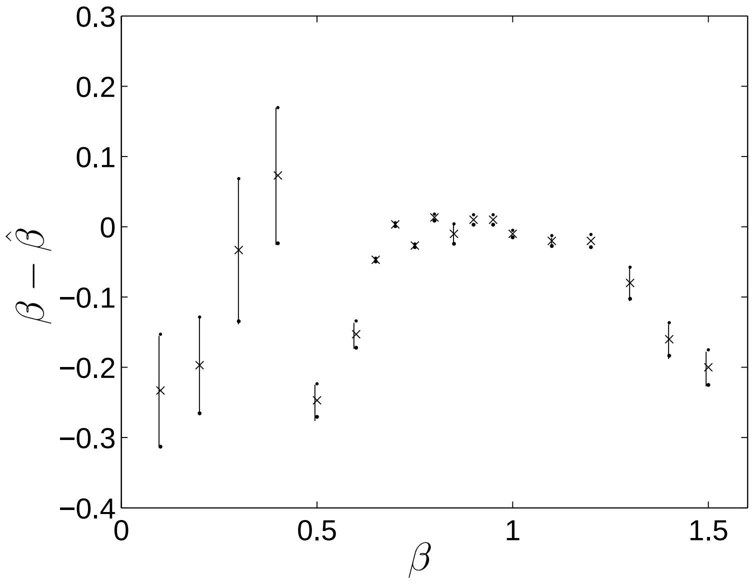

Figure 5.10: Graphs of $\beta - \hat{\beta}$ for real audio signals as a function of β. The crosses represent the mean, and the lines the standard deviation in both directions. $DWR = 15db$, and $WNR = 0db$. The assumption for the estimator is $\widetilde{X} \sim \mathcal{L}(0, \sigma^2_{\widetilde{X}} + \sigma^2_{N_1} + \sigma^2_{N_2})$. The filter transfer function is shown in Fig. 5.2.

5.5 Inverting the Effect of Two-Band Amplitude Attack

Fig. 5.11 shows the behavior of the watermark decoder when the attacked signal is passed through the corrector depicted in Fig. 5.4. The host signal is white noise, the DWR is $15db$, the number of signal samples is 80000, and $\beta = 0.8$. The error probability for reception of attacked signal and the error probability for reception of corrected signal using the corresponding estimates are compared. Fig. 5.11 illustrates how inverting the effect of two-band amplitude attack leads to significant performance improvements. The error probability increases as WNR decreases, since the estimation accuracy decreases due to the strong noise.

5.6 Discussion

In this chapter, we have presented a Maximum Likelihood estimation procedure for estimating a two-band amplitude scaling factor. The estimation technique performs well using only a small number of filter coefficients - those with the largest magnitude. The disadvantage of the estimation procedure is the need for large number of signal samples and the high computational complexity.

Due to the duality between convolution in the time domain and multiplication in the frequency domain, it is possible that linear filtering in the time domain can be modeled by a multiplication in the frequency domain. However, this duality is valid only if the Fourier transform is of infinite length. In practice, due to the finite Fourier transform length, convolutions in the time domain can hardly be modeled by pure multiplication in the frequency domain. This requires a new approach in developing robust to linear filtering

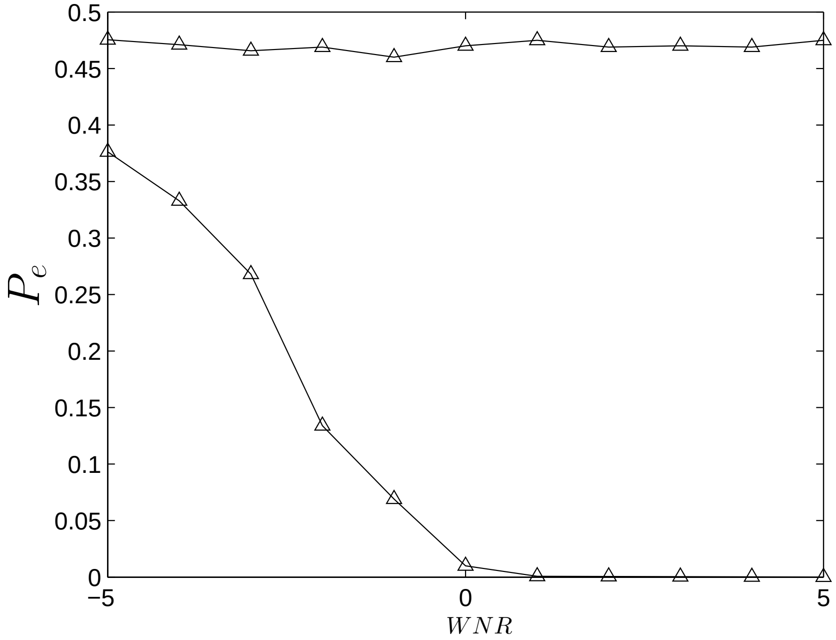

Figure 5.11: Watermark decoder performance. $DWR = 15db$, $\beta = 0.8$. The filter transfer function is shown in Fig. 5.2.

attacks quantization-based watermarking, which we address in the next chapter.

Chapter 6

Robustness to Linear Filtering Attacks *

6.1 Motivation

Quantization-based watermarking is vulnerable to linear filtering attacks. As mentioned in subsection 2.7.4, each sample of the filter output is a linear combination of watermarked data samples. This linear combination of watermarked samples interferes with the watermark in the current sample, and is thus seen as a noise. For high DWR this noise is very powerful and overwrites the watermark even for short filter lengths. Moreover, as mentioned in chapter 2, filtering operations are common in many applications and signal processing systems.

The filtering operation is shown in Fig. 6.1, where $X(k)$ is the input to the filter, $Y(k)$ is the filter output, and $h(\tau)$ denotes the filter impulse response. In security applications, an adversary can apply linear filtering operation with optimally designed $h(\tau)$ to disrupt communication between the transmitter and the intended receiver. Since in this chapter we concentrate on developing invariant to linear filtering quantization-based watermarking schemes, the term *malicious attacker* is irrelevant.

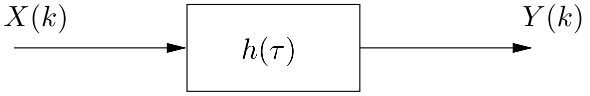

Figure 6.1: LTI filtering attack.

In this chapter we construct a watermarking scheme that is robust against a linear filtering attack. This watermarking scheme was first developed in [99]. The construction principle is based on noting that theoretically every filtering operation can be modeled as multiplication of the input with the filter transfer function in the frequency domain. Solutions to the amplitude scale attack in the frequency domain were proposed in the previous chapter and in [100, 101].

To achieve invariance to linear filtering, we apply the RDM scheme (which is a scale invariant scheme) in the frequency domain, by watermarking the amplitude components

*This chapter contains recent, still unpublished results.

and leaving the phases unchanged. The host signal is chopped into frames and the discrete Fourier transform is performed on each frame. Then the amplitude is computed from the frequency coefficients. The amplitudes with the same index from different frames form the frequency channel, on which the RDM core is applied. Since the signal frames are not periodic signals, the amplitude scale model in the frequency domain does not hold completely and there is also a residual noise term. The residual noise causes a decoding error. To reduce this error we incorporate a windowing operation on each frame before taking the Fourier transform. The reduced decoding error is at the expense of increased distortion due to the watermark. To eliminate this increase in watermark distortion, we propose to use overlapped windowing of the signal frames with 50% overlap. Experimental results of the probability of error are presented of the basic scheme and its modifications, as well as experiments with practical filters. Finally conclusions are drawn.

6.2 Principles of Frequency RDM

Here we start describing the concept of frequency RDM (FRDM) by looking first at the discrete Fourier transform of the linear filtering system. We can write

$$Y(\omega) \quad = \quad H(\omega)X(\omega), \tag{6.1}$$

where $X(\omega)$ is the Fourier transform of X, $Y(\omega)$ is the Fourier transform of Y, and $H(\omega)$ is the filter transfer function.

From (6.1) we can see that the attack operation is multiplication in the frequency domain. Therefore, if we apply RDM in the frequency domain, theoretically we have to achieve invariance to filtering operations in the time domain.

The watermark encoder is shown in Fig. 6.2. First, the encoder applies the DFT on the signal frame $\widetilde{\boldsymbol{X}}_m$ of length N, where m denotes the frame index. From the frequency coefficients, the amplitudes are computed. The number of frequency channels (amplitudes) for each frame is $N/2+1$. The RDM core is then applied on each frequency channel $i \in \{0, \ldots, \frac{N}{2}\}$, by embedding the watermark bits $W_m(i)$ in the amplitudes $\widetilde{X}_m(i)$. The quantized amplitude is denoted as $\widetilde{X}'_m(i)$. Analogously to (2.73), the g function for the ith amplitude of the mth frame is computed based on the ith quantized amplitudes of the previous $m-L$ frames, i.e.

$$g\big(\widetilde{X}'_m(i), L, p\big) \quad = \quad \Big(\frac{1}{L}\sum_{j=m-L}^{m-1} \big|\widetilde{X}'_j(i)\big|^p\Big)^{\frac{1}{p}}, \tag{6.2}$$

At the end an inverse discrete Fourier transform (IDFT) is performed resulting in the watermarked frame $\boldsymbol{X}_m$.

The attack channel and the watermark decoder are shown in Fig. 6.3. After the filtering operation, the decoder performs DFT on $\boldsymbol{Y}_m$. The decoder frames should be in alignment with the encoder frames. The DFT length is the same as that used in the encoder. Then the RDM decoder is applied on each $Y_m(i)$ and an estimate $\hat{W}_m(i)$ of the embedded bits $W_m(i)$ is made. The g function for the decoding process is calculated as

$$g\big(Y_m(i), L, p\big) \quad = \quad \Big(\frac{1}{L}\sum_{j=m-L}^{m-1} \big|Y_j(i)\big|^p\Big)^{\frac{1}{p}}. \tag{6.3}$$

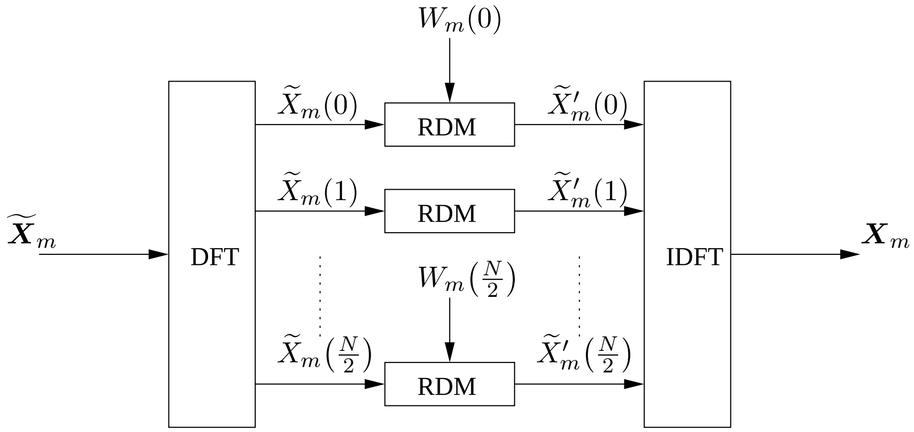

Figure 6.2: FRDM encoder

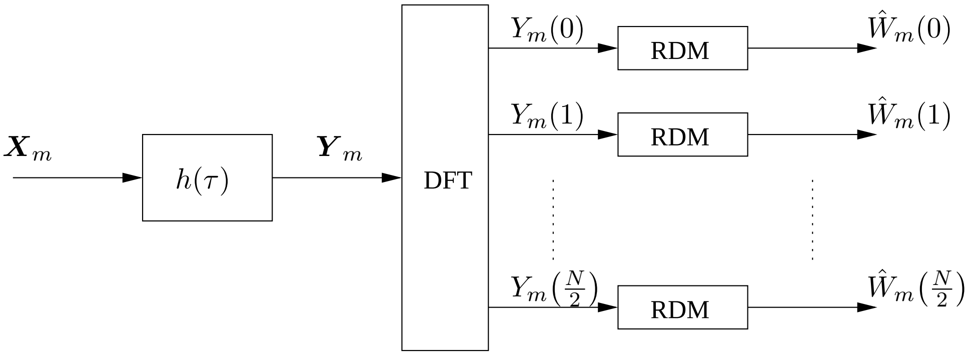

Figure 6.3: Attack channel and FRDM decoder

6.3 Experimental and Theoretical Results

In this section we perform experiments with different filters and synthetic host signals to measure the performance of the watermarking system in terms of probability of error P_e per frequency channel. For ideal low pass filters, we would expect the case $P_e = 0$ in the pass-band zone and $P_e = 0.5$ in the stop-band zone, which is shown in Fig. 6.4. The practical filter transfer functions used in the experiments are shown in Fig. 6.5.

Experimental results of P_e per channel i for different length L_f of the low pass filter are shown in Fig. 6.6 (a). It can be seen that while P_e in the stop-band zone is high as expected, $P_e \neq 0$ in the pass-band zone. Moreover, increasing the filter order increases P_e in the pass-band zone. Experimental results of P_e per channel i for different cutoff frequencies ω_c of the lowpass filter are shown in Fig. 6.6 (b). It can be seen that P_e in the pass-band zone is independent of the cutoff frequency ω_c, showing the validation of the FRDM principle.

Since $P_e \neq 0$ in the pass-band zone, FRDM is not completely invariant to LTI filtering attacks. One way to decrease P_e in the band-pass zone is to increase the DFT length. Fig. 6.7 shows experimental results with $N = 2048$ for different filter orders and different cutoff frequencies. It can be seen that P_e in the pass-band zone is significantly reduced.

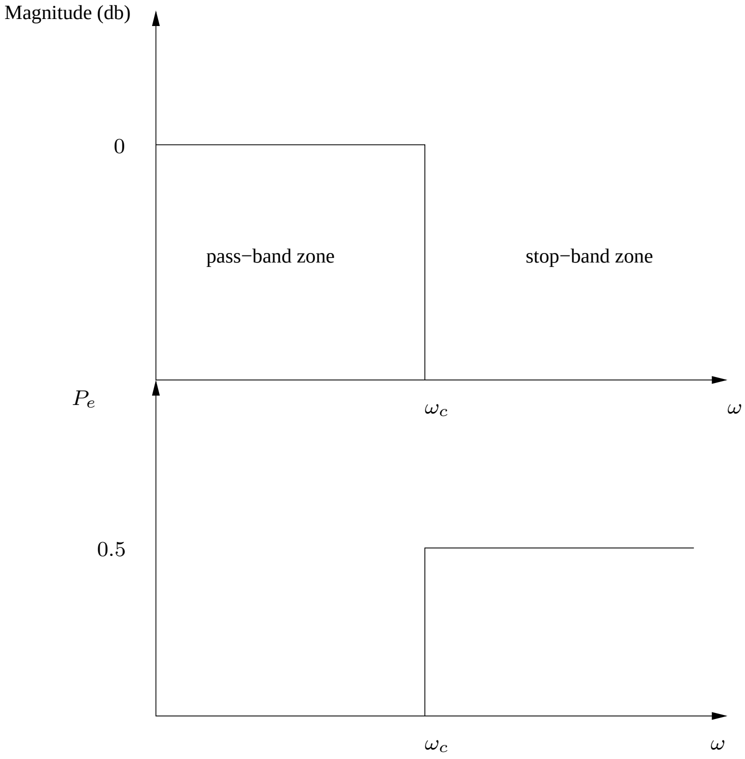

Figure 6.4: Expected FRDM performance.

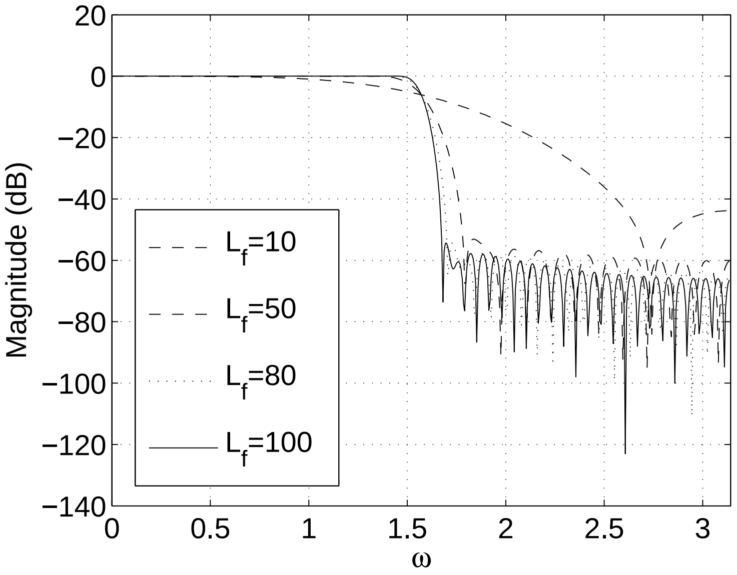

Figure 6.5: Filter transfer functions for different filter lengths L_f.

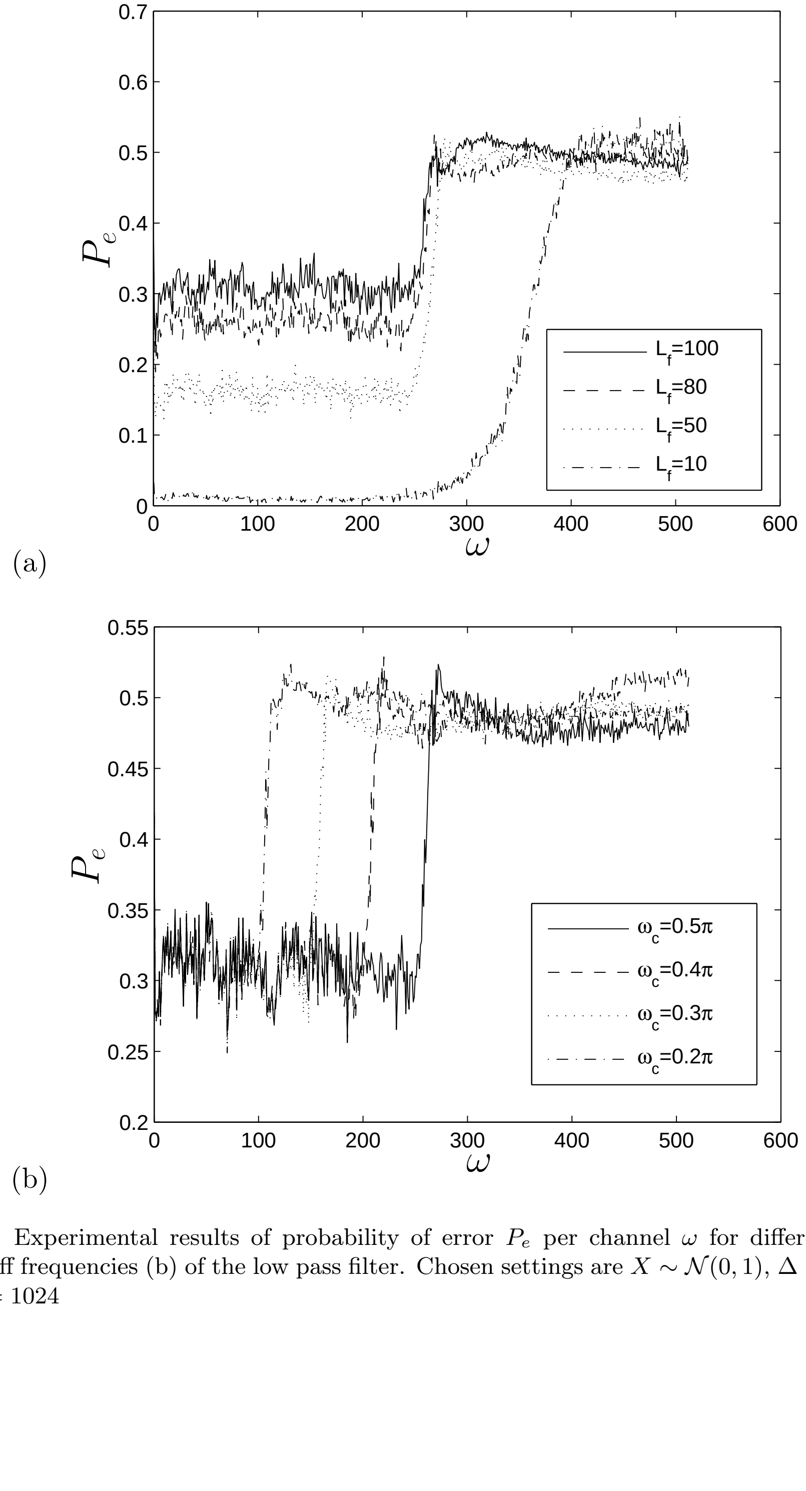

Figure 6.6: Experimental results of probability of error P_e per channel ω for different orders (a) and different cutoff frequencies (b) of the low pass filter. Chosen settings are $X \sim \mathcal{N}(0,1)$, $\Delta = 1$, $L = 10$, $p = 2$, $n = 10^6$, $N = 1024$

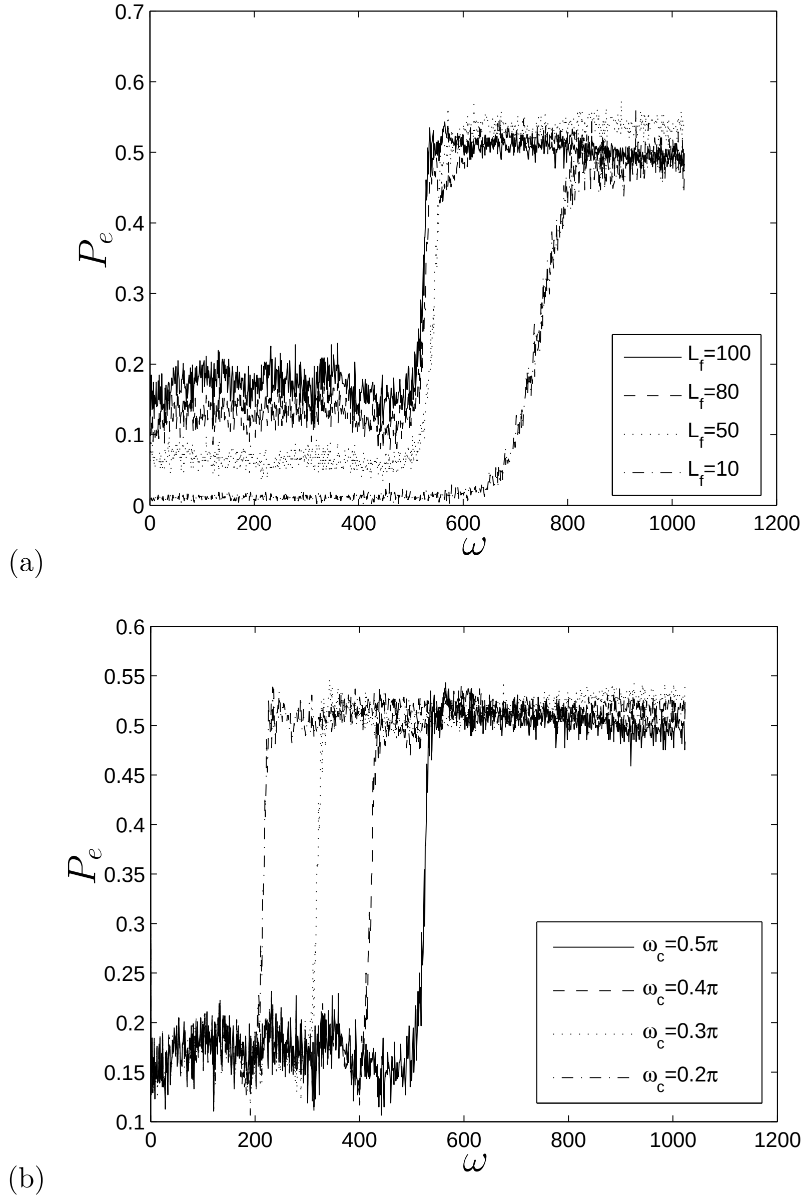

Figure 6.7: Experimental results of probability of error P_e per channel ω for different orders (a) and different cutoff frequencies (b) of the low pass filter. Chosen settings are $X \sim \mathcal{N}(0,1)$, $\Delta = 1$, $L = 10$, $p = 2$, $n = 10^6$, $N = 2048$.

The reason for $P_e \neq 0$ in the pass-band zone is that the relation (6.1) is valid only for an infinite length Fourier transform. In practice we have a finite length discrete Fourier transform (DFT). In such cases, the model is [102]

$$Y(\omega) \;=\; H(\omega)X(\omega) + R_N(\omega), \tag{6.4}$$

where $R_N(\omega)$ is a residual term representing the end effects due to the response on the signal X prior to $t = 0$ and for $t \geq N$. From the model (6.4) it is obvious that the major source for decoding errors is the term $R_N(\omega)$, which is given [102] as

$$R_N(\omega) \;=\; \frac{1}{\sqrt{2\pi N}} \sum_{k=0}^{L_f} h(k) e^{j\omega k} \sum_{t=-k}^{-1} \big(X(t) - X(t+N)\big) e^{j\omega t}, \tag{6.5}$$

where $L_f \in \mathbb{I}^+$ is the filter length (order). The model of the attack channel given by (6.4) is shown in Fig. 6.8. It consists of the multiplicative term $H(\omega)$ and the noise term $R_N(\omega)$.

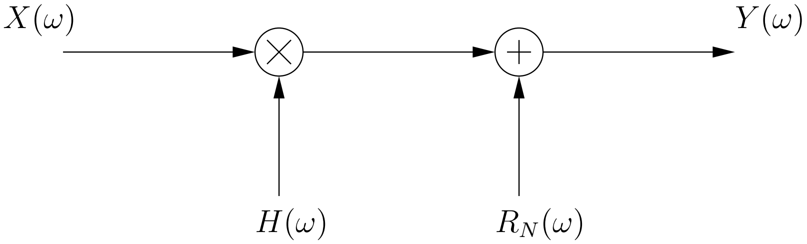

Figure 6.8: Attack channel model of the filtering attack for finite Fourier transform lengths.

To quantify the effect of $R_N(\omega)$ on probability of error, we need to study the dependency between the $R_N(\omega)$ statistics and the parameters of the watermarking scheme (namely N), and the attack channel (namely L_f). First we find the variance of $R_N(\omega)$. We can write the mean of $R_N(\omega)$ as

$$\begin{aligned} E\big[R_N(\omega)\big] \;&=\; E\frac{1}{\sqrt{2\pi N}} \Big[\sum_{k=0}^{L_f} h(k) e^{j\omega k} \sum_{t=-k}^{-1} \big(X(t) - X(t+N)\big) e^{j\omega t}\Big] \\ &=\; \frac{1}{\sqrt{2\pi N}} \sum_{k=0}^{L_f} h(k) e^{j\omega k} \sum_{t=-k}^{-1} \big(E\big[X(t)\big] - E\big[X(t+N)\big]\big) e^{j\omega t} \end{aligned} \tag{6.6}$$

If $X(t)$ is a zero mean process, then we have $E\big[R_N(\omega)\big] = 0$. In this case, the variance can be straightforwardly calculated as

$$
\begin{aligned}
E\big[R_N(\omega)R_N(\omega)^*\big] &= E\Big[\frac{1}{\sqrt{2\pi N}}\Big(\sum_{k=0}^{L_f} h(k)e^{j\omega k}\sum_{t=-k}^{-1}\big(X(t)-X(t+N)\big)e^{j\omega t}\Big) \\
&\times \frac{1}{\sqrt{2\pi N}}\Big(\sum_{m=0}^{L_f} h(m)e^{-j\omega m}\sum_{s=-m}^{-1}\big(X(s)-X(s+N)\big)e^{-j\omega s}\Big)\Big] \\
&= \frac{1}{\pi N}E\Big[\sum_{k=0}^{L_f}\sum_{m=0}^{L_f}\sum_{t=-k}^{-1}\sum_{s=-m}^{-1} h(k)e^{j\omega k}X(t)e^{j\omega t}h(m)e^{-j\omega m}X(s)e^{-j\omega s}\Big] \\
&= \frac{\sigma_X^2}{\pi N}\sum_{k=0}^{L_f} h(k)e^{j\omega k}\Big(\sum_{m=0}^{k-1} m h(m)e^{-j\omega m} + \sum_{m=k}^{L_f} k h(m)e^{-j\omega m}\Big), \qquad (6.7)
\end{aligned}
$$

where $*$ denotes complex conjugate and the second equation follows from the assumption that $X(t)$ is uncorrelated stationary process.

From (6.7) we can see that increasing the Fourier length N decreases the variance $E\big[R_N(\omega)R_N(\omega)^*\big]$, and therefore the error probability. This fact is in accordance with the experimental results. From (6.5) it is easy to see that for periodic signals $X(t)$ with period N, $R_N(\omega) = 0$ and we have the ideal model (6.1), which is also satisfied when $N \to \infty$.

It is difficult from (6.7) to predict how $E\big[R_N(\omega)R_N(\omega)^*\big]$ will change with L_f. The reason is that for different L_f, the filter impulse response $h(\tau)$ is different. Therefore, we resort to numerical computations of (6.7) with the filters used in the experiments. Since (6.7) has imaginary terms, we compute its absolute value. Table 6.1 shows numerical calculations of (6.7) and the corresponding experimental error probabilities P_e in the pass-band zone, for different low-pass filters, and Fourier transform lengths N. It can be seen that increasing L_f increases $\Big|E\big[R_N(\omega)R_N(\omega)^*\big]\Big|$, which leads to increased probability of error in the pass-band zone. Also, increasing N decreases $\Big|E\big[R_N(\omega)R_N(\omega)^*\big]\Big|$ and the probability of error.

N					
1024			2048		
L_f	$\Big\lvert E\big[R_N(\omega)R_N(\omega)^*\big]\Big\rvert$	P_e	L_f	$\Big\lvert E\big[R_N(\omega)R_N(\omega)^*\big]\Big\rvert$	P_e
10	0.0146	≈ 0.01	10	0.0073	≈ 0.01
50	0.3209	≈ 0.15	50	0.1604	≈ 0.07
80	0.805	≈ 0.26	80	0.4025	≈ 0.13
100	1.245	≈ 0.3	100	0.6225	≈ 0.17

Table 6.1: Numerical computation of $\Big|E\big[R_N(\omega)R_N(\omega)^*\big]\Big|$ and the corresponding experimental error probabilities P_e in the pass-band zone, for different low-pass filters, and Fourier transform lengths N. The variance of the watermarked signal is $\sigma_X^2 = 1$.

Based on the experimental and theoretical results in this section we can point out that the main problem to be solved for the FRDM principle is $P_e \neq 0$ in the pass-band zone.

6.4 Improvement using Hamming Windows

In the previous section we saw that one way to decrease P_e in the pass-band zone is to increase the DFT length. However, this will lead to decreased watermark payload per frequency channel. Also the memory and therefore the performance of the of RDM core will be more restricted. Although the number of channels increases, the scheme will have a low payload in the presence of filters with large stop-band zone.

One way of reducing P_e in the pass-band zone while at the same time keeping the watermark payload per channel constant is to apply windowing operation on the host signal, before taking the DFT. However, such operation introduces additional distortion to the host signal.

Experimental results with Hamming window are shown in Fig. 6.9. It can be seen that P_e in the pass-band zone is significantly reduced and independent of the filter. The reason is that the windowing operation forces the signal frame to have more or less a periodic structure, thus reducing spectral leakage [103]. Other windows can also be applied. Generally, the more the windowing operation approximates a periodic signal, the lower the probability of error in the pass-band zone.

6.5 Modification using Cosine Squared Windows

Since the use of Hamming windows causes additional distortion to the host signal, in this section we propose a modification to the original FRDM concept, using overlapped signal frames multiplied by a cosine squared window. An illustration of this principle is shown in Fig. 6.10, where n is the total signal length, 2^r is the overlap, and N is the frame length, with $r \in \mathbb{I}^+$. The number of frames is $\lfloor \frac{n-N}{2^r} \rfloor + 1$, where $\lfloor x \rfloor$ denotes the closest integer smaller or equal to x. The assumption is that $N \geq 2^r$, with equality if there is no overlap.

The modified encoder is shown in Fig. 6.11. Each frame is multiplied by a cosine squared window in the following way:

$$\widetilde{X}_w(i) \quad = \quad \widetilde{X}(i) \cos^2\left(\frac{2i-N}{2N}\pi\right) = \widetilde{X}(i) \sin^2\left(\frac{i}{N}\pi\right). \tag{6.8}$$

Cosine squared windows have perfect reconstruction property. When added back with the same overlap, the windowed frames $\widetilde{X}_w$ perfectly reconstruct the original signal $\widetilde{X}$ and hence there is no additional distortion due to the windowing operation.

The FRDM scheme is applied on each frame $\widetilde{X}_w$. At the output of the FRDM encoder the frames are added back with the same overlap 2^r.

The modified decoder is shown in Fig. 6.12. To obtain a close estimate to the watermarked frames, the attacked frames have to be multiplied with the same window (6.8) prior to FRDM decoding. The Fourier transform of the product can be written as a convolution

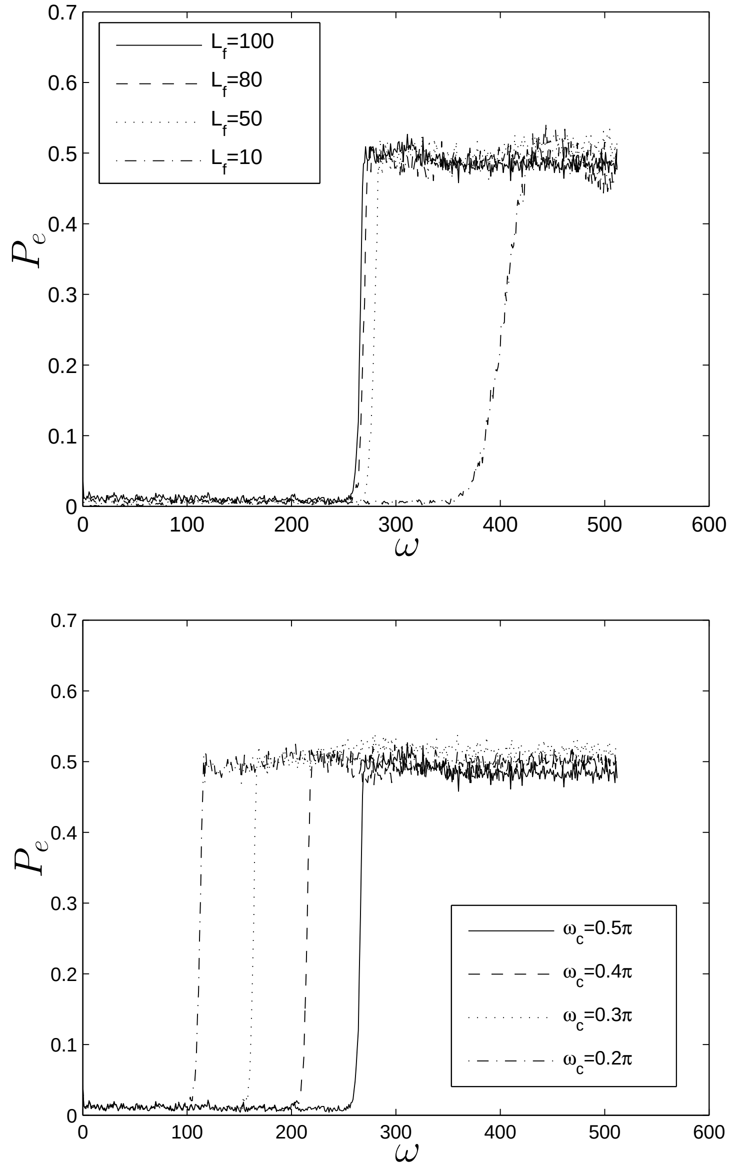

Figure 6.9: Experimental results of P_e per ω for different orders (a) and different cutoff frequencies (b) of the low pass filter, applying Hamming window before FRDM encoding. Chosen settings are $X \sim \mathcal{N}(0,1)$, $\Delta = 1$, $L = 10$, $p = 2$, $n = 10^6$, $N = 1024$.

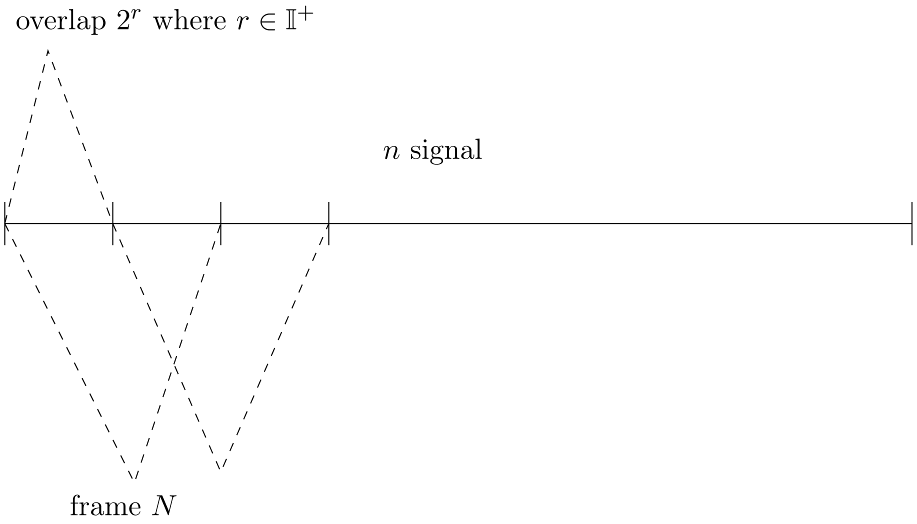

Figure 6.10: Overlapped signal frames.

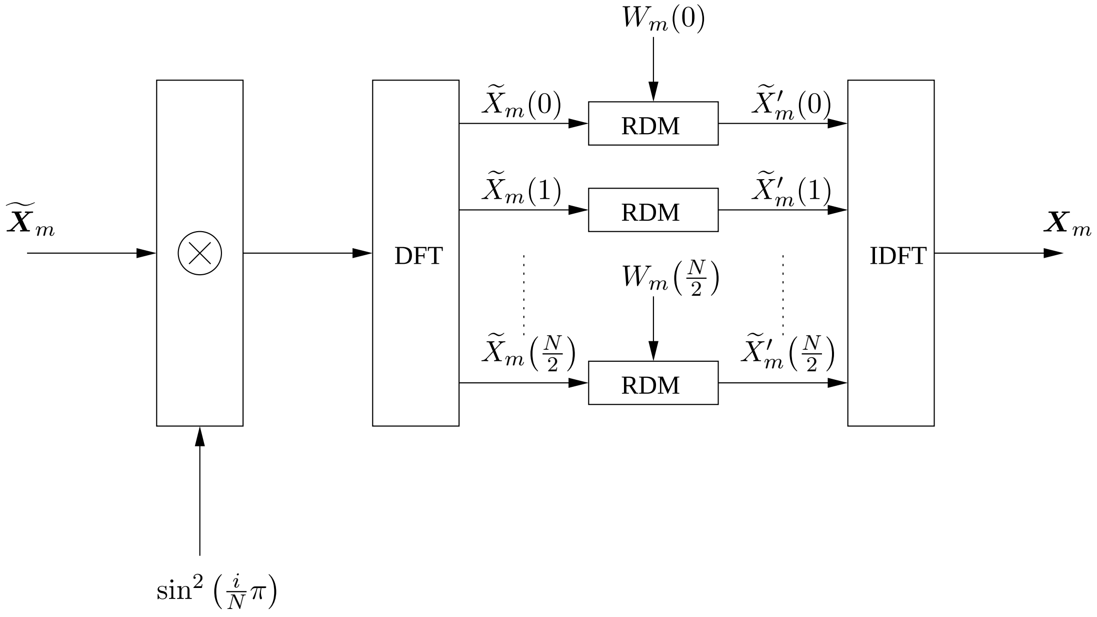
$W_m(0)$
$\widetilde{X}_m(0)$
RDM
$\widetilde{X}'_m(0)$
$\widetilde{X}_m(1)$
RDM
$\widetilde{X}'_m(1)$
$\widetilde{\boldsymbol{X}}_m$
DFT
$W_m\left(\frac{N}{2}\right)$
IDFT
$\boldsymbol{X}_m$
$\widetilde{X}_m\left(\frac{N}{2}\right)$
RDM
$\widetilde{X}'_m\left(\frac{N}{2}\right)$
$\sin^2\left(\frac{i}{N}\pi\right)$

Figure 6.11: Cosine-squared FRDM encoder

of the Fourier transforms of each term, i.e.

$$\begin{aligned}\mathcal{F}_f\Big[Y(i)\sin^2\big(\frac{i}{N}\pi\big)\Big] &= \mathcal{F}_f\big[Y(i)\big] * \mathcal{F}_f\Big[\sin^2\big(\frac{i}{N}\pi\big)\Big] \\ &= \phi_Y(f) * \Big[\frac{\delta(f)}{2} - \frac{1}{4}\delta\big(f-\frac{1}{N}\big) - \frac{1}{4}\delta\big(f+\frac{1}{N}\big)\Big] \\ &= \phi_Y(f) * \frac{\delta(f)}{2} - \frac{1}{4}\phi_Y(f) * \delta\big(f-\frac{1}{N}\big) - \frac{1}{4}\phi_Y(f) * \delta\big(f+\frac{1}{N}\big) \\ &= \frac{1}{2}\phi_Y(f) - \frac{1}{4}\phi_Y\big(f-\frac{1}{N}\big) - \frac{1}{4}\phi_Y\big(f+\frac{1}{N}\big), \qquad (6.9)\end{aligned}$$

where $\omega = 2\pi f$.

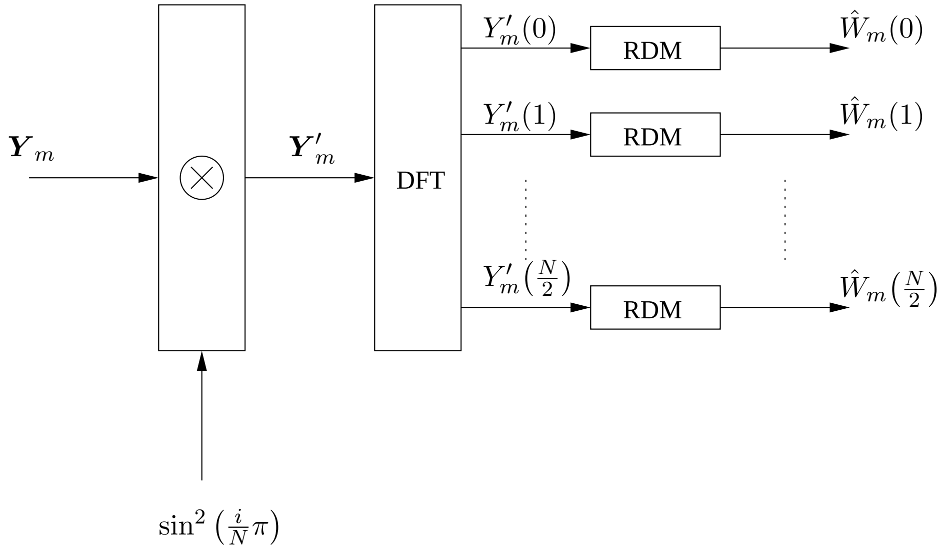

Figure 6.12: Cosine-squared FRDM decoder

From (6.9), it can be seen that the last two terms are actually scaled and shifted versions of $\phi_Y(f)$. It is interesting to observe that these two terms actually *recover* part of the frequencies of the attacked signal that are lost due to the stop-band of the low-pass filter. Therefore, we should expect reduced probability of error for some frequencies in the stop-band zone.

The advantage of the modification with squared-cosine windows is that the watermark payload is increased and is precisely $N/2^r$ the payload of the normal FRDM scheme.

Experimental results of the proposed modification are presented in Fig. 6.13. We can see that due to the overlap, the error is increased and more or less uniformly distributed over all frequencies. The reduced probability of error for some frequencies in the stop-band zone is due to the multiplication with the squared-cosine window before watermark decoding, since the window *recovers* part of the lost spectrum of the attacked signal, according to (6.9).

6.6 Experiments with Practical Filters

The previous sections contained experiments with simple low pass filters, designed to study the capabilities and inner workings of the FRDM principle. The low pass filter is pretty

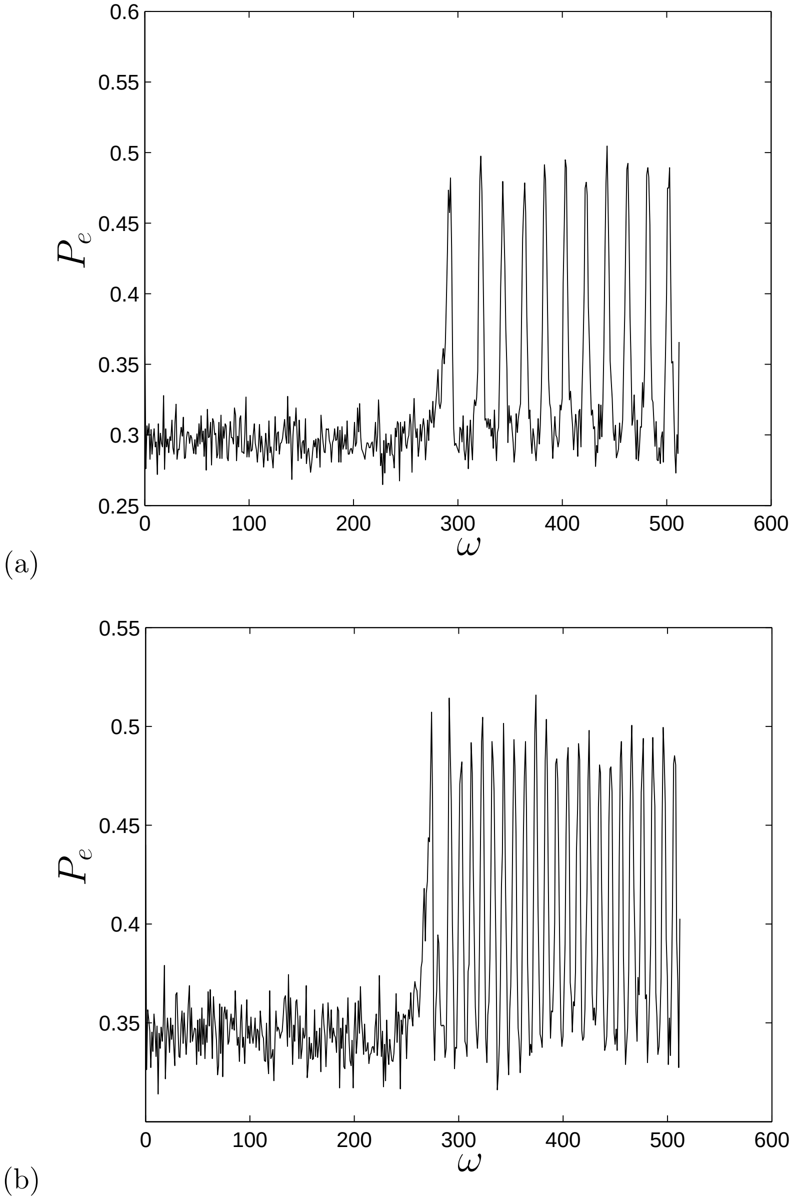

Figure 6.13: Experimental results of P_e per ω for low-pass filter with $\omega_c = 0.5\pi$. Chosen settings are $X \sim \mathcal{N}(0,1)$, $\Delta = 1$, $L = 10$, $p = 2$, $n = 10^6$, $N = 1024$, $2^r = 512$, (a) $L_f = 50$, (b) $L_f = 100$.

devastating, because in the stop band zone all watermark information is lost. In this section we consider equalizers, which usually do not have a (large) stop band zone and therefore are less severe than the low pass filter. Such equalizers have application in music tuning, like the WinAmp graphical equalizer.

The transfer function of a 10-band equalizer is shown in Fig. 6.14. It can be seen that the global transfer function of the equalizer does not contain stop-band zone. The filter in each sub-band contains only three taps, so we will expect that short Fourier transform lengths will be sufficient to achieve low probability of error.

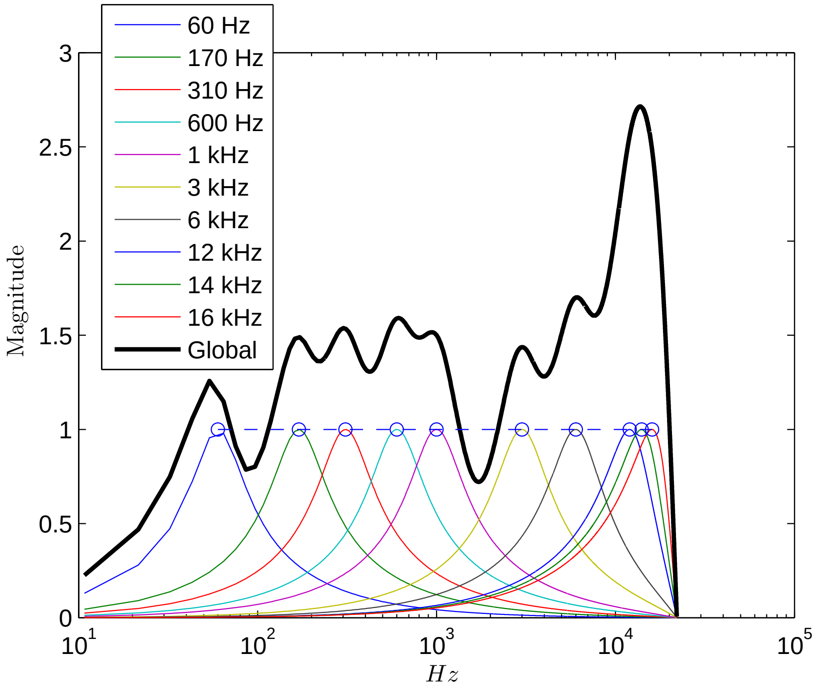

Figure 6.14: Equalizer transfer function. This figure is taken from [99].

Experimental results are shown in Fig. 6.15. It can be seen that for small N, P_e is already pretty low for all frequencies. This is due to the fact that the global transfer function of the equalizer does not contain a stop-band zone. Another, more important for the FRDM principle, reason is that each band contains a filter with short length ($L_f = 3$ in this case). This experiment confirms the fact that the filter length L_f has a major influence on the performance of the FRDM principle, as pointed in section 6.3.

6.7 Discussion

In this chapter we discussed and proposed techniques to combat linear filtering attacks by applying RDM in the frequency domain. The new scheme is called FRDM. We studied the performance of the pure FRDM scheme experimentally. However, there are still decoding errors in the pass-band zone due to the finite length of the Fourier transform. It was shown that increasing the FFT length decreases the errors in the pass-band zone. Furthermore, we

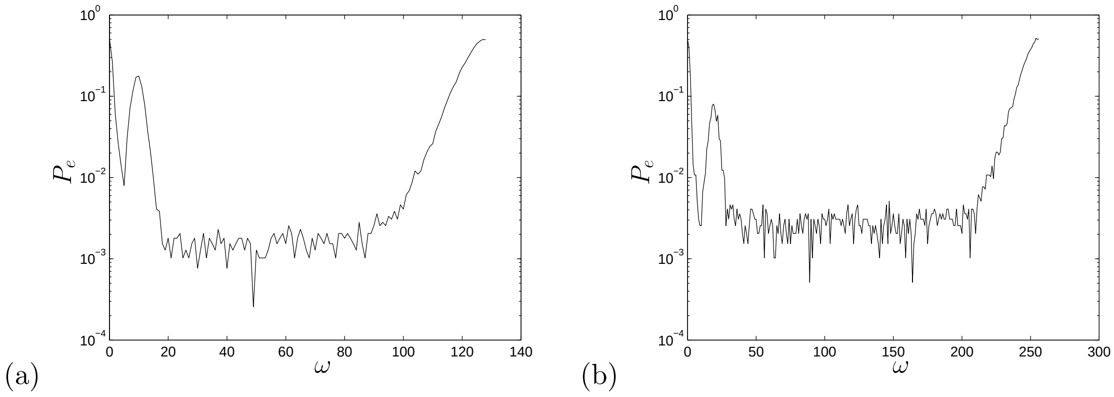

Figure 6.15: Experimental results of P_e for equalizer attacks, (a) $N = 256$ and (b) $N = 512$.

showed that windowing helps to reduce the decoding errors at the expense of increased distortion. To completely cancel the distortion due to the windowing operation, we employed overlapped windows with perfect reconstruction properties (cosine squared windows). However this is achieved at the expense of increased errors in the pass-band zone due to the overlap.

The FRDM scheme, including its proposed modifications, are not completely invariant to arbitrary filtering attacks. The main problem that remains to be solved is to eliminate or at least further reduce the probability of error in the pass-band zone. This can be achieved for example by using specially designed perfect reconstruction windows that avoid the need for overlap. With respect to this direction, the research could concentrate on studying different windows and their properties, and on designing special purpose windows. Another, more promising approach would be to employ a different embedding mechanism, instead of the RDM core.

Chapter 7

Conclusions and Future Research

7.1 Discussion

In this thesis we studied and developed quantization-based watermarking techniques that are robust to nonadditive attacks. Such techniques are suitable in multimedia applications, where the attacks are standard signal processing operations.

The general class of quantization-based watermarking schemes was developed from information theoretic principles and results. It is the class of watermarking schemes that achieves the highest capacity in terms of additive noise attacks. However, a realistic attacker can choose to apply more sophisticated attacks that can hardly be modeled as additive. Moreover, common signal processing operations like amplitude scaling, linear filtering, compression, etc. are clearly nonadditive. Furthermore, it turned out that classical quntization-based watermarking is extremely vulnerable to the above mentioned examples of signal processing operations.

To improve the robustness of quantization-based watermarking against such signal processing operations, while at the same time keeping their good performance with respect to additive noise attacks, we developed statistical estimation procedures for estimating amplitude scale factors in the time and frequency domains.

The amplitude estimation procedure based on Fourier analysis is computationally efficient and gives accurate results for high watermark-to-noise ratios. The estimation principle is based on detecting peaks created due to the encoding process, in the characteristic function of the attack data. The procedure does not require any prior knowledge of the host signal. The performance of the procedure degrades at low watermark-to-noise ratios. In the case of embedding *zeros* and *ones* with equal probability, the estimation technique fails when the power of watermark and attack noise become equal, since then the discontinuities in the density, and the peaks in the characteristic function of the watermarked data disappear.

For watermarking applications, where the allowed variance of the attacker's noise is larger than that of the watermark, we developed maximum likelihood estimation of amplitude scaling. The improvement of estimation accuracy over the Fourier based approach is at the expense of increased complexity and the need for an accurate density model of the host signal. The performance of the maximum likelihood approach is not affected when the embedded messages are *zeros* and *ones*, because the approach does not rely on any discontinuities in the density of the watermarked signal.

For applications that require *secrecy* watermarking, i.e. when the attacker should not be able to decode the watermark, we developed maximum likelihood estimation of amplitude scale factors in the presence of subtractive dither. The dither sequence was introduced to ensure the security of the watermarking system. The dither statistics were derived such that an attacker without having the dither realization is not able to decode the watermark with probability different than 0.5. To perform accurate estimation, we imposed restrictions on the dither variance, but these restrictions did not affect in any way the security of the system.

For applications that require robustness against linear filtering and additive noise attacks, we developed maximum likelihood estimation of amplitude scaling factors in the frequency domain (multiband scaling), by noticing the duality between linear filtering in the time domain and amplitude scale in the frequency domain. It turned out that only a few filter coefficients, those with the largest magnitudes are sufficient for constructing accurate density models for the estimation procedure.

For applications that require only robustness to linear filtering attacks, we studied the application of rational dither modulation in the frequency domain, to develop an invariant to linear filtering attacks scheme. The performance of the proposed modification is not ideal due to the finite length of the Fourier transform, and there are errors in the pass-band zone. We saw that there are two ways to suppress these errors, by increasing the length of the Fourier transform, and by pre-multiplying the signal frames with non-overlapped windows. Both ways have pros and cons. By increasing the Fourier length, we decrease the payload, while by applying a windowing operation we introduce additional distortion to the host signal. We also investigated the application of overlapped windows to eliminate the aforementioned additional distortion.

Overall, we can state that the proposed in this thesis techniques to counter amplitude scale attacks, linear filtering attacks, and security problems, are sufficient for bringing quantization-based watermarking schemes one step closer to practical applications that involve the aforementioned issues.

However, looking at Fig. 2.19 and Fig. 2.20, we see that audio compression and nonlinear operations like voice mixing, A-D conversion, dynamic signal processing, have not been tacked. This is mainly due to the complex nature of these operations, and the difficulty in finding accurate models for them. These problems can be approached by modeling the operations by blocks of simpler well known operations and tackling each individual block separately with the already existing techniques. Since it is difficult to construct a watermarking scheme that is invariant to a large number of different, simple operations, it is probable that a solution to a more complex attack would also require the incorporation of estimation techniques like the ones developed in this thesis.

7.2 Future Research

The future directions are in the lines of the following subjects. The first subject involves our research on the estimation techniques.

Our estimation techniques require brute force searching for the optimal scaling factor, and therefore are computationally inefficient. It is important to find more efficient algorithms for finding the maximum in the likelihood function, but it is unlikely that these will be gradient-based. A possible direction is first to obtain a more convenient analytical expression for the probability density function of the received data, that will allow to find a closed form expression for the maximum likelihood estimate $\hat{\beta}$. Such an expression would allow for considerable reduction in computational complexity, and therefore the applicability of the estimation technique in real-time applications. In this way, we would also be able to do a more careful analysis of the performance via bias and variance, and to compare theoretically with other estimation procedures from the literature.

Furthermore, it is important to theoretically quantify the influence of the inaccuracy of the assumed model on the estimation performance, for example via Kullback-Leibler distance. Thus, we would be able to theoretically quantify the estimation performance for real signals.

It is also important to point out that our current amplitude scaling model includes only a constant scaling factor. Clearly, the attacker may also choose to change the amplitude scale factor with time, in an unpredictable way. One possible way for improvement in this direction is to look into other estimation principles that use smaller amount of signal samples, for example order statistics [104].

The second subject for future research relates to our work on linear filtering invariance. To achieve invariance to linear filtering attacks, it is important to decrease or eliminate the errors in the pass-band zone. We saw that one way of achieving this is by the use of windows. Unfortunately, the used windows introduce additional distortion to the watermarked data. We also saw that perfect reconstruction windows introduce no additional distortion, but cause additional errors in the pass-band zone due to the overlap. Therefore, a possible direction for future research is to study and analyze different windows, and to design special purpose windows that do not cause substantial additional distortion to the watermarked data, and do not require the use of overlapped frames.

The third line for future research relates to countermeasures against more complex attacks like audio compression, voice mixing, A-D conversion, appreciating the estimation techniques in watermarking applications. It would be interesting to see if it is possible to combat such complex operations with the already developed techniques in this thesis. One possible way to deal with such complex attacks is to build detailed models based on simple operations and to attack each building block individually. It is expected that by doing so, the watermarking scheme will result in a system that contains several estimation procedures, since it is difficult (even impossible) to construct a watermarking scheme that is invariant to several operations at the same time.

Finally, we would like to point out that future research on security aspects of watermarking is also of considerable importance. The current security requirements for watermarking are not very high, in comparison to those for cryptography. However, some areas, like military applications, require significantly higher level of security. This would be a stimulation for a future research on secure watermarking algorithms.

Bibliography

[1] D. R. Stinson. *Cryptography: Theory and Practice.* CRC Press Inc., 2002.

[2] A. J. Menezes, P. C. van Dorschof, and S. A. Vanstone. *Handbook of Applied Cryptography.* CRC Press Inc., 1997.

[3] P. Moulin and R. Koetter. Data-Hiding Codes. *Proceedings IEEE*, 93(12):2083–2127, December 2005.

[4] I. J. Cox, M. L. Miller, and J. A. Bloom. *Digital Watermarking.* Morgan Kaufmann, 2001.

[5] S. Katzenbeisser and F. A. P. Petitcolas. *Information Hiding Techniques for Steganography and Digital Watermarking.* Boston, MA: Arthouse Tech, 2000. Computer Security Series.

[6] M. D. Swanson, M. Kobayashi, and A. H. Tewfik. Multimedia Data-Embedding and Watermarking Technologies. *Proceedings IEEE*, 86(6):1064–1087, June 1998.

[7] J. A. Bloom, I. J. Cox, T. Kalker, J. -P. Linnartz, and M. L. Miller. Copy Protection for DVD Video. *Proceedings IEEE*, 87(7):1267–1276, July 1999.

[8] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn. Information Hiding-A Survey. *Proceedings IEEE*, 87(7):1062–1078, July 1999.

[9] M. Barni, C. I. Podilchuk, F. Bartolini, and E. J. Delp. Watermark Embedding: Hiding a Signal Within a Cover Image. *IEEE Communications Magazine*, 39(8):102–108, August 2001.

[10] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. Techniques for Data Hiding. *IBM Systems Journal*, 35(3/4):313–336, 1996.

[11] R. J. Anderson, and F. A. Petitcolas. On the Limits of Steganography. *IEEE Journal on Selected Areas in Communications*, 16(4):474–481, May 1998.

[12] C. E. Shannon. Communication Theory of Secrecy Systems. *Bell System Technical Journal*, 28:656–715, October 1949.

[13] J. L. Massey. *Applied Digital Information Theory.* Lecture Notes, ETH Zurich.

[14] I. D. Shterev and R. L. Lagendijk. Amplitude Scale Estimation for Quantization-Based Watermarking. *IEEE Transactions on Signal Processing*, 54(11):4146–4155, November 2006.

[15] I. J. Cox, M. L. Miller, and A. L. McKellips. Secure Spread Spectrum Watermarking for Multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, December 1997.

[16] R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp. Perceptual Watermarks for Digital Images and Video. *Proceedings IEEE*, 87(7):1108–1126, July 1999.

[17] J. J. Eggers and B. Girod. *Informed Watermarking.* Kluwer Academic Publishers, 2002.

[18] P. Moulin and J. A. O'Sullivan. Information-Theoretic Analysis of Information Hiding. *IEEE Transactions on Information Theory*, 49(3):563–593, March 2003.

[19] A. S. Cohen and A. Lapidoth. The Gaussian Watermarking Game. *IEEE Transactions on Information Theory*, 48(6):1639–1667, June 2002.

[20] B. Chen and G. Wornell. Quantization Index Modulation: A Class of Provably Good Methods for Digital Watermarking and Information Embedding. *IEEE Transactions on Information Theory*, 47:1423–1443, May 2001.

[21] J. J. Eggers, R. Bauml, and B. Girod. Estimation of Amplitude Modifications before SCS Watermark Detection. *SPIE Security and Watermarking of Multimedia Contents IV*, 4675:387–398, January 2002. San Jose, CA, USA.

[22] F. Perez-Gonzalez, C. Mosquera, M. Barni, and A. Abrardo. Rational Dither Modulation: A High Rate Data-Hiding Method Invariant to Gain Attacks. *IEEE Transactions on Signal Processing*, 53(10):3960–3975, October 2005.

[23] I. Cox, M. L. Miller, and A. L. McKellips. Watermarking as Communications with Side Information. *Proceedings IEEE*, 87(7):1127–1141, July 1999.

[24] T. Berger. *Rate Distortion Theory. A Mathematical Basis for Data Compression.* Prentice-Hall, 1971.

[25] R. G. Gallager. *Information Theory and Reliable Communication.* John Wiley and Sons, Inc., 1968.

[26] A. J. Viterbi. *Principles of Digital Communication and Coding.* McGraw-Hill Inc., 1979.

[27] H. V. Poor. *An Introduction to Signal Detection and Estimation.* Springer-Verlag, second edition, 1994.

[28] C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[29] C. E. Shannon. Communication in the Presence of Noise. *Proceedings IEEE*, 86(2):447–457, February 1998.

[30] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, INC., 1991.

[31] P. Moulin and A. Ivanovich. The Zero-Rate Spread-Spectrum Watermarking Game. *IEEE Transactions on Signal Processing*, 51(4):1098–1117, April 2003.

[32] C. E. Shannon. Probability of Error for Optimal Codes in a Gaussian Channel. *The Bell System Technical Journal*, 38(3):611–656, May 1959.

[33] A. Dembo and O. Zeitourni. *Large Deviations Techniques and Applications.* Springer, second edition, 1998.

[34] N. Merhav. On Random Coding Error Exponents of Watermarking Systems. *IEEE Transactions on Information Theory*, 46(2):420–430, March 2000.

[35] R. J. Barron, B. Chen, and G. W. Wornell. The Duality Between Information Embedding and Source Coding With Side Information and Some Applications. *IEEE Transactions on Information Theory*, 49(5):1159–1180, May 2003.

[36] D. Karakos and A. Papamarcou. A Relationship Between Quantization and Watermarking Rates in the Presense of Additive Gaussian Attacks. *IEEE Transactions on Information Theory*, 49(8):1970–1982, August 2003.

[37] A. S. Baruch and N. Merhav. On the error exponent and capacity of private watermarking systems. *IEEE Transactions on Information Theory*, 49(3):537–562, March 2003.

[38] A. S. Baruch and N. Merhav. On the capacity of public watermarking systems. *IEEE Transactions on Information Theory*, 50(3):511–524, March 2004.

[39] S. I. Gel'fand and M. S. Pinsker. Coding for Channel with Random Parameters. *Problems of Control and Information Theory*, 9:19–31, 1980.

[40] C. E. Shannon. Channels with Side Information at the Transmitter. *IBM Journal*, pages 189–293, October 1958.

[41] Y. Steinber and N. Merhav. Identification in the Presence of Side Information with Application to Watermarking. *IEEE Transactions on Information Theory*, 47(4):1410–1422, May 2001.

[42] M. H. Costa. Writing on Dirty Paper. *IEEE Transactions on Information Theory*, 29(3):439–441, May 1983.

[43] I. Csiszar and J. Korner. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Akademiai Kiado, Budapest, 1981.

[44] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices and Groups.* Springer-Verlag, 3 edition, 1999.

[45] J. H. Conway and N. J. A. Sloane. Fast Quantizing and Decoding Algorithms for Lattice Quantizers and Codes. *IEEE Transactions on Information Theory*, 28(2):227–821, March 1982.

[46] J. H. Conway and N. J. A. Sloane. A Fast Encoding Method for Lattice Codes and Quantizers. *IEEE Transactions on Information Theory*, 29(6):820–824, November 1983.

[47] T. Liu and P. Moulin. Error Exponents for One-Bit Watermarking. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 2003. Hong Kong.

[48] U. Erez and R. Zamir. Achieving $\frac{1}{2}\log(1+SNR)$ on the AWGN Channel with Lattice Encoding and Decoding. *IEEE Transactions on Information Theory*, 50(10):2293–2314, October 2004.

[49] H. L. Van Trees. *Detection, Estimation, and Modulation Theory*, volume 1. John Wiley and Sons, Inc., 1968.

[50] J. J. Eggers, R. Baulm, R. Tzchoppe, and B. Girod. Scalar Costa Scheme for Information Embedding. *IEEE Transactions on Signal Processing*, 51(4):1003–1019, April 2003.

[51] M. Kesal, M. K. Michak, R. Koetter and P. Moulin. Iteratively Decodable Codes for Watermarking Applications. *Proc. 2nd Int. Symp. on Turbo Codes and Related Topics, Brest, France*, September 2000.

[52] J. C. Oostveen and T. Kalker. Local Adaptivity for the Scalar Costa Scheme. *SPIE Security, Steganography, and Watermarking of Multimedia Contents VI*, January 2004. San Jose, CA.

[53] A. Maor and N. Merhav. On Joint Information Embedding and Lossy Compression in the Presence of a Stationary Memoriless Attack Channel. *IEEE Transactions on Information Theory*, 51(9):3166–3175, September 2005.

[54] A. Maor and N. Merhav. On Joint Information Embedding and Lossy Compression. *IEEE Transactions on Information Theory*, 51(8):2998–3008, August 2005.

[55] D. Kundur. Implications for High Capacity Data Hiding in the Presence of Lossy Compression. *IEEE Proceedings on Information Technology: Coding and Computing*, pages 16–21, 2000.

[56] V. Licks, F. Ourique, F. Jordan, and F. Perez-Gonzalez. The Effect of the Random Jitter Attacks on the Bit Error Rate Performance of Spatial Domain Watermarking. *IEEE International Conference on Image Processing*, September 2003. Barcelona, Spain.

[57] C. Baggen and J. Wolf. On Band-Limited Additive Gaussian Noise Channels in the Presence of Sampling Jitter. *IEEE International Symposium on Information Theory*, June-July 1994. Trondheim, Norway.

[58] A. Papoulis. Error Analysis in Sampling Theory. *Proceedings IEEE*, 54(7):947–955, July 1966.

[59] Y. Jiang, F. W. Sun, and J. S. Baras. On the Performance Limits of Data-Aided Synchronization. *IEEE Transactions on Information Theory*, 49(1):191–203, January 2003.

[60] P. Moulin, A. Briassouli, and H. Malvar. Detection-Theoretic Analysis of Desynchronization Attacks in Watermarking. *IEEE International Conference on Digital Signal Processing*, July 2002. Santorini, Greece.

[61] A. Briassouli and P. Moulin. Detection-Theoretic Analysis of Warping Attacks in Spread-Spectrum Watermarking. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 2003. Hong Kong.

[62] P. Moulin and A. Ivanovic. The Fisher Information Game for Optimal Design of Synchronization Patterns in Blind Watermarking. *IEEE International Conference on Image Processing*, September 2001. Thessaloniki, Greece.

[63] R. Bauml, J. J. Eggers, and J. Huber. A Channel Model for Watermarks Subject to Desynchronization Attacks. *SPIE Security, Steganography, and Watermarking of Multimedia Contents IV*, January 2002. San Jose, CA.

[64] V. Licks and F. Perez-Gonzalez. Performance Bound on Optimal Watermark Synchronizers. *SPIE Security, Steganography, and Watermarking of Multimedia Contents VI*, January 2004. San Jose, CA.

[65] Q. Li and I. Cox. Rational Dither Modulation Watermarking Using a Perceptual Model. *IEEE Workshop on Multimedia Signal Processing*, October 2005. Shanghai, China.

[66] A. Abrardo, M. Barni, F. Perez-Gonzalez, and C. Mosquera. Trellis-Coded Rational Dither Modulation for Digital Watermarking. *Fourth International Workshop on Digital Watermarking*, September 2005. Siena, Italy.

[67] S. Lin and D. J. Costello. *Error Control Coding: Fundamentals and Applications.* Prentice-Hall, 1983.

[68] E. R. Berlekamp. *Algebraic Coding Theory.* McGraw-Hill, 1968.

[69] W. W. Peterson and E. J. Weldon. *Error-Correcting Codes.* MIT Press, 1972.

[70] P. Moulin. The Role of Information Theory in Watermarking and Its Application to Image Watermarking. *Signal Processing*, 81(6):1121–1139, 2001.

[71] B. Chen and G. Wornell. An Information-Theoretic Approach to the Design of Robust Digital Watermarking Systems. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 1999. Phoenix, AZ.

[72] F. P. Gonzalez, F. Balado, and J. R. Hernandez. Performance Analysis of Existing and New Methods for Data Hiding with Known-Host Information in Additive Channels. *IEEE Transactions on Signal Processing*, 51(4):960–980, 2003.

[73] B. Chen and G. Wornell. Achievable Performance of Digital Watermarking Schemes. *IEEE International Conference on Multimedia Computing and Systems*, June 1999. Florence, Italy.

[74] F. Bartolini, M. Barni, and A. Piva. Performance Analysis of Spread Transform Dither Modulation (ST-DM) Watermarking in Presence of Non-additive Attacks. *IEEE Transactions on Signal Processing*, 52(10):2965–2974, October 2004.

[75] J. J. Eggers and B. Girod. Quantization Effects on Digital Watermarks. *Signal Processing*, 81(2):239–263, 2001.

[76] P. Moulin. Embedded-Signal Design for Channel Parameter Estimation Part I: Linear Estimation. *IEEE International Workshop on Statistical Signal Processing*, September 2003. St. Louis, MO.

[77] P. Moulin. Embedded-Signal Design for Channel Parameter Esimation Part II: Quantization Embedding. *IEEE International Workshop on Statistical Signal Processing*, September 2003. St. Louis, MO.

[78] K. Lee, D. S. Kim, and K. A. Moon. Amplitude-Modification Resilient Watermarking Based on A-Law Companding. *IEEE International Conference on Image Processing*, September 2003. Barcelona, Spain.

[79] M. L. Miller and G. J. Doerr and J. Cox. Dirty-Paper Trellis Codes For Watermarking. *IEEE International Conference On Image Processing*, 2:129–132, September 2002. Rochester, NY.

[80] Q. Li and I. Cox. Using Perceptual Models to Improve Fidelity and Provide Invariance to Valumetric Scaling for Quantization Index Modulation Watermarking. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2005. Philadelphia, PA.

[81] A. B. Watson. DCT Quantization Matrices Visually Optimized for Individual Images. *SPIE Human Vision, Visual Processing, and Digital Display IV*, February 1993. Bellingham, WA.

[82] J. J. Eggers, J. K. Su, and B. Girod. A Blind Watermarking Scheme based on Structured Codebooks. *IEE Colloquium: Secure Images and Image Authentication*, April 2000. London, UK.

[83] I. Shterev, R. Lagendijk, and R. Heusdens. Statistical Amplitude Scale Estimation for Quantization-based Watermarking. *SPIE Security, Steganography, and Watermarking of Multimedia Contents VI*, 5306, January 2004. CA, USA.

[84] T. Kawata. *Fourier Analysis in Probability Theory.* Academic Press, 1972.

[85] M. L. Miller, G. J. Doerr, and I. J. Cox. Applying Informed Coding and Embedding to Design a Robust, High Capacity Watermark. *IEEE Transactions on Image Processing*, 13(6):792–807, June 2004.

[86] R. Zamir and M. Feder. On Lattice Quantization Noise. *IEEE Transactions on Information Theory*, 42:1152–1159, July 1996.

[87] R. Zamir, S. Shamai (Shitz), and U. Erez. Nested Linear/Lattice Codes for Structured Multiterminal Binning. *IEEE Transactions on Information Theory*, 48(6):1250–1276, June 2002.

[88] B. Bradley. Improvement to CDF Grounded Lattice Codes. *SPIE Security, Steganography, and Watermarking of Multimedia Contents VI*, 5306, January 2004. CA, USA.

[89] I. D. Shterev and R. L. Lagendijk. Maximum Likelihood Amplitude Scale Estimation for Quantization-Based Watermarking in the Presence of Dither. *SPIE Security, Steganography, and Watermarking of Multimedia Contents VII*, January 2005. San Jose, CA.

[90] R. Koetter, A. C. Singer, and M. Tuchler. Turbo Equilization. *IEEE Signal Processing Magazine*, 21(1):67–80, January 2004.

[91] F. Balado, K. M. Whelan, G. C. M. Silvestre, and N. J. Hurley. Joint Iterative Decoding and Estimation for Side-Informed Data Hiding. *IEEE Transactions on Signal Processing*, 53(10):4006–4019, October 2005.

[92] M. Whelan, F. Balado, G. C. M. Silvestre, N. J. Hurley. Iterative Estimation of Amplitude Scaling on Distortion-Compensated Dither Modulation. *SPIE Security, Steganography, and Watermarking of Multimedia Contents VII*, January 2005. San Jose, CA.

[93] K. Whelan, G. Silvestre, and N. Hurley. Iterative Decoding of Scale Invariant Image Data-Hiding. *IEEE International Conference on Image Processing*, September 2005. Genoa, Italy.

[94] L. Schuckman. Dither Signals and Their Effect on Quantization Noise. *IEEE Transactions on Communication Technology*, 12(4):162–165, December 1964.

[95] K. Chandrasekharan. *Introduction to Analytic Number Theory.* Springer-Verlag, 1968.

[96] S. Gazor and W. Zwang. Speech Probability Distribution. *IEEE Signal Processing Letters*, 10(7):204–207, July 2003.

[97] R. L. Lagendijk and I. D. Shterev. Estimation of Attacker's Noise and Variance for QIM-DC Watermark Embedding. *IEEE International Conference on Image Processing*, October 2004. Singapore.

[98] F. Perez-Gonzalez. M. Barni, A. Abrardo, and C. Mosquera. Rational Dither Modulation: A Novel Data Hiding Method Robust to Valumetric Attacks. *IEEE International Workshop on Multimedia Signal Processing*, September 2004. Siena, Italy.

[99] F. P. Gonzalez, C. Mosquera, M. Alvarez, and R. Lagendijk. High-Rate Quantization Data Hiding Robust to Arbitrary Linear Filtering Attacks. *SPIE Security, Steganography, and Watermarking of Multimedia Contents VIII*, January 2006. San Jose, CA.

[100] J. Wang, I. D. Shterev, and R. L. Lagendijk. Scale Estimation in Two-Band Filter Attacks on QIM Watermarks. *SPIE Security, Steganography, and Watermarking of Multimedia Contents VIII*, January 2006. San Jose, CA.

[101] J. Wang, I. D. Shterev, and R. L. Lagendijk. Two-Band Amplitude Scale Estimation for Quantization-Based Watermarking. *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, December 2005. Hong Kong.

[102] P. M. T. Broersen. A Comparison of Transfer Function Estimators. *IEEE Transactions on Instrumentation and Measurement*, 44(3):657–661, June 1995.

[103] F. J. Harris. On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform. *Proceedings IEEE*, 66(1):51–83, January 1978.

[104] A. Stuart and K. Ord. *Kendall's Advanced Theory of Statictics: Distrubution Theory*, volume 1. Oxford University Press, sixth edition, 1994.

Acknowledgements

The author wishes to thank Prof. R. L. Lagendijk, many of whose suggested improvements have been incorporated into the thesis.

Thanks to J. Wang for productive cooperation and Y. Zhao, A. Rares, O. A. Niamut, and J. Erkelens for helpful technical discussions.

This research was supported by the Technology Foundation STW, Applied Science Division of NWO, and the Technology Program of the Ministry of Economic Affairs.

Author's Biography

Ivo Shterev was born in Plovdiv, Bulgaria, on 30th of April, 1976. In September, 1994 he joined the Technical University of Sofia at Plovdiv. He graduated in July, 1999 with a Master's of Science in Electronics. After finishing military service in October, 2000 he applied at Delft University of Technology, the Netherlands. In October, 2001 he started his graduate studies in the Faculty of Electrical Engineering, Mathematics, and Computer Science. In the summer of 2004 he did an internship in the Electrical and Computer Engineering Department of University of Illinois at Urbana-Champaign, USA.