ABSTRACT

ON THE DATA HIDING THEORY AND MULTIMEDIA CONTENT SECURITY APPLICATIONS

by
Litao Gang

This dissertation is a comprehensive study of digital steganography for multimedia content protection. With the increasing development of Internet technology, protection and enforcement of multimedia property rights has become a great concern to multimedia authors and distributors. Watermarking technologies provide a possible solution for this problem.

The dissertation first briefly introduces the current watermarking schemes, including their applications in video, image and audio. Most available embedding schemes are based on direct Spread Sequence (SS) modulation. A small value pseudo random signature sequence is embedded into the host signal and the information is extracted via correlation. The correlation detection problem is discussed at the beginning. It is concluded that the correlator is not optimum in oblivious detection. The Maximum Likelihood detector is derived and some feasible suboptimal detectors are also analyzed. Through the calculation of extraction Bit Error Rate (BER), it is revealed that the SS scheme is not very efficient due to its poor host noise suppression. The watermark domain selection problem is addressed subsequently. Some implications on hiding capacity and reliability are also studied. The last topic in SS modulation scheme is the sequence selection. The relationship between sequence bandwidth and synchronization requirement is detailed in the work. It is demonstrated that the white sequence commonly used in watermarking may not really boost watermark security.

To address the host noise suppression problem, the hidden communication is modeled as a general hypothesis testing problem and a *set partitioning* scheme is proposed. Simulation studies and mathematical analysis confirm that it outperforms the SS schemes in host noise suppression. The proposed scheme demonstrates improvement over the existing embedding schemes.

Data hiding in audio signals are explored next. The audio data hiding is believed a more challenging task due to the human sensitivity to audio artifacts and advanced feature of current compression techniques. The human psychoacoustic model and human music understanding are also covered in the work. Then as a typical audio perceptual compression scheme, the popular MP3 compression is visited in some length. Several schemes, amplitude modulation, phase modulation and noise substitution are presented together with some experimental results. As a case study, a music bitstream encryption scheme is proposed. In all these applications, human psychoacoustic model plays a very important role. A more advanced audio analysis model is introduced to reveal implications on music understanding. In the last part, conclusions and future research are presented.

ON THE DATA HIDING THEORY AND MULTIMEDIA CONTENT SECURITY APPLICATIONS

by
Litao Gang

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Electrical Engineering

Department of Electrical and Computer Engineering

January 2002

APPROVAL PAGE

ON THE DATA HIDING THEORY AND MULTIMEDIA CONTENT SECURITY APPLICATIONS

Litao Gang

Dr. Ali N. Akansu, Dissertation Advisor — Date
Professor of Electrical and Computer Engineering, NJIT

Dr. Nirwan Ansari, Committee Member — Date
Professor of Electrical and Computer Engineering, NJIT

Dr. Richard Haddad, Committee Member — Date
Professor of Electrical and Computer Engineering, NJIT

Dr. Mahalingam Ramkumar, Committee Member — Date
Chief Technology Officer, AVWAY, IDT Corp. Newark, NJ

Dr. Yun-qing Shi, Committee Member — Date
Associate Professor of Electrical and Computer Engineering, NJIT

BIOGRAPHICAL SKETCH

Author: Litao Gang

Degree: Doctor of Philosophy

Date: January 2002

Undergraduate and Graduate Education:

- Doctor of Philosophy in Electrical Engineering,
  New Jersey Institute of Technology, Newark, NJ 2002
- Master of Science in Electrical Engineering,
  Beijing Institute of Technology, Beijing, P.R.China 1996
- Bachelor of Engineering in Electrical Engineering,
  Beijing Institute of Technology, Beijing, P.R.China 1993

Major: Electrical Engineering

Publications and Presentations:

Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar
"Linear and nonlinear modulation in oblivious data hiding", submitted to the *IEEE Transactions on Signal Processing*, Dec. 2001

Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar
"Set partitioning in oblivious steganography", submitted to the *IEEE Signal Processing Letters*, Nov., 2001.

Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar
"Robust data hiding in audio signals", submitted to the *IEE Electronics Letters*, Oct., 2001.

Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar
"Security and synchronization in watermark sequence", submitted to *IEEE ICASSP'2002*, Florida.

Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar
"Periodic signaling scheme in oblivious data hiding", *Proc. 34th Asilomar Conference on Signals, Systems, and Computers*, California, pp. 1851-1855, November, 2000.

Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar
"MP3 resistant oblivious steganography", *Proc. ICASSP'2001*, Utah, pp. 1365-1368, May, 2001.

Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar
"Set partitioning in oblivious data hiding", *Proc. ICASSP'2001*, Utah, pp. 1985-1988, May, 2001.

Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar
"Nonlinear modulation in oblivious steganography", *Proc. IEEE-ERUASIP Nonlinear Signal and Image Processing'2001*, Maryland, June, 2001

Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar
"Performance analysis of spread spectrum modulation in data hiding", accepted in *SPIE 2001*, California, July, 2001.

Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar
"Transform selection in steganography", *Proc. Annual Conference on Information Sciences and Systems*, Maryland, pp. 450-453, March, 2001.

Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar
"On-line music protection and MP3 compression", *Proc. International Symposium on Intelligent Signal Information Multimedia Processing'2001*, Hongkong, pp. 13-16, May, 2001.

Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar
"Random sequence selection in watermarking", accepted in *International Conference on Imaging Science, Systems, and Technology*, Neveda, July, 2001.

To the watermarking and steganography community

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my advisor, Prof. Ali N. Akansu. During my Ph.D. study, Prof. Ali N. Akansu has guided my study and research in this very interesting and fast-developing field. With his guidance, I did not lose my way; and with his encouragement, I did not lose heart. I also thank Prof. Akansu as Director for NJCMR of creating such a good environment to work.

I would like to thank Dr. Ramkumar who is a pioneer and expert in the watermarking community. He gave me lots of help through my research. Discussion with him is always constructive, and saved me lots of time groping in darkness.

I extend my sincere appreciation to Professors Richard Haddad, Nirwan Ansari and Yun-Qing Shi for their time and stimulating suggestions. Actually, it is Dr. Haddard's signal processing classes EE640, EE740 that built solid foundation for my PhD work. Drs. Ansari and Shi gave me much consistent help, administratively and academically, throughout my study in NJIT.

I also thank the fellow researchers in our laboratory, for all the generous help they have offered towards completion of this dissertation. I would especially like to thank Taha, Feihong, Xiaodong, Xuefei, Surong, Jie Yang, Bin He, Zhen Zhang, Jian Ye, Jiongkuang, Burak for their support.

Finally, I would like to thank Dr. Ronald Kane and Ms. Clarisa Gonzalez-Lenahan in the Graduate Studies in NJIT for their time and assistance.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter Page

TABLE OF CONTENTS
(Continued)

Chapter Page

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES
(Continued)

CHAPTER 1

INTRODUCTION

1.1 Digital Steganography and its Application

With the rapid ever-growing Internet technology, multimedia content protection has raised great concerns in multimedia creation and delivery community. The traditional data encryption protection methods, such as RSA, DES [76, 80, 85] do not meet all the requirements. Recent years have seen lots of research and application of watermarking or steganography technology to address this issue.

Steganography is the art to embed some message in a *host* (or *cover*) signal without noticeable perceptual degradation. The technologies provide a potential solution to the content protection problem. The message could be copyright information or product information, for ownership proof or copyright infringement tracking.

Two basic requirements for staganography is *transparency* and *robustness*. The former means the perceptual value of the cover signal should not degrade after information embedding. Robustness implies the watermark should not be removed easily. It is particularly important to achieve robustness against the popular compression schemes because of their ubiquitous applications in network transmission and storage. To provide a convincing proof for proprietary copyright, the message should be smartly integrated with the content signal and sufficiently robust against compressions and other signal processing (even malicious pirate attacks). Watermark is supposed to be highly tamper-proof in these applications.

Based on the different application scenarios, steganography can be divided into two categories, *oblivious* applications where the original cover signal is absent at the information extraction, and *escrow* applications where decoding is performed with

the assistance of the original cover signal. Because the original cover signal is not available in most scenarios, it is not strange to see that the current research and development is focused on oblivious applications.

Watermarking for content protection is a specific application of hidden communication. Steganography is usually a more general term which refers to delivering a message in a manner that the very existence of the message itself is kept secret. Watermarking can be considered a subset of steganography [46]. In this dissertation, "data hiding" refers to the general information embedding in the cover signal whereas "watermarking" refers to the hiding with emphasis on robustness and security.

As a kind of digital steganography technology, watermarking is different from data encryption. One prominent difference between data encryption and steganography is that the latter does not prevent unauthorized access. Encryption, on the other hand, does not permit unauthorized access to the contents. Once the encrypted contents are correctly unscrambled, the protection is completely removed. In contrast, a robust watermark is integrated with the content signal all the time and is very difficult (if not impossible) to remove.

Steganography has long been modeled as a communication problem. Figure 1.1 and Figure 1.2 depict the channel models in escrow and oblivious applications. The transmission channel noise is compression noise, or other noises incurred in signal processing procedures.

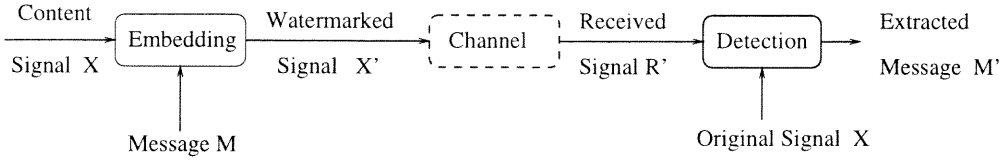


Figure 1.1 Escrow steganography

Watermarking used for ownership proof is often referred to as *robust watermarking* which is aimed at copyright protection. Another form of watermark is called

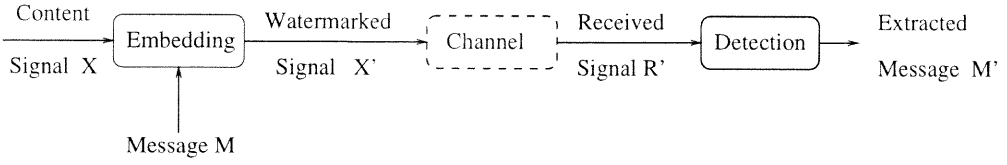


Figure 1.2 Oblivious steganography

fragile watermarking (or *semi-fragile watermarking*) which targets at multimedia authentication scenarios [25, 57, 66, 102]. In this application, a watermark is embedded into a host signal as a signature of authentication for the content. If the watermark can not be recovered correctly at the receiver side, that means this content has been manipulated. A typical application is proposed to authenticate the photo image in a digital filmless camera [26].

Besides authentication and copyright protection, data hiding also finds its way in other applications. NEC, IBM and other companies have proposed and implemented schemes for DVD copy (not copyright) protection control [8, 59, 60]. Lucent, Philips etc. are deploying their watermarking product for broadcast monitoring. Some companies, for example, RealPlayer plans to integrate data hiding into their multimedia players.

In the digital steganography community, lots of attention is paid to image data hiding. One of the earliest watermarking schemes is LSBM (Least Significant Bit Manipulation). In this scheme, the LSB bits are modified according to a predefined pattern to carry messages (for example, modify its parity). Obviously, this scheme is quite crude and does not resist compression and other unintentional attacks, let alone pirate attacks.

In current watermarking schemes, a most influential one is Spread Spectrum (SS) modulation approach. The idea is borrowed from spread-spectrum radio communications. Cox *et al.* [14, 16] are among the earliest to apply the scheme in image watermarking. In one of his schemes, a Gaussian distributed PN sequence

is inserted into the block DCT coefficients. This idea can be applied to different watermark domains. Some work in the whole image DCT domain instead of block DCT domain [15], Discrete Fourier Transform (DFT) phase domain [77], DFT amplitude domain [78], wavelet domain [49, 96] etc. Various Human Visual System (HVS) models are explored to minimize the visual artifacts. Some models work in the pixel domain [90, 99], while many others work in transform domains [18, 92].

There are few publications on video watermarking applications. Most watermarking schemes regard video as a consecutive sequence of still images. The image steganography schemes could be applied to video, where embedding and detection is done on a frame by frame basis. This is the direct extension of image watermarking to the uncompressed (raw) video. An alternative is to embed information in the compressed domain. Jordan *et al.* [24] proposed a scheme to embed the watermark signature into the motion vectors. The author claims survivability to the MPEG compressions. Hartung *et al.* [36, 37, 39] suggested a SS method extension in video. The message extraction computation complexity in video should not be too high due to the real-time processing requirement. Another requirement is the robustness against MPEG-x compressions, frame dropping, frame averaging and other attacks.

There are even fewer publications on audio steganography which is regarded as the more challenging task. Generally speaking, audio signals have much less samples than video. Although this reduces the processing complexity, it limits the hiding capacity.

It is believed that Human Audible System (HAS) is much more sensitive to the artifacts than the Human Visual System (HVS). The embedding distortion inaudibility is more difficult to control. The general audio signal can not be regarded stationary (at most be assumed semi-stationary). Fortunately, through subjective tests and theory studies, people have accumulated extensive knowledge on HAS and human perception models. Several models have been explored and successfully

employed in perceptual compression, for instance MPEG-1 MP3 compression [44], Dolby Audio Encoder AC-3 [19]. This knowledge should be employed in data hiding to minimize artifacts.

In audio steganography, compression-resilience is also more difficult to achieve as the current audio coding algorithms are more advanced. In image and video compression, the algorithms are mainly based on waveform approximation. The compression schemes usually target at MSE (Mean Square Error). i.e., maximizing SNR (Signal-Noise Ratio). The popular compression schemes, SPIHT (Set Partitioning Image Hierarchical Tree) [79], EZW (Embedded Zero Wavelet) [82] focus on quantization and encoding technique after transform or subband filtering. It is expected that the compression noise is not high, the compressed waveform is still close to the uncompressed original signal. In the perceptual audio compression schemes, the purpose is to minimize the perceptual degradation, not MSE. Since it is known that human beings do not measure audio quality in MSE sense. The compression noise could be much higher. In the more advanced audio parametric coding schemes, such as MPEG-4 HILN [1] and the model proposed in [56, 94, 95], audio signal is analyzed and some important parameters are extracted to "describe" the original signal. The reconstructed output at decoder with these parameters may not be close to the original signal. Compression quality can not measured in common sense SNR (Signal-Noise Ratio).

In current audio data hiding schemes, some are the direct extension of the basic SS scheme to audio signal. A random signature sequence is embedded in subband domain [4, 42], cepstra domain [55] or time domain [6] etc. The human perceptual model is explicitly used to shape the watermark signal spectrum according to the masking curve [9]. This is an effective method in controlling the watermark audibility.

1.2 Organization of the Dissertation

In this research, steganography is studied as a general hidden communication problem. In this dissertation, data hiding schemes are investigated from the perspective of signal processing and new algorithms with improved performance over the existing ones are studied. Other data hiding issues are also comprehensively explored, with some concentration on audio applications. Other related topics in multimedia watermarking application are also discussed.

In Chapter 2, the history and development of digital steganography is briefly reviewed, some influential data hiding algorithms in still image, video and audio signals are visited. In practice, the Spread Spectrum (SS) schemes are widely employed. In Chapter 3, a thorough study of the current SS algorithms in data hiding is presented. The hiding capacity is analyzed. The problem of watermark domain selection and its impact on compression robustness is also addressed. Optimum detection in oblivious applications is explored. It is found that the correlation detector is not optimal in oblivious applications. A new scheme is derived and its performance is compared with the existing ones. The random sequence security and watermark signature generation is also covered in Chapter 3.

From the analytical and simulation results, it is concluded that the spread spectrum modulation although effective in escrow applications, is not quite successful in oblivious applications. In Chapter 4, a new scheme *set partitioning* is proposed in oblivious applications. In the mathematical analysis and simulation studies, improvement is demonstrated over the existing schemes, especially in very noisy environment.

Chapter 5 is dedicated to the audio compression-resistant data hiding. Initially several different audio compression schemes are visited, including waveform approximation, perceptual coding and parametric coding. Because of the popular deployment of MPEG-1 layer III (MP3) compression in Internet transmission and

storage, as an example, this compression and human perceptual psychoacoustic model is briefly introduced. Based on this, three hiding schemes, amplitude modulation, phase modulation and noise substitution are proposed. A music encryption scheme is proposed. The last section in Chapter 5 is devoted to the advanced audio analysis model and its impacts on audio watermarking. Chapter 6 covers some outstanding problems in watermarking and data hiding applications. Conclusions are drawn and the future work is proposed.

CHAPTER 2

CURRENT TECHNIQUES AND STATUS

In this chapter, the existing technologies used in still image, video and audio content signals are briefly visited.

2.1 Steganography Applications in Images

The first application of digital watermarking is on still images. One of the earliest data hiding schemes embeds information in pixel's Least Significant Bits (LSB), called LSB Manipulation (LSBM) [93, 98]. It is obvious that LSBM is not robust against compression and other attacks. Spread Spectrum (SS) modulation algorithms embed a small value PN sequence in the selected components of the content signals [7, 16]. It provides much improvement over LSBM on security and robustness. Most of the current watermarking schemes are SS-based, including both escrow and oblivious applications. These schemes can be put into two categories, spatial domain embedding and transform domain embedding.

2.1.1 Pixel Domain Embedding

Bender *et al.* [7] propose a data hiding scheme called "Patchwork". In this scheme, pixel values a_i and b_i in a randomly selected two-pixel pair are increased and decreased respectively by a very small value δ. For an unmarked original image, it can be assumed that the pixel value x_i is a random variable with zero mean. Therefore,

$$\sum_{i=0}^{N-1} (a_i - b_i) \approx 0. \tag{2.1}$$

This makes intuitive sense since the number of times a_i is greater than b_i should be offset by the number of times the reverse is true.

After watermarking, the detection output is

$$\sum_{i=0}^{N-1}[(a_i+\delta)-(b_i-\delta)] = 2\delta N + \sum_{i=0}^{N-1}(a_i-b_i). \tag{2.2}$$

In a watermarked image, the mathematical expectation value of (2.2) deviates from 0. If the pixel pair number N is sufficiently large, a reliable decision based on the statistical sum can be made.

This method can be easily extended to data hiding. Embedding procedure is

$$\begin{cases} a_i' = a_i + \delta, \quad b_i' = b_i - \delta; & \text{bit value 1 embedded} \\ a_i' = a_i - \delta, \quad b_i' = b_i + \delta; & \text{bit value 0 embedded,} \end{cases} \tag{2.3}$$

where a_i' and b_i' are the pixel values after bit embedding.

Detection is based on sum of these pixels,

$$q = \sum_{i=0}^{N-1}(\hat{a}_i - \hat{b}_i), \tag{2.4}$$

where $\hat{a}_i$ and $\hat{b}_i$ are the received pixel values after channel transmission. If $q > 0$, the decision is bit value 1; Otherwise, the decision is bit value 0 instead.

Bender *et al.* [7] analyzed the extraction bit error probability and the impact of different patch shapes on robustness. Pitas *et al.* [62] proposed a quite similar method. Further extension can be found in [53, 54]. The advantage of spatial domain scheme is its efficiency and low computation cost. The shortcoming is that the pixel number N should be sufficiently large which limits the hiding capacity.

To reduce the watermark visibility, extra work should be done to control the visual artifacts. Macq *et al.* [20] proposed a scheme to make the watermark adaptive to the Human Visual System (HVS). For a color image, it is well known that human being are most insensitive to the blue component. Kutter *et al.* [50, 51] took use of this property and embed information into the blue component. It is claimed that the visible distortion is thus minimized.

2.1.2 Transform Domain Embedding

Many Spread Spectrum modulation schemes are applied in transform domains. As a case study, the approach proposed by Cox *et al.* [16] is revisited in the following discussion.

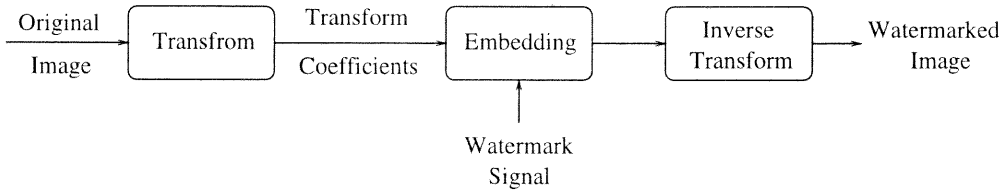


Figure 2.1 Transform domain embedding framework

After transform, some appropriate coefficients x_i in this domain are selected. They are usually the medium frequency coefficients to which humans are not so sensitive. A randomly generated signature sequence s_i is embedded into x_i,

$$x_i' = x_i + s_i. \tag{2.5}$$

In [16], the signature sequence used is Gaussian distributed due to its enhanced security over the bipolar sequence (−1 or +1 bi-value sequence).

Denote the received sequence after channel transmission as $\mathbf{r}$. With the assistance of the original sequence $\mathbf{x}$, the correlation detector output is

$$q = \sum_{i=0}^{N-1} (r_i - x_i) \cdot s_i = \sum_{i=0}^{N-1} (s_i + n_i) \cdot s_i, \tag{2.6}$$

where n_i is the channel noise.

Correlation detector is optimal only if n_i is Gaussian distributed. It is worth noting here that often the watermarking channel is far from Gaussian type. However, correlation is a feasible method and mostly used in watermark verification due to its simplicity.

To control the artifacts introduced in this approach, several formulae are suggested to determine the embedding extent [16],

$$x_i' = x_i + \alpha x_i, \tag{2.7}$$

$$x_i' = x_i(1 + \alpha x_i), \tag{2.8}$$

$$x_i' = x_i(e^{\alpha x_i}), \tag{2.9}$$

where α is a scaling factor.

For different frequency bins, the value of α should be adaptive to the human sensitivity. Very similar HVS model was suggested in [67]. The visual threshold value may be obtained from a visual perceptual model or empirical experiments. This principle is also applied in other transform domains. For example, [49, 101, 103] apply the wavelet domain embedding as a direct variation of this scheme.

In transform domain steganography, the transform selection problem has not been answered. An ideal watermark transform should be superior in performance and low in computation complexity. This problem is addressed in Chapter 3.

2.1.3 DFT Domain Embedding

In image data hiding applications, it is important to achieve robustness against geometric distortions, for example, translation, rotation, and scaling. Scanning procedure can be modeled as a combination of these distortions. There does not exist an ideal solution to countermeasure these attacks. A heuristic approach is to embed a fixed pattern into images and at decoder try to estimate the values of rotation, scaling and translation by pattern match and then compensate for these changes. This is usually done via the simple brute force exhaustive search and therefore a quite computation extensive procedure.

Ruanaidh *et al.* [78] proposed a novel scheme to hide information in DFT amplitude domain. It is translation resistant because the spatial shift is only reflected

in the DFT phase, which is not used in embedding. Its effectiveness has been demonstrated in practice.

Another important feature of DFT amplitude embedding is that the DFT amplitude is perceptually insignificant. It was pointed that the DFT phase contains much more information than the DFT amplitude [64]. As the current popular image compression schemes aim at waveform approximation, the compression noise is relatively small. In the perceptually insignificant DFT amplitude domain, more watermark energy is permitted without much visible artifacts. This results in relative higher SNR. Ramkumar and Akansu [71, 73, 75] pointed out the advantage in DFT domain embedding. Their simulation results demonstrate the robustness in face of various compression schemes.

The above conclusion seems to contradict the long-held view that the watermark should be embedded into the perceptually *significant* components [14]. That is true embedding in DFT amplitude is not tamper resistant. A smart attacker can also use the property of DFT amplitude insignificance to inject more attack noise. The DFT domain embedding just takes advantage of the current compression schemes. Ruanaidh *et al.* [77] also suggested to embed information in the perceptually significant DFT phase. It claims enhanced security against malicious attacks.

The pixel domain and transform domain selection have different impacts on the robustness and complexity. Needless to say, the spatial domain embedding is less computation extensive. Some studies show that the transform domain approaches are more robust to geometric distortion.

2.2 Video Steganography

2.2.1 Embedding in Raw Uncompressed Video

If a video signal is regarded as a continuous still image sequence, the image data hiding scheme can be employed frame by frame to a video sequence. In fact, most of the current approaches are the direct extension of image embedding schemes.

Hartung *et al.* [37, 38, 39] directly extended the image data hiding in the video signal. A pseudo-noise bipolar sequence $\mathbf{p}$ (p_i is either $+1$ or -1) is embedded into the selected 8x8 DCT coefficients v_i. In their approach, a random sequence is embedded into these coefficients.

$$v_i' = v_i + p_i \alpha_i, \tag{2.10}$$

where α_i is the locally adjustable amplitude factor which varies according to the local properties of the video signal. The spatial and temporal masking phenomena of HVS can be applied in embedding. The message is retrieved via correlation at decoder. Their experiments demonstrate the typical hiding capacity is up to 50 bits/sec.

Swanson *et al.* [89, 91] proposed a multi-scale watermarking method. First, the video sequence is segmented into scenes. Then a temporal wavelet transform is applied to each video scene, and temporal low-pass and high-pass frames are obtained. The watermark signal then is embedded into both frames. After inverse transform the watermarked video is obtained. Note the watermark is also embedded in the low-frequency components. To minimize artifacts visibility, a HVS model is exploited in this approach. An efficient watermark embedding is "Millennium" proposed by Digimarc, Philips and Macrovision [13] for DVD copy protection control. Its advantage lies in simplicity and translation invariance.

2.2.2 Embedding in Compressed Video

Instead of embedding directly in the raw video, data can also be embedded in the compressed domain. In both of MPEG-1 and MPEG-2 compression standards, the block Discrete Cosine Transform (DCT) is used. The video sequence is composed of I, P, and B frames. In I (Intra-coded) frames, the picture is split into 8x8 blocks. Then the DCT coefficients are quantized, zig-zag reordered and Huffman encoded in a similar fashion used in JPEG compression. In the inter-coded frames (P or B), the pictures are encoded using forward or backward prediction. The motion vectors and residual prediction error are quantized and encoded.

Hartung *et al.* also experimented their embedding scheme (2.10) in compressed domain [39]. This procedure is applied to every frame, including I, P, B frames. For each compressed frame, the watermark signal is added to the 8x8 DCT coefficients in the video bitstream. Experimental results demonstrate its robustness against standard signal processing.

Jordan *et al.* [24] suggested a very interesting approach to embed information in the motion vectors in the compressed bitstream. As motion vectors are significant perceptually, only areas with less activity are selected for embedding. The authors claims artifact invisibility. The information is directly retrieved from the motion vectors in the compress video. The greatest advantage of this scheme is its low complexity.

2.3 Audio Steganography

In comparison, there are few publications on audio data hiding. Usually, the audio steganography is assumed a more challenging task. One reason is the human beings are more sensitive to watermark distortion. Another reason is the current audio coding technique is much more advanced than the schemes used in image coding, making the robustness to those compressions more difficult.

2.3.1 Echo Hiding

Bender *et al.* [7] suggested an innovative scheme to hide information in audio. This method adds a decayed version of the original signal to itself. The echo is determined by three parameters: initial amplitude, decay rate, and offset (the delay when the echo appears). Usually, the decaying curve used is exponential.

Informal tests show that with appropriate parameters, the echo added is inaudible, only making the original "richer". The information bit can be embedded by selecting different offset values. In Figure 2.2, to embed bit value 0, the offset value is x while the offset value is $x + \delta$ if bit value 1 is embedded.

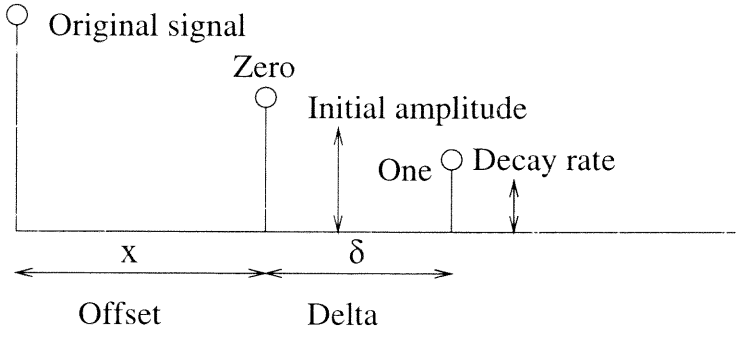


Figure 2.2 Audio echo hiding

Data extraction is via measuring different offset delay value. First, the cepstrum of the embedding output is calculated. Then the autocorrelation of the ceptrum is obtained. With the echoes spaced periodically every x or $x + \delta$, a peak at x or $x + \delta$ in the cepstrum can be obtained. The decision rule is to examine the power level at x and $x + \delta$ and choose whichever bit corresponds to a higher power level.

2.3.2 Other Schemes

The SS modulation can be extended to audio applications. Pitas *et al.* [6] repeated the PN sequence embedding in the time domain. Tewfik *et al.* [9] proposed an embedding scheme in frequency domain. One contribution of the algorithm is

that the explicit human psychoacoustic model is employed to shape the watermark spectrum. The watermark distortion is kept under the masking for distortion inaudibility. In embedding, a watermark signature spectrum is shaped by a filter which is derived from the psychoacoustic model analysis. Some similar schemes were proposed in [4, 42]. The detail of the human psychoacoustic model is given in length in Chapter 5.

There are some other SS versions in other domains. In [55], the embedding and detection was performed in the cepstrum domain. The shortcoming is that in that domain, distortion inaudibility is more difficult to control. Other approaches include data hiding by time-domain modification [58], or by compressed domain manipulation. An example of the latter is the simple scheme modifying the scale factors in MP3 bitstream. [69].

2.4 Application and Product Deployment

Since 1994, the digital steganography technologies have attracted lots of attention both in the industry and academia. Nowadays there are several commercial products on market and some have been deployed in practice.

Digimarc Corp. is a leader in watermarking technology. Products it provides include **ImageBridge** and **MediaBridge** for image watermarking. Besides, Digimarc **MarcSpider** image tracking can crawl the World Wide Web searching for digitally watermarked images to find illegal publications of copyright images. Some corporations have already entered into contracts with the Digimarc company for the use of **PictureMarc** and **MarcSpider**, to protect their interests in digital images. Even Digimarc Corp. itself admits the watermark is vulnerable to common signal processing attacks. There are some other companies providing similar steganography products. For example, Signum Technologies offers *SureSign* watermark product for

content protection. Their product works together with Adobe Photoshop. MediaSec Technologies provides MediaLabel and SysCop products.

Cognicity Corp. provides audio watermarking product **AudioKey**. It claims its robustness to all popular audio compression schemes. Up to 2 layer information can be embedded, reaching total hiding rate of 42 bits per second. The watermark is imperceptible due to the use of psychoacoustic models of human hearing. It also claims robustness to music editing, format conversion (including D/A and A/D conversions), compression, streaming, broadcast, etc. This product was adopted to integrate with the AT& T perceptual audio encoder and RealPlayer's Producer Plus G2 encoder.

Secure Digital Music Initiative (SDMI) is initiated by several label companies to prevent illegal CD copying. They want to standardize the SDMI-compliant players which can play unprotected music and new SDMI-protected music that has been legitimately acquired. A proposal by Verance Corp., a company aiming exclusively at audio watermarking, was adopted by SDMI as Phase I standard. In DVD copy protection applications, there exists two major proposals, one is *Galaxy Group* proposed by Hitachi, IBM, NEC, Pioneer Electronic, and Sony, while Philips, Macrovision, Digimarc unify and offer their *Millennium Group* products. In the near future, more products will be seen on the market together with more applications of data hiding technology.

CHAPTER 3

SPREAD SPECTRUM MODULATION IN STEGANOGRAPHY

In the previous chapters, the Spread Spectrum (SS) modulation scheme in steganography is reviewed. In this chapter, this technique is discussed in detail, including its information extraction and watermark domain selection. A new algorithm is proposed and compared with the existing ones. Its effectiveness is demonstrated in the analysis and simulation studies.

3.1 Hiding Capacity and Watermark Domain Selection

In the SS hiding schemes as indicated by (2.5) and (2.6), the principle is very simple, viz , to superimpose a small value random sequence into the original coefficient sequence. This idea can be applied in different domains, wavelet domain, DCT or spatial domains. What decomposition should be used in watermarking? Which transform, high G_{TC} or low G_{TC} (Gain of Transform Coding) is more advantageous? To be resilient to a specific compression, is it necessary to match the decomposition used in the compression?

3.1.1 Hiding Capacity in Different Watermark Domains

In the following analysis, the simple superposition algorithm is studied,

$$x_i' = s_i + x_i, \tag{3.1}$$

where $\mathbf{s}$ is the watermark signal sequence.

In the SS modulation, it is obvious that escrow applications will reach higher capacity than oblivious ones. Moulin *et al.* [61] pointed out at least in theory oblivious application could achieve the same capacity as in escrow ones. Nevertheless this is not realizable in practice.

In escrow applications, if the attack noise w is Gaussian distributed, $w \sim N(0, \sigma^2)$. The hiding capacity on one coefficient is simply given by

$$C = \frac{1}{2}\log(1 + \frac{\sigma_s^2}{\sigma^2}), \tag{3.2}$$

where σ_s^2 is the watermark signal energy. The capacity is achieved when s is Gaussian distributed, $s \sim N(0, \sigma_s^2)$.

Besides processing attack noise, the cover signal itself is also regarded as noise (*host noise*) in oblivious cases. This noise is much larger than the channel noise. Both of these two noises should be considered in the capacity calculation. Host noise is usually non-Gaussian distributed. Ramkumar and Akansu [74] used an information transformer to convert the noise to Gaussian distribution. The parallel channel model depicted in Figure 3.1 is used in their capacity calculation.

Suppose a watermark signal coefficient s_i is embedded in the cover signal coefficient x_i, and the processing noise is p_i. Assume all the noise is Gaussian distributed. In the ith channel, the capacity is calculated as

$$C_i = \frac{1}{2}\log(1 + \frac{\sigma_{si}^2}{\sigma_{xi}^2 + \sigma_{pi}^2}), \tag{3.3}$$

where σ_{pi} is the channel noise variance, $s_i \sim N(0, \sigma_{si}^2)$ and $x_i \sim N(0, \sigma_{xi}^2)$.

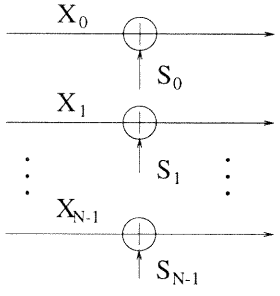


Figure 3.1 Parallel channel model

Because $\sigma_{xi} \gg \sigma_{pi}$, the channel processing noise effect is neglected for simplicity in the following capacity calculation.

The total capacity is the sum of the capacities in these N parallel channels

$$C \approx \frac{1}{2}\sum_{i=0}^{N-1}\log(1+\frac{\sigma_{si}^2}{\sigma_{xi}^2}). \tag{3.4}$$

For embedding in spatial (or time) domain, or other low G_{TC} transform domains, the parallel channel assumption may not be appropriate. "Parallel" implies the noise x_i is independent. Whereas in spatial or time domain, there is strong correlation between x_i and x_{i-1} (for images, correlation coefficient ρ is usually larger than 0.9). The correlation means that the host noise is, more or less, "predictable". Therefore, the correlated channel is not as "harmful" as the parallel channel. More information may be transmitted through the correlated channel. For example, based on (3.3), if 0.1 bit can be hidden on one pixel, on a 256x256 image more than 256x256x0.1 bits can be hidden.

In fact transform does not change the entropy of the host noise $\mathbf{x}$. Suppose $\mathbf{x}$ is Gaussian distributed, but not necessarily independent. Its entropy is

$$H_1(\mathbf{x}) = -\int f(\mathbf{x})\log[f(\mathbf{x})]d\mathbf{x}, \tag{3.5}$$

where $f(\mathbf{x})$ is the pdf of $\mathbf{x}$,

$$f(\mathbf{x}) = \frac{\mathbf{1}}{(\mathbf{2}\pi)^{\mathbf{N/2}}|\mathbf{C}|^{\mathbf{1/2}}}\exp(-\frac{\mathbf{1}}{\mathbf{2}}\mathbf{x^T C^{-1}x}), \tag{3.6}$$

where $\mathbf{C} = \mathrm{var}(\mathbf{x})$ is the variance matrix of $\mathbf{x}$.

If the transform kernel is $\mathbf{A}$, transform output is obtained as $\mathbf{y} = \mathbf{Ax}$. The variance matrix of $\mathbf{y}$ is

$$\mathbf{D} = \mathrm{var}[\mathbf{y}] = \mathbf{ACA^T}. \tag{3.7}$$

The entropy of $\mathbf{y}$ is obtained as

$$H_2(\mathbf{y}) = -\int \mathbf{g(y)}\log[\mathbf{g(y)}]\mathbf{dy}, \tag{3.8}$$

where $g(\mathbf{y})$ is the pdf of vector $\mathbf{y}$

$$g(\mathbf{y}) = \frac{\mathbf{1}}{(\mathbf{2}\pi)^{\mathbf{N/2}}|\mathbf{D}|^{\mathbf{1/2}}}\exp(-\frac{\mathbf{1}}{\mathbf{2}}\mathbf{y^T D^{-1}y}). \tag{3.9}$$

After the transformation, it can be derived that $\mathbf{g(y)dy} = |\mathbf{J}|\mathbf{f(x)dx}$ where $\mathbf{J}$ is the Jacobi matrix. In this linear transform, it is easy to see $\mathbf{J} = \mathbf{A}$, $\mathbf{AA^T} = \mathbf{I}$, therefore,

$$\mathbf{y}^T D^{-1}\mathbf{y} = (A\mathbf{x})^{\mathbf{T}}(\mathbf{ACA^T})^{-\mathbf{1}}(\mathbf{Ax}) = \mathbf{x^T C^{-1} x}, \tag{3.10}$$

Note $|C| = |D|$, $d\mathbf{z} = d\mathbf{u}$. Therefore,

$$H_1(\mathbf{x}) = H_2(\mathbf{y}). \tag{3.11}$$

The entropy conservativeness holds even for non-Gaussian pdf. Comparing (3.5) and (3.8), it is easy to see $g(Ax) = f(x)$ as long as $AA^T = I$.

In oblivious watermarking, host signal is considered noise. The larger the noise entropy, the more "harmful" the noise is. The negative effect of host noise can not be reduced by taking a transform. Intuitively, observing an image I in different domain should give us same information. To embed an information source $\mathbf{s}$ in $\mathbf{x}$ is equivalent to embedding a source $\mathbf{As}$ in $\mathbf{Ax}$ where $\mathbf{A}$ is the transform kernel. Same information is transmitted. This correlation should not be neglected in capacity calculation, watermark embedding and detection. Depovere *et al.* [21] proposed a better detection approach by whitening filter before correlation.

Taking transform should not affect the hiding capacity theoretically. Actually hiding capacity calculation does not shed much light on the decomposition selection in practice. Even the higher capacity does not guarantee more reliable information transmission using the current hiding schemes. For example, it is believed Laplacian channel is of higher capacity compared with the Gaussian channel (Gaussian is regarded as the "worst" noise). However, using the antipodal or M-ary modulation, signal transmission through Gaussian channel could be more reliable. The Bit Error Rate (BER) is determined by the "tail" of the noise pdf curve. Gaussian pdf decreases at the order of e^{-x^2}, faster than the Laplacian pdf e^{-x}. That results in more reliable transmission through Gaussian channel at higher SNR.

Transform selection is a subtle issue in practice. In face of a specific compression, is it favorable to match the decomposition used in the compression? This question is addressed in the following section.

3.1.2 Decomposition Selection in Presence of Compression

The channel noise is mainly introduced in compression in escrow cases. The compression noise is expressed as

$$e_i = x_i - Q(x_i), \tag{3.12}$$

where $Q(.)$ is the quantization operator.

The coefficient x_i is approximately Laplacian distributed before quantization operation. If data embedding and extraction is done in the same compression transform domain, the compression noise is quite close to uniform distribution, i.e. $e_i \sim U(-\delta/2, \delta/2)$ where δ is the quantization step size.

In a transform domain other than the compression domain, the compression noise d_i is not uniformly distributed. Experiments reveal that it is close to Laplacian distribution. It is reasonable to assume the noise is i.i.d. The SS modulation approaches in both of these mismatch and match domains are equivalent to PN sequence signal transmission through two channels, one is uniform channel and the other is Laplacian channel. If the mostly used correlation is employed at receiver, it is observed that the Bit Error Rate (BER) in the Laplacian channel is superior to that in the uniform channel (Fig. 3.2). In simulation the antipodal signal is transmitted in channels with different noise statistic properties.

It is true that the correlation detection is not optimal in a uniform channel case. Further analysis shows that the optimum detector needs to know the quantizer step size δ which is usually unavailable in practice. Correlation, although not optimal, is still widely used in practice. The noise energy is not changed in the Laplacian and

uniform channels. Still a transform changes the channel noise property. Correlation detection is more reliable in Laplacian channel than the uniform channel.

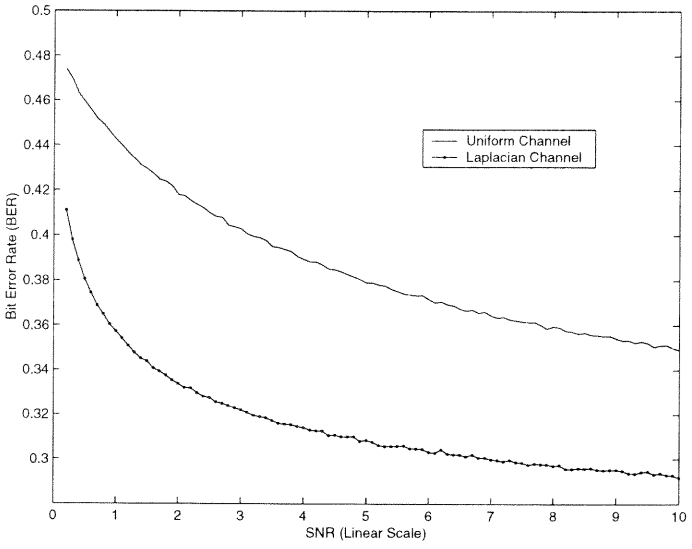


Figure 3.2 BER-SNR in Gaussian, Laplacian and Uniform channels

The SS scheme performance in oblivious applications in different G_{TC} domains is studied next. The following deep embedding scheme [16] is used in simulation,

$$x_i' = \begin{cases} x_i + w_i|x_i|\alpha, & \text{to hide bit value 1} \\ x_i - w_i|x_i|\alpha, & \text{to hide bit value 0} \end{cases} \tag{3.13}$$

where $\mathbf{w}$ is the random bipolar sequence (w_i is $+1$ or -1) and α is the distortion threshold ratio.

Given a received sequence $\mathbf{r}$, the decoder used is also of correlation type

$$q = \sum_{i=0}^{N-1} r_i w_i = \sum_{i=0}^{N-1} x_i w_i + \sum_{i=0}^{N-1} n_i w_i \pm \sum_{i=0}^{N-1} \alpha|x_i|. \tag{3.14}$$

If $q > 0$, bit value 1 is decided; Otherwise bit value 0 is decided instead.

First, the host coefficients in the time domain are generated. The cover signal $\mathbf{x}$ is a highly correlated AR(1) sequence with correlation ratio $\rho = 0.9$. Second, the embedding and extraction is repeated in the time and DCT domain. The distortion

ratio is selected as $\alpha = 0.1$ in experiments. Simulation result in Figure 3.3 demonstrates that time domain embedding is more reliable than that in the DCT domain embedding.

An intuitive explanation for the above result is that in time domain, the coefficients, although highly correlated, is evenly distributed. Whereas in the DCT domain, much energy goes to the low frequency coefficients. These high amplitude host noise coefficients exert much negative effect on the decoding. Further studies demonstrate improvements if these coefficients are skipped in embedding.

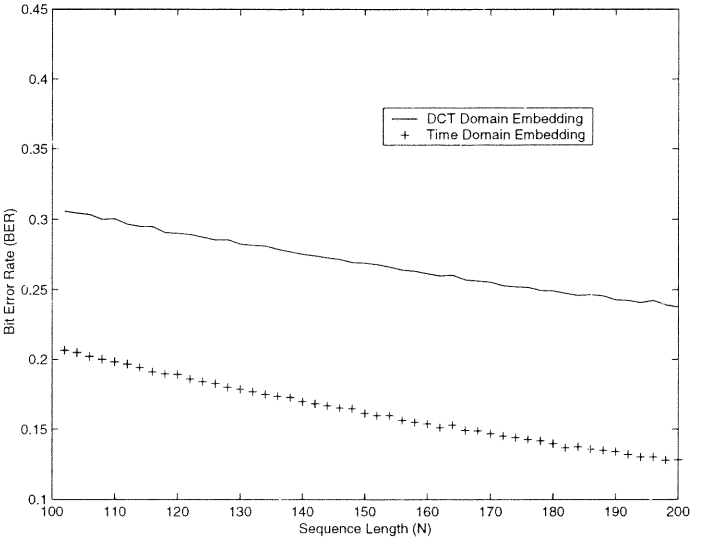


Figure 3.3 Performance in DCT and time domain embedding

The great concern in the oblivious case is host noise suppression. The linear SS algorithms do not suppress the host noise very effectively. Some methods, such as set partitioning [31, 27], Quantization Index Modulation [12], etc. are more successful. It is believed that a mismatched transform is more favorable in these schemes due to the same reason.

3.1.3 Taking Advantage of Compression – an Example

From the above discussion, it is concluded that selecting higher G_{TC} transform does not increase capacity and matching compression transform (usually high G_{TC}) does

not enhance its resilience to compression. Then how to select am advantageous watermark domain?

Bender *et al.* [7] are among the first to point out, "The key to successful data hiding is the finding of holes that are not suitable for exploitation by compression algorithm". A good hiding scheme should take advantage of the compression. The watermark domain is not necessarily a transform domain. Some authors argue the message should be inserted in a domain which is compression-conservative. A novel scheme [2] was proposed to hide information in the eigen vectors of the correlation matrix of a subimage. Because these values are almost unchanged after compression. Just as mentioned in Chapter 2, embedding in DFT amplitude domain provides some advantages due to its perceptual insignificance [71, 73].

In Chapter 5, a noise substitution algorithm in audio data hiding is proposed. The well advanced audio model *sine* + *transient* + *noise* [56, 94, 95] is explicitly used in this scheme. This scheme modifies the noisy components without changing the noisy perception. More details can be found in Section 5.6. In this scheme, matching the decomposition brings some benefits due to the compression coefficient sign conservative property.

3.1.4 Summary

Decomposition used in compression is to de-correlate the signal. High de-correlation is desirable in compression, although may not be suitable in steganography. For SS schemes application in oblivious cases, mismatch the decomposition used in compression is more favorable. Different transform selection does not have any effect on hiding capacity. The extra computation in high G_{TC} decomposition in compressions may not be well justified in data hiding.

The decomposition selection is a complicated issue in practice. For example, audio watermarking in frequency domain is advisable because the psychoacoustic

model in transform domains can be conveniently used. In the design of a compression resistant watermarking scheme, decomposition selection is related with embedding scheme used. Mismatch decomposition is advisable in SS. For the noise substitution scheme in audio data hiding, it is more favorable to match the compression decomposition.

3.2 SS Modulation and Correlation Detection

Although the SS scheme is first introduced in escrow applications, it works in the oblivious cases where the original content signal is not available. In the data hiding scheme (3.13), s_i and x_i are independent. It is assumed that the first term in (3.14)

$$t = \sum_{i=0}^{N-1} s_i x_i \approx 0. \tag{3.15}$$

Compared with (2.6), the extra disturbing term degrades the detector performance. It is true that $t \approx 0$ if the sequence length N is sufficiently large. Obviously, this significantly reduces hiding capacity. In the following part, the degradation is measured quantitatively [30].

3.2.1 Correlation Performance in SS Modulation

In the following discussion, the deep data hiding scheme mentioned in Section 3.1 is studied. Its embedding and extraction is given by (3.13) and (3.14), respectively.

The host noise power is much larger than that of the channel noise in oblivious cases. The channel noise is neglected for simplicity in the following discussion. Then (3.14) is reduced to

$$q = \sum_{i=0}^{N-1} r_i w_i = \sum_{i=0}^{N-1} (x_i w_i + \alpha |x_i|). \tag{3.16}$$

Denote

$$p_i = x_i w_i + \alpha |x_i|, \tag{3.17}$$

Because w_i is either $+1$ or -1, $p_i = x_i + \alpha|x_i|$ or $p_i = x_i - \alpha|x_i|$. Due to the symmetry property of the pdf of x_i, the statistical distribution of p_i does not depend on the value of w_i. It has same mean value and variance as

$$y_i = x_i + \alpha|x_i|. \tag{3.18}$$

Suppose the original coefficient x_i is i.i.d. and Gaussian distributed, $x_i \sim N(0, \sigma^2)$. The mathematical expectation of y_i is

$$E[y_i] = 2\alpha \int_0^\infty \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx = \sqrt{\frac{2}{\pi}}\sigma\alpha. \tag{3.19}$$

The variation of y_i is obtained as

$$E[(y_i - E[y_i])^2] = E[(y_i - \sqrt{\frac{2}{\pi}}\sigma\alpha)^2]. \tag{3.20}$$

After some algebraic steps, the final result yields as

$$E[(y_i - E[y_i])^2] = (1 + \alpha^2)\sigma^2. \tag{3.21}$$

The test statistic q (3.16) can be assumed a summation of random variables y_i (3.18). For a large value of N, the distribution of q is approximately Gaussian

$$q \sim N(\sigma\alpha N\sqrt{\frac{2}{\pi}}, N(1 + \alpha^2)\sigma^2). \tag{3.22}$$

In a similar fashion, while the bit value 0 is embedded the distribution of the test statistic is given as

$$q \sim N(-\sigma\alpha N\sqrt{\frac{2}{\pi}}, N(1 + \alpha^2)\sigma^2). \tag{3.23}$$

If the decision threshold is selected $\gamma = 0$, the Bit Error Rate

$$BER = Q(\alpha\sqrt{\frac{2N}{(1 + \alpha^2)\pi}}), \tag{3.24}$$

where $Q(.)$ is the Gaussian pdf tail integral function.

This analytical result matches the simulation output (Figure 3.4). α is selected as 0.1 in simulation. Figure 3.5 demonstrates the BER versus sequence length N.

With sequence length $N = 200$, $BER = 0.1308$. To achieve the reliability $BER < 10^{-6}$, the sequence length should be $N > 3700$.

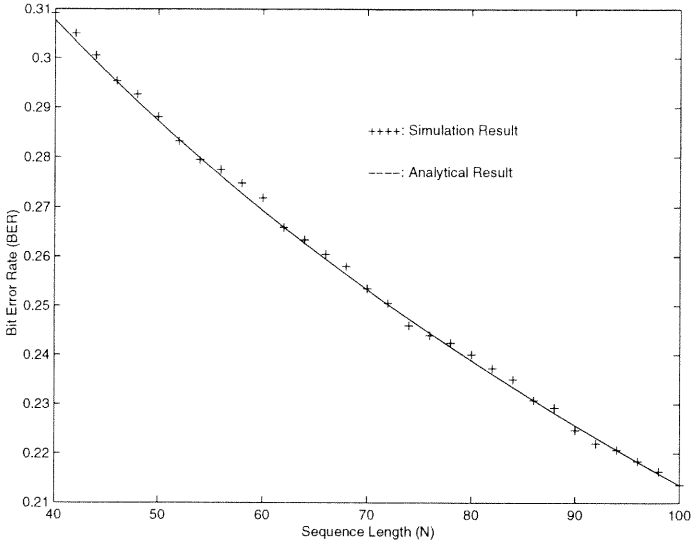


Figure 3.4 Correlation performance in SS modulation

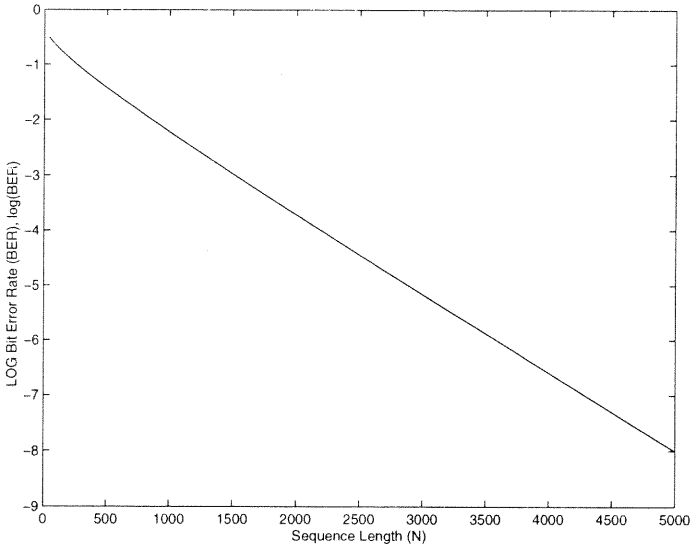


Figure 3.5 Analytical BER-N in SS modulation

3.3 Optimum Detection and Linear Modulation

The above results demonstrate that the performance of the SS modulation schemes is not satisfactory. The SS modulation performance limit in oblivious applications is analyzed in this section.

3.3.1 Maximum Likelihood Detection

The correlation detector is only optimal in white Gaussian noise environment but not in the SS modulation scheme discussed above.

With bit embedding scheme (3.13), decoding is a hypothesis testing problem

$$\begin{cases} \text{H1:} & r_i = x_i + |x_i| \cdot k_i \quad \text{, bit value 1 is embedded} \\ \text{H0:} & r_i = x_i - |x_i| \cdot k_i \quad \text{, bit value 0 is embedded} \end{cases} \tag{3.25}$$

where $k_i = w_i\alpha$.

Given a received sequence $\mathbf{r}$, the Maximum Likelihood (ML) ratio is

$$R = \frac{P(H1|\mathbf{r})}{P(H0|\mathbf{r})}. \tag{3.26}$$

If $R > 1$, the bit value 1 is decided; Otherwise bit value 0 is decided.

Assume the original coefficient x_i is Gaussian distributed, its pdf

$$f(r_i|H1) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma(1+k_i)} \cdot \exp[\frac{-r_i^2}{2(1+k_i)^2\sigma^2}], & (r_i > 0) \\ \frac{1}{\sqrt{2\pi}\sigma(1-k_i)} \cdot \exp[\frac{-r_i^2}{2(1-k_i)^2\sigma^2}], & (r_i < 0) \\ \frac{1}{\sqrt{2\pi}\sigma}, & (r_i = 0) \end{cases} \tag{3.27}$$

In a similar fashion, $f(r_i|H0)$ can be calculated.

Assume $P(H0) = P(H1)$ and neglect the rare case where $r_i = 0$, the ML ratio is

$$\frac{P(r_i|H1)}{P(r_i|H0)} = \begin{cases} (\frac{1-k_i}{1+k_i}) \cdot \exp[-\beta \cdot s(k_i)r_i^2], & (r_i > 0) \\ (\frac{1+k_i}{1-k_i}) \cdot \exp[+\beta \cdot s(k_i)r_i^2], & (r_i < 0) \end{cases} \tag{3.28}$$

where $s(.)$ is the sign function

$$s(x) = \begin{cases} +1, & x > 0 \\ -1, & x < 0 \end{cases} \tag{3.29}$$

and

$$\beta = \gamma\frac{1}{\sigma^2}, \tag{3.30}$$

$$\gamma = \frac{1}{2(1+\alpha)^2} - \frac{1}{2(1-\alpha)^2}. \tag{3.31}$$

If one bit is embedded in a sequence $\mathbf{x}$, the ML ratio (3.26) is obtained as

$$R = \prod_{i=0}^{N-1} (\frac{1-k_i}{1+k_i})^{s(r_i)} \cdot \exp[\sum_{i=0}^{N-1} -s(r_i) \cdot s(k_i) \cdot r_i^2 \beta]. \tag{3.32}$$

The above is the ML optimum detector in the oblivious case and it yields better performance. Yet the computation is quite extensive. Secondly, the calculation of β involves the value of σ^2, which is usually unavailable in practice. Some approximation is needed to derive a suboptimal detector applicable in practice.

In a white sequence $\mathbf{s}$, it is reasonable to assume it has the equal number of $+1$ and -1. One obvious observation in (3.32) is that for sufficiently large sequence length N,

$$\prod_{i=0}^{N-1} (\frac{1-k_i}{1+k_i})^{s(r_i)} \approx 1. \tag{3.33}$$

Under this approximation, a suboptimal detector is obtained as

$$q = \gamma \sum_{i=0}^{N-1} --s(r_i) \cdot {r_i}^2 \cdot s(k_i). \tag{3.34}$$

If $q > 0$, it is decided that the bit value is 1; Otherwise the bit value 0 is decided.

The suboptimal detector has a quite simple form and comparable complexity as the correlation detector (3.14). Figure 3.6 shows the simulation result with visual threshold ratio value $\alpha = 0.1$. The suboptimal detector has lower BER compared with correlation. Still it is inferior to the optimum detector due to the approximation in (3.33).

The suboptimal detector (3.34) is in a form of variation difference distinction. Any hiding scheme changes the statistical property of the original cover signal. From the embedding operation (3.25), it is clear the main impact of hiding operation is the change of variation of x. Intuitively, the detector based on the distinction of

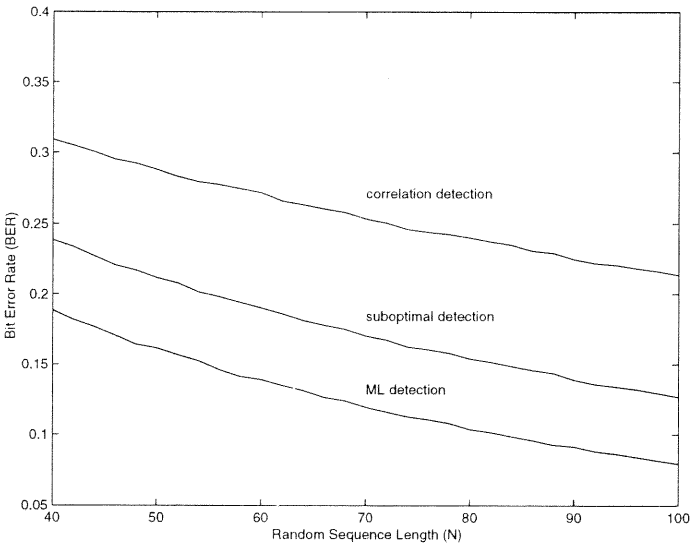


Figure 3.6 Correlation, suboptimal and optimum detection

variation outperforms the correlation detector where the decision is based on the mean value.

The channel noise is not considered in the above discussion. Even taking it into consideration, further simulation studies show the ML and suboptimal detector still outperforms the commonly used correlation detector.

3.3.2 Linear Modulation and Detection

It is demonstrated in simulation studies and mathematical analysis that the suboptimal detector is inferior to the optimum detector. How can the performance be further improved?

The data hiding procedure (3.13) can be slightly modified by removing the absolute value operator. The data hiding hypotheses testing becomes

$$\begin{cases} \text{H1:} & r_i = x_i + x_i \cdot k_i, & \text{to embed bit value 1} \\ \text{H0:} & r_i = x_i - x_i \cdot k_i, & \text{to embed bit value 0} \end{cases} \tag{3.35}$$

where $k_i = w_i\alpha$ (k_i is either $+\alpha$ or $-\alpha$).

After embedding, the variance is modified to

$$\sigma_1^2 = (1+\alpha)^2\sigma^2, \tag{3.36}$$

or

$$\sigma_0^2 = (1-\alpha)^2\sigma^2. \tag{3.37}$$

In a way similar to the analysis in the last section, the ratio on r_i is given by

$$\frac{P(r_i|H1)}{P(r_i|H0)} = (\frac{1-k_i}{1+k_i}) \cdot \exp[\sum_{i=0}^{N-1} -s(k_i) \cdot {r_i}^2\gamma]. \tag{3.38}$$

And the final ML detector output is

$$R = \prod_{i=0}^{N-1} \frac{P(r_i|H1)}{P(r_i|H0)}. \tag{3.39}$$

All the coefficients can be divided into 2 sets. The variances of x_i's in **Set A** are increased while those in **Set B** are decreased.

Statistically, the coefficient number in these two sets is equal. The generation of the white sequence can be controlled so that the coefficient count of $k_i = \alpha$ is equal to the count of $k_i = -\alpha$. That yields

$$\prod_{i=0}^{N-1} \frac{1-k_i}{1+k_i} = 1. \tag{3.40}$$

By simplifying (3.38), the detection test statistic is obtained as

$$q = \gamma \sum_{i=0}^{N-1} s(k_i) \cdot {r_i}^2. \tag{3.41}$$

Remove the factor γ, the test statistic is

$$q' = \sum_{r_i \in \text{Set A}} {r_i}^2 - \sum_{r_i \in \text{Set B}} {r_i}^2, \tag{3.42}$$

If $q' > 0$, the bit value 1 is decoded; Otherwise, bit value 0 is decided instead. That is the optimum detector in the linear embedding.

The performance can be further analyzed if only the host noise is considered. Suppose r_i in the Set A is Gaussian distributed with variance equals to σ_1^2 (3.36); While r_i in Set B is distributed with variance σ_0^2 (3.37).

Denote

$$t_1 = \sum_{r_i \in \text{Set A}} r_i^2 \tag{3.43}$$

and

$$t_0 = \sum_{r_i \in \text{Set B}} r_i^2, \tag{3.44}$$

These two variables are square sum of Gaussian distributed random variables which share the same probability property. It can be proved their distribution is of $M = N/2$ freedom degree Γ distribution [40].

$$f(t_i) = \frac{t_i^{M/2-1} \cdot e^{-\frac{t_i}{2\sigma_i^2}}}{\sigma_i^M \cdot 2^{M/2} \cdot \Gamma(M/2)}. \tag{3.45}$$

For notation simplicity, denote

$$A_i = \frac{1}{\sigma_i^M \cdot 2^{M/2} \cdot \Gamma(M/2)} \tag{3.46}$$

and

$$C_i = \frac{1}{2\sigma_i^2} \tag{3.47}$$

Equation (3.45) can be rewritten as

$$f(t_i) = A_i \cdot t_i^{n-1} \cdot e^{-C_i t_i}, \tag{3.48}$$

where $n = M/2 = N/4$.

Suppose the bit value 1 is transmitted, the Bit Error Rate (BER) is

$$\begin{aligned} BER &= P(t_1 < t_0) = \int_0^{+\infty} f(t_0)dt_0 \cdot \int_0^{t_0} f(t_1)dt_1 \\ &= \int_0^{+\infty} f_0(t_0) \int_0^{t_0} A_1 t_1^{n-1} e^{-C_1 t_1} dt_1 dt_0. \end{aligned} \tag{3.49}$$

For an integer n, using

$$\int x^n e^{-ax} dx = -\frac{e^{-ax}}{a^{n+1}} \cdot [(ax)^n + n(ax)^{n-1} + n(n-1)(ax)^{n-2} + ... + n!] \tag{3.50}$$

and

$$\int_0^{+\infty} x^n e^{-ax} dx = \frac{n!}{a^{n+1}}, \tag{3.51}$$

after some algebraic steps, the final result is obtained as

$$BER = [(1 + \frac{C_0}{C_1})(2n-2)! + \sum_{i=2}^{n} \frac{(n-1)!}{(n-i)!}(1 + \frac{C_0}{C_1})^i] \cdot \frac{-A_0 A_1}{C_0 + C_1^{2n}} + \frac{A_0 A_1 [(n-1)!]^2}{(C_0 C_1)^n}. \tag{3.52}$$

The same result is obtained if bit value 0 is transmitted. Therefore, the average Bit Error Rate (BER) is given by (3.52).

Equation (3.52) is the achievable performance in the linear modulation approach. Figure 3.7 depicts BER calculated by (3.52) and the simulation output. The distortion threshold ratio α is selected as 0.1 and x_i is Gaussian distributed with $\sigma = 50$. The analytical result is a perfect match of the simulation results. Compared with the embedding and extraction scheme using absolute value operation, this scheme achieves the same performance as the optimum detector and outperforms the suboptimal detector. This detector is also easy to implement.

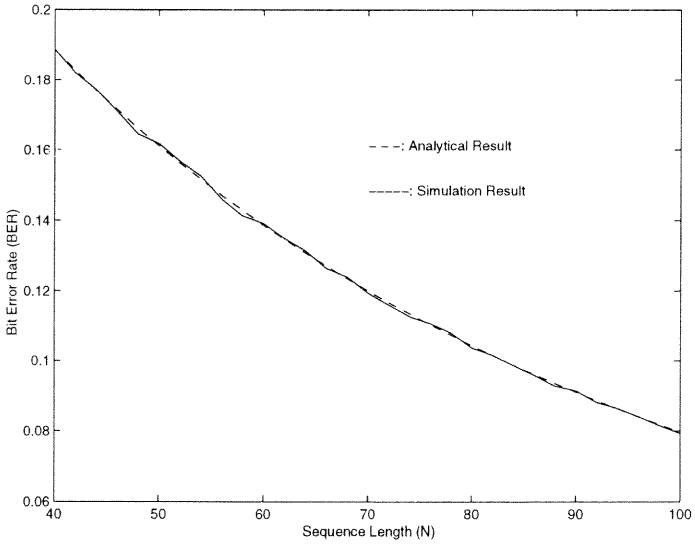


Figure 3.7 Performance comparison in the linear modulation

It is assumed the original coefficient x_i is Gaussian distributed in the above discussion. In many cases x_i is a transform coefficient whose pdf is approximately Generalized Gaussian or Laplacian distribution. The above ML detector (3.42) is not optimum in this case. However, it still outperforms the correlation detector commonly used.

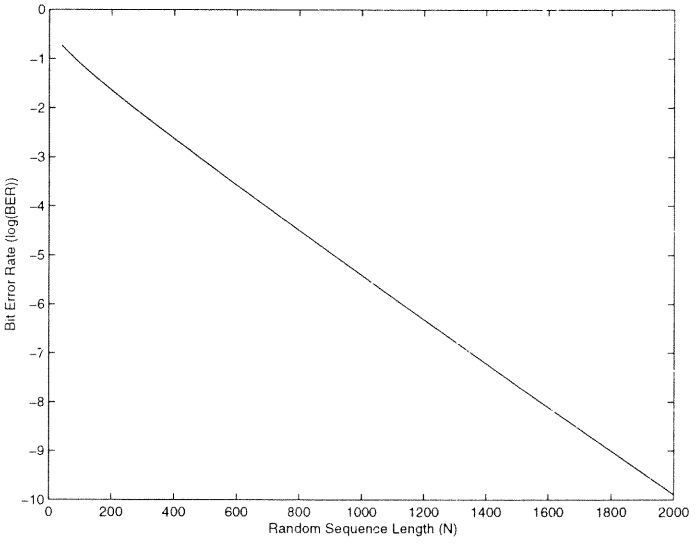


Figure 3.8 Analytical result in the linear modulation

The SS schemes are not quite effective in oblivious cases. Figure 3.8 depicts the BER at different sequence length with the distortion ratio $\alpha = 0.1$ (corresponding to $-20dB$ distortion, very deep embedding). At sequence length $N = 1000$, $BER = 3.91 \cdot 10^{-6}$. To achieve up to $BER \leq 10^{-9}$, the sequence length should be $N > 1800$. Please note the performance is even poorer for correlation detection. This is the limitation of SS schemes.

3.3.3 Image Data Hiding Experiments

The above linear detection scheme can replace the existing SS hiding and extraction schemes. In practice, the value of the distortion ratio α could be obtained from empirical experiments or some more accurate perceptual models. For example, the

distortion threshold ratio in audio signals can be calculated from a psychoacoustic model.

In the experiments with image data hiding applications, first the image is decomposed into 64 subbands. Second, the medium bands are selected and the hiding scheme in (3.35) is employed. At decoder side, (3.42) is used to extract the information. In those experiments, 32 bit is embedded in a 256x256 image. All bits are extracted error-free. Experiments also show its robustness against JPEG compression and other attacks.

(a) Original Lena (b) Marked Lena

Figure 3.9 Lena image before and after embedding

3.4 Signature Sequence: Security and Synchronization

Similar to an encryption system, it is believed a mature watermark system should be employed with a public algorithm and a private key. The key is the seed to generate a random sequence. An attacker can try his best to "guess" a sequence close to the watermark sequence and remove it. This attack is referred to as "guessing" attack. Besides security, another important factor is synchronization requirement on the PN sequence. Both the security of PN sequence and its synchronization requirement are investigated in this section. A random phase sequence generation is proposed later.

3.4.1 Security of Gaussian PN Sequence

White Gaussian Sequence

The white Gaussian sequence is widely used in various watermarking schemes [7, 16]. The sequence $\mathbf{x}$ is of white spectrum, $x_i \sim N(0, \sigma_x^2)$ and x_i is i.i.d. In a public scheme scenario, an attacker knows the parameters σ_x and sequence length N, only does not know the random seed. In the guessing attack, a random sequence $\mathbf{y}$ is generated. The "closeness" between $\mathbf{x}$ and $\mathbf{y}$ is measured by correlation

$$q =< \mathbf{x}, \mathbf{y} >= \sum_{i=0}^{N-1} x_i y_i. \tag{3.53}$$

If the correlation output is larger than a fixed threshold γ, the attacker assumes that $\mathbf{y}$ is sufficiently close to $\mathbf{x}$. By subtracting the sequence $\mathbf{y}$, a good proportion of the watermark energy could be removed.

As a linear combination of $\mathbf{y}$, the output q is Gaussian distributed

$$q \sim N(0, \sigma_x^2 \sum_{i=0}^{N-1} x_i^2). \tag{3.54}$$

The exact value of q is dependent on the individual signature sequence $\mathbf{x}$. For a large value of sequence length N,

$$E[q] = 0, \tag{3.55}$$

and

$$E[q^2] = \sigma^2 \sum_{i=0}^{N-1} x_i^2 \approx N{\sigma_x}^4. \tag{3.56}$$

The successful attack probability is

$$P(q > \gamma) = Q(\frac{\gamma}{\sqrt{N}{\sigma_x}^2}). \tag{3.57}$$

Some numerical result for white sequences (corresponding to $\rho = 0.0$) is shown in Table 3.1. It can be seen that the white sequences are quite secure against this guessing attack. This conclusion is justified by the intuition that the white sequence

is the most "unpredictable". The white sequence has a flat spectrum whereas most cover signals are of low-pass type. Most high frequency energy can be removed by low-pass filtering. For example, in a typical audio signal, most energy is concentrated between 0 – 6kHz. For an audio signal sampled at 48kHz, by suppressing frequency components over 6kHz, a smart attacker can remove 75% of the white watermark signal energy without much noticeable distortion. The white sequence although secure, is less energy efficient.

Low-pass (LP) type random signature can keep most energy after low-pass filtering attack. A simple colored PN sequence – AR(1) random process is analyzed in the next paragraphs.

AR (1) PN Sequence

The first order AR(1) sequence $\mathbf{x}$ is expressed as

$$x_i = \rho x_{i-1} + u_i, \tag{3.58}$$

where

$$\rho = \frac{E[x_i x_{i-1}]}{\sigma_x^2}, \tag{3.59}$$

and $u_i \sim N(0, (1-\rho^2){\sigma_x}^2)$, u_i is i.i.d.

The attacker tries to generate a matching sequence $\mathbf{y}$ randomly based on the same AR(1) model (suppose the value of ρ is public). Correlation output (3.53) measures the success of this attack. It can be easily shown that

$$E[q] = < \mathbf{x}, \mathbf{y} > = 0. \tag{3.60}$$

The variation of correlation output q is

$$E[q^2] = E[(x_0 y_0 + x_1 y_1 + ... + x_{N-1} y_{N-1})^2] \tag{3.61}$$

Using

$$E[y_i y_{i-j}] = \sigma^2 \rho^j \tag{3.62}$$

and

$$E[x_i x_{i-j}] \approx \sigma^2 \rho^j, \tag{3.63}$$

Equation (3.61) is reduced to

$$E[q^2] = N + 2(L-1)\rho^2 + 2(L-2)\rho^4 + ... + 2\rho^{L-1}. \tag{3.64}$$

After some algebraic steps, the final result is obtained as

$$\sigma_q^2 = E[q^2] = 2[\frac{N-\rho^{2N}}{1-\rho^2} - \frac{\rho^2-\rho^N}{(1-\rho^2)^2}] - N. \tag{3.65}$$

For a sufficiently large value of sequence length N, the above can be further expressed as

$$\sigma_q^2 \approx \alpha N, \tag{3.66}$$

where

$$\alpha = \frac{2}{1-\rho^2} - 1. \tag{3.67}$$

Similarly, the successful guessing attack probability yields as

$$P(q > \gamma) = Q(\frac{\gamma}{\sigma_q}) = Q(\frac{\gamma}{\sqrt{\alpha N}\sigma_x^2}). \tag{3.68}$$

Compared with the white sequence, the AR(1) sequence length should be increased by a factor α to reach the same security level. For example, for the case where $\rho = 0.8$, $\alpha = 4.56$, a 456-coefficient AR(1) sequence is of the same robustness as a 100-coefficient white sequence.

Table 3.1 shows the successful attack probability for different ρ and N values. The result reveals that the LP type signal is more vulnerable to the guessing attack due to the correlation between x_i. However, it has some desirable properties, one is the relaxed synchronization requirement at decoder.

	N=30	N=100	N=400
$\rho = 0.0$	$2.16 \cdot 10^{-8}$	$7.62 \cdot 10^{-24}$	$2.75 \cdot 10^{-89}$
$\rho = 0.5$	$3.98 \cdot 10^{-4}$	$4.57 \cdot 10^{-10}$	$8.66 \cdot 10^{-35}$
$\rho = 0.8$	$1.00 \cdot 10^{-2}$	$1.10 \cdot 10^{-5}$	$1.08 \cdot 10^{-17}$

Table 3.1 Sequence security comparison ($\gamma = N\sigma_x^2$)

3.4.2 Synchronization Effect on Detection

In SS modulation, it is well known the decoder is extremely sensitive to synchronization [87]. As it will be seen, the LP type sequence is less sensitive to synchronization, which is often a desirable property in practice.

White Gaussian Sequence

Suppose a signal $\mathbf{x}$ is transmitted through a Gaussian channel,

$$r_i = x_i + z_i, \tag{3.69}$$

where z_i is the channel noise, $z_i \sim N(0, \sigma_z^2)$.

If the sequence is perfectly matched, the decoder output SNR can be shown to be

$$SNR = \frac{S^2}{E[Z^2]} = \frac{(\sum x_i^2)^2}{\sigma_z^2 \sum x_i^2} \approx \frac{N\sigma_x^2}{\sigma_z^2}. \tag{3.70}$$

If the received sequence $\mathbf{r}$ and $\mathbf{x}$ is not perfectly matched, $SNR \approx 0$. The watermark verification completely fails.

AR(1) Random Sequence

For an AR(1) sequence generated by (3.59), although the output SNR degrades if $\mathbf{r}$ and $\mathbf{x}$ are not perfectly synchronized, there is some signal energy residue in the correlation detector output. If $\mathbf{x}$ and $\mathbf{r}$ is synchronized, AR(1) sequence performs as well as the white sequence. In the case where it is misaligned by M sample slip shift, the filter output SNR is given by

$$SNR_M = \frac{(N-M)^2\sigma_x^2\rho^2}{N\sigma_z^2}. \tag{3.71}$$

Figure 3.10 depicts the SNR output value versus misalignment. The parameters selected are $N = 100$, $\rho = 0.8$, $\sigma_x = \sigma_z$. Obliviously, the AR(1) sequence is less sensitive to synchronization than the white sequence.

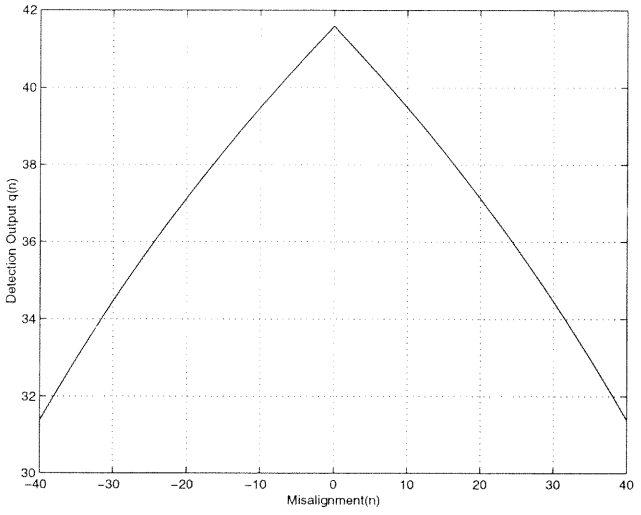


Figure 3.10 AR(1) correlation output vs. misalignment

3.4.3 Sequence Spectrum Shaping

The white sequence and LP type sequence are studied above. The AR(1) sequence is just a special case of a colored sequence. A more general colored sequence can be generated by an AR(M) model

$$x_i = \sum_{j=1}^{M} h_j x_{i-j} + w_i, \tag{3.72}$$

where w_i is white Gaussian noise and x_i is Gaussian distributed, independent with w_i. The above AR(M) colored sequence can be interpreted as the white sequence shaped by a LP filter.

Although the white sequence is more secure than a LP type sequence, this is only true when no attack is present. The low-pass filtering attack can remove the watermark energy in high frequency bands without much artifacts. A smart attacker may combine the low-pass filtering and guessing attack therefore compromise its security down to the level in the LP sequence. The watermarking energy spanning the whole spectrum is not well spent, resulting in energy inefficiency.

It is pointed out in the face of Wiener filtering, the spectrum of the watermark signal should be proportional to that of the cover signal [86]. In this case, the

filtering is nothing but a scaling operation. This implies no gains achieved in this attack. Actually, watermark signal power spectrum $X(\omega)$ need not to be exactly proportional to the cover signal spectrum $N(\omega)$, but should be close to $N(\omega)$. Since the cover signal spectrum $N(\omega)$ is public, the randomness mainly lies in the phase.

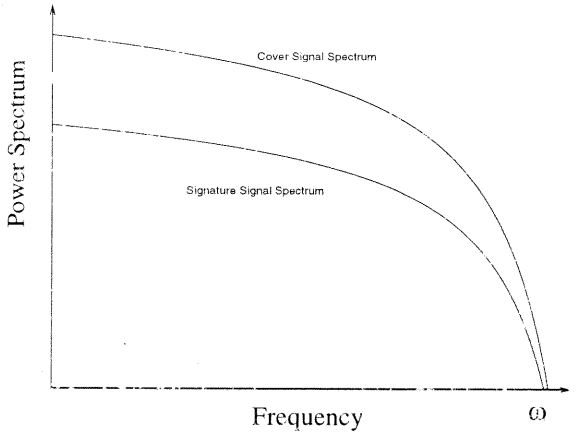


Figure 3.11 Watermark spectrum against Wiener filtering

3.4.4 Random Phase Sequence

The random phase sequence can be easily generated in DFT domain. Suppose the N-point DFT transform of the watermark signal $x(n)$ is $|X(k)|$. A random phase sequence θ_i is generated by a private key. θ_i is i.i.d. $\theta_i \sim U(0, 2\pi)$ and satisfies the odd symmetry property

$$\theta_k = \begin{cases} 0 \text{ or } \pi & k = 0, \frac{N}{2} \\ \theta_k & k = 1, 2, .. \frac{N}{2} - 1 \\ -\theta_{k-N/2} & k = \frac{N}{2} + 1, \frac{N}{2} + 2, ..., N - 1 \end{cases} \tag{3.73}$$

The embedding and extraction operations may be in time or DFT domain.

The watermark sequence in FFT domain is generated as

$$x(k) = \exp(j\theta_k). \tag{3.74}$$

In the security analysis against the guessing attack, the cover signal spectrum is assumed brick-shape for simplicity (Figure 3.12).

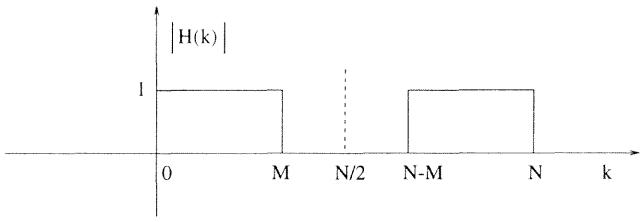


Figure 3.12 Brick-shape LP watermark spectrum

Suppose the attacker randomly generates a phase sequence β, the correlation between these two vectors are

$$q = \sum_{k=0}^{M-1} [e^{j(\beta_k - \theta_k)} + e^{-j(\beta_k - \theta_k)}] = 2 \sum_{k=0}^{M-1} cos(\beta_k - \theta_k), \tag{3.75}$$

Both θ_k and β_k are uniformly distributed in the range $[0, 2\pi)$. The mathematical expectation of $t_k = cos(\beta_k - \theta_k)$ is

$$E[t_k] = \int_0^{2\pi} cos(\beta_k - \theta_k) \frac{1}{2\pi} d\beta_k = 0. \tag{3.76}$$

The deviation of t_k is

$$E[t_k^2] = \int_0^{2\pi} cos^2(\beta_k - \theta_k) \frac{1}{2\pi} d\beta_k = \frac{1}{2}. \tag{3.77}$$

For a large number of M, q is approximately Gaussian distributed. Its distribution can be shown to be $q \sim N(0, M)$. The successful guessing attack probability is

$$P(q > \gamma) = Q(\frac{\gamma}{\sqrt{M}}). \tag{3.78}$$

For different values of M=30, 60 and 100, with the threshold value selected as $\gamma = 2M$, the successful attack probabilities are 2.16×10^{-8}, 4.74×10^{-15} and 7.62×10^{-24}, respectively.

The signal component in the correlation detection output is $2M$ without misalignment. The mathematical analysis shows that with $p(p \neq 0)$ sample shift, the correlation output is

$$y_p = 1 + \frac{cos\frac{2\pi p}{N}(M-1) - cos\frac{2\pi p}{N}M}{1 - cos\frac{2\pi p}{N}}. \tag{3.79}$$

Figure 3.13 shows the output vs. misalignment. The sequence length is $M = 30$ and $N = 200$.

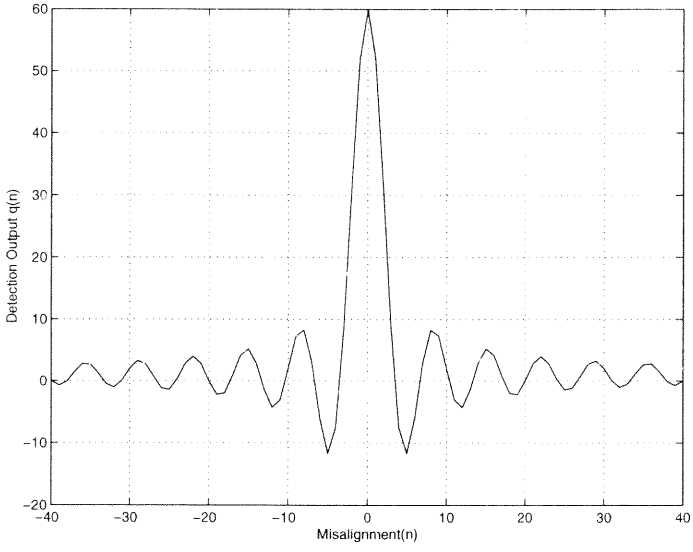


Figure 3.13 Correlation output vs. misalignment

The sequence length N and sequence bandwidth (indicated by M) are two important parameters in sequence generation. Larger N implies enhanced security against guessing attack; while smaller M lowers the security level and relaxes the synchronization requirement. Compared with Gaussian sequence, this signature provides a trade-off between security and synchronization requirement.

In the random phase sequence, the sequence frequency shape is fixed, only the phase is random. Every sequence has exactly the same energy. In practice, it may not be necessary to keep the sequence spectrum strictly brick-shape. Various visual models could be applied to control the distortion visibility.

3.4.5 Summary

The security and synchronization of the white and colored sequence is analyzed in this section. Although white sequence is often used in the literature, the colored sequence is superior to a white sequence due to its energy efficiency. The wider the bandwidth of the sequence, the more secure it is against guessing attack and more sensitive to synchronization. The random phase sequence whose security lies in its DFT phase is robust against Wiener filtering attacks.

CHAPTER 4

NONLINEAR MODULATION IN OBLIVIOUS STEGANOGRAPHY

In Chapter 3, it is concluded that the Spread Spectrum modulation algorithms have limitations in oblivious applications due to its poor host noise suppression. In this chapter some nonlinear embedding schemes are investigated [27, 29, 31]. These schemes are more effective in oblivious cases.

4.1 Set Partitioning in Oblivious Data Hiding

4.1.1 Hypothesis Testing and Set Partitioning

Watermark is motivated to verify the disputed copyright ownership. Given a multimedia content cover signal, the decoder needs to answer the question Yes/No (watermarked or original) or bit value 1/0 depending on a set of received coefficients. It is a hypothesis testing problem in essence.

Suppose c is an original coefficient in some watermark domain, 1 bit is embedded in c. The received coefficient is denoted as r. Two hypotheses are

$$\begin{cases} \text{H0:} & \text{bit value 0 is embedded in } r \\ \text{H1:} & \text{bit value 1 is embedded in } r \end{cases} \tag{4.1}$$

Obviously, H0 and H1 have different statistical properties. Any steganography scheme modifies the original signal properties in one way or another. Based on the property distinction, the decoder decides whether the bit value is 1 or 0.

A good watermarking (data hiding) algorithm should modify the statistical property of a cover signal without much perceptual degradation. There are several approaches to modify the statistical property. Ramkumar and Akansu *et al.* [70] proposed an innovative approach to flip the signs of some small value coefficients in an image. Statistically speaking, an unmarked original image has approximately equal number of positive and negative coefficients, while the watermarked image has

noticeable count difference between positive and negative coefficients. The decision is made based on this difference. Patchwork [7, 16] and many other schemes change other statistical properties.

To answer the above hypothesis testing problem, a natural question is, how can the decoder make a reliable decision H1/H0 merely on a given r? Begin with the simplest case where no noise exists, answer is simple and straightforward, make H0 and H1 have *no* element in common. Thus, decoder can always make a correct decision.

This embedding works well in a noise-free scenario. Yet in a practical noisy environment, detection is not as reliable as in noise-free cases. To increase its robustness to noise, the element in H0 and H1 should be simply kept some distance apart. That is the simplest way to "separate" them.

This simple idea is extended to the following heuristic data hiding scheme. It is simple yet effective. Two separate sets are constructed on the real axis (Figure 4.1). The embedded coefficient value x should be kept in a set according to the bit value to be hidden. To embed bit value 1, the output coefficient x should be kept in set 1. If the original value c is already in set 1, no modification needed. Otherwise it is replaced by the nearest element in set 1. Similarly after embedding bit value 0, x should be kept in set 0.

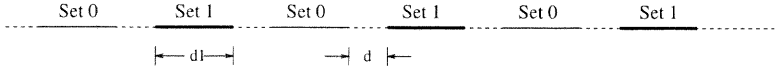


Figure 4.1 Set partitioning scheme

To enhance embedding and extraction reliability, usually one bit information is embedded in a coefficient sequence **c**. To do that, it is need to define a deter-

ministic pattern to represent bit values. For example, to embed 1 bit in a 5-coefficient sequence, patterns similar to antipodal signaling can be defined as

$$\begin{cases} \text{Pattern} \quad \text{A (bit 1):} \quad [\text{set 1, set 0, set 1, set 0, set 1}] \\ \text{Pattern} \quad \text{-A (bit 0):} \quad [\text{set 0, set 1, set 0, set 1, set 0}] \end{cases} \tag{4.2}$$

To hide an information bit, the modified sequence $\mathbf{x}$ should comply with Pattern A (to hide bit 1) or Pattern -A (to hide bit 0). For example, to hide bit value 1, after embedding the output coefficients should be

$$x_0 \in \text{set 1},\ x_1 \in \text{set 0},\ x_2 \in \text{set 1},\ x_3 \in \text{set 0 and } x_4 \in \text{set 1}.$$

This method is named *set partitioning*. It does not hide a specific watermark signal in a cover signal, but try to modify its statistical property to facilitate the detection at decoder. Watermarking is a game played between robustness and distortion. The more distortion it introduces, the more reliable it could be.

4.1.2 Average Distortion

In the calculation of the distortion energy, for simplicity, it is assumed the original coefficient c is uniformly distributed in the region $(-a, a)$. This assumption is true for the data in spatial or time domains, although may not accurate for coefficients in transform domains. It is reported the coefficients are approximately Laplacian distributed [5]. Simulation studies show that distortion difference due to the pdf is negligible.

Denote the error introduced in embedding as $e = x - c$. As depicted in Figure 4.2, suppose the bit value 1 is to be embedded, consider the typical region AD:

If c is in the range AB, no modification needed, $e = 0$.

If c is in the range BD, e is uniformly distributed in $(-d - d1/2, d + d1/2)$.

The corresponding conditional probabilities can be expressed as

$$P(c \in AB | c \in AD) = \frac{d1}{2d1 + 2d} \tag{4.3}$$

and

$$P(c \in BD | c \in AD) = \frac{2d + d1}{2d1 + 2d}. \tag{4.4}$$

Therefore, the average distortion energy introduced is

$$D = \frac{(2d + d1)}{(2d1 + 2d)} \cdot \frac{(2d + d1)^2}{12} = \frac{1}{12} \frac{(2d + d1)^3}{(2d + 2d1)}. \tag{4.5}$$

In a similar fashion, the distortion calculation yields the same result if bit value 0 is embedded. Th average is just given by (4.5).

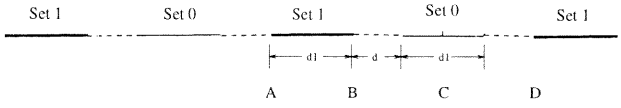


Figure 4.2 Average distortion calculation

4.1.3 Hard Decision Detection

In one bit per coefficient embedding, the hard decision is based on the distance between the received coefficient r and the two sets (here distance is defined as the minimum distance between r and any element in the set, it is zero if r belongs to the set).

In practice, it is rare to embed one bit in one coefficient. Consider the case where 1 bit is embedded in an N-coefficient sequence **c**. The simplest detection is majority vote. This is the hard decision based on individual coefficient. Real axis is divided into two decision regions (Figure 4.3). If the received coefficient r falls in Region 1, it is decided the transmitting signal x comes from set 1. Otherwise it is assumed it comes from set 0. In the example discussed in Section 4.1.1, if a received sequence pattern is [set 0, set 0, set 1, set 0, set 0], which is more similar to pattern A (2 coefficient difference) than to Pattern -A (3 coefficient difference), the decision is made in favor of bit value 1.

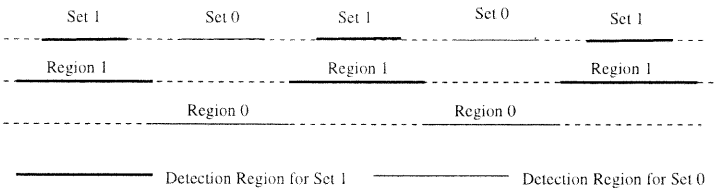


Figure 4.3 Hard decision region

4.1.4 Maximum Likelihood Detection

The above simple detector makes decision based on the individual coefficients. Detection reliability can be enhanced using a soft-decision detector.

Denote r as the received coefficient after Gaussian channel transmission, noise $n \sim N(0, \sigma^2)$. The ML likelihood ratio [48] is

$$R = \frac{P(x \in \text{set } 1|r)}{P(x \in \text{set } 0|r)}. \tag{4.6}$$

In those two sets, there are infinite transmitting signals. Denote any element in these two sets as ξ (set 1) and τ (set 0), and rewrite the above equation,

$$R = \frac{\sum_{\xi \in \text{set } 1} P(\xi|r)}{\sum_{\tau \in \text{set } 0} P(\tau|r)}. \tag{4.7}$$

Using

$$P(\xi|r) = \frac{P(\xi) f(r|\xi)}{f(r)}. \tag{4.8}$$

and

$$P(\tau|r) = \frac{P(\tau) f(r|\tau)}{f(r)}. \tag{4.9}$$

Equation (4.6) becomes

$$R = \frac{\sum_{\xi \in \text{set } 1} P(\xi) f(r|\xi)}{\sum_{\tau \in \text{set } 0} P(\tau) f(r|\tau)}. \tag{4.10}$$

Gaussian noise probability density function is

$$f(r|\xi) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp[\frac{-(r-\xi)^2}{2\sigma^2}]. \tag{4.11}$$

The original coefficient c is uniformly distributed, its probability density function $f(c) = \frac{1}{2a}$, $c \sim U(-a, a)$. After embedding information bit value 1, the calculation of the probability of the transmitting signal $P(\xi)$ is depicted in Figure 4.4. Note the probability pulses at the end points. They are transmitted with greater probability because any c out of the set 1 is replaced by the end points.

$$\begin{aligned} \sum_{\xi \in \text{set } 1} P(\xi) f(r|\xi) \quad &= \tfrac{1}{2a} \int_{r-l_1-d1}^{r-l_1} \tfrac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(\xi-r)^2}{2\sigma^2}} d\xi \\ + \frac{1}{\sqrt{2\pi}\sigma} \frac{d + d1/2}{2a} e^{\frac{-l_1^2}{2\sigma^2}} \quad &+ \tfrac{1}{2a} \int_{l_1-2d-3d1}^{l_1-2d-2d1} \tfrac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(\xi-r)^2}{2\sigma^2}} d\xi + \ldots \end{aligned} \tag{4.12}$$

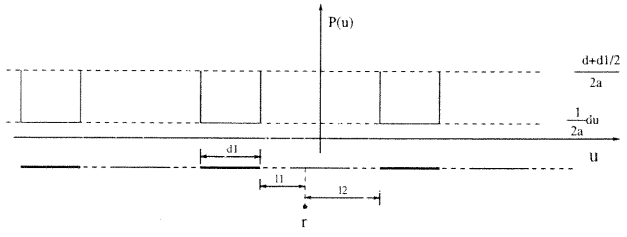


Figure 4.4 Calculation of ML ratio

In the similar manner, $\sum_{\tau \in \text{set } 0} P(\tau) f(r|\tau)$ can be calculated and yields a result similar to (4.12). Still a closed-form result of ML ratio can not be obtained. The ML detector is also too computation expensive. Besides, the detector needs the value of the noise power σ^2, which is usually unavailable at decoder. The ML optimum detector is infeasible in practice.

The challenge in the decoding is that the transmitting signals could be any value belonging to these two sets. The ML ratio calculation thus involves all elements in set 1 and set 0. A way to simplify detection is to assume transmitting signals finite. In the following suboptimal methods, the transmitting signals are assumed discrete instead of continuous.

4.1.5 Suboptimal Detection

In the first suboptimal detection, the transmitted signals are simply assumed at the center of the continuous segments, the signaling is a pattern like xoxo as depicted in Figure 4.5. Signal points x and o are transmitted with equal probability.

After approximation, the ML ratio can be expressed as:

$$R = \frac{P(x \in \text{set } 1|r)}{P(x \in \text{set } 0|r)}. \tag{4.13}$$

Still there are many x and o points to be considered.

Simulation studies demonstrate that it can be further simplified by merely considering the nearest x and o points (See the Section 4.2). Thus, (4.13) reduces to

$$R = \frac{P(r|x = u)}{P(r|x = v)}, \tag{4.14}$$

where u/v is the nearest transmitting points x/o in set 1 and set 0.

Now the suboptimal detector is in a form of minimum distance detector. Only the nearest transmitting points are considered due to their higher transmitting probabilities.

Some other assumptions result in a different form of suboptimal detectors. In Figure 4.4, it is observed that the endpoints are transmitted with higher probabilities because those original coefficients not in the two sets are replaced by the end points. Another reasonable approximation assumes the transmitted signals have xxoo pattern as shown in Figure 4.6.

In that case, it is reasonable to assume that only the nearest end points are considered as transmitting signals, that yields the same results as (4.14).

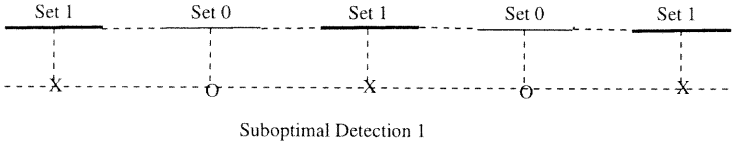


Figure 4.5 Suboptimal detection 1

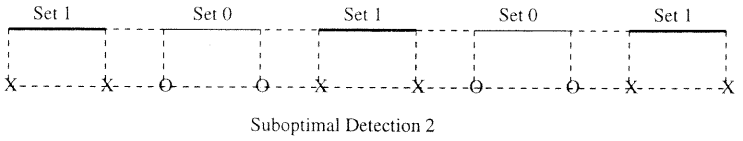


Figure 4.6 Suboptimal detection 2

In the example mentioned in Section 4.1.1, suppose a 5-coefficient sequence **r** is received. The nearest x and o points to r_i are located and denoted as u_i (in Set 1) and v_i (in Set 0). According to the given patterns (A and -A), two corresponding sequence candidates are constructed,

$$\begin{cases} \text{Pattern} \quad \text{A type:} & \mathbf{a} = [u_0, v_1, u_2, v_3, u_4] \\ \text{Pattern} \quad \text{-A type:} & \mathbf{b} = [v_0, u_1, v_2, u_3, v_4] \end{cases} \tag{4.15}$$

If $||\mathbf{r} - \mathbf{a}|| < ||\mathbf{r} - \mathbf{b}||$, the received sequence is more similar to the Pattern A, bit value 1 is decided; Otherwise, bit value 0 is decided.

The two suboptimal detectors demonstrate different performance.

4.1.6 Experiments and Results

To evaluate this set partitioning scheme, the extracted Bit Error Rate (BER) in Gaussian noise environment versus the distortion introduced is measured. The Signal-Noise Ratio (SNR) is redefined as the ratio of the distortion energy S over the noise power σ^2.

$$SNR = \frac{S}{\sigma^2} \tag{4.16}$$

The three detectors are compared. One information bit is embedded into an 11-coefficient sequence (Fig. 4.7). The ratio is selected $d/d1 = 1$. The result shows that suboptimal detector 2 outperforms suboptimal detector 1. Further simulation shows decoding performance in suboptimal detector 2 is almost the same as the ML optimum detector. Both suboptimal methods far outperform the hard decision decoder.

It is observed that BER-SNR is different for different d and $d1$ values. The determining factor is the ratio $d/d1$, not individual values of d or $d1$. Figure 4.8 is the result of embedding 1 bit in an 8-coefficient sequence.

In practice, an accurate prediction of the channel noise property may not be known in advance. However, data hiding seldom works at higher SNR, usually $SNR < 1$. Embedding distortion is not expected to be larger than the moderate or severe compression distortion. Therefore, smaller $d/d1$ is more favorable in applications. That implies the smaller $d/d1$ is more reliable in noisy scenarios.

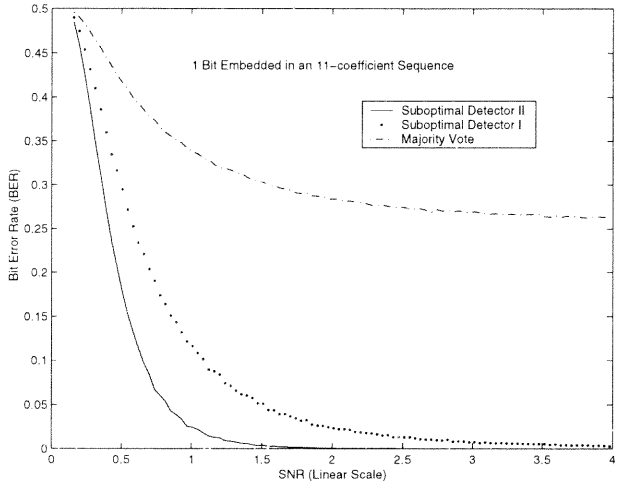


Figure 4.7 Detection performance comparison

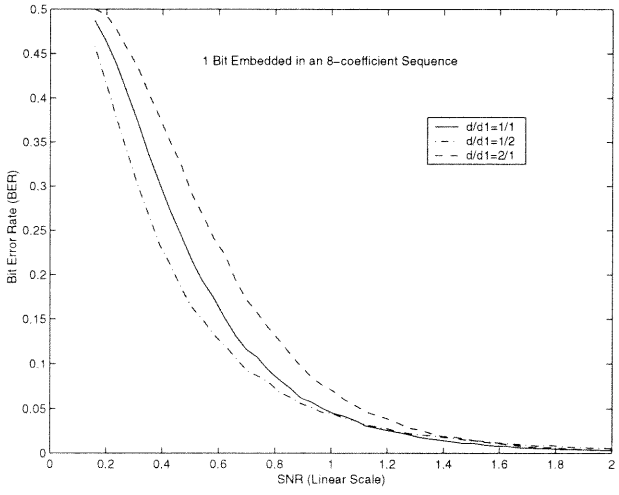


Figure 4.8 Performance with different $d/d1$

The set partitioning scheme may be used in place of the SS modulation in various watermark domains. It shows great advantage in host noise suppression. In image data hiding experiments, information bits are embedded in the DFT amplitude domain. A pattern is embedded in the medium frequency coefficients. Total 64 bits are hidden in a 256x256 images. Experiments demonstrates its robustness against common compression and other attacks.

4.2 QIM Embedding and Detection

Chen and Wornell *et al.* [12, 11] applied dither modulation technique as a special case of Quantization Index Modulation (QIM) for oblivious watermarking. It can achieve more reliable extraction without referring to the original cover signal.

4.2.1 QIM in Oblivious Data Hiding

In the SS modulation schemes, a fixed watermark signal is superimposed on the original signal. The set partitioning scheme modifies a coefficient only when necessary thus minimizing the distortion. There are several approaches to hiding information in the oblivious applications. The greatest challenge is that the original signal is unknown. If a good estimate of the original signal is obtained, the detection reliability will be boosted.

Given a received coefficient r, what is the original value? It is reasonable to assume the original value must be close to r. A good estimate of the unknown original is its quantized version $Q(x, \delta)$ where δ is the quantization step size. The difference between the "estimated cover signal" and the received coefficient x is the small value signal embedded which could be extracted as

$$s = Q(x, \delta) - x. \tag{4.17}$$

After fixing the signal extraction method, the corresponding invertible embedding operator is not difficult to find as

$$x = Q(c + s, \delta) - s, \tag{4.18}$$

where c is the original coefficient, x is the marked coefficient after embedding and s is the watermark signal.

The invertibility of encoding and decoding can be demonstrated by the following example. Suppose an antipodal signal s or $-s$ is to be embedded to hide information bit value 1 or 0. If the original value $c = 26.40$, quantization step size $\delta = 1.0$ and watermark signal $s = 0.25$, the marked coefficient $x = Q(c + s, \delta) - s = Q(26.65, 1.0) - 0.25 = 26.75$. In a noise-free scenario, the extracted signal is again $s' = Q(x, \delta) - x = 27.00 - 26.75 = 0.25$.

4.2.2 Maximum Likelihood Detection in QIM

The embedding operator (4.18) and extraction operator (4.17) are invertible in noise free scenario. The final decision could be based on the correlation value of the extracted signal and the watermark signal. Yet it is far from optimum in noisy case.

Continue with the above example, suppose the bit value 1 is embedded, the marked coefficient $x = 26.75$. After noise channel, if received value $r1 = 26.51$, the extracted signal $s' = Q(r1, \delta) - r1 = 0.49$. If $r2 = 26.49$ is received instead, $s' = Q(r2, \delta) - r2 = -0.49!$ $r1$ and $r2$ are quite close, nevertheless results in two totally different extracted signals. The reason is that the quantization operation is nonlinear and has discontinuity around the points xxx.50.

The quantization operator is not necessary. In this scheme, it is needed to decide a received coefficient r comes from x points or from o points. The Maximum Likelihood (ML) ratio is [48]

$$R = \frac{P(x \in \text{Set } 1|r)}{P(x \in \text{Set } 0|r)}, \tag{4.19}$$

If $R > 1$, the bit value 1 is decided; Otherwise bit value 0 is decided.

The probability calculation is a little complicated. There exist many signal points corresponding to one information bit. On a received coefficient r, it is known that the transmitted signal x could be xxx.75 or xxx.25. Suppose $r = 6.30$ is received, all the possible transmitted signals can be divided into two sets.

$$\begin{cases} \text{Set 1}: & \{\mathbf{6.75}, 5.75, 7.75, 8.75, 4.75, ...\} \\ \text{Set 0}: & \{\mathbf{6.25}, 7.25, 5.25, 8.25, 4.25, ...\} \end{cases} \tag{4.20}$$

Set 1 represents information bit value 1; Set 0 represents bit value 0.

If the noise is Gaussian distributed, its pdf is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{-x^2}{2\sigma^2}). \tag{4.21}$$

The probability $P(x \in \text{set } 1, r)$ can be calculated as

$$\begin{aligned} P(x \in \text{set } 1|r)P(r) &= P(r = 6.30|x = 6.75)P(x = 6.75) \\ &+P(r = 6.30|x = 7.75)P(x = 7.75) \\ &+P(r = 6.30|x = 5.75)P(x = 5.75) \\ &+... \end{aligned} \tag{4.22}$$

$P(x \in \text{set } 0|r)P(r)$ can be obtained as well.

Assume the probabilities for all transmitting signals are equal

$$P(x = 6.75) = P(x = 6.25) = P(x = 5.75) = ... \tag{4.23}$$

Equation (4.19) can be reduced to

$$R = \frac{P(r = 6.30|x = 6.75) + P(r = 6.30|x = 5.75) + ...}{P(r = 6.30|x = 6.25) + P(r = 6.30|x = 7.25) + ...} \tag{4.24}$$

The above equation involves many terms, no closed-form result can be obtained. The dominating element in each set is defined as *leader*. In the above example, the leaders in Set 1 and Set 0 are $u = 6.75$ and $v = 6.25$. They are the most likely candidates. If all the remaining terms are neglected, the ML ratio (4.24) becomes

$$R \approx \frac{P(r|u)}{P(r|v)} = \frac{\exp[\frac{-(r-u)^2}{2\sigma^2}]}{\exp[\frac{-(r-v)^2}{2\sigma^2}]}. \tag{4.25}$$

Simulation shows the above is a good approximation for Gaussian noise environment. The result is similar to that obtained above in set partitioning calculation.

In fact, the idea of QIM has long been used in some watermarking algorithms. It is similar to the parity manipulation schemes. Some schemes in this category modify the parity of an integer coefficient c. For example, c can be modified to an even number to embed bit value 1, or to an odd number to embed bit value 0 [2]. In fragile watermarking, the DCT coefficients are modified in a similar way for image authentication [100]. Its embedding procedure is, in essence, the same as QIM scheme. The detection used usually is hard decision detector, i.e. a majority vote. In the above example, if even integers out-count odd ones, bit value 1 is decided; Otherwise it is decided 0. It is inferior to the above soft decision suboptimal detector.

Figure 4.9 depicts the comparison results of the detectors: majority vote detector, correlation using quantization operation and soft decision detector. The original coefficient c_i is Gaussian generated with variance $\sigma = 80$. The soft decision detector yields the better result over the other two.

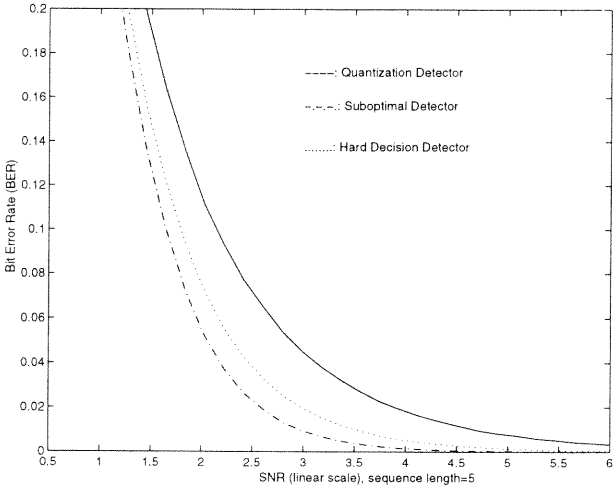


Figure 4.9 Detection performance in QIM

4.2.3 Performance Analysis

From the above analysis, it is found that the QIM scheme is superior to the SS modulation in oblivious applications. Chen *et al.* [11, 12] pointed out that BER in QIM is calculated as

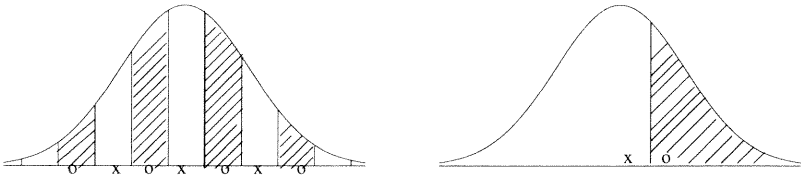
$$BER = Q(\frac{d}{2\sigma}). \tag{4.26}$$

where d is the distance between the x and o points, σ is the noise variance value.

The analyzed BER value is the same as the non-periodic antipodal communications cases. This is true only when the SNR value is very large. In most data hiding applications where SNR is usually low, this periodic scheme is far from the non-periodic scheme. BER in the non-periodic antipodal case is simply given by (4.26).

The BER in QIM case is the shadowed area in Figure 4.10,

$$BER = \int_{-d}^{0} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x+d/2)^2}{2\sigma^2}} dx + \int_{d}^{2d} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x+d/2)^2}{2\sigma^2}} dx + ... \tag{4.27}$$



(a) Periodic Signaling (b) Non-periodic Signaling

Figure 4.10 BER calculation in QIM and antipodal case

The gap between QIM and antipodal cases is depicted in Figure 4.11. In the antipodal case, the transmitting signals are fixed, s or $-s$. While in the QIM case, there exist many transmitting signals, the decoder never knows for sure which is the transmitted signal and has to "guess" one (the nearest one in the suboptimal detectors). The performance degradation can be regarded as a price paid for the "uncertainty" at decoder.

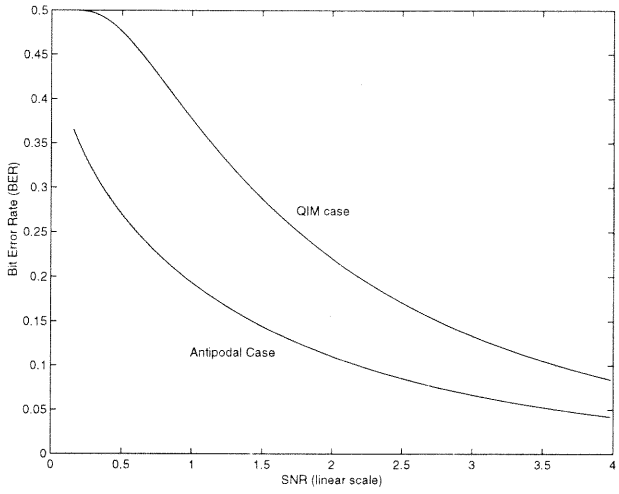


Figure 4.11 BER in periodic and non-periodic signaling

4.2.4 Comparison With Set Partitioning

The QIM scheme can be viewed as a special case of the proposed set partitioning scheme in Section 4.1. The latter provides the flexibility to choose different d and $d1$ values. The ratio value $d/d1$ has different implications in practice. The performance with different $d/d1$ ratio values is compared with the QIM scheme (Figure 4.12 and Figure 4.13).

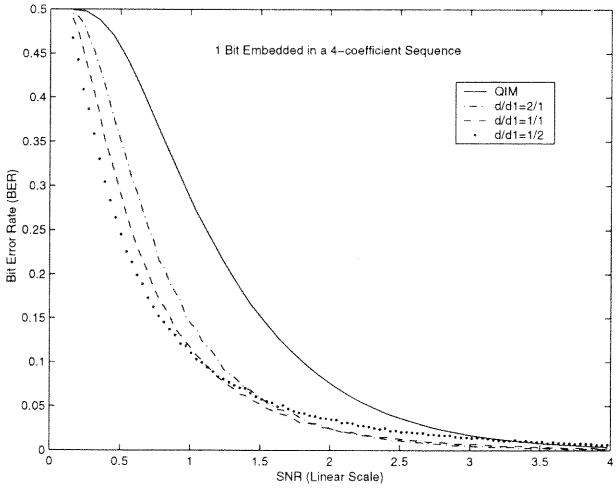


Figure 4.12 Performance at lower SNR

Observe the BER-SNR curves for $d/d1 = 1$, at lower SNR where most data hiding applications are employed the improvement over QIM is noticeable. One may

notice that with fixed average distortion energy, the set partitioning scheme has larger maximum distortion amplitude. Even considering this, at a given maximum distortion (that implies higher distortion energy in QIM), simulation studies show the set partitioning scheme is still superior to the QIM. This is a quite effective oblivious hiding scheme.

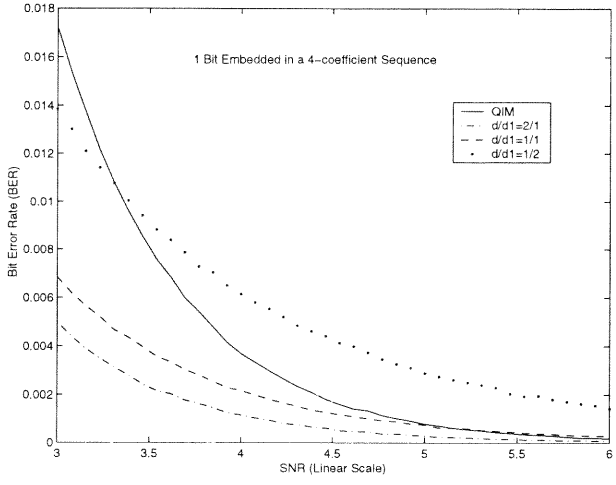


Figure 4.13 Performance at higher SNR

Note that at different SNR the comparison result is different. Figure 4.12 and Figure 4.13 show that the smaller $d/d1$ performs better at lower SNR. At higher SNR, larger $d/d1$ is more advantageous.

4.3 Limitations of Set Partitioning

As seen above, the set partitioning scheme is powerful in oblivious data hiding. The QIM periodic scheme can be regarded as a special case of this approach. Nevertheless, it has several limitations in watermarking applications. The latter just divides the data into two sets and there is no other constraint for the set signaling. For example, usually the relative distortion is much more important than the absolute distortion for human perception. The set can be design with more distortion at the high value end.

Both the encoder and decoder know the set signaling. Once the signaling of the two sets is known, it is not difficult to remove the watermark. To enhance the security, the signaling scheme should be kept secret. Ramkumar and Akansu *et al.* [72] proposed a QIM like scheme with a known signaling and a random transform. Without knowing the transform, it is more difficult to remove the watermark. One side effect is that in a random transform domain, it is difficult (if not impossible) to control the artifacts.

An alternative is to introduce randomness into the signaling by shifting the set signaling. For example, instead of modifying the original coefficient c to comply with a set pattern, the value of $c + p$ can be modified so that it complies with the same pattern where p is a random variable. This may in some degree enhance its robustness against attacks.

Even with the suitable transform, it is still a little difficult to control the artifacts. In the cases where the signaling is adaptive to some perception control (for audio signals, the permitted distortion in every subband is different), the decoder should know the signaling change as well. Since the encoder has no way to notify the change, the decoder has to "guess" the signaling used at encoder. This greatly increases the computation complexity at decoder. A more severe problem is that after sever compression, the signaling scheme estimated by decoder may be different from the one used in data embedding. The message extraction is sensitive to this "set signaling error".

CHAPTER 5

CONTENT PROTECTION IN AUDIO SIGNALS

In the previous chapters, multimedia data hiding in general is studied. In this chapter, data hiding applications in audio signals are explored, and several algorithms for compression-resistant hiding schemes are investigated.

5.1 Introduction to Audio Compression

In compression resistant data hiding design, it is important to understand the compression algorithms. Better understanding of the compression can lead to a more robust and effective scheme. In this section, popular audio compression schemes are reviewed. The current multimedia compression methods can be roughly categorized into three groups, *waveform approximation, perceptual coding* and *parametric coding.*

5.1.1 Waveform Approximation Coding

The goal of the waveform approximation is to construct a compressed version closest to the original waveform at a given bit rate. In other word, it aims at the highest SNR. Usually, none or only very little perceptual knowledge is employed. One example is the Adaptive Differential Pulse Code Modulation (ADPCM). Other examples include speech type narrow-band audio signal compression G.721 and G.723.

The schemes are also used in image compression where most schemes are after the highest SNR. In current popular schemes, wavelet, sub-band or other various filtering techniques are exploited to reduce the redundancy without much consideration of the Human Visual System (HVS). Lots of research is focused on quantization procedure after filtering. Zerotree [82], SPIHT [79] are some of the recent achievements. In JPEG compression, basic human perpetual knowledge (human eyes are more sensitive to low frequency components than to the high frequency ones) is considered in Q-table design. Still it is very simple and crude.

This kind of compression is based on information rate-distortion theory. MPEG-1 and MPEG-2 video compression and various image compressions fall into this category. The compression noise introduced is usually not very large.

5.1.2 Perceptual Coding

Perceptual coding is widely used in well-advanced audio compression. An outstanding example is the advanced audio compression MP3 which is ubiquitously used in Internet transmission and storage.

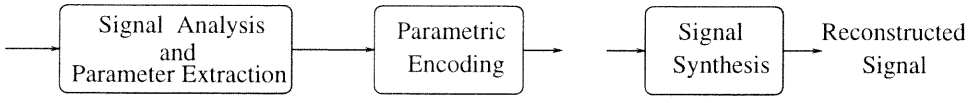
It was realized that the waveform approximation coding schemes were not successful in audio applications. The reason is multi-fold. The audio signal is "hard" to compress. It is a kind of non-stationary signal, or at most a *quazi-stationary* signal. Secondly, people do not measure "audio quality" by "square error" and Human Audio System (HAS) is much more sensitive than the Human Visual System (HVS).

To obtain high coding gain, popular audio compression schemes such as MP3, MPEG-2 AAC-2, and Dolby AC-3, NTT Twin-VQ, etc., all explicitly make full use of the human psychoacoustic model. There have already existed several effective psychoacoustic models. These models which are best described in transform domain are used in quantization operation to shape the compression noise. Compared with the waveform approximation, the compression in this group is much more efficient. Although compression in this group is based on human perception, it uses quantization error as a distortion measure. In contrast, *parametric coding* does not use the error as a quality benchmark.

5.1.3 Parametric Coding

Parametric coding is a complicated technique in audio compression. Some schemes of this kind are standardized, for instance, in MPEG-4 parametric coding [1]. The compression might achieve very high compression rate.

The basic principle in the compression is not to "encode" a signal but to "describe" the signal. The original signal is analyzed and the parameters describing the signal are extracted. These parameters are then encoded and transmitted. At the decoder side, the audio signal is reconstructed with these parameters. Its waveform may not be close to the original, and SNR may be very low. Still it presents the same perceptual effects as the original.



(a) Parametric Encoder (b) Decoder

Figure 5.1 Audio parametric coding

The encoder is composed of two parts, parameter extraction and parameter encoding. The most complicated part is model-based parameter estimation. There are several models underlying the compressions.

In the audio analysis, several different models are studied for parameter estimation. In MPEG-4 parametric coding [1], the Harmonic and Individual Lines plus Noise (HILN) tool is adopted for audio parametric coding. In this model, signal is regarded as a combination of harmonic component (one fundamental frequency and a couple of harmonic components), individual frequency lines and noise component. It is claimed suitable for less complicated audio signals. An advanced audio signal is modeled as [56, 94, 95]

$$\text{Audio=Sines + Transients + Noise}$$

The sinusoidal waves are the most significant components in the audio signals. The transients are broadband signals that do not have tonal peaks. They are also referred to as *attacks* in audio compression. The last significant component is noise. It is claimed this compression achieves the same perceptual effect at same bit rate compared with most complicated perceptual coder MPEG-2 AAC. This model will be revisited in some length in Section 5.7.

5.2 MP3: A Typical Perceptual Audio Compression

MP3 (MPEG-1 Layer III) is a widely used audio compression algorithm. It has become the standard in audio signal transmission and storage.

As a typical perceptual coding scheme, MP3 is composed of several blocks (Figure 5.2). The time-domain waveform is transformed to the subband domain to remove the redundancy. The input signal also goes through psychoacoustic analysis. The output is used to shape the quantization noise according to the masking threshold curve. The final stage is bitstream formation.

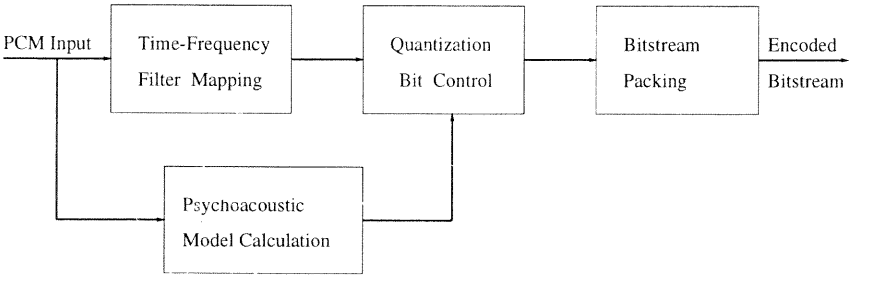


Figure 5.2 A typical audio perceptual compression block

5.2.1 Sub-band Filtering and MDCT

The MPEG-1 layer III is an extension of the MPEG-1 layer I and II. In the layer I and II, a polyphase filter bank is employed for time-frequency mapping. The filter bank is composed of 32 filters, each with equal bandwidth. In Layer III, each output channel is further subdivided into 18 bands via a windowed Modified Discrete Cosine Transform (MDCT) for a better frequency resolution.

A fine frequency resolution is preferred for signal redundancy reduction, which favors the long transform length selection. On the other hand, for the attack signals (transients), quantization using long transform length tends to produce some "spreading" effect in time domain. This makes the attack signal not so "crisp". It is well known that the fine time resolution and frequency resolution can not be achieved simultaneously. The transform length should be adaptive. For a stationary

signal segment, long block is used. If an attack is present, short block should be used instead. The decision to switch between long and short transform is based on perceptual entropy first proposed in [47].

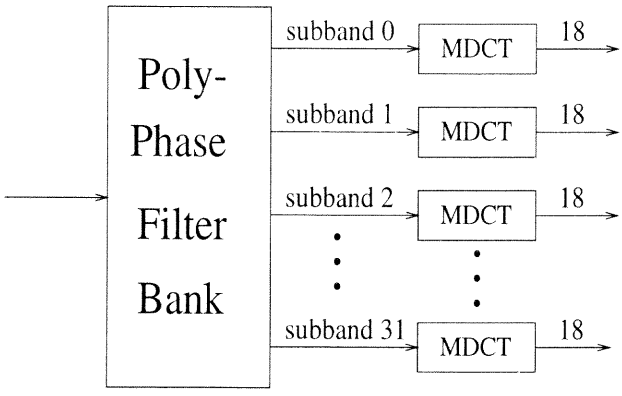


Figure 5.3 Subband filtering and MDCT

5.2.2 Frequency Masking

Human psychoacoustic model plays a very important part in perceptual coding. The psychoacoustic studies have made significant progress in characterizing human auditory perception and several perceptual models have been developed and applied in audio coding.

Both subjective experiments and studies show that human ears perceive the audio signal within an interval of time. The perception procedure is analogous to the short-time spectral analysis. Distinctive regions in the cochlea perceive different frequency components. These frequency partitions are called *critical bands.*

People tend to "mix" the effect of the frequency components in one critical band. The subjective response to the components out of the critical band is abruptly changed. Empirical work shows the human audible frequency range is divided into $23 \sim 27$ critical bands, each with different bandwidths. The critical bandwidths are increasing towards the high frequency end. The distance of 1 critical band is

referred to as 1 *bark*, which is a nonlinear measure scale used in psychoacoustics. An empirical formula to convert from Hz to Bark scale is

$$z(f) = 13 \cdot arctan(0.00076f) + 3.5 \cdot arctan[(\frac{f}{7500})^2] \ \ (Bark) \qquad (5.1)$$

Human frequency masking takes place for inter-bands and intra-bands. Simply put, a component (masker) can "mask off" another component (maskee), rendering it less audible. Above some threshold value, the masker completely masks the maskee, making it inaudible. The audibility masking value is called *masking threshold*, which is not only related to the loudness of the masker, but also to the "tonality" of the masker. Psychoacoustic experiments reveal that the SMR (Signal-to-Mask Ratio) of pure sinusoids is much larger than that of the white noise signal. In other words, a noisy component is a better masker than a tonal component. If the compression noise is completely under the "masking threshold curve", it is inaudible. The psychoacoustic output is used to control the quantization procedure.

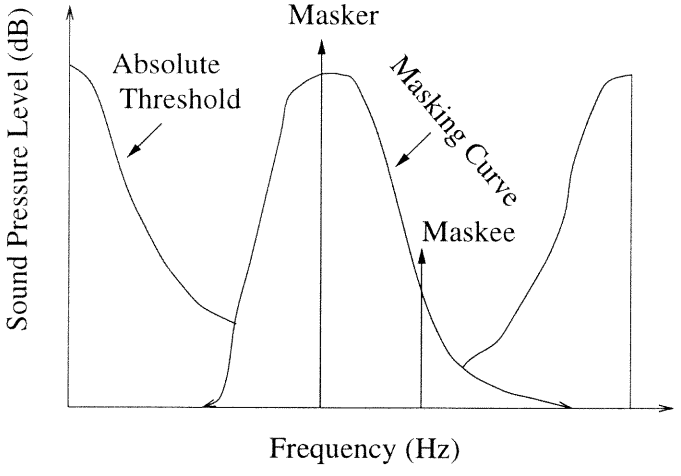


Figure 5.4 Frequency masking effect

The psychoacoustics analysis in MP3 compression is a quite complicated procedure. The Hann-windowed FFT complex spectrum of the input signal is calculated. Then the unpredictability, which is a rough tonality measure is calculated. The frequency bins are grouped into *threshold calculation partitions*

which are approximately one-third Bark band scale. The signal energy in each partition is summed-up. With the obtained unpredictability measure, the partitioned energy is convolved with the spreading function which models the natural excitation spreading along the basilar membrane in the cochlea. Subsequently, the actual energy threshold in one band is calculated and spread over all FFT lines. Considering absolute thresholds, the final energy threshold of audibility is obtained. Regrouping the threshold values into *scale factor bands* (frequency bins in a same band share a same scale factor, resulting in an equal quantization resolution), and the distortion threshold ratio is therefore obtained.

$$r = \frac{\text{allowed distortion energy}}{\text{scale-factor band energy}}. \tag{5.2}$$

Figure 5.5 is a psychoacoustic analysis output in one granule. It is used in the subsequent quantization and quantization iteration processing.

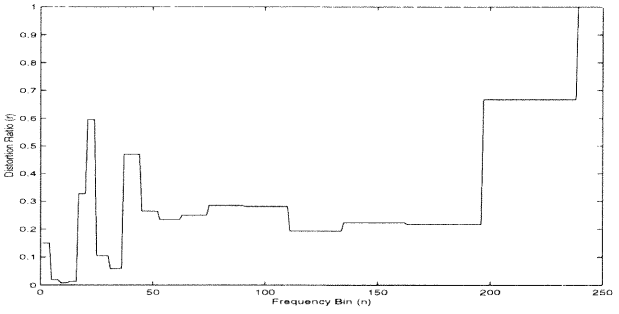


Figure 5.5 Scale-factor band distortion ratio

5.2.3 Temporal Masking

Temporal masking is the masking effect taking place in the time domain. HAS system perceives the audio signal in an interval of time. The perceived effect is a "sum-up" effect during the interval. The masker has both pre-echo and post-echo effects in the time domain as depicted in Figure 5.6. The masker masks off the signal whose londness is under the audibility threshold curve.

Please note the masking effect is different in pre-masking and post-masking regions. The pre-masking only lasts half a dozen milliseconds while post-masking can extend to 50 $\sim$ 300 milliseconds depending on the londness and duration of the masker. Usually, only "pre-echo" effect is considered in compression.

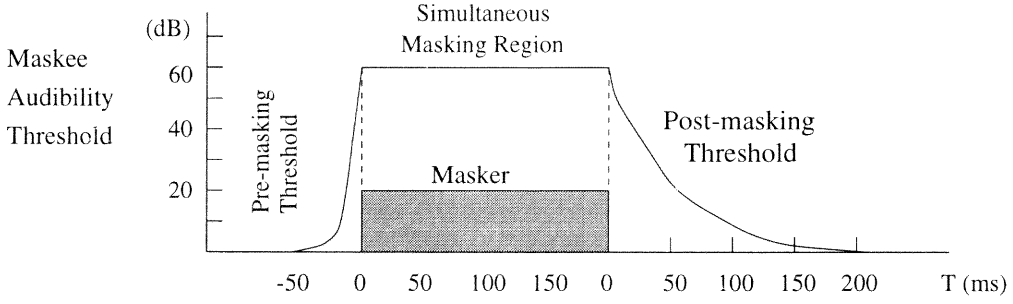


Figure 5.6 Temporal masking effect

The temporal masking effect is studied in audio coding to address the so-called "pre-echo" problem. Although the compression noise in frequency domain can be well shaped by employing the frequency masking property, it is difficult to control the quantization noise in time domain. In the transform coding the quantization noise spreads in the time domain within the transform block, which could result in audible compression noise. This only happens at attack (transient) signals, for instance, a castanet, or the beginning when a key is stricken. Based on temporal masking property, the common remedy is the selection of sufficiently short transform. MP3 adopts transform length switch to solve the problem. The transform in MP3 is adaptive to the signal property. Most of the time the signal is regarded as stationary. Long transform length is used for fine frequency resolution (thus higher coding gain). At the time of the signal abrupt change (transients), a short transform length should be used to prevent pre-echo artifacts. The switch decision depends on *perceptual entropy* [47] and it must be gradual for the perfect reconstruction purpose [44].

5.2.4 Quantization and Distortion Control

The MDCT coefficients are small floating-point numbers. All is multiplied by a global gain. Besides, every MDCT coefficient in one scale factor band is multiplied by a common scale factor. The MP3 quantization procedure is composed of two loops —- inner loop for bit rate control and outer loop for distortion control.

- *Inner Iteration Loop*: After quantization, the coefficient values are Huffman-coded. If the bit consumption is larger than the bits available, the global gain is decreased by one step size so that quantization noise is increased and the bit consumption is decreased. The operation repeats several times until the bit consumption is less than the bits available.

- *Outer Iteration Loop*: After the inner loops, the quantization noise is calculated in each scale factor band. If the noise is larger than the permitted distortion obtained out of the psychoacoustic analysis, this scale factor in this band is increased, resulting in finer quantization and less distortion. The re-quantization goes on until the noise is completely masked off.

This quantization operation is quite complicated. Usually, it takes 12-17 loops to finish it. It is possible that the rate and distortion requirements might not be met at the same time. The iteration should be terminated according to other conditions, for example, after a given maximum loops. The coding bits needed are not the same for different segments in an audio signal. To absorb bit consumption imbalance, "bit reservoir" technique is employed. The current frame is permitted to "borrow" bits saved from past frames, if necessary.

Figure 5.7 depicts the flow chart of MP3 compression. The real compression operation is quite complicated and is composed of the block and window length switches and other algorithms for effective stereo signal coding. For more detailed description on MP3, refer to [10, 23, 65].

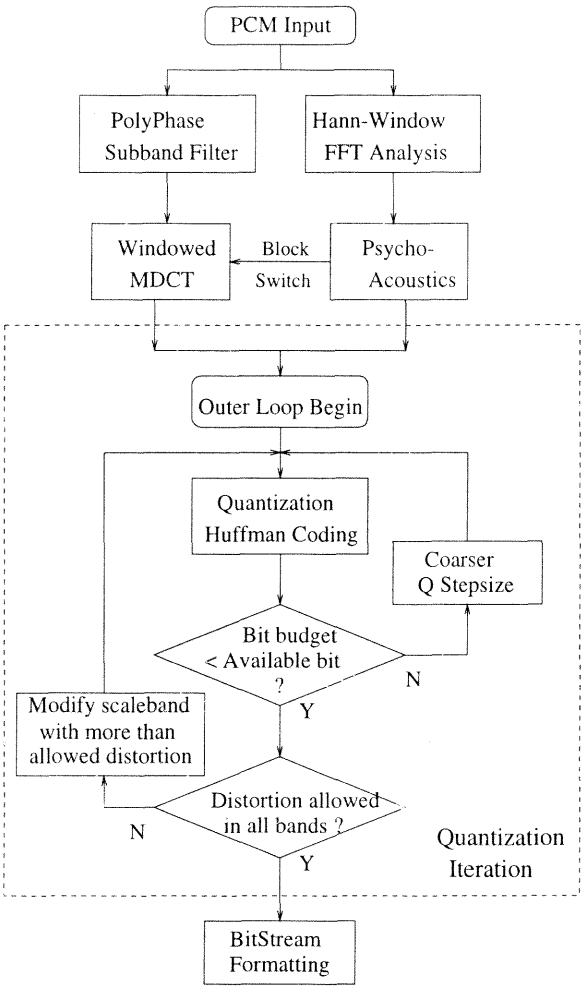


Figure 5.7 MP3 encoding flow chart

5.3 Amplitude Modulation Data Hiding

The spread spectrum modulation can be extended to audio applications as well [28]. In this section, this technique in audio data hiding is employed and some results are presented.

5.3.1 Hiding and Extraction

The data hiding scheme is the direct extension of the SS modulation widely used in image and video applications. A small valued PN sequence is embedded in the

original signal. Suppose one information bit is to be embedded in a coefficient sequence $\mathbf{x}$. The embedding procedure used is deep embedding

$$x_i' = \begin{cases} x_i + w_i|x_i|\alpha_i, & \text{to embed bit value 1} \\ x_i - w_i|x_i|\alpha_i, & \text{to embed bit value 0} \end{cases} \tag{5.3}$$

where $\mathbf{w}$ is a random bipolar sequence (w_i is either -1 or +1) and α_i is the threshold ratio.

The watermark domain could be the subband domain or other transform domains. It is advised that the watermark domain should facilitate distortion control. If the subband domain is used, the same psychoacoustic model used in MP3 compression can be applied in the artifact control.

The output of the psychoacoustic model analysis is the energy threshold ratio r_i in one scale factor band. The amplitude threshold ratio should satisfy

$$\alpha_i \leq \sqrt{r_i}. \tag{5.4}$$

It is realized that the psychoacoustic model used in the compression is not quite accurate in data hiding. The energy distortion control does not imply that artifact is inaudible as long as the energy in the critical band is unchanged. Although people tend to "mix" the frequency components in one critical band, distortion may still be perceived. That means the amplitude modification must be sufficiently small. In practice, the selection of the value α_i should be smaller than $\sqrt{r_i}$. In the sensitive low frequency range, α_i should be further tuned.

Given the random seed, the decoder generates the random sequence $\mathbf{w}$. The correlation detector output is

$$q = \sum_{i=0}^{N-1} r_i w_i = \sum_{i=0}^{N-1} x_i w_i + \sum_{i=0}^{N-1} x_i n_i \pm \sum_{i=0}^{N-1} |x_i|\alpha_i, \tag{5.5}$$

where r_i is the received coefficient and n_i is the channel noise. If $q > 0$, the decision is bit value 1; Otherwise bit value 0 is decided instead.

It should be noted that the watermark domain selection is not limited to the subband decomposition used in MP3. The same principle could be applied in DFT domain or other transform domains, for instance, the transforms used in MPEG-2 AAC or Dolby AC-3.

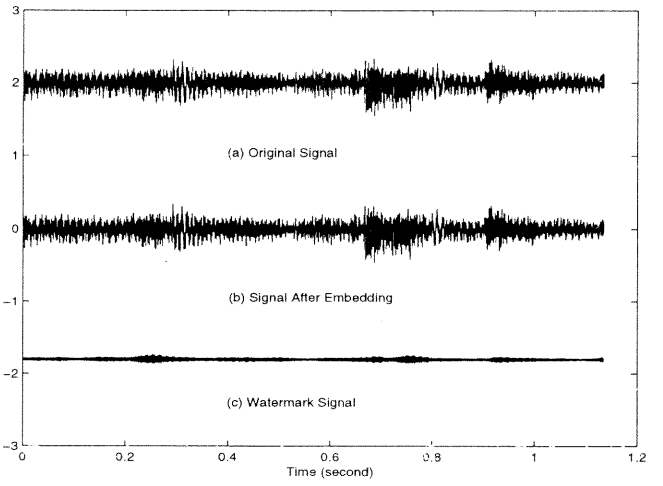


Figure 5.8 Amplitude modulation data hiding

5.3.2 Experimental Results

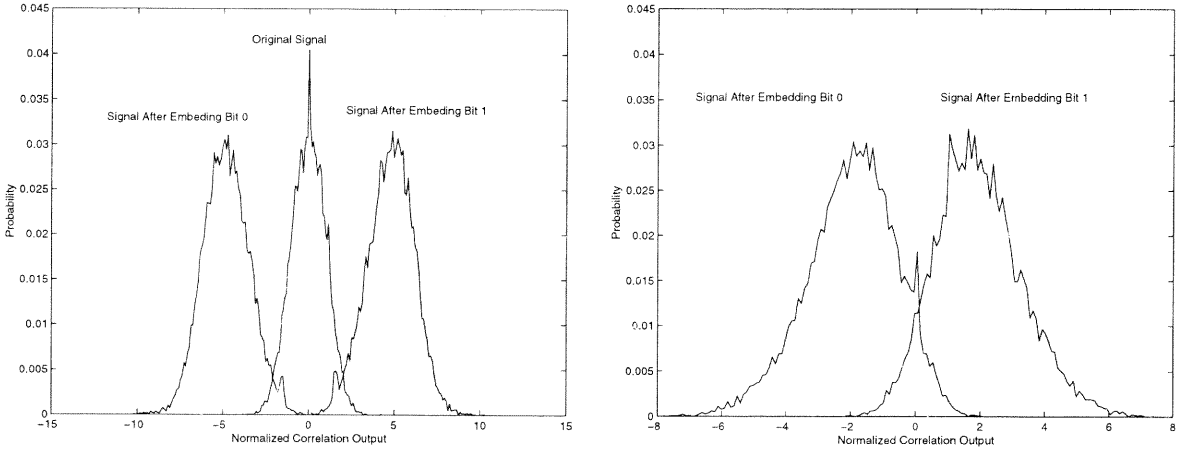
Mono music clips sampled at 44.1kHz are used in the experiments. The information bit is embedded into the MDCT coefficients from scale factor bands 6 to 18 which correspond to frequency range from 1kHz to 10kHz. To decrease the artifacts, the threshold α_i is selected smaller than $\sqrt{r_i}$ in the same scale factor band. In the sensitive bands from 1kHz to 3kHz, α_i is further tuned to reduce artifacts.

Given a received coefficient $\mathbf{r}$, the normalized detection output is obtained as

$$q = \frac{< \mathbf{r}, \mathbf{w} >}{||\mathbf{r}||}. \tag{5.6}$$

One information bit is embedded every granule (576 samples) of a mono audio clip. Figure 5.9 depicts the different q distribution after embedding bit value 1 and 0. The message extraction may not be sufficiently reliable due to the host noise interference. Some ECC code can be used to enhance its reliability.

The main advantage of this scheme is that the psychoacoustic model can be explicitly employed to control the artifact.



(a) Before 64kbits/s MP3 compression

(b) After 64kbits/s MP3 compression

Figure 5.9 Normalized detector output distribution in amplitude modulation

5.4 Phase Modulation Data Hiding

In music signal perception, it is well known that human ears are more sensitive to the amplitude than to the phase. The most significant components are signal frequencies and amplitudes. The signal amplitude spectrum contains more significant information than the phase spectrum. Data can be hidden in phases with less artifacts [28].

5.4.1 Hiding and Extraction

Bender *et al.* [7] proposed a scheme to hide information into the DFT phase. First, the audio signal is divided into frames and Discrete Fourier Transform (DFT) is applied to each frame. Second, the DFT phases in the first frame are modified while the phases in the following frames are modified, respectively. Nevertheless, the phase difference (*relative phase*) is kept unchanged. This procedure is repeated on the audio stream.

Although the phase is believed to be less significant perceptually, people are still sensitive to the phase continuity between frames. In the above scheme, the frame continuity is destroyed when the next bit is to be embedded. This may result in a beat pattern. The abrupt phase change modifies the signal spectrum. Informal listening tests show that small modifications in DFT phase are inaudible. This property is exploited for data embedding.

To directly employ SS modulation in the phase domain is not successful as the original host noise is quite large. The nonlinear schemes discussed in Chapter 4 are more effective. For instance, the Quantization Index Modulation (QIM) [12] can be applied in phase domain. Figure 5.10 depicts the phase QIM signaling.

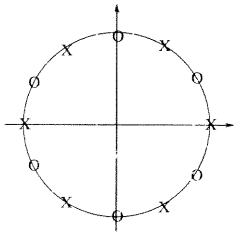


Figure 5.10 QIM in phase modulation

In this scheme, the original DFT phase value θ_i at one frequency bin is replaced by the nearest x point (to hide bit 1) or the nearest o points (to hide bit 0) on the unit circle. To embed one bit in a phase sequence, deterministic patterns are defined to represent bit values. For example, for a 4-coefficient sequence embedding, two patterns similar to antipodal signaling can be defined:

$$\begin{cases} \text{Pattern} \quad \text{A: [x o x o]}, & \text{represent bit value 1} \\ \text{Pattern} \quad \text{-A: [o x o x]}, & \text{represent bit value 0} \end{cases} \tag{5.7}$$

To hide a bit, θ_i is modified to comply with pattern A or -A.

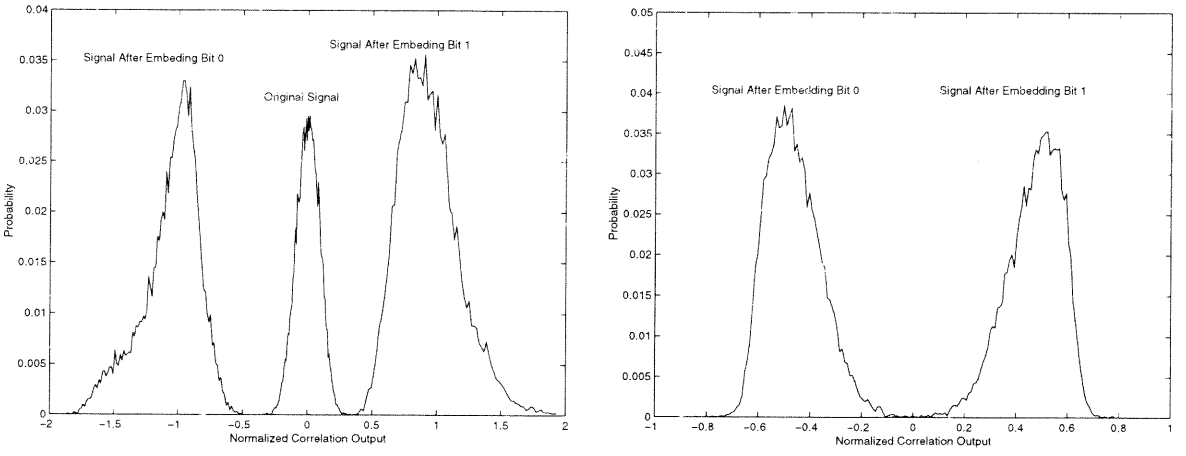
Obviously, the DFT phase noise is much larger if the corresponding amplitude is smaller. A simple suboptimal detector is a weighted minimum distance detector. Denote the received DFT amplitude and phase as r_i and ϕ_i, respectively. Find the

nearest x and o points α_i and β_i to ϕ_i and construct two sequences according to these two patterns:

$$\begin{cases} \mathbf{u} = & [\alpha_0, \beta_1, \alpha_2, \beta_3] \\ \mathbf{v} = & [\beta_0, \alpha_1, \beta_2, \alpha_3] \end{cases} \tag{5.8}$$

If $\sum r_i(\alpha_i - \phi_i)^2 < \sum r_i(\beta_i - \phi_i)^2$, decision is bit value 1; Otherwise bit value 0 is decided.

The distortion introduced is determined by the phase difference d between x and o points. Smaller value of d is selected at the sensitive frequency bands while larger value of d may be used at high frequency bands. After embedding, the DFT phases are fixed at x or o points. To introduce randomness, the value of $\theta_i + a_i$ is replaced by the x or o points where a_i is a random shift value.



(a) Before 64k MP3 compression

(b) After 64k MP3 compression

Figure 5.11 Normalized output distribution in phase modulation

5.4.2 Experimental Results

In the experiments, DFT length is 512 and the DFT phases from 1kHz to 8kHz are changed. The value of d_i varies from $\pi/12$ to $\pi/4$. To measure the embedding performance, the normalized correlation output is defined as

$$q = \frac{< \mathbf{r}, ||\mathbf{u} - \mathbf{w}||^2 - ||\mathbf{v} - \mathbf{w}||^2 >}{||\mathbf{r}||}, \tag{5.9}$$

where $\mathbf{w}$ is the received phase sequence and $\mathbf{u}$ and $\mathbf{v}$ are given by (5.8).

The above can be regarded as the normalized distance detector. The statistical distribution of q is shown in Figure 5.11. Experimental results indicate that this scheme is effective in oblivious applications. The audio quality is preserved at relatively lower SNR than that in amplitude modulation. The main disadvantage is that accurate distortion is difficult.

5.5 Noise Substitution Data Hiding

Message can also be hidden in the noisy component in audio signals [28, 33].

5.5.1 Perception of Noise Components

The audio signal had long been regarded as combination of sine waves in the computer music studies. X. Sierra [84] was among the first to introduce noise component in computer music. Lack of noise component makes the music "unnatural" (A good example of noise is the breathiness of a flute). Noise component is also perceptually significant.

In the advanced audio analysis model, noise component is indispensable. In the HILN model [1], signal is modeled as **harmonic+individual sines+noise**. Another influential model is **sines+transients+noise** [56]. Some studies argue that for noise components, what is significant is not the fine frequency structure in noisy bands, but the noise energy shape. The noise energy shape can be described by its DCT coefficients [1] or by a source filter model. A commonly used model is the Linear Predictor (LP) filter.

Goodwin *et al.* [35] proposed the Equivalent Rectangular Band (ERB) noise modeling. The authors claimed that the energy in ERB is more important than the noise spectral shape. People do not resolve the fine frequency structure in a noisy band, only a "mixing effect" is felt.

Recently, Levine *et al.* [56] and Verma *et al.* [94] proposed a similar approach Bark Band Noise Modeling. In noisy bark bands, only the noise energy gain is coded and transmitted. The reconstructed noise spectrum is flat over the frequency range of each bark band. The authors claimed higher quality compared with DCT spectral envelope and LPC-smoothed representations.

This idea can also be used in audio compression. Fine spectral structure in noisy bands need not to be encoded. The noise energy gain is adequate. In MP3, all the MDCT coefficients are encoded, including the higher frequency ones. This property can be employed to embed messages in these noisy bands. This is *noise substitution.*

5.5.2 Experimental Results

The noisy components can be modified in message embedding, while the energy gain in noisy bands is kept constant. There are many approaches meeting this requirement. One simplest method changes the sign of those coefficients x_i's in noisy bands by a random pattern. The hiding procedure is

$$x_i' = \begin{cases} p_i|x_i|, & \text{to hide bit value 1} \\ -p_i|x_i|, & \text{to hide bit value 0} \end{cases} \tag{5.10}$$

where $\mathbf{p}$ is a bipolar random sequence and p_i is either -1 or $+1$.

The information bit is extracted via correlation. Given a received sequence $\mathbf{r}$, the decoder output statistic is

$$q = \sum_{i=0}^{N-1} r_i p_i. \tag{5.11}$$

If $q > 0$, the extracted bit value is 1; Otherwise bit value 0 is decided.

To accurately distinguish the noisy bands from non-noisy ones is not an easy job. Not all high frequency coefficients are noisy, some may be the high frequency components of a transient signal. It is reported that over 80% of the high frequency coefficients are "non-edged". Several complicated algorithms are proposed in [81] to

make a distinction between noisy and non-noisy bands. In experiments, for simplicity, the frequency bands over 5kHz are regarded as noisy. Informal listening test shows it is a reasonable assumption.

This method could be applied in different watermark domains. It is advantageous to match the compression decomposition. Because in compression quantization procedure, a small value coefficient might be quantized to zero, nevertheless it never inverts its sign. This sign conservative property promises its robustness against the very compression.

The noise substitution is not robust against low-pass filtering, and it may not survive the next-generation compression where noise substitution technique is used. Nevertheless it survives current perceptual compression schemes, such as MP3. Experiments show those methods can reach around $20 \sim 60$ bits/second hiding capacity.

5.6 MP3 Compression and Encryption

Besides watermarking, encryption is also widely studied and deployed in multimedia protection. It is often used in multimedia email, teleconference to prevent unauthorized access to the multimedia contents. Multimedia signal scrambling is different from the general data encryption that involves extensive computation [22, 63]. Two important considerations are *efficiency* and *security*. The former requires real-time operation of the decryption process. This is different from the data unscrambling where off-line operation is acceptable. The security requirement is not as rigorous as that in data encryption. Feasible solutions are trade-offs considering these factors.

Current media encryption algorithms fall into two categories. One integrates scrambling with source coding, viz., to scramble media content before quantization and coding. The other scrambles compressed bitstream. Usually, it is desired that the encrypted output is bitstream syntax compatible. Some algorithms have been

proposed and applied in video and image scenarios [68, 83]. In this section the encryption schemes [32] in audio signals are discussed.

5.6.1 Encryption Integrated with Source Coding

In this approach, the encryption is performed before quantization and encoding. A widely used signal scrambling method is time-frequency permutation [17].

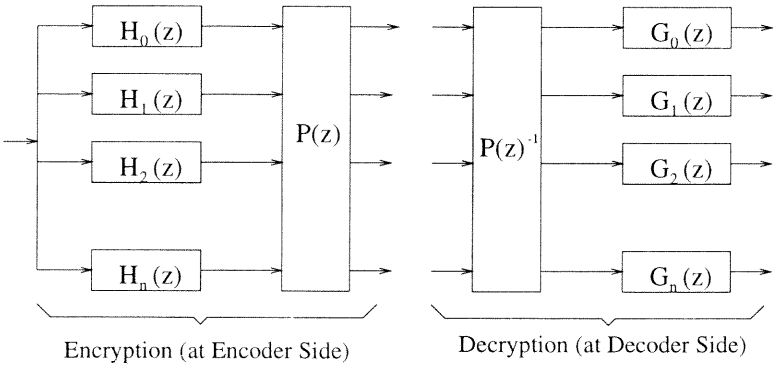


Figure 5.12 Time-frequency permutation

Figure 5.12 shows the block diagram. $P(z)$ is a permutation function and its inverse function is $P(z)^{-1}$. Its effectiveness has been proved in practice and can be applied directly in the MP3 MDCT domain. The side effect is that the random permutation changes the coefficient distribution property and renders the Huffman table non-optimal. The scrambling also destroys the correlation between contiguous granules. These result in lower compression rate.

There does not exist an easy solution. A possible remedy to enhance Huffman coding efficiency is to divide the frequency range into several bands, only permute coefficients within a band. It keeps the coefficient distribution property to some degree at the price of compromised security.

In addition, for a stereo signal, the coefficients in one granule can be further permuted between two channels. For most music materials, it is reported that the left and right channels in a stereo source have little correlation [41]. Thus, swapping

the data in these channels completely destroys the content. This increases encoding (decoding) latency and memory requirement.

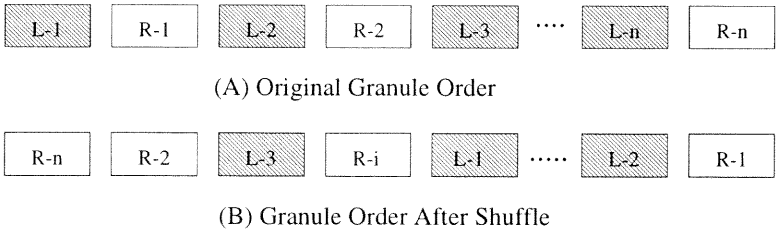


Figure 5.13 Stereo signal granule shuffle

5.6.2 MP3 Bitstream Syntax

In MP3 compression, the sign and amplitude of MP3 MDCT coefficients are coded separately. The total coefficients are divided into 3 regions: big-value region, small-value region and zero region. The big value region at low frequency end is further divided into 3 sub-regions where different Huffman tables are used. The small-value region is composed of coefficient values of $+1$, -1 or 0. Each codeword represents a pairs of contiguous coefficients in the big-value region or 4 coefficients (quadruple) in the small-value region. The remaining coefficients are implicitly set to zeros (Figure 5.14).

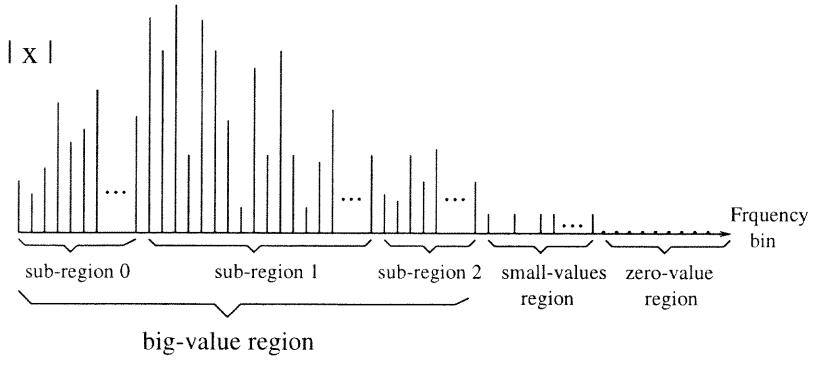


Figure 5.14 Partitioning of MDCT coefficients

An MP3 frame is an independent decoding unit. Its *header()* specifies important parameters for decoding operation, bit rate, sampling frequency, coding mode, etc. The encryption should not change these fields.

```
audio_data()
{
    main_data_begin
                :
                :
    for (gr=0; gr<2; gr++)
       for (ch=0; ch<nch; ch++) {
            part2_3_length[gr][ch]
            big_values[gr][ch]
            global_gain[gr][ch]
            scalefac_compress[gr][ch]
            window_switching_flag[gr][ch]
            if(window_switching_flag[gr][ch])
               block_type[gr][ch]
               mixed_block_flag[gr][ch]
               for (region=0; region<2; region++) {
                   table_select[gr][ch][region]
               for (window=0; window<3; window++)
                   subblock_gain[gr][ch][region]
            } else {
               for (region=0; region<3; region++)
                   table_select[gr][ch][region]
               region0_count[gr][ch]
               region1_count[gr][ch]
            }
            preflag[gr][ch]
            scalefac_scale[gr][ch]
            count1table_select[gr][ch]
    }
    main_data()
}
```

Figure 5.15 Side information in MP3 syntax

The *audio_data()* field provides the decoding control parameters and MDCT data (main_data()). The first half of audio_data() specifies the side information and *main_data()* is composed of the codewords and signs of the MDCT coefficients (Figure 5.15).

Although the length of a frame is constant at a given bit rate, bit consumption for samples in one granule (576 samples) is variable. The "bit reservoir" technique permits the current frame to "borrow" bits saved from past frames to absorb the imbalance. The current frame data may locate in previous frames. The location where the main_data() begins is determined by *main_data_begin*, a 9-bit offset value.

5.6.3 Encryption in Compressed Domain

In this section, discussion is focused on encryption directly on an MP3 bitstream. To avoid confusing the decoder, the encrypted bitstream should comply with the MP3 bitstream syntax. According to different sensitivity requirements, three different protection levels are provided: 1) *slight protection*, where the encrypted bitstream presents a satisfactory music quality for a casual listener, although not good enough for Hi-Fi reproduction. This can be used to generate different music versions for casual users and professionals; 2) *moderate protection*, where the scrambled content is meaningful and the main music features are kept with obvious degradation. This can be used for customer evaluation. After test listening, customers could pay and obtain a decryption key to recover the quality. 3) *maximum protection*, where the music content is completely destroyed, rendering the MP3 bitstream meaningless. To be MP3 syntax compatible, the bitstream can not be simply scrambled, since this generates an invalid bitstream and confuses decoder. And the file size should be kept unchanged.

The selected Huffman table should not be changed. The minimum unit that can be manipulated is a codeword (of a pair of coefficients in the big-value region or quadruple of 4 coefficients in the small-value region). The encryption can work at the following levels: codeword level, sub-region level and granule level. The scrambling is one or combination of these strategies.

At the codeword level, because the coefficient amplitudes and signs are separately encoded, the codes can be permuted and (or) the signs be flipped by a random pattern. While sign flipping can happen on any non-zero coefficient, the permutation should be limited to codes using the same Huffman table for syntax compatibility. For the coefficients whose amplitudes are greater than 15, the *linbits* field can be scrambled without any constraints. At the sub-region level, the sub-regions can be permuted. Respective permutation is also required for related side information parameters, such as code counts and Huffman table index. At the granule level, the granules inside a frame can be reordered randomly. The corresponding parameters, such as *part2_3_length*, etc. should be shuffled for the integrity of the granule. For different applications, special attention should be taken to meet the degradation requirements.

Encryption with Slight Distortion

In the MP3 time-frequency decomposition, a fine frequency resolution is applied at low and high frequency bands. At high frequency end, it is not quite necessary. That gives us some room for manipulation. The noise perception property should be used in scrambling. For example, the signs of the MDCT coefficients over 5kHz can be flipped. The frequency shape in a critical band is unchanged. In addition, these coefficients within one scale factor band can be permuted as the noise energy gain is still kept. The modification is almost transparent for a casual listener. If more distortion is permitted, the lower frequency coefficients can even be permuted or sign-flipped. This operation can be further tuned for specific requirements.

Encryption with Moderate Degradation

It is believed the frequency amplitude is more important than phase in audio signal. However, sign-flipping of the non-noisy coefficients introduces obvious degradation. The permutation and sign-flipping can be used in this case. To scramble medium frequency coefficients introduces obvious degradation. The audio signal

spectrum has a wide dynamic range. To keep features of music clips, the large value coefficients should be skipped, and only the relatively smaller ones are manipulated. Experiments reveal that the components under 3kHz are perceptually significant and should not be manipulated much.

Encryption with Maximum Protection

To provide maximum protection, it is desired to completely destroy the audio content while keeping the bitstream syntax. Sign-flipping and codeword permutation can be employed at codeword level. The order can be shuffled at the sub-region level. The side information parameters, such as Huffman table index *table_select*, codeword count *region_count* (Figure 5.15) etc. should also be permuted accordingly. The permutation can also happen at higher level. For instance, in a stereo signal, two granules in each channel can be shuffled in one frame. To abide by the syntax, the order of the side information parameters should be changed respectively.

5.7 Music Perception and Audio Model Analysis

In this section, the music perception and audio parametric coding is revisited. Discussion is focused on audio spectral model analysis and how music phenomenon is explained and described by the model.

5.7.1 Audio Signal Models

In parametric coding, sound signal is analyzed and a parametric representation is constructed. Different models can be used to describe a sound signal. There are three kinds of models to describe the sound signal. One is *abstract model*. For example, FM modulation [97] approach represents sound signal as

$$y(t) = Asin(c(t) + [Isin(M(t))], \tag{5.12}$$

where A is the peak amplitude, $c(t)$ is varying carrier frequency, I is modulation index and $M(t)$ is the modulated signal. A general audio waveform is approximated

by an FM modulated signal, although this model does not have apparent physical interpretation.

Physical model is based on source modeling. The parametric representation describes the mechanical and acoustic behavior of a music instrument. The parameters are different for different instruments. Some instrument physical models have been studied extensively [88].

A very powerful tool in audio analysis is based on *spectrum model.* Human psychoacoustic studies reveal the signal short time spectrum is extremely important for human perception procedure. The perceptual difference is negligible as long as the short time spectrum of the reconstructed signal is sufficiently close to the spectrum of the original signal.

Some models have been developed in signal spectrum "description". The model *individual sines+harmonic sines+noise* is used in MPEG-4 parametric coding [1]. In this model, spectrum is described by some individual sines, harmonic sines and noise component. Some algorithms just describe signals by harmonic components and noisy component [43].

Much of the computer music generation is based on the spectrum model. Music has long been regarded as combination of sines. Sierra [84] is among the first to introduce *noise* component in it. All the above audio analysis algorithms model music signal as a combination of deterministic part and stochastic parts. The following part describes the influential model proposed in [56, 95].

5.7.2 Spectral Model: Transient+Sines+Noise

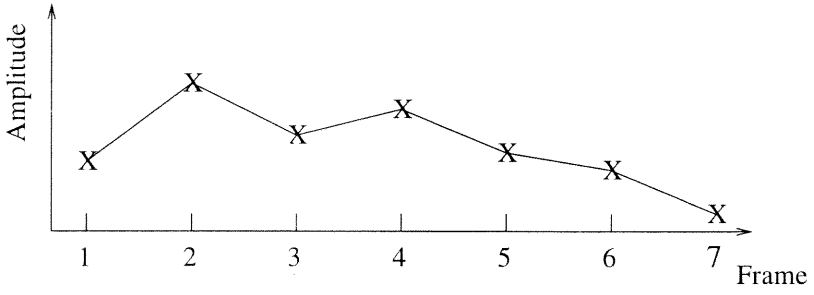
The most significant part perceptually is the sines in a piece of music. Earlier music models decompose music into sinusoids and stochastic component *noise.* Model analysis extracts the time-varying parameters of amplitude, phase, and frequency to describe sinusoids.

Audio analysis is usually done frame by frame. In each frame, most perceptually significant (usually those with larger amplitude) sines are picked up and parameters are estimated to describe them. These parameters include amplitude, frequency and initial phase. Therefore, suppose in the l^{th} frame, R_l sinusoids are picked and extracted. The signal is thus represented as a combination of these sines.

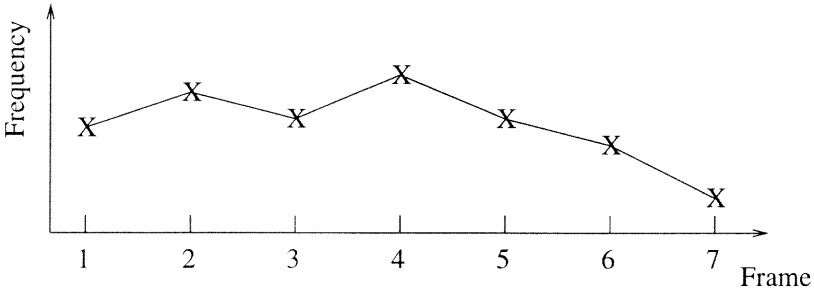
$$s(m) = \sum_{r=1}^{R_l} A_{r,l} cos[m\omega_{r,l} + \phi_{r,l}]. \tag{5.13}$$

The r^{th} sinusoid is described by a triple $A_{r,l}, \omega_{r,l}, \phi_{r,l}$.

It is natural to find these parameters are different from frame to frame. Further studies reveal that although those sines are changing, most sines in the current frame are close to one of those sines in the next frame. So it is reasonable to assume the sines in the next frame are the continuation of the sines in the previous frame. The sinusoids are not interpreted as individual sines, but as a sinusoid evolving slowly from frame to frame. The evolution of the evolving sine is called a *track*.



(a) Track Amplitude Evolution



(b) Track Frequency Evolution

Figure 5.16 Track of sinusoid

Further analysis demonstrates it is reasonable to assume the frequency and amplitude of those sinusoids are evolving linearly. What people feel is the changing sinusoids. In a track, the initial phase is not significant and not encoded. At decoder side, an arbitrary phase is used. Humans may perceive the phase discontinuity at frame borders. Figure 5.16 shows the evolution of a typical sinusoids. It is seen that the track gives birth in Frame 1, and "dies" in Frame 7. Sinusoidal modeling is not sufficient to represent audio signals as this model can not track non-stationary (rapid-changing) signals. The rapid-changing signals are referred as "attacks" in the previous discussion.

In a clip of music generated by an instrument, the phenomenon usually match the pattern Attack-Decay-Sustain-Release [45]. Figure 5.17 shows the music waveform envelope when a key is stricken.

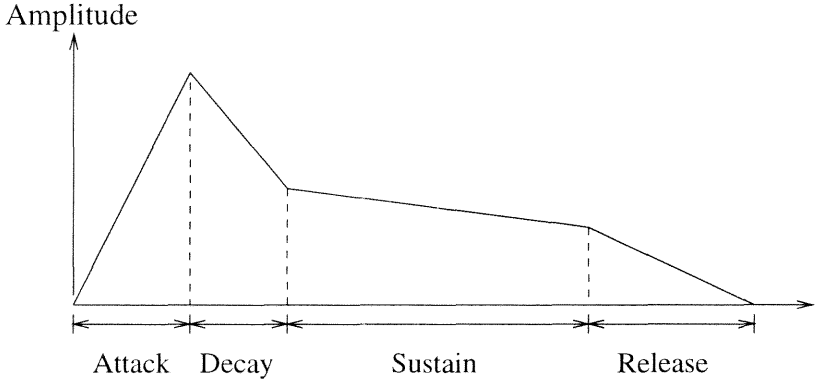


Figure 5.17 Attack-Decay-Sustain-Release pattern

The attack phase which is a rapid-changing part shows non-stationary property of the signal. It is the most difficult part for coding. In MP3, lots of complexity is involved dealing with attacks. In audio model analysis, analysis and experimental studies proved it is not appropriate to model the attack as the sum of sinusoids. Basically, transients are broadband signals that can not be well represented by sinusoids.

After subtracting the sum of sinusoids from the original signal, the error residue is analyzed to determine whether transients are present. If so, the attack is described by transform coefficients [56] or by other methods [94]. Subsequently, the remaining parts are stochastic which can be modeled using an appropriate noise models [34, 35, 56, 94]. In Section 5.5, noise perception is covered. The audio parameters are perceptually significant, to be robust against the parametric compression, data hiding should modify these parameters in a transparent way. It is a very challenging task.

CHAPTER 6

CONCLUDING REMARKS AND FUTURE RESEARCH

In this research work, some watermarking topics have been covered. The detection in data hiding is studied and new algorithms are proposed. Much of the focus is on audio data hiding and human psychoacoustic model. The effectiveness of the algorithms proposed has been demonstrated in the analytical work and applications.

It is important to achieve watermark compression resilience while meeting transparency requirement. Compression reduces the redundancy in the signal without losing perceptual value (*transparency*). Its function is to remove the perceptually insignificant components, while the steganography embeds some perceptually insignificant information. Note that it does not mean information can not be embedded in the perceptual significant components. Nevertheless, watermark should be insignificant to meet the transparency requirements. Obviously, compression and steganography are in a kind of "arms race" [52]. Petitcolas *et al.* [3] pointed out, steganography is almost impossible to survive ideal compression.

The above conclusion is easy to understand. In a signal space, perceptually equal signal points should be compressed to one point by an ideal compression. If two different points in the space are of same perceptual value, that implies the compression is not efficient enough (thus not ideal). Ideal compression does not exist in reality. It can be concluded that a more efficient compression scheme makes data hiding more difficult. Some researchers suggest to integrate watermarking with compression design. This makes the watermarking robust to *this* compression, although not guarantee of its survival against other compression algorithms. Wang *et al.* presented a watermarking scheme in their proposal to JPEG 2000 [96]. Recently, Cognicity Corp. has already integrated their hiding technology AudioKey with Lucent Perceptual Audio Coder (PAC).

It is desired to take advantage of the compression in data hiding [7]. Some factors addressed in compression such as quantization and coding, need not to be considered in steganography. However, today's compression robust scheme may not survive the more advanced compression of the next generation.

In order to verify multimedia copyright ownership, watermarking must be compression and tamper resistant. In the SS modulation scheme, the decoder has a very rigorous requirement on synchronization. The watermark verification fails if the synchronization can not be kept. In image and video applications, watermarking should be robust against geometric distortions, rotation, and translation. In audio applications, it should survive the more advanced parametric compression and model-based compression. An even more challenging task is to be robust against *time-scale modification* and *pitch-scale modification.*

Many watermarking problems are still unresolved. The ongoing research work will focus on audio data hiding, especially on study of the hiding schemes resistant to MPEG-4 compression. Future research topic is the design of compression with content protection features. A more accurate psychoacoustic model should be developed for steganography applications based on music signal perception and understanding.

REFERENCES

1. ISO/IEC FDIS 14496-3 Sec 2. "Information technology–Coding of audio-visual objects, Part 3:Audio, Section 2: Parametric audio coding". 1999.

2. Masoud Alghoniemy and A. H. Tewfik. "Progressive quantized projection watermarking scheme". *Proc. 7th ACM International Multimedia Conference*, pages 295–298, Nov. 1999.

3. Ross J. Anderson and Fabien A. P. Petitcolas. "On the limits of steganography". *IEEE Trans. Journal of Selected Areas in Communications*, 16(4):474–481, May 1998.

4. Michael Arnold and Sebastian Kanka. "MP3 robust audio watermarking". *Proc. DFG VIIDII Watermarking Workshop'99, Erlangen, Germany*, 1999.

5. M. Barni, F. Bartolini, and F. Rigacci. "Statistical modelling of full frame dct coefficients". *Proc. EUSIPCO 98, 9th European Signal Processing Conference*, 6(8-11):1512–1516, Sept. 1998.

6. P. Bassia and I. Pitas. "Robust audio watermarking in the time domain". *Proc. EUSIPCO'98, 9th European Signal Processing Conference*, pages 25–28, Sept. 1998.

7. W. Bender, D. Gruhl, N. Morimoto, and A. Lu. "Technique for data hiding". *IBM System Journal*, 35(3-4):313–336, 1996.

8. J. A. Bloom, I. J. Cox, T. Kalker, J-P. Linnartz, M. L. Miller, and B. Traw. "Copy protection for DVD video". *Proc. IEEE Special Issue on Identification and Protection of Multimedia Information*, 87(7):1267–1276, 1999.

9. L. Boney, A. H. Tewfik, and K. N. Hamdy. "Digital watermarks for audio signals". *Proc. IEEE International Conference on Multimedia Computing and Systems*, pages 473–480, June 1996.

10. Karlheinz Brandenburg and Gerhard Stoll. "ISO-MPEG-1 Audio: A generic standard for coding of high-quantity digital audio". *Journal Audio Engeering Society*, 42(10):780–792, Oct. 1994.

11. B. Chen and G. W. Wornell. "Digital watermarking and information embedding using dither modulation". *Proc. of IEEE 2nd Workshop on Multimedia Signal Processing'98*, pages 273–278, Dec. 1998.

12. B. Chen and G. W. Wornell. "Dither modulation: A new approach to digital watermarking and information embedding". *Proc. of SPIE: Security and Watermarking of Multimedia Contents*, 3657:344–353, Jan. 1999.

13. Digimarc Corp., Macrovision Corp., and Philips Research. "The Millennium Group: DVD copy-protection system, A Technical overview". Aug. 1999.

14. I. Cox and M. L. Miller. "A review of watermarking and the importance of perceptual modeling". *Proceeding of Electronic Imaging*, Feb. 1997.

15. I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon. "Secure spread spectrum watermarking for images, audio, and video". *NEC Research Technical Report*, 10, 1995.

16. I. J. Cox, Joe Kilian, Tom Leighton, and Talal Shamoon. "A Secure, robust watermark for multimedia". *Workshop on Information Hiding'96*, May 1996.

17. Charles D. Creusere and Sanjit K. Mitra. "Efficient image scrambling using polyphase filter banks". *Proc. ICIP'94*, 2:81–85, 1994.

18. B. L. Yeo D. Benham, N. Memon and M. Yeung. "Fast watermarking of DCT-based compressed images". *Proc. International Conference on Image Science, Systems, and Technology (CISST'97), Las Vagas*, pages 243–253, June 1997.

19. Grant A. Davidson, Louis D. Fielder, and Brian D. Link and. "Parametric bit allocation in a perceptual audio coder". *http://www.dolby.com/tech/highqual.html.*

20. J. F. Delaigel, D. De Vleeschouwer, and B. Macq. "Low cost perceptive digital picture watermarking method". *Proc. ECMAST'97, Milan, Italy*, pages 153–167, May 1997.

21. Geert Depovere, Ton Kalker, and Jean-Paul Linnartz. "Improved watermark detection reliability using filtering before correlation". *Proc. ICIP'98*, 1:430–434, 1998.

22. W. Diffie and M. E. Hellman. "New directions in cryptography". *IEEE Trans. on Information Theory*, 22:644–654, Nov. 1976.

23. E. Eberlein, H. Popp, B. Grill, and J. Herre. "Layer III a flexible coding standard". *Audio Engineering Society preprint 3493, 94th Convention, Berlin, Germany*, March 1993.

24. M. Kutter F. Jordan and T. Ebrahimi. "Proposal of a watermarking technique for hiding/retrieving data in compressed and decompressed video". *ISO/IEC Doc. JTC1/SC29/WG11/MPEG97/M2281*, June 1997.

25. J. Fridrich. "Image watermarking for tamper detection". *Proc. ICIP'98*, 2:404–408, Oct. 1998.

26. G. Friedman. "The trustworthy digital camera: Restoring credibility to the photographic image". *IEEE Trans. Consumer Electronics*, 39:905–910, Nov. 1993.

27. Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar. "Periodic signaling scheme in oblivious data hiding". *Proc. 34th Asilomar Conference on Signals, Systems, and Computers 2000*, pages 1851–1855, November 2000.

28. Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar. "MP3 resistant oblivious steganography". *Proc. ICASSP'2001*, pages 0000–0000, May 2001.

29. Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar. "Nonlinear modulation in oblivious steganography". *Accepted in Proc. IEEE-ERUASIP Non-linear Signal and Image Processing'2001*, pages 0000–0000, June 2001.

30. Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar. "Performance analysis of spread spectrum modulation in data hiding". *Proc. SPIE'2001, California*, pages 0000–0000, July 2001.

31. Litao Gang, Ali N. Akansu, and Mahalingam Ramkumar. "Set partitioning in oblivious data hiding". *Proc. ICASSP'2001*, pages 0000–0000, May 2001.

32. Litao Gang, Ali N. Akansu, Mahalingam Ramkumar, and Xuefei Xie. "On-line music protection and MP3 compression". *Proc. International Symposium on Intelligent Signal Informaton Multimedia Processing'2001, Hongkong*, pages 13–16, May 2001.

33. Litao Gang, Ali N. Akansu, and Taha H. Sencar. "Transform selection in steganography". *Proc. Annual Conference on Information Sciences and Systems, Maryland*, March 2001.

34. M. Goodwin. "Residual modeling in music analysis-synthesis". *Proc. ICASSP'96*, 1996.

35. M. Goodwin. "Adaptive signal models: Theory, algorithms, and audio applications". *Ph.D. thesis, University of California, Berkley*, 1997.

36. F. Hartung and B. Girod. "Digital watermarking of raw and compressed video". *Proc. European EOS/SPIE Symposium on Advanced Imaging and Network Technologies, Berlin, Germany*, Oct. 1996.

37. F. Hartung and B. Girod. "Fast public-key watermarking of compressed video". *Proc. IEEE ICIP'97, Santa Barbara, CA*, Oct. 1997.

38. F. Hartung and B. Girod. "Digital watermarking of uncompressed and compressed video". *Trans. of Signal Processing — Special Issue on Copyright Protection and Access Control for Multimedia Services*, 66(3):283–301, 1998.

39. F. Hartung and B. Girod. "Watermarking of uncompressed and compressed video". *Signal Processing*, 66(3):283–301, May 1998.

40. C. W. Helstrom. "Probability and Stochastic Processes for Engineers". *Macmillan, New York*, 1991.

41. Jurgen Herre, Ernst Eberlein, and Karlheinz Brandenburg. "Combined stereo coding". *AES preprint 3369, 93rd Convention, Calif, USA*, Oct. 1992.

42. Mikio Ikeda, Kazuya Takeda, and Fumitada Itakura. "Audio data hiding by use of band-limited random sequences". *Proc. ICASSP'99*, 4:2315–2318, 1999.

43. Rafael Angel Irizarry. "Statistics and Music: Fitting a Local Harmonic Model to Music Sound Signals". *PhD. Thesis, University of California, Berkeley*, 1998.

44. ISO/IEC. "IS 11172-3: Coding of moving pictures and associated audio for digital storage media at up to about 1.5Mbits/s". *ISO/IEC*, 1993.

45. Crystal Semiconductor Corp. Jim Heckroth. "Tutorial on MIDI and music synthesis". *http://home.earthlink.net/ mma/Tutorial/Tutor.htm.*

46. Neil F. Johnson and Sushil Jajodia. "Steganalysis of images created using current steganography software".

47. J. D. Johnston. "Transform coding of audio signals using perceptual noise criteria". *IEEE Trans. Journal of Selected Areas in Communications*, 6:314–323, Feb. 1988.

48. Steven M. Kay. "Fundamentals of Statistical Signal Processing". *Volume 2*, Prentice-Hall PTR, 1993.

49. D. Kundur and D. Hatzinakos. "Digital watermarking using multiresolution wavelet decomposition". *Proc. ICASP'98*, 5:2969–2972, May 1998.

50. M. Kutter, F. Jordan, and F. Bossen. "Digital signature of color images using amplitude modulation". *Proc. Electronic Image'97, San Jose, CA*, May 1997.

51. M. Kutter, F. Jordan, and F. Bossen. "Digital signature of color images using amplitude modulation". *Journal of Electronic Imaging*, 7(2):326–332, April 1998.

52. Jack Lacy, Schuyler R. Quackenbush, Amy R. Reibman, and James H. Snyder. "Intellectual property protection systems and digital watermarking". *Optics Express*, 3(12), Dec. 1998.

53. C. Langelaar, J. C. A. ven der Lubbe, and R. L. Lagendijk. "Robust labeling methods for copy protection of images". *Proc. Electronic Imaging'97, San Jose, CA*, 3022·298–309, Feb. 1997.

54. G. C. Langelaar, J. C. A. van der Lubbe, and J. Biemond. "Copy protection for multimedia data based on labeling techniques". *Proc. 7th Symposium Information Theory in Benelux, Enschede, The Netherlands*, May 1996.

55. Sang-Kwang Lee and Yo-Sung Ho. "Digital audio watermarking in the cepstrum domain". *IEEE Trans. Consumer Electronics*, 46(3):334–335, Aug. 2000.

56. S. Levine. "Audio representations for data compression and compressed domain processing". *Ph.D. thesis, Stanford University*, 1998.

57. Eugune T. Lin and Edward J. Delp. "A review of fragile image watermark". *Proc. ACM Multimedia'99, Multimedia and Content Security Workshop*, pages 25–29, Oct. 1999.

58. M. Mansour and A. Tewfik. "Audio watermarking by time-scale modification". *Proc. ICASSP'2001*, May 2001.

59. M. L. Miller, I. J. Cox, and J. A. Bloom. "Watermarking in the real world: An application to DVD". *Proc. 33rd Asilomar Conference on Signals, Systems, and Computers*, pages 1496–1502, 1999.

60. Norishige Morimoto. "Digital watermarking technology with practical applications". *Informing Science (Special Issue on Multimedia Informing Technologies)*, 2(4):107–111, 1999.

61. Pierre Moulin and Joseph A. O'Sullivan. "Information-theoretic analysis of watermarking". *Proc. ICASSP'2000, Istanbul*, 5:3630–3633, Jun. 2000.

62. N. Nikolaidis and I. Pitas. "Copyright protection of images using robust digital signatures". *Proc. ICASSP'96, Atlanta, GA*, 4:2168–2171, May 1996.

63. National Institute of Standards and Technology (NIST). "FIPS Publication 46-1: Data Encryption Standard". Jan. 1998.

64. A. V. Oppenheim and J. S. Lim. "The importance of phase in signals". *IEEE processing*, 69(5):512–541, May 1981.

65. Davis Pan. "A tutorial on MPEG/audio compression". *IEEE Trans. Multimedia Journal*, 1995.

66. S. Pankanti and M. Yeung. "Verification watermarks on fingerprint recognition and retrieval". *Proc. IS&T/SPIE Conference on Security and Watermarking of Multimedia Contents*, pages 67–78, Jan. 1999.

67. A. Piva, M. Barni, E. Bartoloni, and V. Cappellini. "DCT-based watermarking recovering without resorting to the uncorrupted original image". *Proc. ICIP'97, Santa Barbara, CA*, 1:520–523, 1997.

68. Lintian Qiao and Klara Nahrstedt. "A new algorithm for MPEG video encryption". *Proc. International Conference on Imaging Science, Systems, and Technology (CISST'97), Las Vegas*, pages 21–29, 1997.

69. Lintian Qiao and Klara Nahrstedt. "Non-invertible watermarking methods for MPEG encoded audio". *Proc. SPIE Conference on Security and Watermarking of Multimedia Contents*, 3657:194–202, Jan. 1999.

70. M. Ramkumar and A. N. Akansu. "A robust scheme for oblivious detection of watermarks/data hiding in still images". *Proc. SPIE, Symposium on Voice, Video and Data Communication*, 3528:474–481, Nov. 1998.

71. M. Ramkumar and A. N. Akansu. "Self-noise suppression schemes for blind image steganography". *SPIE Special Session on Image Security*, 3845, Sept. 1999.

72. M. Ramkumar and A. N. Akansu. "Some design issues for robust data hiding systems". *Proc. 33rd Asilomar Conference on Signals, Systems, and Computers*, Oct. 1999.

73. M. Ramkumar, A. N. Akansu, and A. A. Alatan. "A robust data hiding scheme for digital images using DFT". *Proc. ICIP'99*, II:211–215, Oct. 1999.

74. M. Ramkumar, A. N. Akansu, and A. A. Alatan. "On the choice of transforms for data hiding in compressed video". *Proc. ICASSP'99*, VI:3049–3052, March 1999.

75. Mahalingam Ramkumar. "Data Hiding in Multimedia – Theory and Applications". *Ph.D. Thesis, New Jersey Institute of Technology*, Jan. 2000.

76. R. L. Rivest, A. Shamir, and L. M. Adleman. "A method for obtaining digital signatures and public-key cryptosystems". *Communications of the ACM*, 21(2):120–126, Feb. 1978.

77. J. J. K. Ruanaidh, W. J. Dowling, and F. M. Boland. "Phase watermarking of digital images". *Proc. ICIP'96*, pages 239–242, Sept. 1996.

78. J. J. K. Ruanaidh and T. Pun. "Rotation, scale and translation invariant spread spectrum digital image watermarking". *Signal Processing*, 66(3):303–317, May 1998.

79. A. Said and W. A. Pearlman. "A new fast and efficient implementation of an image codec based on set partitioning in hierarchical trees". *IEEE Trans. Circuits and Systems for Video Technology*, 6(3):243–250, June 1996.

80. B. Schneier. "Applied Cryptography". New York: Wiley, 1996.

81. D. Schulz. "Improving audio codecs by noise substitution". *Journal Audio Engineering Society.*, 44(7/8):593–598, Jul./Aug. 1996.

82. J. M. Shapiro. "Embedded image coding using zerotrees of wavelet coefficients". *IEEE Trans. Signal Processing*, 41(12):3445–3462, 1993.

83. Changgui Shi and Bharat Bhargava. "A fast MPEG video encryption algorithm". *Proc. ACM Multimedia'98, Bristol, UK*, pages 81–88, 1998.

84. X. Sierra and J. Smith. "Musical sound modeling with sinusoids plus noise". Webpage http://www.iua.upf.es/ sms/.

85. D. Stinson. "Cryptography Theory and Practice". Boca Raton, FL: CRC Press, 1995.

86. J. K. Su and B. Girod. "Power-spectrum condition for energy-efficient watermarking". *Proc. ICIP'99*, Oct. 1999.

87. J. K. Su, F. Hartung, and B. Girod. "A channel model for a watermark attack". *Proc. SPIE, Security and Watermarking of Multimedia Contents, Electronic Imaging '99, San Jose, CA*, 3657:159–170, Jan. 1999.

88. C. R. Sullivan. "Extending the Karplus-Strong filter to synthesize electric guitar timbres with distortion and feedback". *Computer Music Journal*, 14:3, 26 37.

89. M. Swanson, B. Zhu, and A. H. Tewfik. "Data hiding for video-in-video". *Proc. ICIP'97, Santa Barbara, CA*, 2:676–679, Oct. 1997.

90. M. D. Swanson, B. Zhu, and A. H. Tewfik. "Robust data hiding for images". *Proc. IEEE Digital Signal Processing Workshop*, pages 37–40, Sept. 1996.

91. M. D. Swanson, B. Zhu, and A. H. Tewfik. "Multiresolution scene-based video watermarking using perceptual models". *IEEE Trans. Journal on Selected Area in Communications*, 16(4):540–550, May 1998.

92. M. D. Swanson, Bin Zhu, and A. H. Tewfik. "Transparent robust image watermarking". *Proc. ICIP'96*, Sept. 1996.

93. R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne. "A digital watermark". *Proc. ICIP'94*, 2:86–90, 1994.

94. T. Verma. "A perceptually based audio signal model with application to scalable audio compression". *Ph.D. Thesis, Stanford University*, 2000.

95. T. Verma, S. Levine, and T. Meng. "Transient modeling synthesis: a flexible transient analysis/synthesis tool for transient signals". *Proc. of International Computer Music Conference*, pages 164–167, Sept. 1997.

96. Houng-Jyh Wang, Yi-Liang Bao, C.-C. Kuo, and Homer H. Chen. "Multithreshold wavelet codec (MTWC), Document No. WG1N805, Geneva, Switzerland". *http://costard.usc.edu/ cckuo/*, March 1998.

97. B. Winduratna. "FM analysis/synthesis based audio coding". *AES 104th Convention, Preprint 4746,*, May 1998.

98. R. B. Wolfgang and E. J. Delp. "A watermark for digital images". *Proc. ICIP'96*, pages 219–222, Sept. 1996.

99. RayMond B. Wolfgang, Christine I. Podilchuk, and Edward J. Delp. "Perceptual watermarks for digital images and video". *Proceedings of IEEE*, 7:1108–1126, July 1999.

100. M. Wu and B. Liu. "Watermarking for image authentication". *Proc. ICIP'98, Chicago*, 1998.

101. X. Xia, C. Boncelet, and G. Arce. "A multiresolution watermark for digital images". *Proc. ICIP'97, Santa Barbara, CA*, 1:548–551, Oct. 1997.

102. M. Yeung and F. Mintzer. "Invisible watermarking for image verification". *Journal of Electronic Imaging*, 7:578–591, July 1998.

103. W. Zhu, Z. Xiong, and Y. Q. Zhang. "Multiresolution watermarking for images and video: a unified approach". *Proc. ICIP'98, Chicago, Illinois*, 1:465–468, 1998.