

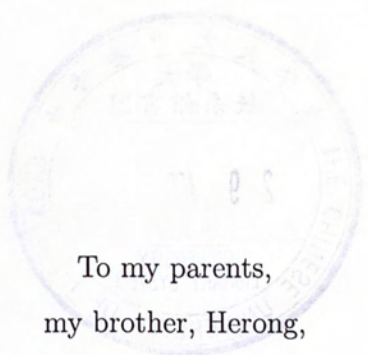
A NEW APPROACH FOR IMPROVING
TRANSPARENCY OF
AUDIO WATERMARKING

CHEN BENRONG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF PHILOSOPHY
IN
ELECTRONIC ENGINEERING

©THE CHINESE UNIVERSITY OF HONG KONG
AUGUST 2003

The Chinese University of Hong Kong holds the copyright of this thesis. Any person(s) intending to use a part or whole of the materials in the thesis in a proposed publication must seek copyright release from the Dean of the Graduate School.



To my parents,
my brother, Herong,
my wife, Jessica Shen Hua,
and to the memory of my grandmother.



Acknowledgements

First, I would like to express my sincere gratitude to my supervisor, Prof. P.C. Ching, for his insightful guidance and invaluable support over the course of my research. I learnt from Prof. Ching knowledge as well as approaches to complicated problems.

I would also like to express my sincere appreciation Prof. Tan Lee, for his encouragement and suggestion on my research. Thanks are due to Prof. X.G. Xia, Prof. Y.T. Chan, and Dr. F. Soong for their precious advices.

I would like to remember all the colleagues and friends who gave me help and shared with me lots of happiness during my study. To name some of them: Dr. W.K. Lo, Dr. W.K. Ma, W. Lau, Y.W. Wong, K.F. To, Y.J. Li, L.Y. Ngan, P. Kam, C.H. Yau, C. Yang, Y. Qian, S.W. Lee, Y.Y. Tam. Y. Zhu, N.H. Zheng, M. Yuan, W. Zhang, Y.C. Chan, L. Tsang, W. Lam. The technical assistance from Mr. K.O. Luk is much appreciated.

Finally, I wish to express my deepest gratitude to my parents for their disciplines and continuous encouragement. Special thanks are due to my dearest Jessica for her endurance, understanding and support.

Abstract of thesis entitled:
**A New Approach For Improving Transparency Of
Audio Watermarking**
Submitted by **CHEN BENRONG** (陳本榮)
for the degree of **Master of Philosophy**
in **Electronic Engineering**
at **The Chinese University of Hong Kong** in
August 2003.

Digital Audio Watermarking, same as other watermarking techniques, has found a wide variety of applications in recent years, in particularly for copyright protection and other digital rights management(DRM), due to the outstanding progress of digital technologies facilitate reproducing and retransmitting of digital data and thus the increasing concern over copyright protection of digital audio.

Digital audio watermarking embeds information related to the cover audio Work without any noticeable trace of degradation of perceptual quality and guarantees the information embedded to be detected again with proper approach.

There are many techniques that can be used for audio watermarking implementation. Watermarking system based on *spread spectrum* technique exploits the properties of pseudorandom sequence thereby providing effective ways for embedding and detecting watermark information bits. However, one of performance indicator for watermarking system, the transparency of the embedding processing, can't be satisfied. Requirement for transparency improvement is demanded. This requirement can be satisfied by filtering the watermark signal with a psychoacoustic filter, which is rather complicated and may lose the robustness of the embedded watermark.

In this thesis, we first propose a *content adaptive embedding method* so as to

make the spread spectrum technique based watermarking system more effective.

To improve the transparency of the watermarking system, we propose a sample selection scheme in the frequency domain together with a frame selection scheme in the time domain.

By experimental and theoretical analysis of the frequency properties of audio signals, we derive a relationship between the system performance and the frequency band selected for watermarking. Under the guidance of this relationship, sample selection can be applied to fulfill requirements of different applications.

Each frame along the time axis has its own properties different from others, therefore by choosing those suitable frames for watermark embedding, the overall performance of the system can be further improved. By analyzing the experimental results, we propose different frame selection criteria, which can improve the transparency and have their own merits.

To validate the usefulness of the proposed method for audio watermarking, psychoacoustic model analysis on the watermark embedded Work was carried out and the robustness of the approach has been verified.

摘要

近年來，數字音頻水印技術，和其他水印技術一樣，呈現廣闊的應用前景，特別是在版權保護和其他數字化版權管理等方面有重要的應用。這是由於數字技術的發展以及網絡的普及，數字多媒體可以被方便地複製和傳輸，因此對數字多媒體，包括數字音頻信號的版權保護得到了廣泛關注和重視。

數字音頻水印技術，是在完全不影響音頻信號的音質的情況下，將與該音頻文件有關的版權信息嵌入音頻信號中，同時保證嵌入的信息能夠通過適當的方法再次檢測出來。

許多技術都能夠應用到音頻水印的實現過程中。基於擴頻技術的數字音頻水印系統充分利用了偽隨機碼的特性，進而能夠提供有效的方法進行嵌入和檢測水印信息。然而，水印嵌入系統的一個性能表現，嵌入過程的透明性，在這樣的系統中沒有得到實現從而導致了音頻文件的失真。因此，有必要對嵌入過程的透明性做進一步提高。這可以通過採用由心理聲學模型推導出來的濾波器對水印信號進行濾波處理來實現，但這是一個複雜而且有可能破壞水印信號的穩健性的過程。

在本篇論文中，我們首先提出了內容自適應嵌入法，用來提高基於擴頻技術的水印系統的有效性和穩健性。爲了提高水印系統嵌入過程的透明性，我們提出了一個在頻域上進行樣點選取結合在時域上作時幀選取的方案。

通過對音頻信號的頻域特性進行實驗和理論分析，我們推導出了水印嵌入系統性能和被選取用來嵌入水印頻段之間的內在聯繫。在這種內在聯繫的指引下，我們能夠通過在頻域上進行樣點選取來滿足各種實際應用的需要。

時域上每個不同的時幀都具有其獨特的屬性，因此通過選取適當的時幀來進行水印嵌入，系統的整體性能將會得到更進一步的提高。通過實驗分析，我們提出了不同的時幀選取標準，這些標準都有助於提高系統嵌入過程的透明度，同時這些標準也有各自的優點。

爲了驗證我們爲數字音頻水印提出的方法的有效性，我們利用心理聲學模型對已經嵌入水印的音頻文件進行了分析，並證實了我們提出的方法的穩健性和有效性。

Contents

1	Introduction	1
1.1	What's Watermarking	1
1.2	Information Hiding, Steganography, and Watermarking	3
1.3	History of Watermarking	5
1.4	Importance of Digital Watermarking	8
1.5	Objectives of the Thesis	9
1.6	Thesis Outline	10
2	Applications and Properties of Audio Watermarking	12
2.1	Applications	13
2.1.1	Ownership Identification and Proof	13
2.1.2	Broadcast Monitoring	16
2.1.3	Other Applications	18
2.2	Properties	19
2.2.1	Transparency	20
2.2.2	Robustness	20
2.2.3	Other Properties	21
3	Possible Methods for Audio Watermarking	24
3.1	Overview of Digital Audio Watermarking System	25
3.2	Review of Current Methods	27
3.2.1	Low Bit Coding	27
3.2.2	Phase Coding	28
3.2.3	Echo Coding	29

3.2.4	Spread Spectrum Watermarking	30
3.3	Other Related Approaches	31
3.4	Outline of Proposed New Method	33
4	Audio Watermarking System Based on Spread Spectrum	36
4.1	Introduction	36
4.2	Embedding and Detecting Information Bit	39
4.2.1	General Embedding Process	39
4.2.2	General Detection Process	43
4.2.3	Pseudorandom Bit Sequences (PRBS)	45
4.3	An Optimal Embedding Process	48
4.3.1	Objective Metrics for Embedding Process	48
4.3.2	Content Adaptive Embedding	52
4.3.3	Determination of Frame Length L	57
4.4	Requirement For Transparency Improvement	58
5	Sample and Frame Selection For Transparency Improvement	60
5.1	Introduction	60
5.2	Sample Selection	61
5.2.1	General Sample Selection	62
5.2.2	Objective Evaluation Metrics	65
5.2.3	Sample Selection For Transparency Improvement	66
5.2.4	Theoretical Analysis of Sample Selection	87
5.3	Frame Selection	90
5.3.1	General Frame Selection	91
5.3.2	Frame Selection For Transparency Improvement	94
5.4	Watermark Information Retrieve	103
6	Psychoacoustic Model For Robustness Verification	105
6.1	Introduction of Human Auditory System	106
6.1.1	Absolute Hearing Threshold	106
6.1.2	Critical Bands	108
6.1.3	Masking Effect	111

6.2	Psychoacoustic Model of Human Auditory System	112
6.3	Robustness Verification by Psychoacoustic Model Analysis . . .	117
7	Conclusions and Suggestions For Future Research	121
7.1	Conclusions	121
7.2	Suggestions For Future Research	123
	Bibliography	125

List of Tables

1.1	Four Categories of Information Hiding	5
4.1	Noise Energy $EN_i(dB)$ Induced while Modifying Samples with Different Index. $L = 1024; \alpha = 2.6$	53
4.2	Statistical Result of Noise Energy EN and Frame SNR $fSNR$, With Different Embed Index	54
4.3	Frame Length Versus Signal SNR and Capacity (The sample to be embedded is the sample with second largest sample; $\alpha = 2.6$);	57
4.4	SNR of A Set of Audio File Samples. $L=1024; \alpha = 2.6$; All Frames are Embedded with Information Bit '1'; (These testing files can be found in audiofile1.pdf in the attached CD. To listen these sound files, please click the file names.)	59
5.1	Sample Selection Result On a Audio Work With Different $BeginSample$; The parameter are set as: $L=1024; SeLe=512; \alpha = 2.6$	67
5.2	Mean of ED (Energy of D); Noise Energy EN ; Frame SNR $fSNR$; Sample Portion SNR, $pSNR$; Sample Portion Energy EP and Frame Energy EF , Vary With Different $BeginSample$; The parameter are set as: $L=1024; SeLe=512; \alpha = 2.6$	71
5.3	Standard Deviation(StD) of ED (Energy of D); Noise Energy EN ; Frame SNR $fSNR$; Sample Portion SNR, $pSNR$; Sample Portion Energy EP and Frame Energy EF , Vary With Different $BeginSample$; The parameter are set as: $L=1024; SeLe=512; \alpha = 2.6$	71

5.4	Relationship Between Frame Length L and Evaluation Metrics, which are in Statistical Mean Value and with Standard Deviation in the Bracket. All Samples are Used For Watermarking, $SeLe = L$; $pSNR = fSNR$; $EP = EF$;	79
5.5	Relationship Between $SeLe$ and Other Metrics, Which are in Mean Vale with Standard Deviation in the Bracket. The <i>BeginFrequency</i> fixed to 4KHz, in term of <i>BeginSample</i> =186; $L = 1024$; $\alpha = 2.6$	82
5.6	The Statistical Relationship Between $SeLe$ and Other Metrics, Which are in Mean Vale with Standard Deviation in the Bracket. The <i>CenterFrequency</i> fixed at 8KHz; $L = 1024$; $\alpha = 2.6$;	83
5.7	$SNRs$ and $BERs$ of A Set of Audio File Samples. $L=1024$; $\alpha = 2.6$; $SeLe = 256$; $B = 186$ (<i>BeginFrequency</i> =4kHz). All Frames are Embedded with Information Bit '1'; (These testing files can be found in <i>audiofile1.pdf</i> in the attached CD. To listen these sound files, please click the file names.)	86
5.8	$SNRs$ and $BERs$ of A Set of Audio File Samples. $L = 1024$; $\alpha = 2.6$; $SeLe = 512$; $B = 186$ (<i>BeginFrequency</i> =4kHz). All Frames are Embedded with Information Bit '1'; (These testing files can be found in <i>audiofile1.pdf</i> in the attached CD. To listen these sound files, please click the file names.)	86
5.9	Performance of Different Frame Selection Method, with the <i>BeginFrequency</i> fixed to 4KHz, in term of <i>BeginSample</i> =186; $L = 1024$; $SeLe=256$. (In each experiment, half of the frames are selected for embedding with information bit '1'.)	99
5.10	SNR, Noise Energy of 15 Audio Files, When Subject to Different Frame Selection Criteria; Based on D, EFEP, EP and NoSelection respectively; With $SeLe=256$; <i>BeginFrequency</i> fixed to 4kHz, in term of <i>BeginSample</i> =186; $L = 1024$;	101

5.11	Bit Error Rate of 15 Audio Files, When Subject to Different Frame Selection Criteria; Based on D, EFEP, EP and NoSelection respectively; With $SeLe=256$; $BeginFrequency$ fixed to $4kHz$, in term of $BeginSample=186$; $L = 1024$	102
6.1	Critical Band Upper Boundaries	110

List of Figures

1.1	A General Watermarking System	2
1.2	Classification of Information Hiding Techniques.	3
3.1	A General Digital Audio Watermarking Flowchart (Dotted Line Means Optional).	26
3.2	The Whole Outline of the Proposed Watermarking Method . .	35
4.1	Block Diagram of a Watermarking System Based on Spread Spectrum Technique	37
4.2	Block Diagram of the Watermarking Detection Module	38
4.3	Flowchart of the Embedding Process Based on Spread Spectrum Technique	40
4.4	The Procedure of Embedding One Bit into a Frame.	42
4.5	Flowchart of the Detection Process Based on Spread Spectrum Technique	44
4.6	The Procedure of detecting One bit from a Given Frame.	45
4.7	An Embedding Process By Pseudorandom Number Sequence. .	46
4.8	Fibonacci Implementation of LFSR.	47
4.9	A Typical Example of the Original Frame, Noise Induced and Watermarked Frame.	51
4.10	A Typical Example of the Original Signal, Induced Noise and Watermark Embedded Signal.	51
5.1	Block Diagram of Watermarking Embedding Based on Spread Spectrum Technique Combined With Sample Selection	64

5.2	Probability Density Function of ED_i , the energy of D_i . B denotes the $BeginSample; L=1024; \alpha = 2.6; SeLe=512$	68
5.3	Probability Density Function of EN_i , the Energy of Noise $n_i(n)$. B denotes the $BeginSample; L=1024; \alpha = 2.6; SeLe=512$	70
5.4	Theoretical and Experimental Analysis Relationship Between ED and EN . $L=1024; \alpha = 2.6; BeginSample = 186 SeLe=512$	73
5.5	Probability Density of Frame SNR, $fSNR$ $L=1024; \alpha = 2.6; SeLe=512$	74
5.6	Probability Density of Portion SNR, $pSNR$; $L=1024; \alpha = 2.6; SeLe=512$;	75
5.7	Probability Density of Portion Energy, EP ; $L=1024; \alpha = 2.6; SeLe=512$;	76
5.8	The Tendency Relationship Between the $BeginSample$ and the evaluation metrics; $L=1024; \alpha = 2.6; BeginSample = 186; SeLe=512$;	77
5.9	Statistical Relationship between Frame Length L and Frame SNR $fSNR$; $SeLe = L; fSNR = pSNR$	79
5.10	Tendency Relation Between $SeLe$ and Evaluation Metrics; $SeLe = L; fSNR = pSNR$	80
5.11	Statistical Relationship of $SeLe$ and EN ; With $L = 1024; BeginFrquency = 4kHz; BeginSample = 186$	81
5.12	Tendency Relationship of $SeLe$ with $ED, EP, pSNR$ and $fSNR$; With $L = 1024; BeginFrquency = 4kHz; BeginSample = 186$	82
5.13	The Tendency Relationship Between $SeLe$ and Other Metrics, Which are in Mean Vale with Standard Deviation in the Bracket. The $CenterFrequency$ fixed at $8KHz; L = 1024; \alpha = 2.6$;	84
5.14	PDF of $ED, fSNR pSNR$ and EP , Without Frame Selection. .	91
5.15	The Block Diagram of Watermark Embedding Based on Spread Spectrum Technique Combined With Sample and Frame Selection.	93
5.16	Relationship of D and Frame SNR $fSNR$	95
5.17	PDF of $ED, fSNR pSNR$ and EP , When Frame Selection Based on D	96

5.18 Relationship of D with other Metrics	97
5.19 <i>PDF</i> of <i>ED</i> , <i>fSNR</i> <i>pSNR</i> and <i>EP</i> , When Frame Selection Based on Sample Portion Energy	98
5.20 <i>PDF</i> of <i>ED</i> , <i>fSNR</i> <i>pSNR</i> and <i>EP</i> , When Frame Selection Based on <i>EF-EP</i>	99
5.21 The Block Diagram of Watermark Detection Based on Spread Spectrum Technique Combined With Sample and Frame Selection	104
6.1 Absolute Hearing Threshold as a Function of Frequency	107
6.2 Relationship Between Bark Frequency Scale and Linear Fre- quency Scale	109
6.3 Critical Bands Defined in MPEG-1 Psychoacoustic Model 1. . .	110
6.4 An Example of Frequency Masking	111
6.5 The Primary Components of MPEG 1 Encoder	112
6.6 An Example of Masking Threshold Calculated by Psychoacoustic Model.	116
6.7 Robust Verification On One Frame	118
6.8 Robust Verification On An Whole Audio Sample. Average Result	119
6.9 Robust Verification On An Whole Audio Sample. Statistical Analysis Result	120

Chapter 1

Introduction

1.1 What's Watermarking

Hold a Hong Kong \$100 bill (issued by The Hongkong Shanghai Banking Corporation) up to the light. If you are looking at the side with the lion head, you will see that the lion head is echoed as a watermark on the right. This kind of watermark is embedded directly into the paper during the papermaking process, and is therefore very difficult to forge.

The watermark on the HK\$100 bill, just like most paper watermarks today, has two properties. First, the watermark is hidden from the view during normal use, only becoming visible as a result of special viewing process (in this case, holding the bill up to the light). Second, the watermark carries information about the object in which it is hidden (in this case, the watermark indicates the authenticity of the bill).

In addition to paper, watermarking can be applied to other physical objects and to digital signals. The techniques presented in this thesis focus on the watermarking of digital signals, especially on digital audio signal. We will adopt the following terminology to describe these signals. We refer to a specific song, music, speech file, video or picture as a *Work*, and to the set of all possible Works as *Content*. Thus, audio file is an example of content, and the song “The Power of Love” by Celine Dion is an example of a Work. The original unwatermarked Work is sometimes referred to as the *cover Work*,

in that it hides or “covers” the watermark. We use the term *media* to refer to the means of representing, transmitting and recording content. Thus, the audio CD on which “The Power of Love” is recorded is an example of a medium.

Digital Watermarking is defined as the practice of imperceptibly modifying a Work to embed information about that Work.

We can deduce the definition for digital audio watermarking as: the practice of imperceptibly modifying an audio Work to embed information about that audio Work. In general, a watermarking system of the type we will discuss consists of an *embedder* and a *detector*, as illustrated in Figure 1.1. The embedder takes at least two inputs. One is the information we want to encode as a watermark, and the other is the cover Work in which we want to embed the watermark. The output of the watermark embedder, watermarked Work, is typically transmitted or recorded. Later, that Work (or some other that has not been through the watermark embedder) is presented as an input to the watermark detector. The detector tries to determine whether a watermark is present, and if so, decipher the information encoded by that watermark. The key to the embedder and the detector is not a necessary element of the watermarking system. But it is usually used to improve the security of the system.

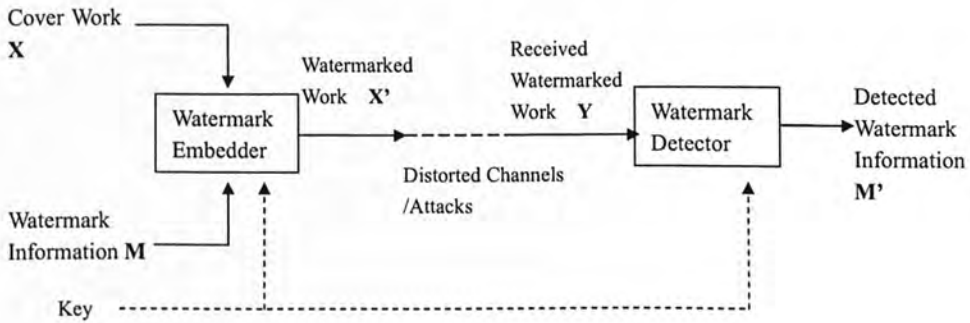


Figure 1.1: A General Watermarking System

In the late 1990s there was an explosion of interest in digital systems for watermarking of various content. The main focus has been on photographs, audio and video entertainment. The proposed applications of these methods are many and varied, including copyright ownership identification, proof of ownership, transaction tracking, indication to recording equipment, verification that content has not been modified since the watermark was embedded, and the monitoring of broadcast. The details of these applications will be discussed in chapter 2

1.2 Information Hiding, Steganography, and Watermarking

Watermarking is closely related to the fields of information hiding and steganography. These three areas have a lot of overlaps and share many technical approaches. However, there are fundamental philosophical differences that affect the requirements, and thus the design, of a technical solution. We will discuss these differences in this section.

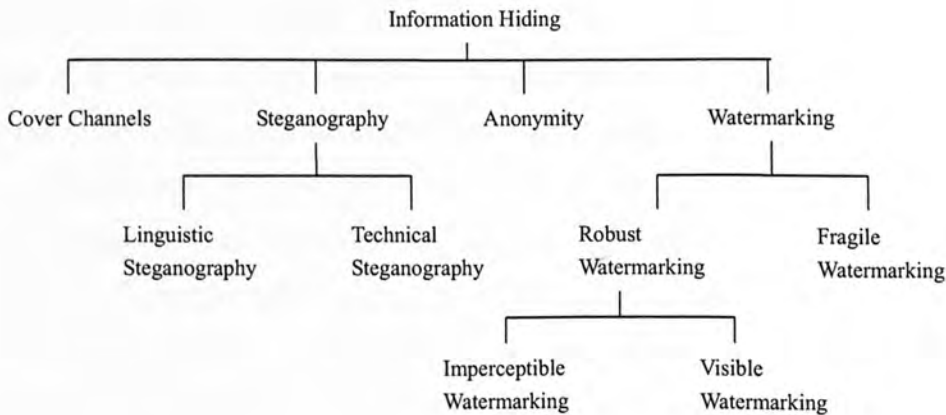


Figure 1.2: Classification of Information Hiding Techniques.

From Figure 1.2, which is given in [1], we can see that *information hiding* is a general term encompassing a wide range of problems beyond that of embedding information in content. Steganography and Watermarking are two subdisciplines of information hiding.

Steganography is a term derived from the Greek word *steganos*, which means “Covered”, and *graphia*, which means “writing” [1]. It is the art of concealed communication. Steganography is about concealing the very existence of the information. An often cited examples of steganography is a story from Herodotus [2], who tells of a slave sent by his master, Histiaëus, to the Inonian city of Miletus with a secret message tattooed on his scalp. After tattooing, the slave grew his hair back in order to conceal the message. He then journeyed to Miletus, upon arriving, shaved his head to reveal the message to the city’s regent, Aristagoras. The message encouraged Aristagoras to start a revolt against the Persian King. In this case, the message is of primary value to Histiaëus and the slave is simply the carrier of the message.

This example can be used to highlight the difference between steganography and watermarking. Imagine that the message on the slave’s head read, “This slave belongs to Histiaëus”. In that this information refers to the slave (cover work), this would meet our definition of watermark. If someone else claimed possession of the slave, Histiaëus could shave the slave’s head and prove ownership of the slave. In this case, the slave is of primary value to Histiaëus, and the message provides useful information about the cover work.

Systems for hiding information in Work can thus be divided into watermarking systems, in which the information is related to the cover Work and non-watermarking systems, in which the message is unrelated to the cover Work. They can also be independently divided into steganographic systems, in which the very existence of the information is kept secret, and non-steganographic systems, in which the existence of the information need not be secret. This results in four categories of information hiding systems, which are summarized in Table 1.1

	Cover Work Dependent Information	Cover Work Independent Information
Existence Hidden	Steganographic Watermarking	Cover Communications
Existence Known	Non-steganographic Watermarking	Overt Embedded Communications

Table 1.1: Four Categories of Information Hiding

By distinguishing between embedded data that relates to the cover work and hidden data that does not, we can anticipate the different applications and requirements of the information hiding methods. However, the actual techniques used for watermarking systems may be very similar, or in some cases identical, to those used in non-watermarking systems.

1.3 History of Watermarking

Although the art of papermaking was invented in China over a thousand years earlier, paper watermarks did not appear until about 1282, in Italy. The marks were made by adding thin wire patterns to the paper molds.

The meaning and purpose of the earliest watermarks are uncertain. They may have been used for practical functions such as identifying the molds on which sheets of papers were made, or as trademarks to identify the paper maker. On the other hand, they may have represented mystical signs, or might simply have served as decoration.

In the eighteenth century, watermarks on paper made in Europe and America had become more clearly utilitarian. They were used as trademarks, to record the date the paper was manufactured and to indicate the sizes of original sheets. It was also about this time that watermarks began to be used as anti-counterfeiting measures on money and other documents.

The term *watermark* seems to have been coined near the end of the eighteenth century. About that time, counterfeiters began developing methods of

forging watermarks used to protect paper money. Counterfeiting prompted advances in watermarking technology. William Congreve, an Englishman, invented a technique for making color watermarks by inserting dyed material into the middle of the paper during papermaking [3]. The resulting marks must have been extremely difficult to forge, because of the Bank of England itself declined to use them because they were too difficult to be made. Then a more practical technology was invented. This replaced the fine wire patterns used to make earlier marks with a sort of shallow relief sculpture, pressed into the paper mold. The resulting variation on the surface of the mold produced beautiful watermarks with varying shades of gray. This is the basic technique used today for the lion head on the HK\$100 bill.

Examples of our general notion of watermarks—Imperceptible information about the objects in which they are embedded—probably date back to the earliest civilizations. David Kahn, in his classic book *The Codebreakers*, provides interesting historical notes [4]. An especially relevant story describes a message hidden in the book *Hypnerotomachia Poliphili*, anonymously published in 1499. The first letters of each chapter spell out “Poliam Frater Franciscus Columna Peramavit,” assumed to mean “Father Francesco Columna loves Polia.”¹

Four hundred years later, we find the example of a technology seminar to the digital methods we are now discussing. In 1954, Emil Hembrooke of the Muzak Corporation filed a patent for “watermarking” musical Works. An identification code was inserted in music by intermittently applying a narrow notch filter centered at 1KHz. The absence of energy at this frequency indicated that the notch filter had been applied and the duration of the absence used to code either a dot or a dash. The identification signal used Morse code. The 1961 U.S. Patent describing this invention states[5]:

¹This translation is not universally accepted. The words can be translated into other meanings.

The present invention makes possible the positive identification of the origin of a musical presentation and thereby constitutes an effective means of preventing such piracy, i.e. it may be likened to a watermark in paper.

The system was used by Muzak until around 1984. It is interesting to speculate that this invention was misunderstood and became the source of persistent rumors that Muzak was delivering subliminal advertising message to its listeners.

It is hard to say when *digital watermarking* was first discussed. In 1979, Szepanski [6] introduced a machine-detectable pattern that could be placed on documents for anti-counterfeiting purposes. Nine years later, Holt *et al.* [7] described a method for embedding an identification code in an audio signal. However, it was Komatsu and Tominaga [8], in 1988, who appeared to have first used the term *digital watermarking*. Still, it was probably not until the early 1990s that the term *digital watermarking* really came into vogue.

About 1995, interest in digital watermarking began to increase rapidly. In 1996, the first Information Hiding Workshop(IHW) [9] was held. In this conference, digital watermarking was one of its primary topics. Began from 1999, the Society of Photo-optical Instrumentation Engineers(SPIE) began devoting a conference specifically to *Security and Watermarking of Multimedia Contents* [10].

In addition, about this time, several organizations began considering watermarking technology for inclusion in various standards. The Copy Protection Technical Working Group(CPTWG) [11] tested watermarking systems for protection of video on DVD disks. The Secure Digital Music Initiative(SDMI) [12] made watermarking a central component of their system for protecting music. The European Union, VIVA [13] and Talisman [14] sponsored two projects tested watermarking for broadcast monitoring. The International Organization for Standardization (ISO) took an interest in watermarking technology in the context of designing advanced MPEG Standards.

In the 1990s several companies were established to market watermarking

products. Technology from the Verance Corporation was adopted into the first phase of SDMI and was used by Internet music distributors, such as Liquid Audio².

1.4 Importance of Digital Watermarking

The sudden and rapid increase interest in watermarking is mostly due to the increase in the concern over copyright protection of content. Because of the advancement of digital technologies in the past few years, multimedia data, including audio, picture and video, can be reproduced and retransmitted easily with high quality.

The high-capacity reproduction ability provided by digital recording devices raise the risk of piracy. When the only way for common customer to record a music was on analog tape, pirated copies were usually with a lower quality compared to the originals. and the quality of second-generation pirated copies.(i.e, copies of copy) was generally very poor. However with digital recording devices, music can be recorded with little, if any, degradation in quality.

The Internet had become user friendly with the introduction of web browser in 1993 [15]. People are downloading music, pictures, and videos through the network. The Internet becomes an excellent distribution system for digital content because it is inexpensive, eliminates warehousing and stock, and delivery is almost instantaneous.

By using the digital recording devices and the Internet for distribution, potential pirates can easily reproduce and distribute copyright-protected material without appropriate compensation being paid to the actual copyright owners. Thus, multimedia content owners are facing a high risk of piracy. And they are eagerly seeking technologies that promise to protect their rights and benefit.

The first technology in which content owners turn to is cryptography. Cryptography is probably the most common method of protecting digital content [16]. While the content is encrypted before delivery, and a decryption key is

²<http://www.liquidaudio.com/>

provided only to those who have paid money for legitimate copies of the content. The encrypted file can then be made available via the Internet, but would be useless to a pirate without an appropriate key. However, the content owner can no longer protect the content after decryption. A pirate can purchase the content legitimately, use the decryption key to obtain an unprotected copy of the content, and then reproduce and distribute illegal copies.

Thus, a technology that can protect the content all the time, even after selling to the customer, is strongly demanded. Watermarking has the potential to fulfill this need because it places information within content where it will never be removed under normal usage. A watermark can be designed to survive common signal processing, such as, compression, resampling and filtering.

Watermarking has been considered for many copyright protection and copy prevention applications. In copyright protection applications, the watermark may be used to identify the copyright holder. Whilst for copy prevention, the watermark may be used to inform software or hardware devices that copying should be restricted. Although copyright protection and copy prevention have been the major driving force behind research in this watermarking area, there are a number of other applications for which watermarking has been used or suggested. These include broadcasting monitoring, transaction tracking, authentication, and many others.

1.5 Objectives of the Thesis

Digital audio watermarking is blooming in recent years, but is still new technology to many people. The objective of this thesis is to, systematically and thoroughly, investigate different approaches for efficient and effective digital audio watermarking. This thesis will give reader a comprehension overview of digital audio watermarking from its definition, development history, applications to requirements and possible methods for implementation.

In addition, a new method for improving the transparency and robustness of watermarking system will also be proposed. The method introduces effective

means of sample selection in the frequency domain and frame selection in the time domain by exploiting the properties of audio signal. Compared with other existing methods [17, 18], the proposed approach is an effective and simple way for performance improvement. Furthermore, the specific implementation procedures and requirements for digital audio watermarking will also be elaborated while presenting this new approach.

1.6 Thesis Outline

In Chapter 2, we shall describe several applications that can be implemented with watermarking and examine the advantages watermarking might have over alternative technologies. we will also describe several required properties of digital watermarking systems in this chapter. In Chapter 3, we first describe the overview of a generic audio watermarking system. Then we review on the methods published in the literature. Finally we briefly describe the outline of the new approach we are going to proposed.

In Chapter 4, we shall first introduced a sophisticated audio watermarking system based on Spread Spectrum technique. Then we will define some evaluation metrics and proposed *content adaptive embedding* method for performance improving subjective to these metrics. Subjective and objective performance evaluation on such an digital audio watermarking system will be carried on and the requirement for transparency improvement will be raised.

To fulfill the requirement for transparency improvement, we will proposed effective method in Chapter 5, which relies on sample selection in frequency domain and frame selection in time domain. The general procedures for both selection processes will be first described and the way to realize optimal selection will be elaborated in this chapter. Performance evaluation on the system exploiting sample and frame selection will be carried on and proof the selection processes have improved the transparency of the system significantly.

In Chapter 6, human auditory system(HAS) and psychoacoustic model exploiting temporal and frequency characteristics of HAS will be introduced. Psy-

Chapter 2

Applications and Properties of Audio Watermarking

General speaking, if it is necessary to associate some additional information with an audio Work, this information can be embedded secretly using watermarking techniques without any noticeable trace of degradation of perceptual quality. Audio watermarking has found a wide variety of applications in recently years, in particularly for copyright protection.

It is well known that the associate information can be placed in a Work using simpler ways, such as placing it in the header of the digital file, or even speaking it aloud as an introduction to an audio clip. The questions are: when is watermarking a better alternative? What can watermarking do that cannot be done using other techniques?

Watermarking is distinguished from other techniques in three aspects. First, watermarks can be imperceptible. The watermarks will not be perceived during normally usage. Second, watermarks are inseparable from the Works in which they are embedded. Unlike the header fields, watermarks can not be removed when the Works are displayed or converted to other file formats. Finally, watermarks undergo the same transformation as the Works. This means that it is sometimes possible to learn something about those transformations by looking at the resulting watermarks. It is these three attributes that make watermarking invaluable for many applications.

The performance of a given watermarking system can be evaluated on the basis of a small set of properties. For example, *transparency* describes how imperceptible the watermarks are. Where *robustness* indicates how well watermarks survive common signal processing operations. The relative importance of these properties highly depends on the application of which the system is designed for. For example, in applications where we have to detect the watermark in a copy of Work that has been broadcasted over an analog channel, the watermark must be robust against the degradation caused by the channel. However, if we can reasonably expect that a Work will not be modified at all between embedding and detection, then robustness is problem irrelevant.

In this chapter, we are going to describe several applications that can be implemented with watermarking and examine the advantages watermarking might have over alternative technologies. Then we will describe several required properties of watermarking systems, discussing how their relative importance and interpretation vary with application.

2.1 Applications

2.1.1 Ownership Identification and Proof

Under many legislation systems, the creator of a song, painting, story, or any other original Work automatically holds copyright to it in the instant the Work is recorded in some physical form. Textual copyright notice is usually used to indicate the Work is owned by some body and being protected. There is clearly requirement for the exact form of the copyright notice. For visual Works, it must say either “Copyright *date owner*”, “©*date owner*” or “Copr. *date owner*”. For sound recordings, the copyright notice takes the similar form “℗ *date owner*” and must be placed on the surface of the physical media, the label, or on the packaging so as to “give reasonable notice of the claim of copyright” [19].

Copyright notice is now no longer compulsory required since 1988 [3]. But, if a Work that is protected by copyright is misused, and the courts choose to award the copyright holder damages, that award can be significantly limited

if a copyright notice of acceptable form and placement was not found on the distributed material [19] .

Textual copyright notices have several limitations as a mean for identifying the ownership of a Work. Two major setbacks include: the copyright notice may be easily removed intentionally or unintentionally. So subsequent users may be unaware of the copyright protection. Even if the Work is assumed to be protected, it may be difficult to find the identity of the creator or person whose permission is required. A famous case in which the loss of textual copyright notice on an image causing such problems is a photograph of Lena Sjööblom. It belongs to Playboy Enterprises, Inc. The textual copyright notice was cropped while scanning, and this image has since been distributed electronically around the world, and most researchers use it are probably unaware that they are infringing Playboy's copyright.

Another problem with textual copyright notice for images is that they can be aesthetically ugly and may cover a portion of the image. Although it is usually possible to make them unobtrusive, such practice makes them more susceptible to being cropped. The situation is even worse in audio, where the copyright notice is placed on the physical medium (disk, tape, record, and so on) and on the packaging. Neither of these notices would normally be copied along with the audio content. In fact, for some audio content that may exist only in electronic form, no physical medium or packaging would even exist.

Since watermarks can be made both imperceptible and inseparable from the Work that contains them, they are likely to be superior to text for owner identification. If users of Works are supplied with watermark detectors, they should be able to identify the owner of a watermarked Work, even after the Work has been modified in ways that would remove a textual copyright notice. Digimarc's watermark¹ for images was designed with precisely this application in mind. They achieved widespread distribution of their watermark detector by bundling it with Adobe's popular image processing program, Photoshop. When Digimarc's detector recognizes a watermark, it contacts a central database over

¹<http://www.digimarc.com/>

the Internet, and uses the watermark message as a key to find contact information for the image's owner.

The legal impact of such an application has not yet been tested in court. At present, given that the exact form of a copyright notice holds such legal significance, a copyright notice in a watermark probably would not suffice as an alternative to including the standard “©” or “®” notice. However, the system does make it easier for honest people to find out who they should contact about using a digital Work.

It is enticing to try to use watermarks not just to *identify* copyright ownership but to actually *prove* ownership. This is something a textual notice cannot do, because it can be so easily forged. For example [3], suppose an artist (called Alice) creates an image and public it, with the copyright notice “©2003 Alice”. An adversary (called Bob) then steals the image, uses an image processing program to replace the copyright notice with “©2003 Bob”, and then claims to own the copyright himself. How to resolve such kinds of dispute?

One way of resolving such a dispute is by use of a central repository. Before putting her image on the Web, Alice could register the image with the Office of Copyrights and Patents by sending a copy to them. They archive the image, together with information about the rightful owner. Then, when a dispute between Alice and Bob arises, Alice contacts the Office of Copyrights and Patents to prove that she is the rightful owner.

However, Alice might decline to register her image because it is too costly. Registering with the Office of Copyrights and Patents costs \$30 per document². With many images to be registered, this can add up to a substantial expense. If Alice can't afford this expense, she might find herself prosecuting Bob without the benefit of the Office of Copyrights and Patents on her side.

In such a case, Alice would have to show evidence that she originally created the image. For example, either the file negative or the the early drafts maybe used as the evidence. But all these can be forged by Bob as well.

Can Alice protect her rights, and avoid incurring the cost of registration,

²<http://www.loc.gov/copyright>

by applying a watermark to her image? In the case of Digimarc's watermark, the answer is probably no. The problem with their system is that the detector is readily available to adversaries. Anybody who can detect a watermark can probably remove Alice's watermark and replace it with his own.

To achieve the level of security required for proof of ownership, it is probably necessary to restrict the availability of the detector. When an adversary does not have a detector, removal of a watermark could be difficult if not impossible. Therefore, when Alice and Bob go before the judge, Alice would enter the disputed copy into the detector, and the detector would detect Alice's watermark.

However, even if Alice's watermark cannot be removed, Bob might be able to undermine her. As described by Craver *et al.* [20], Bob, using his own watermarking system, might be able to make it appear as though his watermark were present in Alice's original copy of the image. Thus, a third party would be unable to judge whether Alice or Bob had the true original.

This problem can be solved if we make a change in the problem statement. Instead of trying to directly prove ownership by embedding an "Alice owns this image" watermark message in it, we will instead try to prove that one Work is derived from another. This can be done by combining with cryptographic technique [3].

2.1.2 Broadcast Monitoring

Advertisements, either through television or radio broadcasting, are very common in nowadays society. Advertisers, having paid money for the advertisements would of course like to ensure that they receive all of the air time they purchase from the broadcasters. The performer, in turn want to ensure that they get the royalties due to them from the advertising firms. In addition, owners of the copyrighted Work want to ensure that their property is not illegally rebroadcasted by pirate stations. All these types of organizations and individuals are interested in broadcast monitoring.

A rather low-tech method of broadcast monitoring is to have human observers watching the broadcasts and recording what they see or hear. This

method is costly and error prone. It is therefore highly desirable to replace it with some form of automated monitoring. Techniques for doing this can broadly be broken down into two categories. *Passive* monitoring systems try to directly recognize the content being broadcasted, in effect simulating human observers, though more reliably and at lower cost. *Active* monitoring systems rely on associated information that is broadcast along with the content.

A passive system consists of a computer that monitors broadcasts and compares the received signals with a database of known Works. When the comparison locates a match, the song, film, TV program, or commercial being aired can be identified. This is the most direct and least intrusive method of automatical broadcast monitoring. It doesn't require any change to advertisers' workflow, neither does it require any cooperation between the advertiser or broadcaster.

However, there are several potential problems with implementing passive monitoring systems. First, retrieving the received signal from the database requires very complicated computation. Thus, the system has to process the database and the received signals into smaller signatures that with enough information to differentiate between all possible Works yet small enough to be used as indices in database search. But the difficulty of deriving meaningful signatures and searching in a large database, it is difficult to design a passive monitoring system with 100% reliable. Even if the problem of searching in a large database is solved, storing and managing the database can be expensive.

Passive monitoring does not require the cooperation of those being monitored, it allows the monitoring service to tabulate competitive market research data. For example, such a system can be used to help Pepsi estimate how much CocaCola spends on advertising in a certain market. However, passive monitoring has not been used for verification purpose. This may because that the system is simply not accurate enough. An accuracy of 95% is adequate for acquiring competitive market research data. However, an error rate of 5% is too high for verification services.

To reduce the monitoring cost and obtain a higher accuracy, an active monitoring system is needed. Active monitoring is technically simpler to implement

than passive monitoring. The identification information is straightforward to be decoded reliably, and no database is required.

One way to implement an active system is to place the identification information in a separate area, such as the file header of the broadcast signal. But this separate area may not be guaranteed to be transmitted. Moreover, the appended information is unlikely to survive format changes.

Watermarking is an obvious alternative method of coding identification information for active monitoring. It has the advantage of existing within the content itself, rather than exploiting a particular segment of the broadcasted signal. This allows the information to be detected reliably. The primary disadvantage is that the embedding process is more complicated than placing data in the separate area. There is also a concern, especially on the part of content creators, that the watermark process may degrade the visual or audio quality of the Work. Nevertheless, watermarking is still a promise technique for broadcast monitoring, and there are a number of companies that provide watermark-based broadcast monitoring services.

2.1.3 Other Applications

Digital watermarking has many different widely applications, besides those mentioned above, there are other applications including: Transaction Tracking, Content Authentication, and so on. These applications will be briefly introduced in this section.

Transacting Tracking There is another application of watermarking called *transaction tracking* [3]. In this application, the watermark records one or more transactions that have taken place in the history of the copy of Work in which it is embedded. For example, the watermark might record the recipient in each legal sale or distribution of the Work. The owner or producer of the Work would place a different watermark in each copy. If the Work were subsequently misused (redistributed illegally), the owner could find out who was responsible.

Content Authentication With modern digital technology, it becomes easier and easier to tamper digital Works in ways that are difficult to detect. If the digital Works were a critical piece of evidence of a legal case, such kind of tampering might pose a serious problem. *Content Authentication* is a technique used to authenticate the originality of digital Works.

There are different ways of implementing content authentication. One common approach to this problem is through cryptography, in which a digital signature, essentially an encrypted summary of the message, is created. An asymmetric key encryption algorithm is used, so that the key required to encrypt the signature is different from that required to decrypt it [21]. Only the authorized source of messages knows the key required for creating signatures. Therefore, no adversary can create the same signature. If someone subsequently compares the modified message against the original signature, they will find that the signatures do not match and will know that the message has been modified. But these signatures are metadata that might be removed, specially during format changing.

A preferable solution might be to embed the signature directly into the Work using watermarking[22]. We refer to such an embedded signature as an *authentication mark*. Authentication marks designed to become invalid after even the slightest modification of a Work are called *fragile* watermarks. The use of authentication marks eliminates the problem of making sure the signature stays with the Work. Of course, care must be taken to ensure that the act of embedding the watermark does not change the Work enough to make it appear invalid when compared with the signature.

2.2 Properties

Watermarking system can be characterized by several defining properties [1, 23, 24, 25, 26]. We will describe some of them in this section. The relative importance of each property is dependent on the requirements of the application and the role the watermark will play.

2.2.1 Transparency

We define the *transparency* of a watermarking system as the perceptual similarity between the watermarked Work and the unwatermarked one at the point at which they are presented to a consumer.

There is another term for describing this property, so called *fidelity*. The watermarked audio Work is not necessary of high quality, since the quality of the original Work might be just so so. But the audio Work with watermark embedded should at least be undistinguishable with the original one.

According to the definition of watermarking, the embedding process should be transparent to the consumer while the digital Work is under normal usage. Perceptually, either from vision or hearing, the watermarked Work should be identical to the original one. In this thesis, we consider this property as the one with the highest priority.

2.2.2 Robustness

Robustness refers to the ability to detect the watermark after common signal processing operations. Example of common operations on images include spatial filtering, lossy compression, printing and scanning, and geometric distortions. Audio watermarks should to be robust to processes such as linear and nonlinear filtering, A/D, D/A conversions, noise added, lossy compression(MP3,AAC *etc*), temporal scaling, equalization, removal or insertion of samples [26].

It is not easy to have a watermarking system robust to all possible form of signal processing. And not all watermarking applications require robustness to all signal processing operations. Rather, a watermark needs only to survive the common signal processing operations likely to occur between the time of embedding and the time of detection. This is highly application dependent.

In some cases, robustness may be completely irrelevant, or even undesirable. In fact, an important branch of watermarking research focuses on fragile watermarks. A fragile watermark is one designed so that it is not robust [3] . For instance, a watermark designed for authentication purposes should be fragile.

At the other extreme, there are applications in which the watermark must be robust to every conceivable distortion that does not destroy the value of the cover Work. This is the case when the signal processing between embedding and detection is unpredictable.

When discussing about robustness, there is an assumption that the signal processing won't degrade the perceptual quality of the audio Work. The watermarking system can be designed to survive only such kinds of signal processing.

2.2.3 Other Properties

In addition to the properties mentioned above, there are several important issues that should also be considered while developing a watermarking system. These include: Security, Data Payload, Blind or Informed Detection and Cost .

Security The security of a watermark refers to its ability to survive hostile, malicious attacks. The types of attacks fall into three broad categories: Unauthorized detection, Unauthorized removal and Unauthorized embedding [3]. Detection does not modify the cover Work and is therefore referred to as a passive attack. Removal and embedding are referred to as *active* attacks because these attacks modify the cover Work.

Detection refers to the practice of detecting the presence of the watermarking, or detecting and distinguishing the mark from the others but can't decipher it, or even detect and decipher the embedded watermarking. Removal refers to attacks that prevent a Work's watermark from being detected. This may be done by eliminating or masking the original watermark. Embedding refers to the acts that embed illegitimate watermarks into Works that should not contain them. Based on different application requirements, the watermarking system shall be designed to have different levels of security to resist malicious attacks.

The same as common signal processing, only malicious attacks that can maintain the quality of the audio works are considered as successful attacks. So the watermarking system should be designed so that an attempt to remove or obscure the watermark should damage the quality of the Work, this is called

tamperproof-ness [27].

Data Payload Data Payload refers to the number of bits a watermark encodes within a unit of time or within a Work. For audio, data payload refers to the number of bits encoded per second that are transmitted [3].

Different applications may require very different data payload. In the watermarking research literature, many systems have been proposed in which there is only one possible watermark and the detector determines whether or not that watermark is present. Many other applications require not only to detect the presence of the watermark, but also to decode what exactly the message is. Suppose the watermarking system can embed N information bits, then the system can be used to embed any one of 2^N different messages. So the detector would therefore have $2^N - 1$ possible output values: 2^N messages plus “no watermark present”.

Blind or Inform Detection In some applications, the original Work can’t be accessed. The detector should be able to detect the watermark information without any knowledge of the original Work. Yet, in some other applications, the original Work is available at the detection end. We refer to the detection that does not require any information related to the original as *blind detection* [28]. Conversely, the detection that requires access to the original Work is referred to as an *informed detection* [29]. Whether a watermarking system employs blind or informed detection can be critical in determining whether it can be used for a given application. Informed detection provides extra robustness against intentional and unintentional attacks [30]. But informed detection can only be used in those applications where the original Work is available.

Cost The economics of deploying watermark embedders and detectors can be extremely complicated and depends on the business models involved. From a technological point of view, two principal issues should be concerned: the speed with which embedding and detection must be performed; and the number of embedders and detectors that must be deployed. Some applications may require

real time processing, like broadcast monitoring, both embedders and detectors must work in real time. Some other applications, like identification and proof of ownership, the user may be willing to wait for a while to see the result. And the number of embedders and detectors also varies with applications, and should be taken into consideration while designing a watermarking system.

Chapter 3

Possible Methods for Audio Watermarking

During the past few years, more and more research interests in audio watermarking have been shown and different digital audio watermarking methods have been proposed. The current methods can be divided into the following aspects based on coding technique: low bit coding, phase coding, echo coding and spread-spectrum based coding [30, 31]. And if classified by embedding domain, the methods can be applied in time domain and transformation domain, such as DCT domain, Wavelet domain, *etc.*.

The characteristics of Human Auditory System(HAS) are exploited by many of these methods. Audio watermarking is especially challenging, because HAS operates over a wide dynamic range, and it is sensitive to additive random noise. However, human ears at the same time are not perfect detectors, they can't detect sound below a certain threshold. While the HAS has a large dynamic range, it has a fairly small differential range[30]. As a result, there exists a phenomenon called *masking* and it is exploited in some of the techniques. The knowledge of modern telecommunication theory, such as spread spectrum technique, information entropy, and other theory, such as information hiding, cryptography can also be applied in audio watermarking system.

In this chapter, we will first describe the overview of a general audio watermarking system in Section 3.1. Then we will review on the methods published

in the literature in Section 3.2. Some other related methods will be introduced in Section 3.3. Finally we will describe the outline of the method we are going to propose in Section 3.4.

3.1 Overview of Digital Audio Watermarking System

Like common watermarking system introduced at Chapter 1, digital audio watermarking system has at least two basic components, the embedding and the detection modules. Before the embedding process, there is a very important step namely the watermark generation. A general system of audio watermarking, including watermark generation, embedding and detection, can be illustrated in a diagram as shown in Figure 3.1.

From Figure 3.1, the watermark generation module takes the ownership information or buyer information M , as its basic input. In some applications, the secret key K , the cover audio file may be used by the watermark generation to generate a cover Work depended watermark [17]. For security purpose, the original information M may be encrypted using cryptographic techniques. The watermark generator then produces a binary bit sequence, which is the information to be embedded in the Work. Error correction coding (ECC) technique may be applied to encode the binary stream with redundancy, so that the binary sequence has the ability to self-detect and self-correct the error bits which may happen during storage, transmission or detection process. When using ECC, there is a trade off between robustness and the embedding capacity.

After the binary information bit stream is generated, it will be fed into the Embedding System accompanied with the cover audio Work x and the secret key K . Here is the most important part of the whole system. There are many approaches to embed the cover audio Work with the watermark information bits. These approaches will be discussed in Section 3.2. The output of this module is the watermarked embedded Work x' , which will be stored in some kinds of media or transmitted through certain channel. During these procedures, the

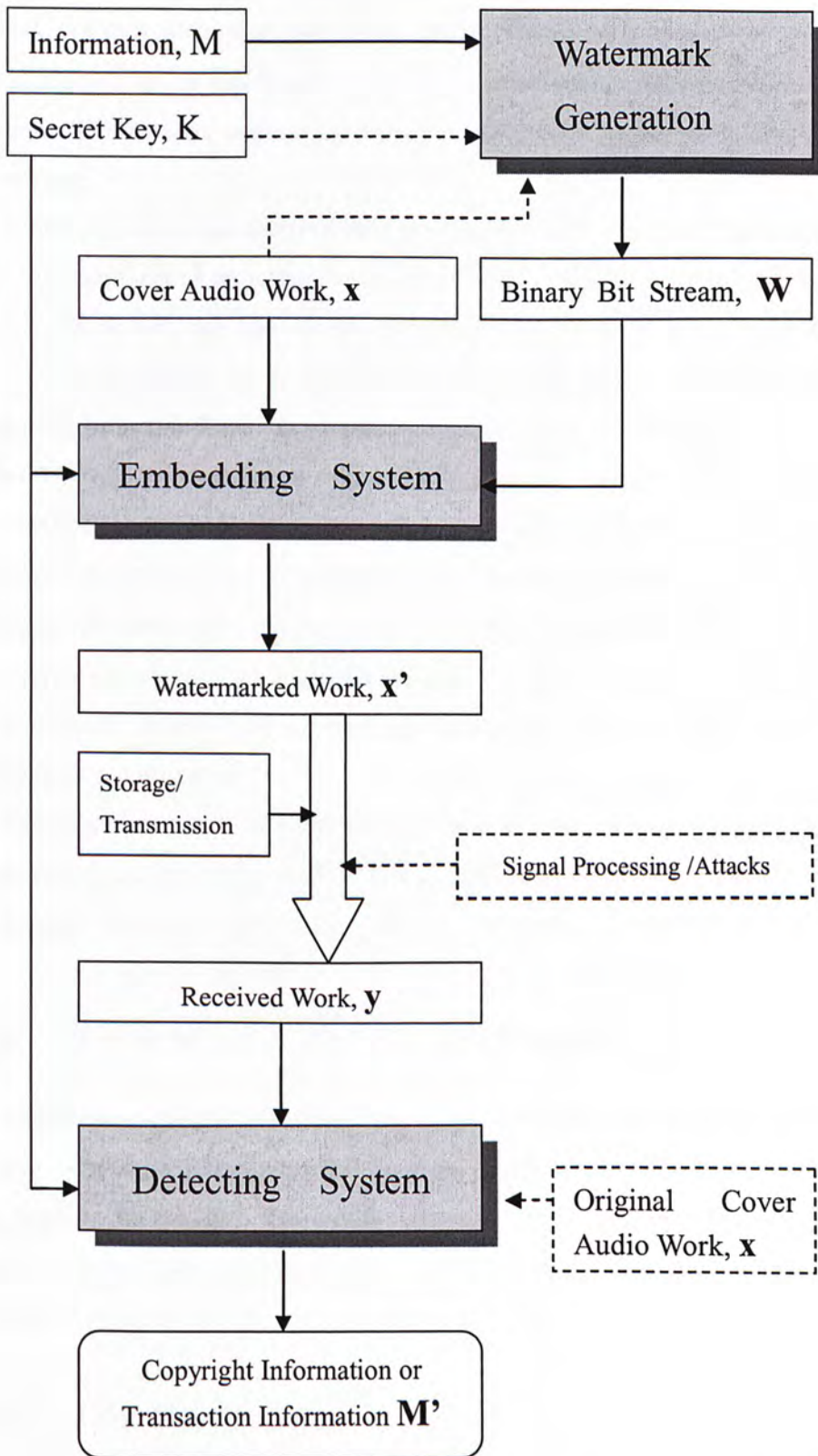


Figure 3.1: A General Digital Audio Watermarking Flowchart (Dotted Line Means Optional).

audio Work may encounter lossy compression, resampling or some other kinds of signal processing. In addition some malicious attacks from the pirates could also happen. The attacks may try to detect, remove or even change the embedded watermark.

When dispute about the ownership of a certain audio Work occurs, or whenever the copyright of an audio Work y needs to be proved, the copyright owner of that Work will use the secret key and some other parameters about the watermarking system, to extract the copyright information or transaction information from the Work. Some methods may require the access to the original audio Work [32], while some others don't have that requirement. According to whether the original Work is needed, watermarking methods are classified as *informed-detected* and *blind-detected*, discussed in Chapter 2, Section 2.2.3. Actually, the detection process is usually the reverse process of the embedding one. After the binary information bit stream is detected and it will go through the watermark information decoder, and decoded into the copyright information or transaction information.

We have just given an overview of a general audio watermarking, the real implementation may vary with different applications. We will now introduce some proposed approaches in the following sections.

3.2 Review of Current Methods

Several previous digital audio watermarking techniques will be reviewed in this section. The techniques presented here can be derived from the general model described in Section 3.1. They don't correspond to the actual implementation of the available commercial products, but rather, constitute the basis for some of them.

3.2.1 Low Bit Coding

This method, also known as *least significant bit (LSB) substitution*, is both common and easy to apply in both steganography and watermarking [33]. The

basic idea of this technique is to embed the watermark information bits in the cover audio file by replacing the least significant bit of each sampling point. Take a 16 bits coded sample as an example, the least four bits can be used for hiding information bits. The retrieval of the hidden data in low bit coding is done by reading out the value from the low bits. The stego key is the position of the altered bits [30]. Low bit coding is the simplest and most intuitive way to embed data into digital audio and can be applied in all ranges of transmission rates with digital communication modes. Suppose the least significant bit of each sample is replaced, then ideally in the noise-free channel, the channel capacity for an 1kHz sampling sequence is 1kbps. Thus 8kbps is the capacity of 8kHz sequence and 44kbps is the capacity of 44kHz sequence.

This large channel capacity in return is an introduction of audio noise. The impact of this noise is a direct function of the content of the host signal. For example, crowd noise during a rock concert would mask some of the noise that would be audible in a string quartet performance. Another disadvantage of this method is its poor immunity to manipulations. Encoded information can be destroyed by channel noise, re-sampling, etc. In order to be robust, these techniques reduce the data rate, usually by one to two orders of magnitude.

Low bit coding embedding can also be done in transformation domain besides time domain. In such an approach, a transformation is applied to the signal, and then the least significant bits of the coefficients representing the audio signal on the transformation domain are modified to embed watermark information bits. The inverse transformation is performed to obtain the watermarked audio file after the embedding process. The possible transformation used for watermarking can be: discrete Fourier transform(DFT), discrete cosine transform(DCT), Mellin-Fourier transform, and wavelet transform [34]. However, they are used more popular in the field of image and video watermarking.

3.2.2 Phase Coding

As we know, human auditory system is less sensitive to phase components of sound than to noise components. And the human auditory system is sensitive to

relative phase differences, but not to absolute phase changes. This property is exploited by some audio compression schemes. Watermarking by phase coding makes use of this characteristic as well [32, 33].

This method is done by substituting the phase of the original audio signal with reference phases. The original signal $\{x(n)\}$ is split into a set of frames s_i . Then a discrete Fourier transform(DFT) is applied to each one of the resulting frames. This transformation generating a matrix of phases Φ and a matrix of Fourier transform magnitudes. The phase coding method works by substituting the phase of the initial audio segment with a reference phase that represents the data. Since the phase shift between consecutive signal frames must be preserved in the watermarked file, the phase of subsequent segment is adjusted in order to preserve the relative phase between them [32].

The embedding process inserts the watermark information in the phase vector of the first frame of $\{x(n)\}$, called $\vec{\Phi}_0$. The new phase matrix Φ' is created using the original phase difference. Along with the original matrix of Fourier transform magnitudes, it is used to construct the watermarked file, by inverse Fourier transform. At this stage, the absolute phases of the signal have been modified, but their relative differences are preserved, which suppose to guarantee the transparency of the watermark embedding. Through the process, the matrix of Fourier amplitudes should be kept constant. Any modification to it could generate intolerable degradation [34].

In the recover stage, the same segmentation and transformation are applied and the value of vector $\vec{\Phi}_0$ can be measured and thereby the encoded watermark information can be found.

Phase coding is sensitive to most audio compression algorithms.

3.2.3 Echo Coding

Echo coding embeds data by introducing an echo[30, 35]. It attempts to embed information on the original audio signal by introducing a repeat version of a component of the audio signal with small enough offset(or called delay). The echo signal can be $\alpha x(t - d)$, and the watermarked signal is expressed as:

$$x(t) = x(t) + \alpha x(t - d) \quad (3.1)$$

In the most basic echo watermarking schemes, the information is encoded by modifying the offset d . This means two different values d_1 and d_2 are used in order to encode either a zero or a one. Both offset values have to be carefully chosen in a way that makes the watermark both inaudible and recoverable [33]. As the offset between the original and the echo decreases, the two signals blend. At a certain point, human ear cannot distinguish between the two signals. The echo is perceived as added resonance[32]. This point is not easy to determine exactly, as it depends on many factors, such as the quality of the original recording, the type of the sound and so on. However, in general, one can expect the value of the offset to be around 1 millisecond. And α is also an amplitude factor that can help to achieve transparent embedding.

The basic scheme can only embed one bit in a signal, a more practical method is to segment the original $\{x(n)\}$ in to various blocks and then each block is used to embed one bit, using the method introduced above.

In the detection stage, a technique known as *cepstrum autocorrelation* is used [36]. This technique produces a signal with two pronounced amplitude humps, or called *spikes*. By measuring the distance between these two spikes, the information bit, one or zero, can be determined. Echo hiding can effectively embed watermarking information with transparency. However, it is easy to be detected and vulnerable to attacks.

3.2.4 Spread Spectrum Watermarking

Spread Spectrum technique is widely used in telecommunication applications, such as CDMA. Spread Spectrum technique is described thoroughly in [37] by Simon et al. and [38] by Pickholtz et al. Watermarking approaches based on spread spectrum exploit theory from the communications community [36]. The basic idea of spread spectrum watermarking is to embed a narrow-band signal (the watermark information W) into a wide-band channel (the cover audio Work x). The characteristics of both W and x seem to suit this model perfectly. In

addition, the pseudorandom sequence is a very important component in spread spectrum technique, the secret key used in the pseudorandom sequence generator offers the possibility of ensuring the security of the watermarking system.

Spread spectrum technique has two characteristics that are important to watermarking. First, the signal energy inserted into any one frequency is very small. This low energy modification reduces the risk of perceptible distortion. However, these low energy modification present in a wide frequency range can be de-spread or concentrated and the output of the detector can with a high energy watermark signal. Second, the fact that the watermark is spread over a large number of frequencies component also provides robustness to many common signal processing [3].

Furthermore, embedding information in the high frequency component of the signal usually result in transparency but no robustness. Whereas low frequency components have the opposite characteristics. Thus the objective of transparent is often fundamentally opposed to the requirement of robustness. One possible solution to this conflict is to spread the watermark over a large frequency range. This is why spread spectrum techniques are valuable not only for robust communication but for watermarking as well.

Boney, et al. [17] suggest to embed a watermark into significant portions of audio files by generating a PN sequence of binary number and filtering it using a 10th order all-pole approximation of the MPEG-1 Lay-1 psychoacoustic masking model. Prior to embedding, the filtered PN sequence is weighed in time domain to prevent pre-echoes. While most of the embedding properties used by Boney resemble those of a transparent, robust and secure watermark model, the process of embedding the watermark into the entire audio frame is not very efficient.

3.3 Other Related Approaches

Cox, et al, [39] suggested a method of embedding watermark information in the frequency domain of selected samples, based on the perceptual entropy of the

image or audio file. Watermark information W is represented using an independently selected random number sequence $W = [w(1), w(2), \dots, w(n)]$ within the range $[0,1]$ with uniform distribution and reasonable but finite precision. The embedding function can be represented as:

$$x'(n) = x(n) * [1 + \alpha(n) * w(n)] \quad (3.2)$$

where $x'(n)$ represents a sample of multimedia Work with watermark embedded, $x(n)$ represents the original sample and $w(n)$ is a single watermark sample embedded using scaling parameter $\alpha(n)$. The scaling parameter can be a constant or a variable value that is determined based on the acceptable distortion introduced by the embedding process. While the idea of embedding information using perceptually determined significant portions of the multimedia Work ensures watermark robustness against the majority of sophisticated attacks. The process of determining a set of scaling parameters $\alpha(n)$ can be a complicated and time-consuming task. Because the process is based on the perceptual properties of each individual sample and should vary with each frame of the file.

A more efficient method is given by Petar Horvatic et al. [18]. In this approach, the author, selected a group of frame samples, totally 64 out of 512 samples representing the majority of the audio energy, as candidates for watermark embedding. Candidates are selected based on power spectral analysis. It was proposed to spread a narrow band watermark over the range of significant frequencies using combination of Fourier transforms and scrambling function. The scrambling function used is a pseudorandom number sequence of 64 real number in the range $[0,1]$. The authors claim that the designed system is transparent and robust. The process of selecting samples based on power spectral analysis requires heavy computation, and it is complex to scramble the frame samples by using a pseudorandom number sequence. The embedding scaling scheme proposed is not efficient enough. A filter formed by the 64 selected samples used to achieve the transparency reduces the robustness of the watermark signal as well.

3.4 Outline of Proposed New Method

As introduced in Section 3.2 and 3.3, there are quite a number of possible methods for audio watermarking. Different approaches are suggested based on different tasks, and the requirements also vary with applications. To propose a certain approach, we should first make clear what kinds of applications the approach aims at. Then the basic requirements of these applications should also be taken into consideration when designing the system.

In this thesis, we propose an approach for audio watermarking, based on spread spectrum technique. The outline of this approach is shown in Figure 3.2. The major applications of this approach can be: Ownership Identification and Proof, Broadcast monitoring and Transaction Tracking of audio Work. As discussed in Chapter 2, the similarity of all these applications is: before transmitting or distributing any copyright protected audio Work, an ownership information or buyer(recipient) information is embedded into that Work as a watermark. The basic and most important requirement is that the watermarking process won't introduce any perceptual distortion to the Work. The watermark is also supposed to be robust to common signal processing.

To apply these applications, people may turn to digital audio watermarking technology. In this thesis, we focus our study on using spread spectrum technique. This method codes the watermark information into binary bit stream and embeds the information bits into the audio Work frame by frame based on *content adaptive embedding* method. This primary approach, without any other supplementary processes, can embed and detect watermark information effectively, however it may introduce audible distortion to the audio Work. The perceptual quality of the audio Work however cannot be guaranteed. These issues will be introduced and elaborated in Chapter 4.

In order to improve the audible quality of the watermarked Work, we investigate an effective sample selection process to implement the embedding task. It allows the best frequency samples to be picked for processing based on experimental trials as well as considering the psychoacoustic model of human auditory system. Psychoacoustic model is derived based on the characteristics of Human

Auditory System (HAS), which can not only provide possibility of transparent embedding, but also make the watermarking process less susceptible to lossy compression.

In all the applications we consider, the number of watermark information bits to be embedded is usually smaller compared to the frame number of the audio Work. Thus in addition to sample selection, there is a possibility to select only some frames, which provide better effect for doing watermarking. The word “better” means the watermark will be more transparent by embedding into these frame. The general implementation procedure of the sample and frame selection processes as well as the ways of realizing transparent embedding will be elaborated in Chapter 5. Finally, the robustness property of this approach will be verified in Chapter 6.

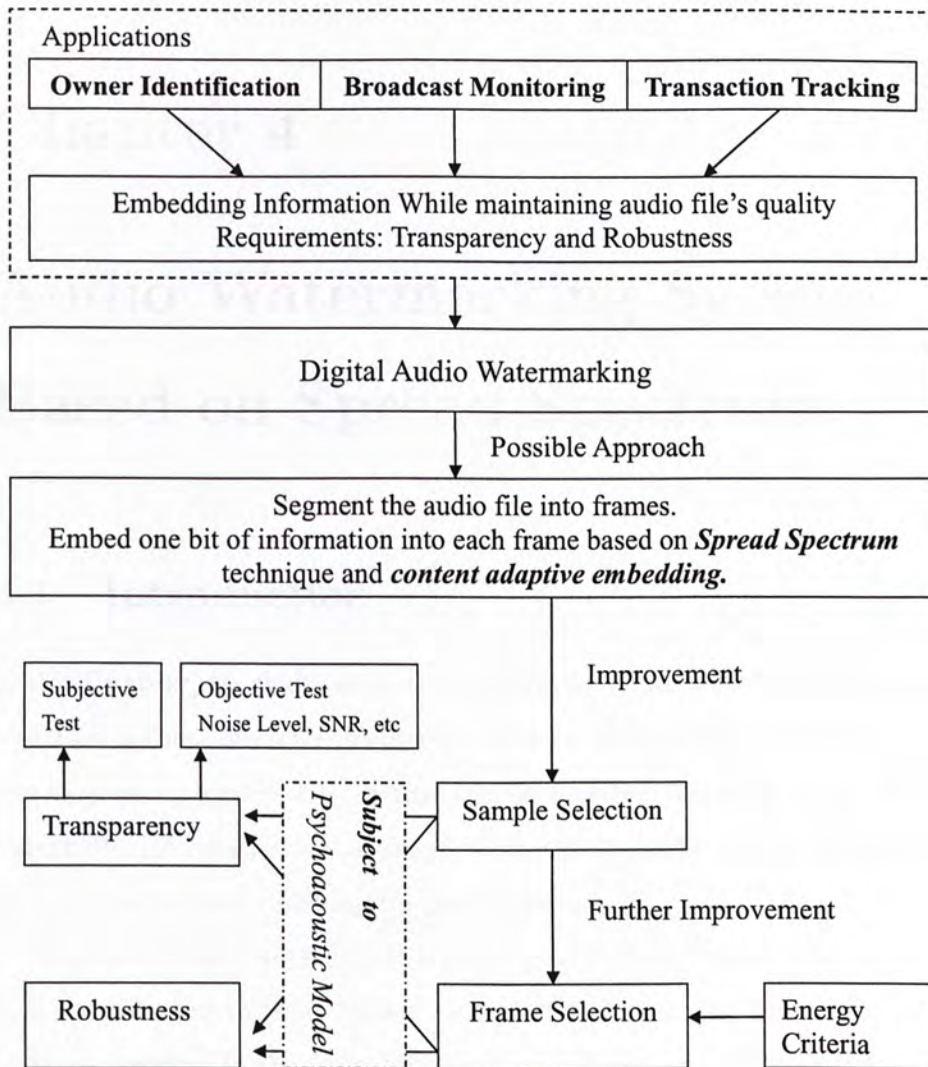


Figure 3.2: The Whole Outline of the Proposed Watermarking Method

Chapter 4

Audio Watermarking System Based on Spread Spectrum

4.1 Introduction

In this Chapter, we shall introduce a sophisticated audio watermarking system based on spread spectrum technique. The basic block diagram of such a system is given in Figure 4.1. In this approach, given the cover audio Work and watermark information, we attempt to embed the audio Work with the hidden information without affecting its perceptual quality.

The cover audio Work $\{x(n)\}$ is first segmented into frames with length of L . So a set of frames in time domain $\{s_i(n)\}$ is obtained after the *Segmentation* block as depicted in the figure. The watermark information M is first encoded into a $\{0,1\}$ binary bit stream, and then processed by error correcting coding (ECC) or/and encryption if applicable. These processes can be summarized as the *Watermark Generation* module shown in the block diagram. The output of this module is a binary information bit stream W . Each of the binary information bits is to be embedded in one frame by modifying the frequency samples of that particular frame. The watermark embedding is performed on a frame by frame basis according to the bit sequence. The embedding process can be described as follows: when the bit to be embedded is '0', then the frame content will be unchanged; whereas when the information bit is '1',

a modification of the frame will be implemented. Such a modification shall introduce as little perceptual distortion as possible; and shall be detectable at the detection end. The embedding process, represented by the **Embedding Block**, is based on spread spectrum technique. The embedding process takes the cover audio frames $\{s_i(n)\}$, the information bits \mathbf{W} and also the secret key \mathbf{K} as its inputs. The output of this module is the time frames $\{s'_i(n)\}$ with the information bits embedded. These frames are used to reconstruct the audio Work $\{x'(n)\}$ with watermark information included. This spread spectrum based embedding method can spread the induced distortion to a wider frequency range effectively [3, 28, 40, 18]. However, it can't guarantee transparency of the watermarking system fully by only using this technique. In other words, $\{x'(n)\}$ may loose its fidelity as compared with the original one $\{x(n)\}$.

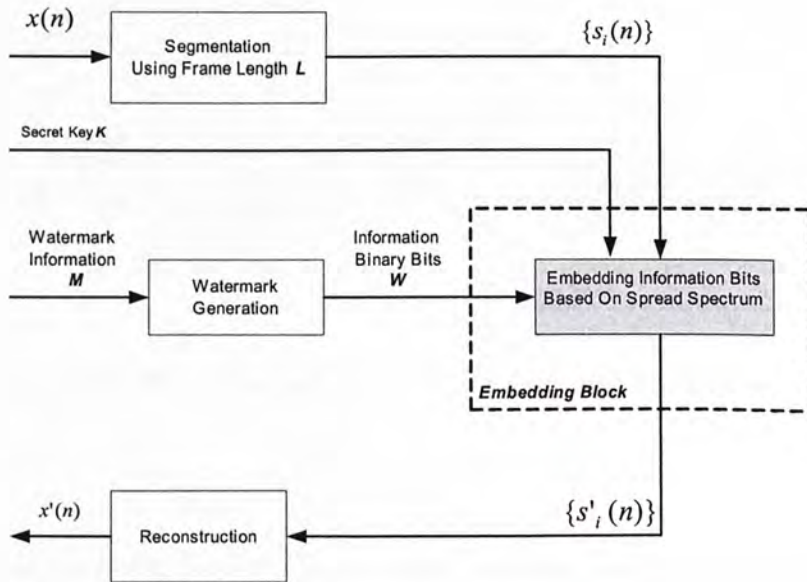


Figure 4.1: Block Diagram of a Watermarking System Based on Spread Spectrum Technique

At the detection end, the process can be explained by the block diagram in

Figure 4.2. The audio Work $\{x'(n)\}$ to be examined is segmented into frames with the same length L as in embedding stage. All these frames are assumed to be with watermark information bits included, either '0' or '1'. They are fed into the **Detection Block** shown in the diagram with the same secret key K used in the embedding stage. The detection process recovers the information bits by de-spreading the watermark signal energy which has been distributed along the frequency domain while embedding. The output of the detection block is binary information bit stream W' . And the information bits are used to decode the watermark information by the **Watermark Information Decoding** module in the diagram. The output of this module is the watermark information M' .

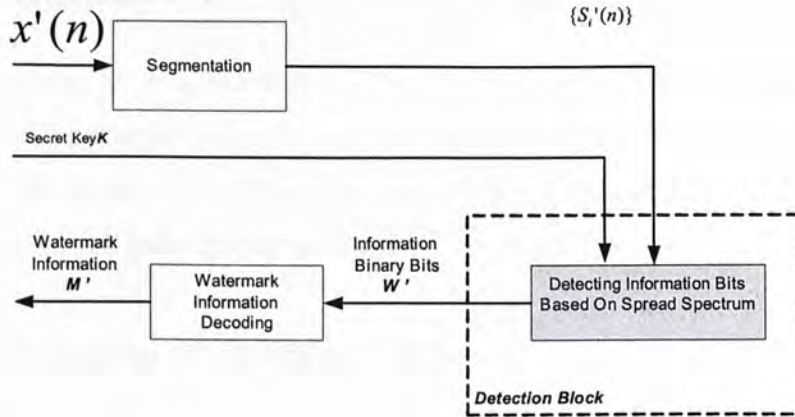


Figure 4.2: Block Diagram of the Watermarking Detection Module

In the following Section 4.2, we shall introduce the general procedure of embedding and detecting one information bit in each frame based on spread spectrum technique. After describing the embedding and detection flowcharts, we will then discuss the generation of the so called *Scramble Function* [18, 17] and the role it plays in the system.

For optimal embedding processing to minimize the degradation, objective comparing metric is needed. The difference between the original frame and the

watermark embedded one, will be defined as the induced noise. An objective metric based on the energy of the induced noise will be proposed for evaluating the system performance. Signal to noise ratio (SNR) of the watermark embedded Work may also be used as an evaluation indicator. These metrics will be explained at the beginning of Section 4.3. Then the so called *Content Adaptive Embedding* method will be introduced and discussed. We will also discuss how to set the frame length L for embedding briefly.

Although watermarking based on spread spectrum technique can embed and detect watermark information bits effectively, perceptual distortion still exists even with the parameters being set to their optimal values. These problems and their possible solutions will be discussed briefly in Section 4.4.

4.2 Embedding and Detecting Information Bit

In this section, we will illustrate how an information bit can be embedded and detected in a particular frame $s_I(n)$. The flowcharts of embedding and detection processes are described in Figure 4.3 and Figure 4.5 separately. They will be elaborated in the following subsections 4.2.1 and 4.2.2 .

4.2.1 General Embedding Process

The flowchart depicted in Figure 4.3 shows how the embedding process is carried out. In this flowchart, the three inputs are: time frame $s_I(n)$, the watermark information bit W and the secret key K . Let the frame $s_I(n)$ be with length of L , then $n = 1, \dots, L$.

Given the secret key K , we can generate a pseudorandom bits sequence (PRBS) $r(n)$ by using Linear Feedback Shift Registers (LFSR), which will be explained in subsection 4.2.3. $r(n)$ is a bipolar bit sequence with length L , the same size as $s_I(n)$. $r(n)$ here is called a *scrambling function*, which is used to scramble $s_I(n)$ into a noise-like signal with flat spectrum in frequency domain. The operation of scrambling is to multiply $r(n)$ and $s_I(n)$ together on a sample by sample basis. We can then obtain the resultant output bit stream,

which can be expressed mathematically as:

$$b_I(n) = r(n) * s_I(n); \quad \text{where} \quad n = 1, \dots, L; \quad (4.1)$$

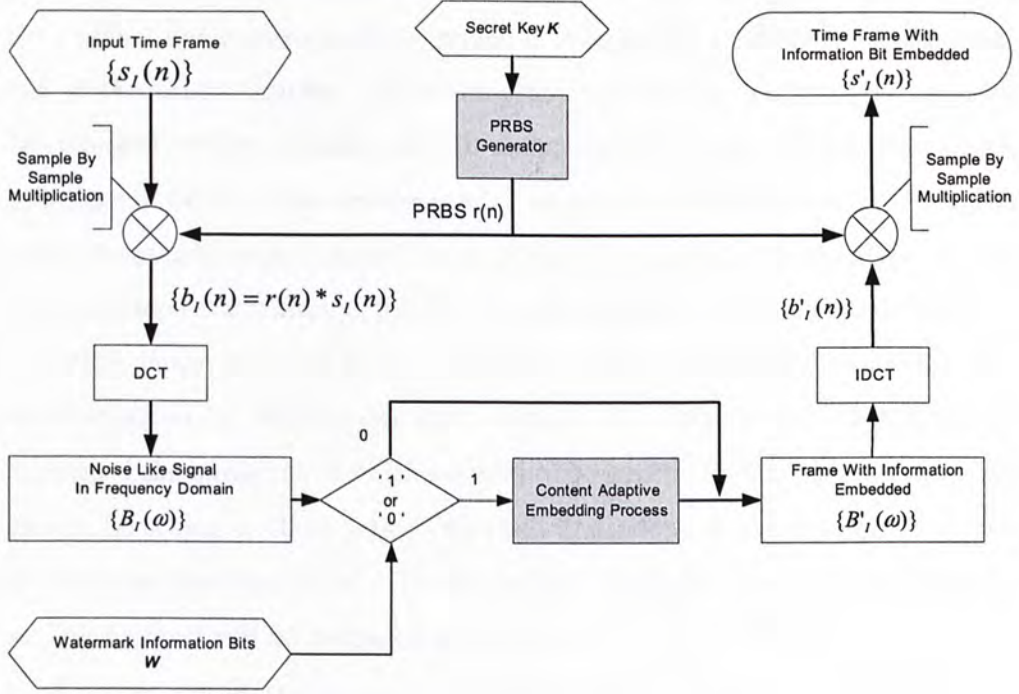


Figure 4.3: Flowchart of the Embedding Process Based on Spread Spectrum Technique

The goal of this multiplication process is to scramble $s_I(n)$ into a random noise-like signal, and we expect the spectrum of $b_I(n)$ to be flat-like as well. It means the frequency spectral samples of $b_I(n)$ are more or less the same, this noise-like characteristics can be utilized for watermarking.

The scrambled signal $b_I(n)$ is transformed into the frequency domain, denoted by $B_I(\omega)$, by using DCT. $B_I(\omega)$ is used to carry one information bit. As

mentioned in the introduction: when the bit to be embedded is '0', the frame content will be unchanged; whereas when the information bit is '1', a modification of the frame will be implemented. So a decision is made to carry out the necessary processing according to the information bit, as shown in the flowchart. $B_I(\omega)$ will be modified to another specified representation called $B'_I(\omega)$ when the information bit is '1'.

In order to embedding information bit '1', one of the frequency samples in the original flat spectrum will be scaled to a larger level, so that it is dominant out of the other samples. (If in the other end such a dominant sample can be detected within a frame, we will claim that this frame carries watermark information bit '1'.) The problems arise here are: to embed an information bit, which frequency sample should be modified? How should it be modified so that the embedded watermark is robust and will introduce the least distortion?

While Petar, et al [18] have proposed to encode a bit of information into the random signal by "*scaling one of its samples to a large value*", they have not addressed the questions of which samples to be scaled and how large the samples should be scaled in their paper. We shall first adopt this scaling method here for finishing the description of the embedding flowchart and a *Content Adaptive* scaling method will be proposed in Section 4.3.

Suppose ***embedIndex*** is the index of the frequency sample in $B_I(\omega)$ to be modified, and this sample is scaled to a new value, denoted by ***newValue***.

According to the pre-definition above, we can modify the frame by:

$$B'_I(\omega) = \begin{cases} \text{newValue} & \omega = \text{embedIndex} \\ B_I(\omega) & \omega \neq \text{embedIndex}; \quad \omega = 1, \dots, L; \end{cases} \quad (4.2)$$

$B'_I(\omega)$ is then the frequency frame with the watermark information bit '1' embedded. According to the flowchart, $B'_I(\omega)$ will be transformed into time domain by applying inverse transform IDCT. That is $b'_I(n) = IDCT\{B'_I(\omega)\}$. Next $b'_I(n)$ will be descrambled by the same scrambling function $r(n)$ as used at the embedding stage. Sample by sample multiplication is applied again,

$$s'_I(n) = b'_I(n) * r(n). \quad \text{where } n = 1, \dots, L; \quad (4.3)$$

$s'_I(n)$ becomes the time sequences of the frame with the embedded information bit. The energy of the enlarged sample will be spread throughout the whole frequency range of the frame by this sample wise multiplication and the DCT process. Each sample in the frame is being changed slightly. The spreading energy can be converged again by the de-spread operation at the detection end.

Figure 4.4 below is an example of the waveform of the time frame and the spectrum of the frame at each step of embedding process.

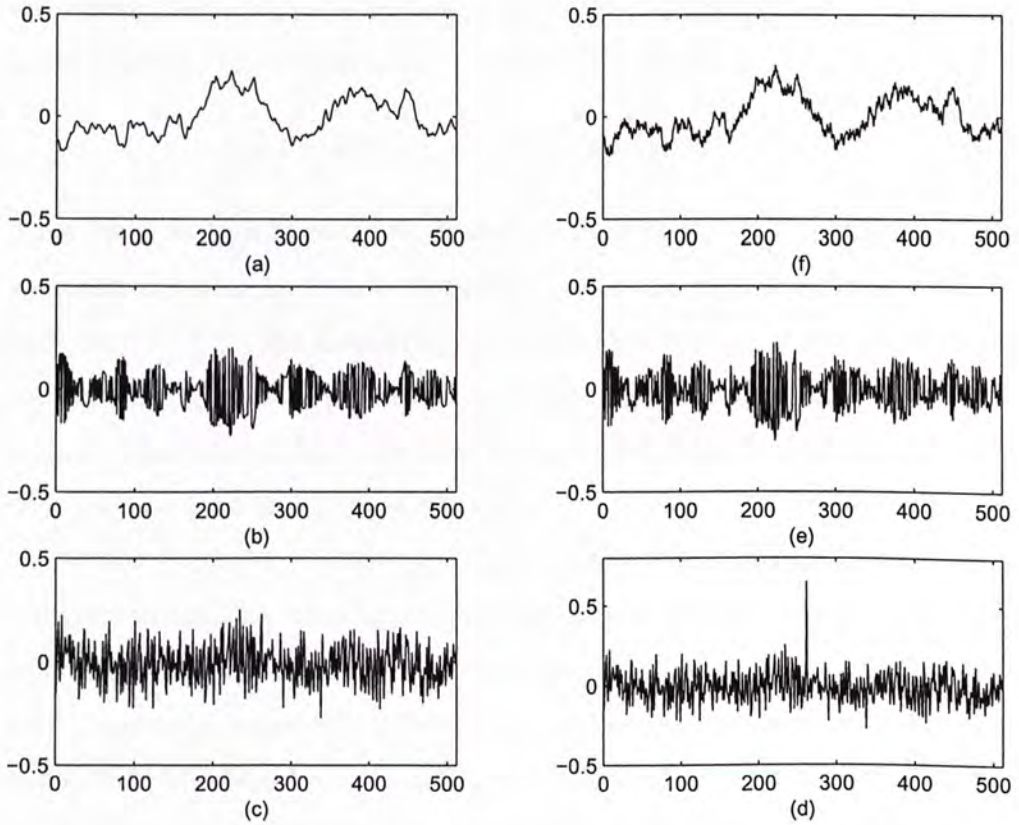


Figure 4.4: The Procedure of Embedding One Bit into a Frame.

Here (a) is the original time frame $s_I(n)$; (b) is the frame after scrambling $b_I(n)$; (c) is the spectrum of the frame $B_I(\omega)$, we can see that it has the noise-

like characteristics; (d) is the frame with a sample enlarged $B'_I(\omega)$; (e) is $b'_I(n)$; (f) is the frame $s'_I(n)$ with watermark information embedded. We can see (f) has maintained the basic waveform of (a) however with noise additive.

4.2.2 General Detection Process

The flowchart of the detection process is displayed in Figure 4.5. The inputs of this process are the testing time frame $s'_I(n)$ and secret Key \mathbf{K} , where the output of this process will be the embedded information bit, either '0' or '1'.

As shown in the flowchart, a time frame $s'_I(n)$ is received by the detector and it is assumed to carry one watermarking information bit with it, either '0' or '1'. The task here is to determine whether '0' or '1' is embedded. The same scrambling function $r(n)$ is applied to decipher $s'_I(n)$ through sample by sample multiplication. The output sequence can be represented by:

$$b''_I(n) = s'_I(n) * r(n). \quad \text{where} \quad n = 1, \dots, L \quad (4.4)$$

Then $b''_I(n)$ is transformed into frequency domain by DCT, giving rise to a spectrum denoting by $B''_I(\omega)$. According to our design, if this frame has been embedded with '1', the spread energy of the watermark signal will be converged together through these processes, then there must be a dominant sample in $B''_I(\omega)$. Otherwise, if a '0' has been incorporated, then there is flat spectrum without such kind of dominant sample.

Thus a threshold which is equal to or less than the dominant sample's value but greater than the other samples will be set as a decision criteria. If there is a sample with amplitude greater than the threshold, then we claim an information bit '1' has been embedded in this frame. Otherwise, we claim an '0' has been embedded. The decision block is shown in the flowchart.

An example of detecting a '1' information bit from a frame was given in Figure 4.6. In the figure, (a) is the frame $s'_I(n)$ to be determined which information bit has been embedded; (b) is $b''_I(n)$, which has been scrambled by $r(n)$; (c) is the $B''_I(\omega)$ with the watermark information embedded. The dotted line in (c) is the threshold.

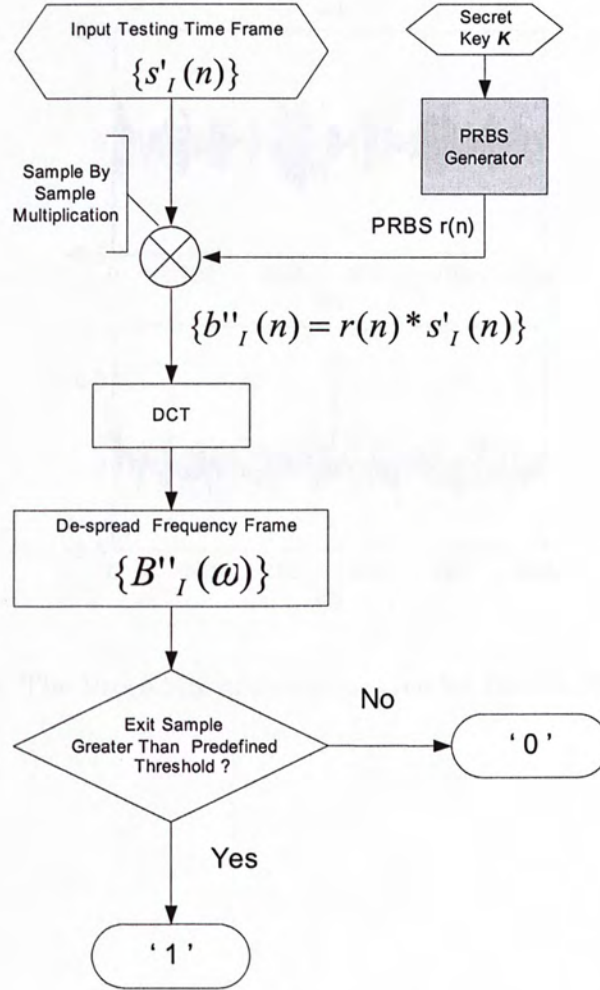


Figure 4.5: Flowchart of the Detection Process Based on Spread Spectrum Technique

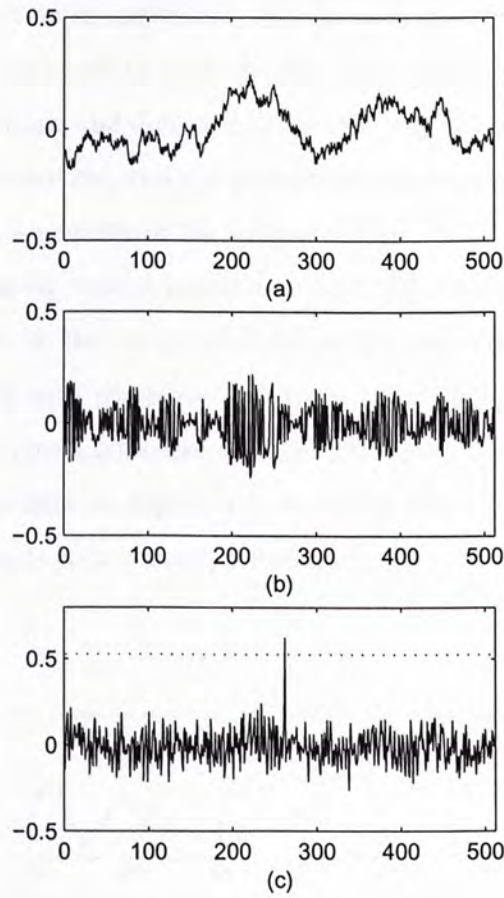


Figure 4.6: The Procedure of detecting One bit from a Given Frame.

4.2.3 Pseudorandom Bit Sequences (PRBS)

In the spread spectrum based watermarking system, the pseudorandom sequence plays a very important role. It forms the basis of the watermarking scheme because of its noise-like characteristics and resistance to interference [17]. It is the pseudorandom sequence that makes it possible to spread the modified frequency point to the entire frequency spectrum.

Spread spectrum signals are resistant to interference such as unintentional

interference, channel noise, multi-path interference, or intentional jammers [38]. A secret key K is required to generate the pseudorandom sequence, which is used in both embedding and detection stage. So only the key holder can detect the watermark information, and the pseudorandom sequence has promised the robustness as well as security of the watermarking.

In [18], Petar. et al. used a random number sequence of 64 unique integers with values chosen in the range of $[0,63]$ as the pseudorandom sequence to scramble the signal into noise-like. But this kind of sequence is unable to scramble the frame into a noise-like segment efficiently. Referring to Figure 4.7, the notions are the same as Figure 4.4, it can be seen that a larger distortion is induced which leads to a poorer performance.

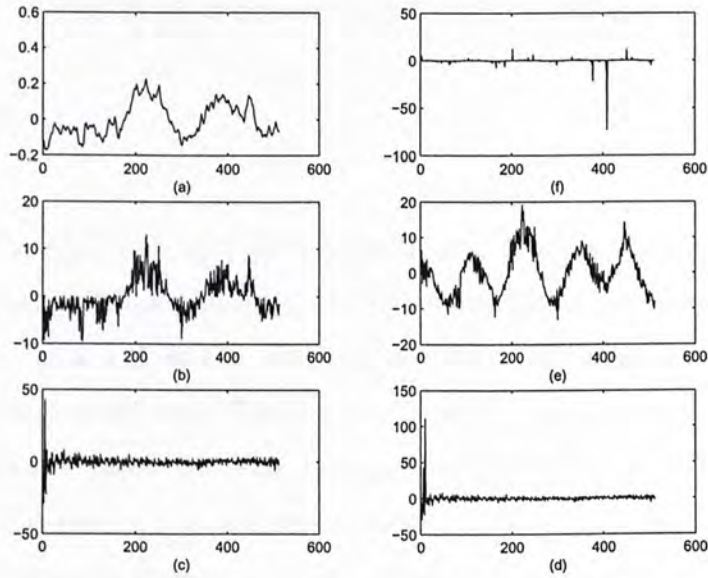


Figure 4.7: An Embedding Process By Pseudorandom Number Sequence.

We propose to use pseudorandom bits sequence (PRBS) as the scrambling function instead. A PRBS can be generated by a Linear Feedback Shift Register (LFSR) [38]. Linear feedback shift registers can be implemented by

the *Fibonacci implementation*. This implementation consists of simple shift registers, as shown in Figure 4.8. A binary weighted modulo-2 sum of the taps is fed back to the input, (modulo-2 sum of two 1-bit binary numbers yields 0 if the two numbers are identical, and 1 if different: $0 + 0 = 0, 0 + 1 = 1, 1 + 1 = 0$)

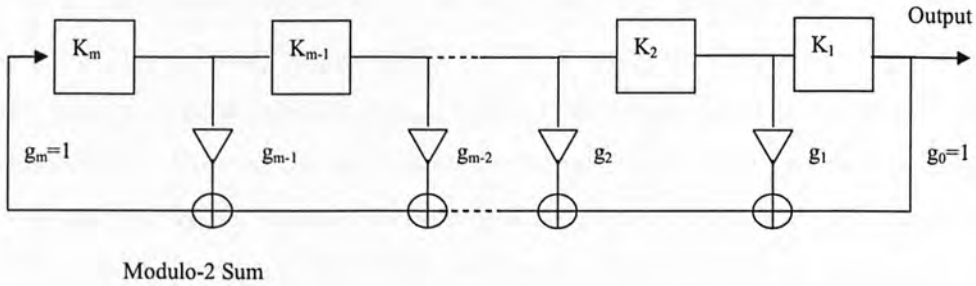


Figure 4.8: Fibonacci Implementation of LFSR.

$K = \{K_m, K_{m-1}, \dots, K_1\}$ is the initial vector. $G = \{g_m, g_{m-1}, \dots, g_1, g_0\}$ is the weight vector. These two vectors are considered as the key of the watermarking system. For any given tap, weight g_i is either 0, meaning “no connection”, or 1, meaning it is fed back. Two exceptions are g_0 and g_m , which are always 1 and thus always connected. Note that g_m is not really a feedback connection, but rather it is the input of the shift register.

LFSR generators produce what are called linear recursive sequences (LRS) because all operations are linear¹ [37, 38]. Generally speaking, the length of the sequence before repetition occurs depends upon two things, the feedback taps and the initial state. An LFSR of any given size m (number of registers) is capable of producing every possible state during the period $2^m - 1$, but will do so only if proper feedback taps, have been chosen. Such a sequence is called a *maximal length sequence*, *maximal sequence*, or less commonly, *maximum length sequence*.

¹<http://www.newwaveinstruments.com/index.htm>

quence. It is often abbreviated as *m-sequence*. In certain industries *m-sequences* are referred to as a pseudonoise (PN) or pseudorandom sequences, due to their optimal noise-like characteristics.

Maximal length generators can actually produce two sequences. One is the trivial one, of unity length, that occurs when the initial state of the generator is all zeros. The other one, useful one, has a length of $2^m - 1$. Together, these two sequences account for all 2^m states of an m -bit state register.

The m -bit binary sequence $\{01 \cdots 10\}$ is generated from the generator, can be used in digital communication directly, since the signal transmitted is also binary bits. However in our system the signal is represented by waveform with real number value. In order to recover the scrambled signal, we will change the '0' in the sequence to '-1'. Thus the pseudorandom bits sequence used in the watermark becomes bipolar pseudorandom bits sequence.

Since the original audio Work is segmented into a set of frames with L length, the $r(n)$ for each frame can set to the same one, or varies with frame. It is a trade-off between security and system complexity.

4.3 An Optimal Embedding Process

4.3.1 Objective Metrics for Embedding Process

While doing the modifying process, there are quite a lot of alternatives to modify the frequency sample to embed one bit of information. How can we compare and analyze the performance of different approaches such that an optimal method can be chosen? The most intuitive way of testing the performance is subjective listening test. But intuitive listening test in many cases just can't tell the slight difference between different approaches, and it is not easy to express different kinds of performance accurately. So it is necessary for us to define some reasonable and effective objective metrics for transparency test of the embedding process.

What we are going to see is how the modifying process will affect the transparency quality of the audio Work. Given a original time frame $s_i(n)$, the energy

of it is fixed, the noise introduced into each frame while doing the embedding process can be used to indicate the distortion. The noise, called $n_i(n)$ can be defined as the difference between the original frame $s_i(n)$ and the watermark embedded one $s'_i(n)$, i.e.

$$n_i(n) = s'_i(n) - s_i(n) \quad \text{where} \quad n = 1, \dots, L \quad (4.5)$$

The energy of the $n_i(n)$ can be used as a quantity measurement of the distortion to the original frame, which can be expressed as:

$$EN_i = 10 * \log_{10} \sum_{n=1}^L |n_i(n)|^2 \quad (dB) \quad (4.6)$$

It is obvious that, to each frame, the less the noise energy, the better the transparency quality the embedding algorithm has achieved.

Another metric takes into consideration the signal to noise ratio of the whole frame, called $fSNR$, which is defined as:

$$fSNR_i = 10 * \log_{10} \frac{\sum_{n=1}^L |s_i(n)|^2}{\sum_{n=1}^L |s'_i(n) - s_i(n)|^2} \quad (dB); \quad (4.7)$$

This can also be written as:

$$fSNR_i = EF_i - EN_i \quad (dB); \quad (4.8)$$

where EF_i is the energy of the whole i^{th} frame.

For evaluation the overall quality of the watermarked embedded file, the SNR can be used and defined as:

$$SNR = 10 * \log_{10} \left\{ \frac{\sum_{n=1}^N |s(n)|^2}{\sum_{n=1}^N |n(n)|^2} \right\} \quad (4.9)$$

where N is the total length of the audio Work

The energy of the induced noise EN , the frame signal to noise ratio $fSNR$ and the SNR of the whole file would be used as objective transparency metrics while embedding watermark information.

Figure 4.9 gives a typical example of the original frame waveform, the noise induced and the watermark embedded frame. In the figure, the energy of the noise equals $EN_I = -6.14$; energy of that frame $EP_I = 9dB$; and the signal to

noise ratio of is $fSNR_I = 15.14dB$. In Figure 4.10. A segment of audio file with 10 second length is shown. (a) is the original signal $\{x(n)\}$ and (b) is the noise introduced $\{n(n)\}$ while doing embedding process. (c) is the watermark embedded signal $\{x'(n)\}$. The SNR of $\{x'(n)\}$ is around $16.21dB$.

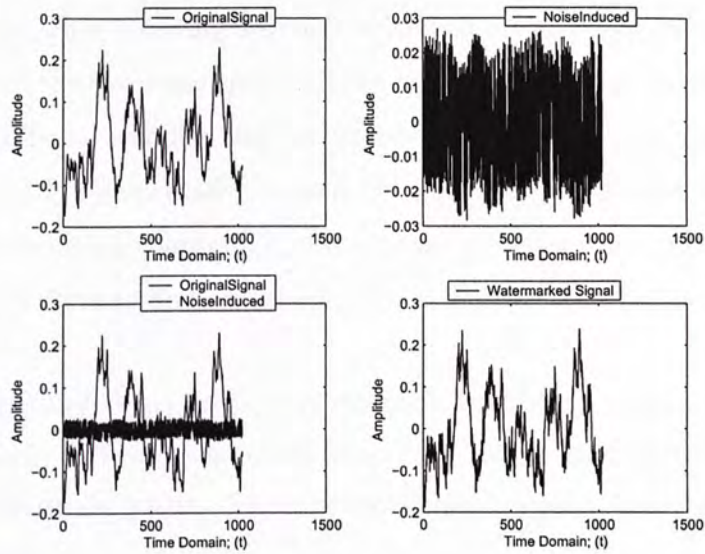


Figure 4.9: A Typical Example of the Original Frame, Noise Induced and Watermarked Frame.

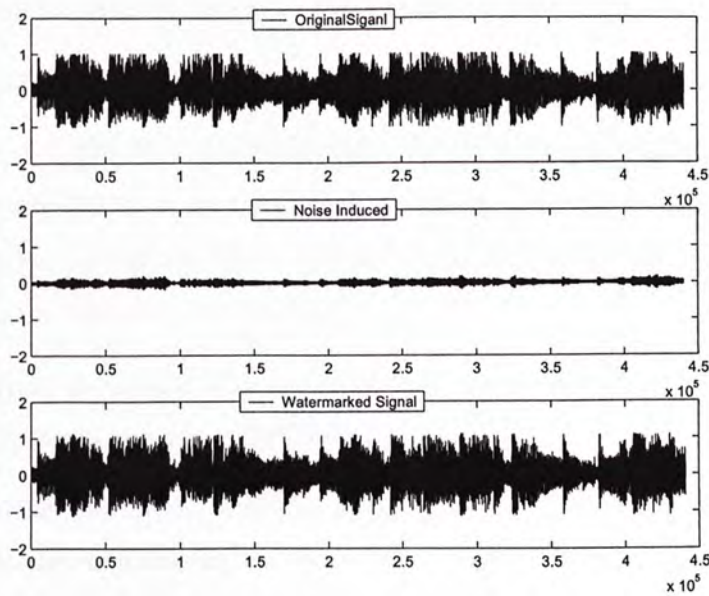


Figure 4.10: A Typical Example of the Original Signal, Induced Noise and Watermark Embedded Signal.

4.3.2 Content Adaptive Embedding

Theoretically, when selecting a sample to embed a watermark information bit, any sample in the frequency spectrum can be the candidate to be enlarged. But modifying different samples may introduce different noise levels. As discussed in Section 4.2.1, the method proposed by [18] is not an efficient one. We will use the induced noise energy EN_i of the frame, as defined in Section 4.3.1, as a comparison criteria to find out which sample shall be modified to produce the best result.

To a particular $B_I(\omega)$, we enlarge different samples independently and compare the energy EN_I of each modification trial. As we know, the less the EN_I , the better the audio quality. The indexes of the frequency samples to be modified including:

- The sample with the smallest amplitude, denoted by *sMallest*
- The first frequency sample in the spectrum, denoted by *First*;
- The last frequency sample in the spectrum; denoted by *lasT*;
- The sample with the second largest amplitude, denoted by *seconD*
- The sample with the largest amplitude, denoted by *Largest*;

In Table 4.1, the first column denotes the 30 different frames. The entries in the table are the noise energies when modifying samples of different indexes, as given by the above list. From the experimental results, we can see that to each frame, the following conclusions hold.

- Modifying the sample with the smallest amplitude will introduce largest noise energy, shown in column *sMallest*;
- Modifying the sample with the largest amplitude will introduce least noise energy, shown in column *Largest*;
- The position of the sample modified has less effect on the noise introduced. The amplitude of the sample modified acts as the plays main role.

Because it can be seen that the energy of noise introduced has no inherent relationship with the position of that sample, as shown in column *First* and

	<i>sMallest</i>	<i>First</i>	<i>lasT</i>	<i>seconD</i>	<i>Largest</i>
1	-10.76	-11.85	-12.06	-16.52	-17.38
2	-10.54	-12.59	-14.92	-16.76	-16.78
3	-10.89	-14.63	-15.41	-17.19	-17.30
4	-7.73	-12.03	-9.65	-13.19	-14.01
5	-8.70	-10.74	-11.48	-15.32	-15.53
6	-3.88	-5.42	-5.30	-9.49	-10.01
7	-7.88	-9.68	-9.83	-14.29	-14.63
8	-10.39	-13.61	-12.70	-16.60	-16.70
9	-10.97	-11.86	-13.55	-16.58	-16.95
10	-15.62	-19.79	-17.40	-21.25	-21.67
11	-15.17	-17.47	-18.21	-21.19	-21.55
12	-15.47	-17.38	-18.96	-20.92	-21.08
13	-17.14	-19.96	-19.75	-23.77	-23.87
14	-14.61	-18.63	-15.82	-20.36	-20.66
15	-10.30	-13.17	-14.80	-15.72	-16.25
16	-2.79	-4.41	-5.70	-9.49	-9.52
17	-1.35	-3.28	-6.12	-7.83	-7.99
18	-0.97	-2.26	-2.96	-6.75	-6.82
19	-2.57	-3.05	-4.75	-8.67	-8.82
20	0.26	-4.86	-4.17	-6.18	-6.56
21	-1.40	-4.65	-4.44	-7.72	-8.35
22	-0.09	-6.68	-0.14	-6.61	-6.68
23	0.52	-1.93	-1.48	-4.77	-5.67
24	0.14	-1.73	-4.00	-6.01	-6.36
25	-5.59	-8.15	-7.93	-10.52	-11.34
26	-12.20	-13.34	-14.71	-18.12	-18.18
27	-10.09	-11.51	-12.64	-14.89	-16.00
28	-9.13	-13.14	-13.43	-15.95	-16.27
29	-11.67	-16.62	-13.56	-17.77	-17.83
30	3.95	-0.67	0.45	-2.12	-2.77
Average	-3.91	-6.82	-6.62	-9.96	-10.40

Table 4.1: Noise Energy $EN_i(dB)$ Induced while Modifying Samples with Different Index. $L = 1024; \alpha = 2.6$

	sM_{allest}	F_{irst}	$lasT$	$seconD$	L_{argest}
EN(dB)	9.20(3.82)	6.23(3.89)	6.31(4.00)	2.34(3.84)	2.01(3.81)
fSNR(dB)	9.61(1.23)	12.58(1.63)	12.50(1.80)	16.47(1.41)	16.80(1.25)
EF(dB)	18.81(3.58)	18.81(3.58)	18.81(3.58)	18.81(3.58)	18.81(3.58)

Table 4.2: Statistical Result of Noise Energy EN and Frame SNR $fSNR$, With Different Embed Index

column $lasT$, either one can be greater than the other with different frames; But the noise energy is inverse proportion to the amplitude of the modified sample energy, as shown in column L_{argest} , $seconD$ and sM_{allest} ;

- Modifying the second largest sample will introduce noise energy close to the one introduced by modifying the largest sample. As shown in column L_{argest} and $seconD$.

So in case the largest sample can't be modified, the second largest one can be the alternative.

These observations are also noted when we do experiments on more frames. This is validated by results shown in Table 4.2, which were based on the analysis of a 10 seconds long audio sample. From the results, it can be seen that the noise energy level is lowest when the sample to be modified is the one with the largest intensity. Also we can see that in this case the $fSNR$ gives the highest value. To modify the second largest sample will give the second best result, but it is very close to the best one.

The reason behind the result is: changing a sample with small amplitude and a sample with large amplitude to a same larger level, the later case will induce relatively less distortion. Suppose $embedIndex$ is the index of the frequency sample in $B_I(\omega)$ to be modified. This sample is scaled to a new large value, denoted by $newValue$. Then the difference D introduced between the watermarked frame and the original one is actually the difference between the $newValue$ and $B_I(embedIndex)$.

$$D = \text{newValue} - B_I(\text{embedIndex}); \quad (4.10)$$

If the *newValue* is fixed, the larger the $B_I(\text{embedIndex})$ is, the less difference D will be. So we will select the largest sample to enlarge whenever possible.

Once we can determine which sample to modify will produce less distortion, the next problem to be solved is to what extent should it be modified to have trade off between transparency and robustness.

If we follow the discussion in Section 4.2, set a fixed *newValue* for every frame, then it will be difficult to find one single value that will give the best result, because the energy of each frame varies within a large dynamic range. A *newValue* that is dominant in a certain frame may be rather small in other frames. A *newValue* that can be dominant in all frames will surely introduce very large distortion to the whole audio Work. So the modification would better depend on the characteristic of each individual frame.

We will propose to use a method, called *Content Adaptive Embedding*, in which the *newValue* depends on the largest sample in that frame and the preset scaling factor α . α is used to tune the trade-off between the robustness and the transparency. If α is fixed then the *newValue* will adapt according to the largest sample of that frame.

To perform content adaptive embedding, a scale factor α is set for all frames. Then in a particular frame $B_I(\omega)$, we first find out the sample with the largest amplitude and the sample with the second largest amplitude. Take their indexes as *L_Index* and *S_Index* respectively. We then do the modification as below:

$$B_I(S_Index) = B_I(L_Index) * \alpha; \quad \alpha > 1; \quad (4.11)$$

In this case, we replace the second largest sample with the largest sample value multiplied by a parameter α . So $B_I(S_Index)$ is considered as the watermark signal. Its amplitude varies with the energy of the largest sample in that frame and thus less distortion would be induced.

If the $B'_I\{\omega\}$ is obtained from $B_I\{\omega\}$ by embedding a '1' information bit,

then the original largest point in $B_I\{\omega\}$ will become the second largest one in $B'_I\{\omega\}$. And the original second largest frequency point in $B_I\{\omega\}$ now becomes the largest in $B'_I\{\omega\}$. Suppose L_Index and S_Index are the indexes of the largest and the second largest sample in $B'_I\{\omega\}$, we know the relationship between these two samples by equation (4.12).

$$B'_I(L_Index) = B'_I(S_Index) * \alpha; \quad (4.12)$$

So we can set a threshold Th , which satisfies:

$$B'_I(\omega) < Th \leq B'_I(L_Index); \quad \omega = 1, \dots, L; \quad \omega \neq L_Index; \quad (4.13)$$

We claim '1' information bit has been embedded if we find $B'_I(L_Index)$ is equal to or greater than the threshold. We make this claim based on our design at the embedding end.

Another assumption we based on is: the scramble function can scramble the frame into a noise-like signal with flat spectrum, so $B'_I(L_Index) > B'_I(S_Index) * \alpha$ will never happen when '0' is embedded with this frame, recall if that with '0' embedded means making no modification to the frame. This assumption may be violated when the $r(n)$ can't scramble the frame to a noise-like one. The consequence of assumption violation will cause the embedded information bit '0' to be detected as '1'.

By experiments, we confirm that modifying the largest or the second largest sample will introduce less distortion to the audio file than modifying other samples within a frame. Modifying frequency sample according to properties of that frame will introduce less distortion and promise the robustness theoretically. Beside the algorithm itself, the only information need to be know in the detection stage is only the scaling factor α . We can set a fixed α for all the audio Works, or vary α according to audio file genre, or even set different α for each Work. Since α has relationship with robustness and transparency thus should be chosen carefully.

Frame Length L	1024	512	256
SNR(dB)	16.2	14.22	12.185
Capacity(bit)	430	860	1720

Table 4.3: Frame Length Versus Signal SNR and Capacity (The sample to be embedded is the sample with second largest sample; $\alpha = 2.6$);

4.3.3 Determination of Frame Length L

While doing segmentation in this system, a parameter needed to be determined is the length L of a frame.

This determination shall be application dependent. Because it is a trade off between the embedding capacity and the quality of the output watermarked Work. By using a longer frame, also a longer $r(n)$ is used, then the frame can be scrambled into a more noise-like signal. So the induced noise $D = newvalue - B_I(embedIndex)$ will also be less since the signal spectrum is more noise-like. In addition, by spreading to a wider frequency range the modification to each sample in the frame is relative smaller compared with a narrow range.

However the embedding capacity of the system, known as **data payload**, decreases with the frames getting longer, since only one bit information will be embedded within each frame. So there is a trade-off between the transparency and the data payload. This trade-off can be adjusted according to the application requirements. With frame length increases, smaller information bits can be embedded, but the total energy of noise introduced will reduce too. It is expected that the SNR of the watermarked Work will increase with the frames become longer, when other parameters, including *embedIndex* and scaling factor α , are fixed. From the experiment results shown in Table 4.3, we can find that SNR decreases and capacity increase with frame getting shorter.

4.4 Requirement For Transparency Improvement

We have tried to minimize the distortion introduced during the embedding process. Based on the metrics we used, a more efficient pseudorandom bits sequence $r(n)$ is used in stead of pseudorandom number sequence; The embedding is based on *content adaptive embedding* method. Other two parameters L and α are determined according experimental trials. Given $L=1024$, the system has a better trade off between transparency and capacity. Given $\alpha=2.6$, the system has a better trade-off between transparency and robustness.

Experiments have been done to a set of audio file segments². Each frame is embedded with information bit '1'; From the result in Table 4.4, we can see that the system can embed and detect the watermark information bits efficiently with the system settings above, the bit error rate (BER) is almost zero for all the file samples. However, the result Work can't fulfill the requirement of transparency. From the waveform in Figure 4.9 and Figure 4.10 we can see noise with large energy has been induced. And from Table, SNR is as low as $16dB$. The subjective listening test can find audile distortion in the watermark embedded Work.

Therefore the next problem we are going to solve is how to improve the transparency of the resulting watermarked Work, while maintaining the efficiency and robustness of the system. The methods we will use are so called *sample selection* with one frame and *frame selection* in the frames set. These two methods used for transparency improvement will be introduced and discussed in the Chapter 5. The robustness of the system combined spread spectrum technique with sample and frame selection will be verified in Chapter 6 by psychoacoustics model analysis.

²Nine of these audio file samples are provided Gordy et al. [41], can be download from http://www-mddsp.enel.ucalgary.ca/People/gordy/audio_watermarking.htm. The other six samples are selected from a CD collection named *The Power of Love*, copyright of these samples belongs to Warner Music Records co.,LTD.

	File1	File2	File3	File4	File5	File6	File7	File8
SNR(dB)	16.21	16.01	15.85	15.45	15.65	15.50	15.92	15.59
BER(%)	0	0	0	0	0	0	0	1.16
	File9	File10	File11	File12	File13	File14	File15	
SNR(dB)	15.81	15.75	16.15	15.80	15.68	15.64	16.06	
BER(%)	0	0.64	2.43	0.58	0	0	0.7	

Table 4.4: SNR of A Set of Audio File Samples.L=1024; $\alpha = 2.6$; All Frames are Embedded with Information Bit ‘1’; (These testing files can be found in audiofile1.pdf in the attached CD. To listen these sound files, please click the file names.)

Chapter 5

Sample and Frame Selection For Transparency Improvement

5.1 Introduction

Digital audio watermarking based on spread spectrum technique can embed and detect watermark information successfully. However, even after the optimal setting during the process, audible distortion still exists. These two issues have been discussed in Chapter 4. In this Chapter, we are going to find ways for reducing the distortion induced during the watermarking process so as to improve the transparency of the whole system.

One possible measure for improving the transparency is to select and modify only some of the frequency samples within a frame for watermarking. The operation of selecting a portion of the frequency samples within a frame for watermarking is called *Sample Selection*. The principles and procedure of sample selection and how to realize optimal selection will be introduced and elaborated in Section 5.2.

The number of information bits to be embedded is usually much less than the number of available frames, and the physical property of individual frame might be very different. Therefore it is possible to select those frames that will produce the least degradation when watermarks are being embedded to further improving the transparency . This operation is called *Frame Selection*.

The frequency points that have been selected for watermarking will form a new data segment, we call this frequency sample portion. It is obvious that the selected frequency points will have immediate effect on the watermarking result. The criteria of frame selection would better be based on these frequency sample portions directly. Thus, it is recommended that the sample selection should be done before frame selection. While improving the transparency, frame selection may have some drawbacks to the watermarking system and the outcome should be attend to.

After *sample selection* and *frame selection*, the candidate sample portions will be embedded with watermark information bits by the spread spectrum technique. The information-carry frequency sample portions are then used to reconstruct the corresponding time frame and the audio Work, with watermark embedded. The necessity, feasibility and compensation, the general process as well as frame selection for transparency improvement will be introduced and elaborated in Section 5.3.

The steps of retrieving the binary information bits at the detection end will also be described in Section 5.4.

5.2 Sample Selection

As discussed in Chapter 4 Section 4.4, transparency improvement is demanded for the usefulness of the proposed watermarking system.

In practice, it is not always appropriate to embed a watermark in all the coefficients of a Work's representation. If the audio signal is transformed into frequency domain, it can be seen that most of the energy of the signal is contained within a certain portion of the frequency points, usually in the lower frequency band.

In the watermarking system introduced in Chapter 4, the energy of the noise signal in each frame is adaptive and is directly proportional to the energy of that frame. To improve the transparency of the watermarked Work, we might use a portion of frequency points which is perceptual insignificant and contain

only a small part of energy of that frame. The promised portion is usually the higher frequency components. But a portion of perceptual insignificant frequency points might be totally discarded when while undergoing signal processing like lossy compression. This will make the watermark unreliable and lost its robustness property. Thus which portion and how large the portion to be selected that will produce transparent watermarked audio Work need to be considered? By taking into consideration of robustness and transparency requirements of the system, we will discuss how to make an appropriate choice of samples for watermarking.

5.2.1 General Sample Selection

In this section, referring to Figure 5.1, we shall explain the general procedure of selecting a sample portion from an original audio frame for watermarking. Given an audio Work $\{x(n)\}$, it will be segmented into frames along the time axis, with frame length L . A time frame set \mathbf{s} is obtained,

$$\mathbf{s} = \{s_1(n), \dots, s_i(n), \dots, s_N(n)\}; \quad n = 1, \dots, L; \quad (5.1)$$

where $N = \lfloor \text{Length}/L \rfloor$ and Length is supposed to be the total length of the audio Work. As shown in the flowchart, each frame in \mathbf{s} is next transformed into the frequency domain by using DCT (Discrete Cosine Transform), and a set of corresponding frequency frames \mathbf{S} is obtained, which can be expressed as,

$$\mathbf{S} = \{S_1(\omega), \dots, S_i(\omega), \dots, S_N(\omega)\}. \quad \text{where} \quad S_i(\omega) = DCT\{s_i(n)\}. \quad (5.2)$$

As discussed earlier on, to fulfill the requirement of transparency, it is suggested to select a certain portion of frequency samples to perform the watermarking task. How to select these frequency samples within one frequency frame? What frequency band should they cover (which is related to where the selection should begin)? How many samples should be selected? We shall first describe a general guideline for sample selection, and then attempt to answer to in Section 5.2.3.

In order to avoid resulting in extra data payload, we will select usable samples consecutively. Suppose *BeginSample* denotes the position of the starting

point and $SeLe$ denotes the length of samples to be selected. Let $U_i(\omega)$ be a portion of frequency point selected from $S_i(\omega)$, $U_i(\omega)$ can be expressed as,

$$U_i(\omega) = S_i(\omega + BeginSample - 1), \quad \text{where} \quad \omega = 1, \dots, SeLe \quad (5.3)$$

It is obvious that the above equation should satisfy the condition of $\omega + BeginSample - 1 \leq L$. The same process applies to all frames which gives rise to a sequence of segments, denoted by U , in the following form:

$$U = \{U_1(\omega), \dots, U_i(\omega), \dots, U_N(\omega)\}. \quad (5.4)$$

Each frame of $U_i(\omega)$ in U will be transformed into a time domain sequence, denoted by $u_i(n)$, using Inverse Discrete Cosine Transform (IDCT). A set u contains all $u_i(n)$ can be formed.

$$u = \{u_1(n), \dots, u_i(n), \dots, u_N(n)\}. \quad (5.5)$$

Subsequently $u_i(n)$ will be fed to the **Embedding Block** to include the information bit using the spread spectrum technique. The output of the **Embedding Block** is a set of $u'_i(n)$. These $u'_i(n)$ will undergo a series of inverse processes as shown in the **Frame Reconstruction Block** of the flowchart. A watermarked audio Work $\{x'(n)\}$ eventually is obtained after the **Reconstruction Block** as depicted in the flowchart.

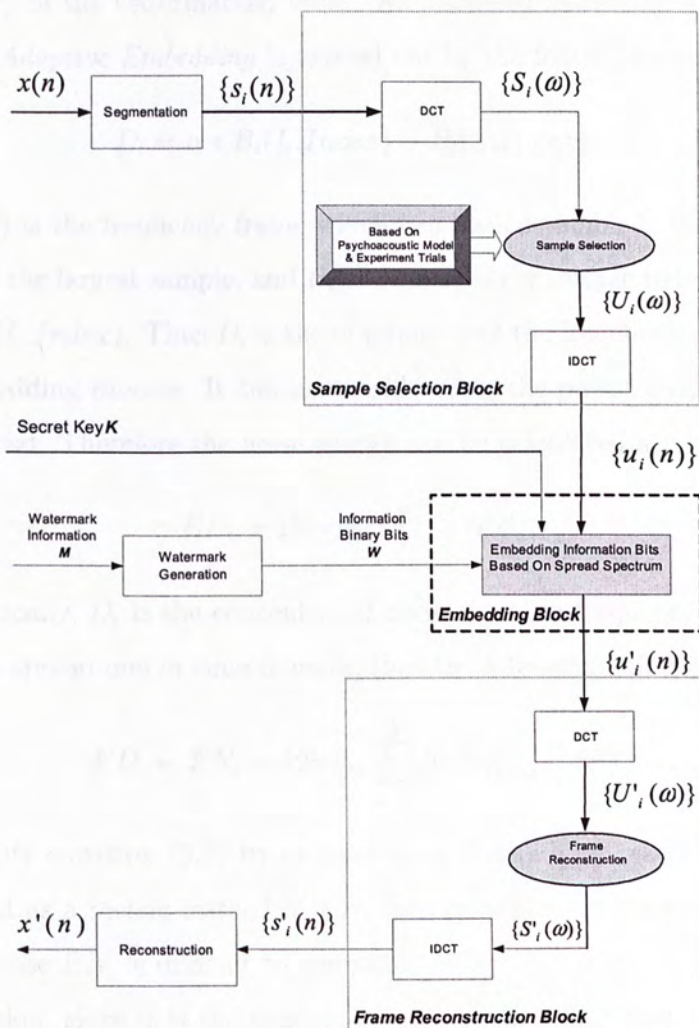


Figure 5.1: Block Diagram of Watermarking Embedding Based on Spread Spectrum Technique Combined With Sample Selection

5.2.2 Objective Evaluation Metrics

The noise signal induced in each frame, $n_i(n)$ and its energy level EN_i , the frame signal to noise ratio $fSNR_i$ as well as SNR for the whole audio file, defined in Chapter 4 Section 4.3.1, can still be used to evaluate the effective transparency of the watermarked Work. As discussed in Section 4.3.2 Chapter 4, *Content Adaptive Embedding* is carried out by the following operation,

$$D_i = \alpha * B_i(L_Index) - B_i(S_Index); \quad (5.6)$$

where $B_i(\omega)$ is the frequency frame which has been scrambled, $B_i(L_Index)$ is the value of the largest sample, and $B_i(S_Index)$ is the sample to be substituted with $\alpha * B_i(L_Index)$. Thus D_i is the of intensity of the frequency point subject to the embedding process. It can be considered as the power level of the noise signal induced. Therefore the noise energy can be calculated as:

$$ED_i = 10\log_{10}(D_i^2) \quad (dB); \quad (5.7)$$

and theoretically, D_i is the concentrated noise signal in frequency domain and $n_i(n)$ is the spread one in time domain, thus the following will hold:

$$ED_i = EN_i = 10\log_{10} \sum_{n=1}^L |n_i(n)|^2 \quad (dB); \quad (5.8)$$

We can verify equation (5.8) by experimental results later on. Thus ED_i can also be used as a metric instead of EN_i for evaluating the performance of the method in case EN_i is difficult to compute.

In addition, since it is the selected sample portion that that will effect the system's transparency performance, we will define a metric based on the selected portion directly. A metric called $pSNR$, which denotes the signal to noise ratio of the selected sample portion is defined as:

$$pSNR_i = 10 * \log_{10} \frac{\sum_{n=1}^{SeLe} |u_i(n)|^2}{\sum_{n=1}^{SeLe} |u'_i(n) - u_i(n)|^2} \quad (5.9)$$

where $u_i(n)$ is the original sample portion in time frame and $u'_i(n)$ is the watermarked embedded one. Suppose EP_i is the energy of the selected sample

portion, and recall that the noise energy within the selected sample portion is the same as the noise in the whole frame, so $pSNR$ can be expressed as:

$$pSNR_i = EP_i - EN_i; \quad (dB) \quad (5.10)$$

All these indicators will be used to measure the transparency performance of the system with different parameters setting. The principle of watermarking is to embed watermark information while introducing the smaller noise signal and achieve the highest SNR .

5.2.3 Sample Selection For Transparency Improvement

The lower frequency components are usually the significant portion of the signal and contain major energy of the whole frame. The higher frequency points are generally insignificant and account for only a small energy of the signal. Our goal is to embed watermarks effectively while keeping distortion to minimum. To achieve this objective, we need to pick out the samples at the high frequency band.

However, we have to consider the robustness of the watermark information. Therefore in order to fulfill the requirement of transparency and the specific signal to noise ratio demanded by the application, the watermark signal should be embedded into the significant components of the signal.

To select a portion of frequency samples from the whole range, the parameters need to be determined include *BeginSample* and *SeLe*, which have been defined in Section 5.2.1. *BeginSample* can be interpreted as the frequency point at which the selection begins:

$$BeginFrequency = (BeginSample/L) * (Fs/2); \quad (5.11)$$

where L is the frame length and Fs is the sampling frequency of the original audio Work.

Both *BeginSample* and *SeLe* may affect the system performance. Thus we shall first fix the frame length L and the length of the selected sample portion *SeLe* and discuss the effect of *BeginSample* by experiment trials. Followed by

varying $SeLe$ and we shall also examine the effect of it on the system performance.

Beginning Frequency Point (*BeginSample*) for Sample Selection

We randomly pick an audio file sample for experimental trial. The file sample is segmented with $L = 1024$; the length of the selected portion is set to be $SeLe = 512$; the scale factor is $\alpha = 2.6$. The *BeginSample* will vary according to the values appeared in the first row of Table 5.1. The beginning frequencies of each *BeginSample* corresponds to are listed in the second row of the table. The third row is the total **SNR** of the watermarked file sample. From the result in table we can see that with a larger *BeginSample*, that is the selected portion is located at higher frequency band, the **SNR** of the audio file will increase, which can be considered as a transparency improvement.

BeginSample	1	28	46	186	372	511
BeginFrequency(Hz)	20	600	1000	4000	8000	11000
SNR(dB)	13.89	25.18	27.09	29.35	33.36	37.09

Table 5.1: Sample Selection Result On a Audio Work With Different *BeginSample*; The parameter are set as: $L=1024$; $SeLe=512$; $\alpha = 2.6$

To further find out what is the relationship between *BeginSample* and the defined evaluation metrics, experiments as described above for one audio sample will be done to a serial of 15 audio file segments¹. These audio segments for experiment cover different genres of music, from Blues, Classical to Pop music, totally 330 seconds long. The sampling frequency is 44.1kHz. To each of these file segment, segmentation with $L=1024$ is first done.

¹Nine of these audio file samples are provided Gordy et al. [41], can be download from [http : //www – mddsp.enel.ucalgary.ca/People/gordy/audio_watermarking.htm](http://www-mddsp.enel.ucalgary.ca/People/gordy/audio_watermarking.htm) The other six samples are selected from a CD collection named *The Power of Love*, copyright of these samples belongs to Warner Music Records co.,LTD.

Thus totally about 14000 frames are obtained. A sample portion with length of $SeLe = 512$ will be selected from each of the frames. With $\alpha = 2.6$ as the scaling factor and PRBS $r(n)$ as the scrambling function, watermark information bits are embedded into the sample portions by method based on the *Spread Spectrum* technique combined with *Content Adaptive Embedding*. All frames are embedded with information bit '1';

We will analyze the experimental results when *BeginSample* varies according to the pattern $\{1, 28, 46, 186, 372, 512\}$ as shown in the first row of Table 5.1 so as to find out the relationship of *BeginSample* and the defined evaluation metrics. Statistical methods will be applied.

The statistical analysis relationship between the *BeginSample* and Energy of D , ED is displayed in Figure 5.2

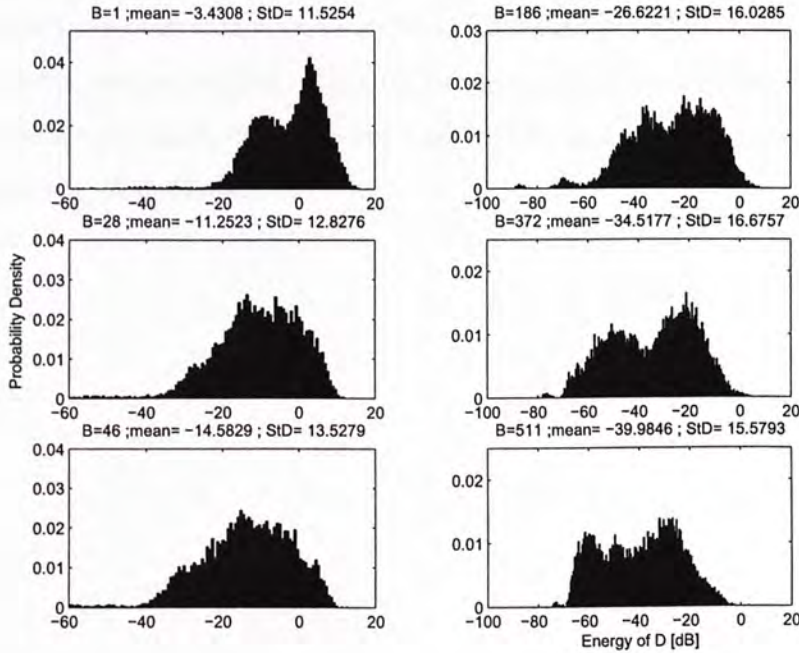


Figure 5.2: Probability Density Function of ED_i , the energy of D_i . B denotes the *BeginSample*; $L=1024$; $\alpha = 2.6$; $SeLe=512$

Each sub-figure in Figure 5.2 is the *probability density function(PDF)* of ED with different *BeginSample*, obtained by histogram analysis of D . The X axis is the energy level of D , which is $ED = 10 * \log_{10}(D^2)$, in dB unit. The whole range shown is divided into 150 sub-sections for analysis. Y axis is the percentage of frames with a noise energy level represented by the center of that sub-section. The mean and standard deviation of each *PDF* are calculated and shown. (This histogram analysis is applied to analyze the other metrics later on). Reminder that the mean and standard deviation here and later on only have statistical meaning but carry no physical interpretation.

By varying the *BeginSample* while keeping other parameters unchanged, we got different *PDFs* for each *BeginSample* values. The *BeginSamples* are also shown in the title of each figure. From the *PDFs* shown in the figure, we can see that the overall noise energy level decreases with the *BeginSample* varies from the lowest frequency to higher one. The mean values of the *PDFs* of the noise signal have been summarized in Table 5.2 third row. Thus from the figure and the table, we can conclude that the energy noise signal induced decreases with *BeginSample* increases. So it is possible to obtain different noise level by using different *BeginSample*.

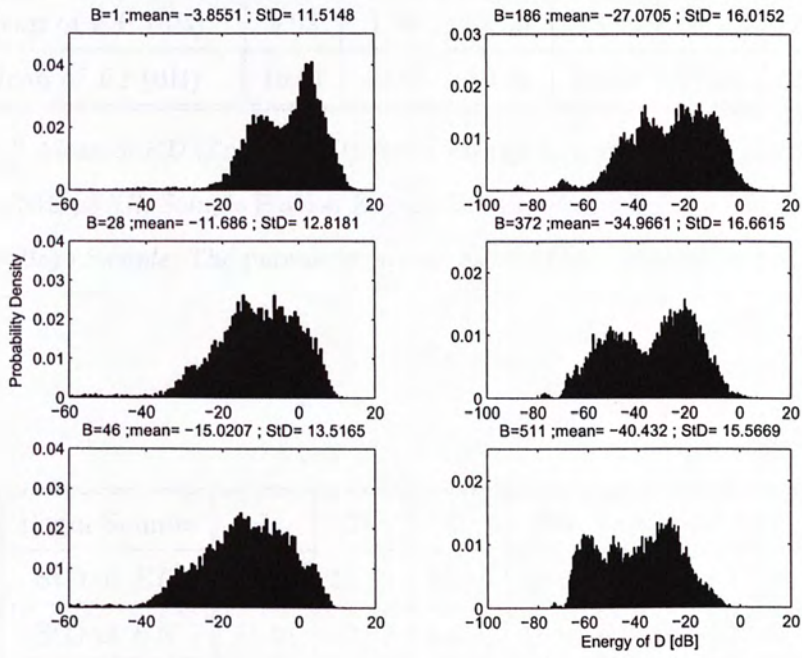


Figure 5.3: Probability Density Function of EN_i , the Energy of Noise $n_i(n)$. B denotes the *BeginSample*; $L=1024$; $\alpha = 2.6$; $SeLe=512$

Begin Sample	1	28	46	186	372	511
Begin Frequency(Hz)	20	600	1000	4000	8000	11000
Mean of ED (dB)	-3.43	-11.25	-14.58	-26.62	-34.52	-39.98
Mean of EN (dB)	-3.86	-11.69	-15.02	-27.07	-34.97	-40.43
Mean of $fSNR$ (dB)	13.87	21.70	25.03	37.08	44.98	50.44
Mean of $pSNR$ (dB)	13.78	13.68	13.63	13.52	13.53	13.53
Mean of EP (dB)	9.92	1.99	-1.39	-13.55	-21.43	-26.90
Mean of EF (dB)	10.01	10.01	10.01	10.01	10.01	10.01

Table 5.2: Mean of ED (Energy of D); Noise Energy EN ; Frame SNR $fSNR$; Sample Portion SNR, $pSNR$; Sample Portion Energy EP and Frame Energy EF , Vary With Different $BeginSample$; The parameter are set as: $L=1024$; $SeLe=512$; $\alpha = 2.6$

Begin Sample	1	28	46	186	372	511
StD of ED	11.53	12.83	13.53	16.03	16.68	15.58
StD of EN	11.51	12.82	13.52	16.02	16.665	15.57
StD of $fSNR$	1.50	5.49	6.21	9.80	11.37	11.06
StD of $pSNR$	1.15	1.18	1.18	1.22	1.32	1.28
StD of EP	11.45	12.75	13.43	15.89	16.52	15.45
StD of EF	10.94	10.94	10.94	10.94	10.94	10.94

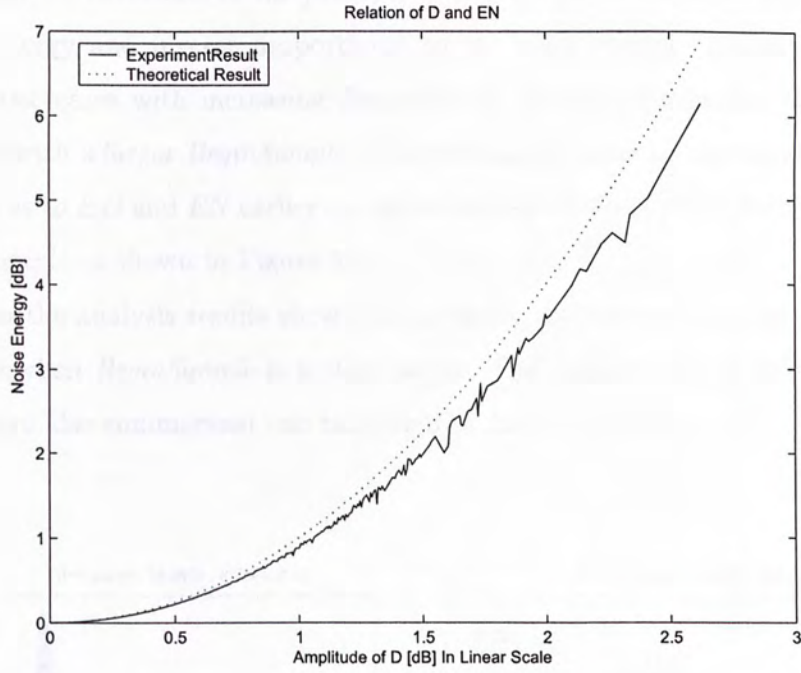
Table 5.3: Standard Deviation(StD) of ED (Energy of D); Noise Energy EN ; Frame SNR $fSNR$; Sample Portion SNR, $pSNR$; Sample Portion Energy EP and Frame Energy EF , Vary With Different $BeginSample$; The parameter are set as: $L=1024$; $SeLe=512$; $\alpha = 2.6$

ED is the noise energy calculated from theoretical prediction. The practical noise energy EN s calculated from the difference between the original frame and the watermark embedded frame, vary with different *BeginSamples* are shown in Figure 5.3 and listed in Table 5.2. From these results we can have observation and conclusion same as the analysis on ED . Further more, we note that, ED and EN are almost equal to each other, as predicted by theory in (5.8), which means the energy of the concentrated signal D_i in frequency domain is equal to the EN_i of the spread signal in time domain. Then we will verify this proposition as follows.

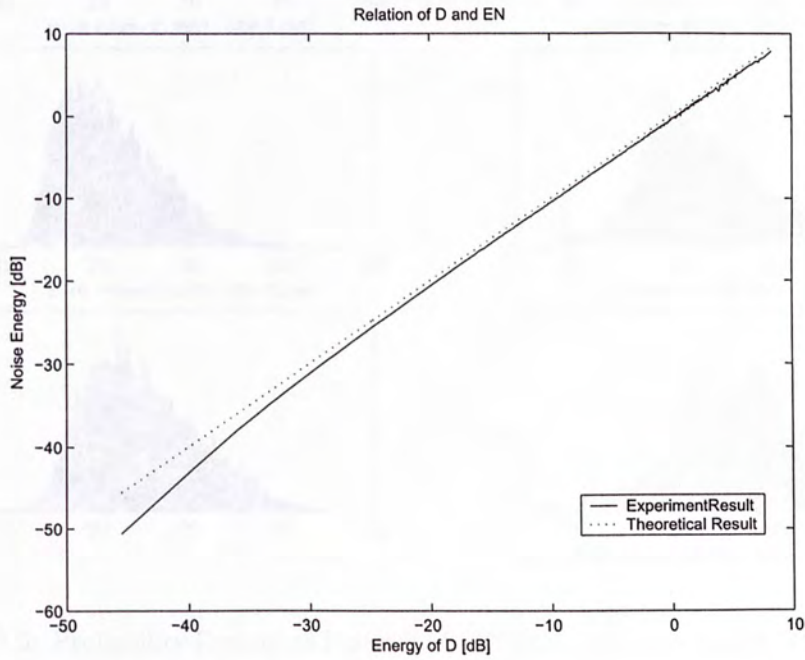
As shown in Figure 5.4, in sub-figure(a), the dotted line is the theoretical relationship of D and ED , which satisfies:

$$ED_i = D_i^2; \quad (5.12)$$

The solid line with fluctuation is obtained by analyzing the experimental data in the following steps: First divided the range of D into sub-sections, then the noise energy of those frames within each sub-section are averaged and the overall result is used to denote the noise energy of that section. From the figure we can tell that the experimental analysis result curve is close enough to the theoretical one. Sub-figure (b) is obtained by taking the log scale of both the X axis and Y axis. From (b) we can also confirm that the noise energy ED calculated from D can be used to represent the noise energy EN accurately.



(a) Relationship Between D , ED and EN , in Linear Scale



(b) Relationship Between ED and EN , in Log Scale

Figure 5.4: Theoretical and Experimental Analysis Relationship Between ED and EN . $L=1024$; $\alpha = 2.6$; $BeginSample = 186$ $SeLe=512$

Recall the definition of the $fSNR$, it is directly proportional to the original signal energy and inverse proportional to the noise energy. Since the noise energy decreases with increasing $BeginSample$, we expect that the $fSNR$ will increase with a larger $BeginSample$. By applying the same histogram statistical method as to ED and EN earlier on, we obtain the $PDFs$ of $fSNR$ with different $BeginSample$ as shown in Figure 5.5.

From the analysis results shown in the figure, we can see that the $fSNR$ do increase when $BeginSample$ is getting larger. The mean values of the $PDFs$ of $fSNRs$ are also summarized and tabulated in Table 5.2 fifth row.

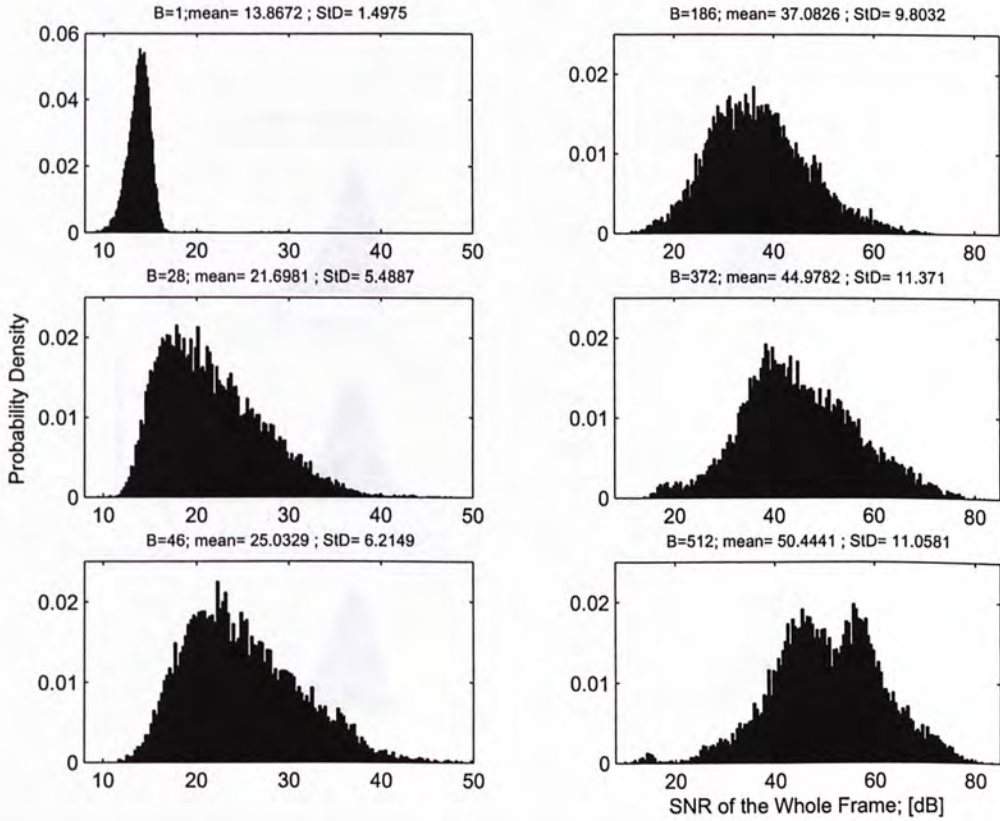


Figure 5.5: Probability Density of Frame SNR, $fSNR$ $L=1024$; $\alpha = 2.6$; $SeLe=512$

The next metric to be analyzed is $pSNR$. The statistical results are displayed in Figure 5.6. From the figures we can see that, with different $BeginSample$, the mean values as well as the standard deviations of $pSNRs$ do not change significantly as compared with the variation of ED and $fSNR$ caused by varying $BeginSample$. The mean values of $pSNRs$ with different $BeginSamples$ are tabulated in Table 5.2 row six, which are all around $13.5dB$. From these results we can draw a conclusion: different from $fSNR$ and D , given fixed $SeLe$, $pSNR$ do not necessary vary with the $BeginSample$ directly. Thus which factor will determine $pSNR$? Would $SeLe$ be the one? The answer is probably 'Yes'. But it will be analyzed in the later experiments.

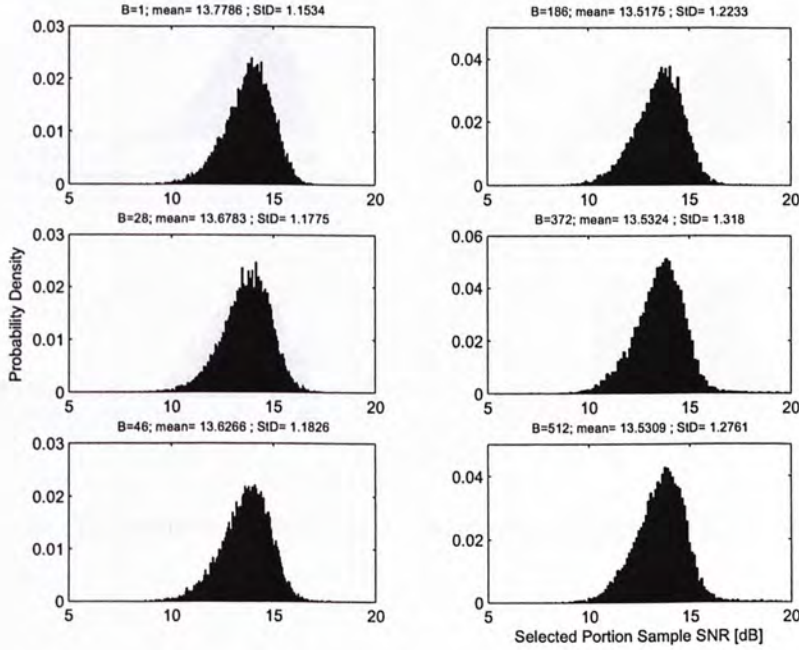


Figure 5.6: Probability Density of Portion SNR, $pSNR$; $L=1024$; $\alpha = 2.6$; $SeLe=512$;

From the analytical result shown in Figure 5.7, we can see that the energy of the selected sample portion decreases monotonically when the *BeginSample* is getting larger. The mean values of portion energy with different *BeginSample* are summarized in Table 5.2 seventh row.

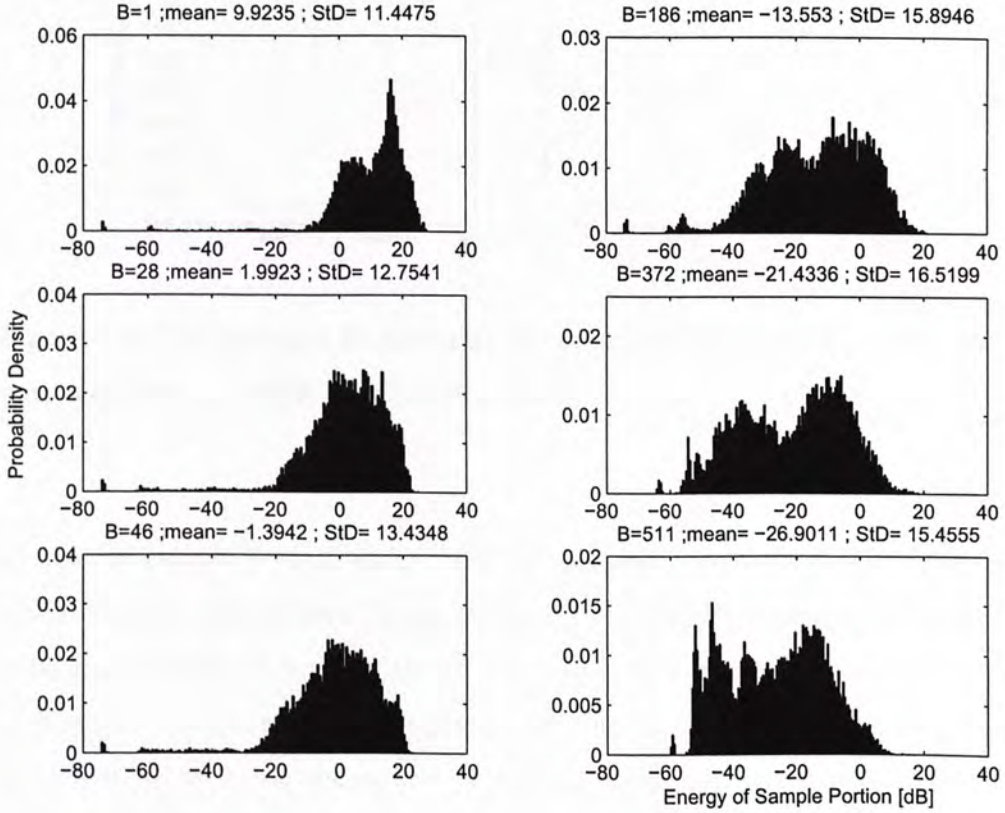


Figure 5.7: Probability Density of Portion Energy,EP; $L=1024$; $\alpha = 2.6$; $SeLe=512$;

In addition to using the table to tabulate and histogram figures to depict the experimental results, the tendency relationship between *BeginSample* and the evaluation metrics can be roughly summarized in Figure 5.8. From the curves shown in the figure, we can conclude that the mean values of *ED*, *EN* as well

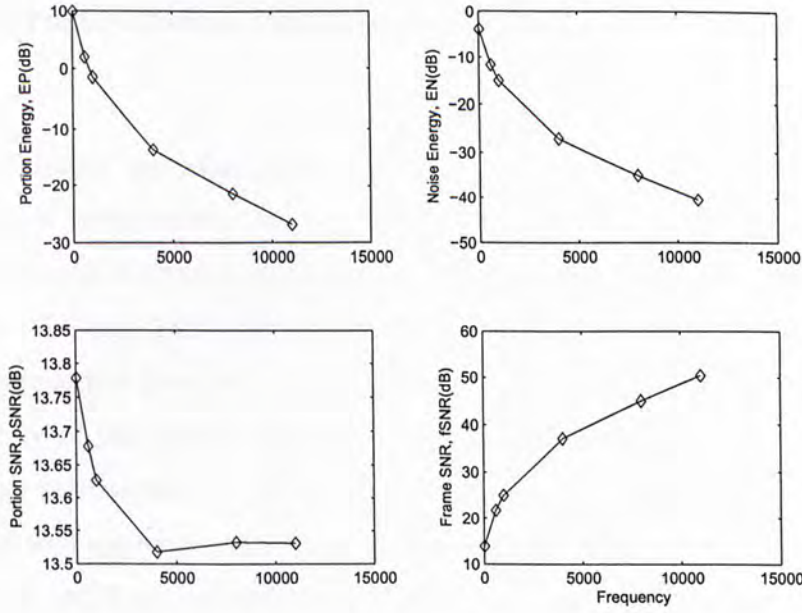


Figure 5.8: The Tendency Relationship Between the *BeginSample* and the evaluation metrics; $L=1024$; $\alpha = 2.6$; *BeginSample* = 186; *SeLe*=512;

as that of sample portion energy *EP* all decrease, while the *fSNR* increases when *BeginSample* becomes larger. Although *pSNR* has a tendency to decrease with *BeginSample* it varies within a range of $0.2dB$, which can be considered as constant compared with the change of other measures. This observation is consistent with the theoretical prediction earlier on, which says selecting frequency portion from the higher frequency band will introduce less distortion to the Work and can achieve high *fSNR* and *SNR*. But bearing the robustness in mind, we see that for transparency purpose it is not necessary to select samples from the highest portion, since experimental results show that selecting 512 samples from 4KHz, and embed each frame with information bit '1'. the mean of *fSNR* of the frames is high enough, at about $37dB$, and subjective testing confirms the embedding watermark signal is transparent to human ears. In addition, for real application, the information bits to be embedded consists of '0' and '1', so that the overall *SNR* will increase further.

Sample Portion Length *SeLe* For Sample Selection

After analyzing the relationship between *BeginSample* and the performance in terms of transparency, the next step is to see how does ***SeLe*** affect the performance of the watermarking system. We have seen from the results shown in Figure 5.6 and Table 5.2 that with *SeLe* keep constant at 512 and varies *BeginSample* will have very limited effect on the *pSNR*. No matter how much the energy of the sample portion is, once the *SeLe* is fixed then the *pSNR* is almost predetermined. From this observation, we can summarize that: in our proposed approach, *pSNR* of a certain frequency point portion depends on *SeLe* only, but has no close relationship with the energy of that portion. Given (5.10) $pSNR_i = EP_i - EN_i(dB)$, We conclude that: selecting a frequency point portion with same *SeLe*, the higher energy of the sample portion, the larger the energy level of noise signal will be introduced. Then how will the performance of the system affected by *SeLe*?

We speculate that *pSNR* is directly determined by *SeLe*. To verify this proposition, a set of experiments have been done. First the original file is segmented with different *L* each time, and the whole frame is used for watermarking, no sample selection process, which implies $SeLe = L$ and therefore $fSNR = pSNR$. From the experimental result shown in Figure 5.9 and in Table 5.4, it is obvious that *pSNR* (equal to *fSNR* with $SeLe = L$) **DOES** increase with a larger *SeLe*. Therefore we conclude that *SeLe* affects *pSNR*, the larger *SeLe*, the higher *pSNR*.

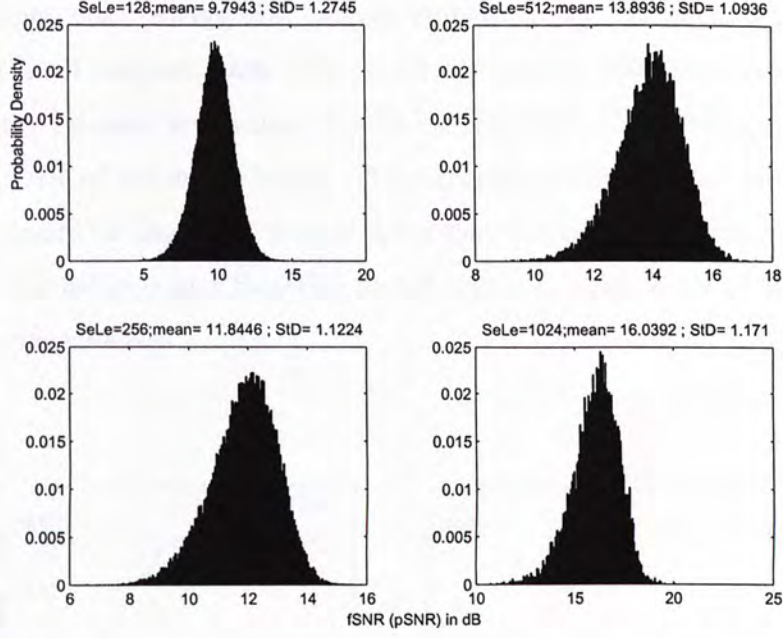


Figure 5.9: Statistical Relationship between Frame Length L and Frame SNR $fSNR$; $SeLe = L$; $fSNR = pSNR$

$SeLe=L$	128	256	512	1024
ED (dB)	-8.24(11.28)	-7.44(11.10)	-6.57(11.06)	-5.78(11.08)
EN (dB)	-9.39(11.25)	-8.12(11.09)	-6.98(11.05)	-6.03(11.06)
$fSNR$ (dB)	9.79(1.27)	11.84(1.12)	13.89(1.09)	16.04(1.17)
EF (dB)	0.41(11.20)	3.73(11.06)	6.91(10.97)	10.01(10.94)

Table 5.4: Relationship Between Frame Length L and Evaluation Metrics, which are in Statistical Mean Value and with Standard Deviation in the Bracket. All Samples are Used For Watermarking, $SeLe = L$; $pSNR = fSNR$; $EP = EF$;

The relationships between $SeLe$ and other evaluation metrics have been tabulated in Table 5.4 and depicted in Figure 5.10. From the analysis result, we note that noise energy EN , sample portion energy EP as well as frame SNR $fSNR$, all increase when $SeLe$ is getting larger. But the percentage of noise energy increase is less than that of EP . Therefore the $fSNR$ increases, so does the SNR of the whole Work. This experimental result also confirm our assertion made in Chapter 4 Section 4.3.3 that the longer the frame, the less noise energy induced and therefore higher signal to noise ratio of the whole Work can be achieved.

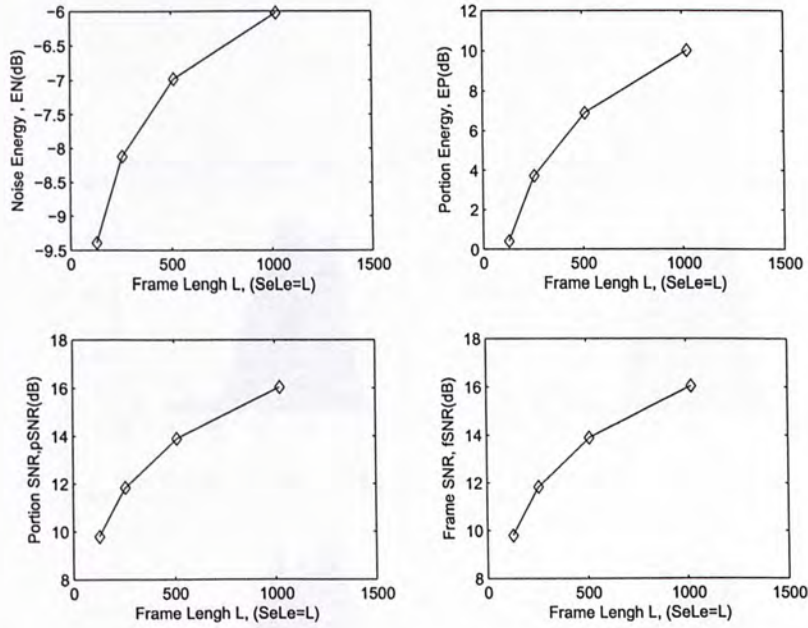


Figure 5.10: Tendency Relation Between $SeLe$ and Evaluation Metrics; $SeLe = L$; $fSNR = pSNR$

The following experiment has been conducted in which we segment the audio file with $L = 1024$ and select sample portion with different length $SeLe$, but begin from the same frequency point, i.e. $BeginSample$ is fixed. The statistical result of the relationship between noise energy ED and different $SeLe$ is shown in Figure 5.11. How other evaluation metrics vary with $SeLe$ are shown in Table 5.5. For better interpretation these tendency relationships are depicted in Figure 5.12. From all of the observations and analyses, we note that when $SeLe$ is getting larger, the energy of sample portion, EP as well as the SNR of sample portion, $pSNR$ will increase monotonically with $SeLe$. The energy of the noise signal, ED decreases while $fSNR$ increases when $SeLe$ is getting larger.

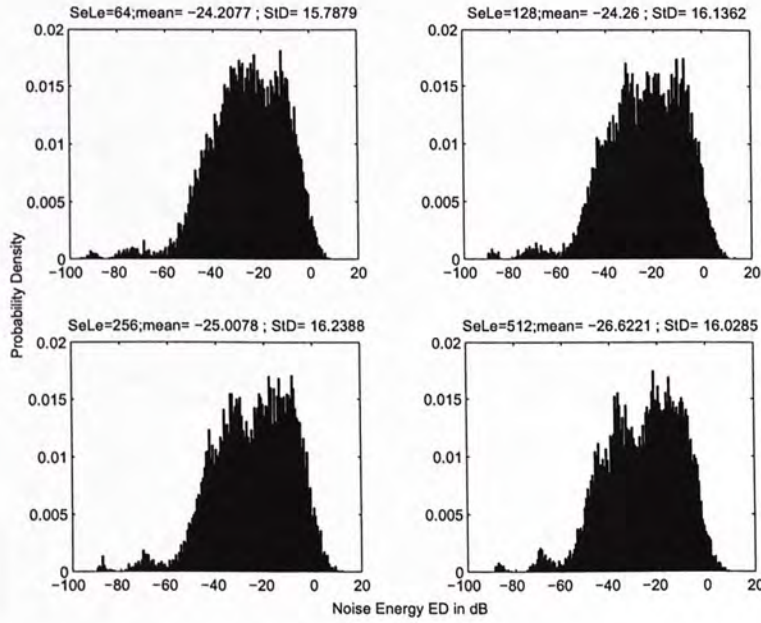


Figure 5.11: Statistical Relationship of $SeLe$ and EN ; With $L = 1024$; $BeginFrquency = 4kHz$; $BeginSample = 186$

<i>SeLe</i>	64	128	256	512
<i>ED</i> (dB)	-24.21(15.79)	-24.26(16.14)	-25.01(16.24)	-26.62(16.03)
<i>EN</i> (dB)	-26.10(15.72)	-25.48(16.09)	-25.76(16.21)	-27.07(16.02)
<i>EP</i> (dB)	-18.19(15.61)	-15.96(15.96)	-14.32(16.08)	-13.55(15.89)
<i>pSNR</i> (dB)	7.92(1.45)	9.52(1.37)	11.44(1.30)	13.52(1.22)
<i>fSNR</i> (dB)	36.12(9.57)	35.49(9.82)	35.77(9.90)	37.08(9.80)
<i>EF</i> (dB)	10.01(10.94)	10.01(10.94)	10.01(10.94)	10.01(10.94)

Table 5.5: Relationship Between *SeLe* and Other Metrics, Which are in Mean Vale with Standard Deviation in the Bracket. The *BeginFrequency* fixed to 4KHz, in term of *BeginSample*=186; $L = 1024$; $\alpha = 2.6$

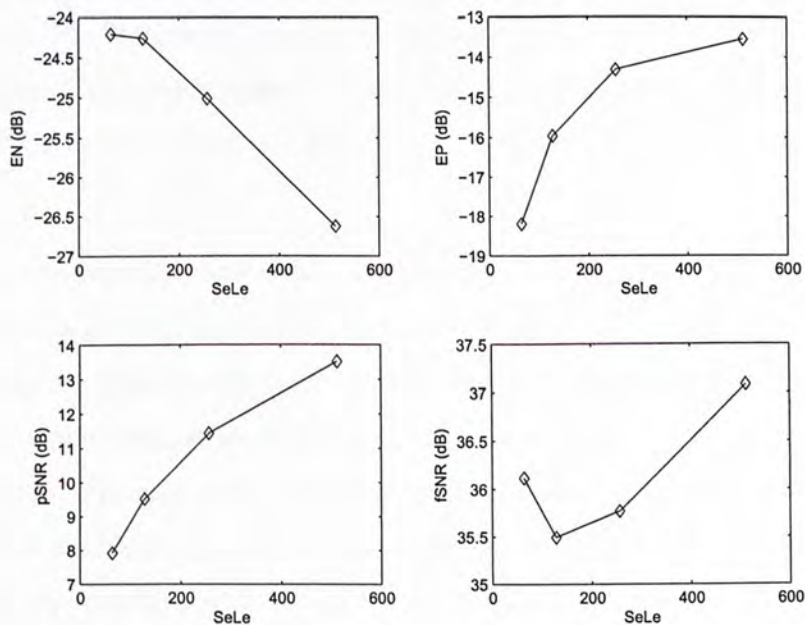


Figure 5.12: Tendency Relationship of *SeLe* with *ED*, *EP*, *pSNR* and *fSNR*; With $L = 1024$; *BeginFrquency* = 4kHz; *BeginSample* = 186

To further investigate the relationship between $SeLe$ and the others evaluation metrics while L is fixed, we have carried out another experiment. In this test, $L = 1024$, $\alpha = 2.6$, and $SeLe$ varies as the pattern of $\{64, 128, 256, 512\}$, but each time, the selected sample portion will be centered at frequency 8kHz. The experimental results are summarized in Table 5.6 and the tendency relationships are depicted in Figure 5.13.

$SeLe$	64	128	256	512
BeginSample	247	215	151	23
ED (dB)	-27.79(17.32)	-26.19(16.77)	-22.53(15.76)	-9.91(12.35)
EN (dB)	-29.68(17.24)	-27.42(16.74)	-23.28(15.74)	-10.35(12.34)
$pSNR$ (dB)	7.94(1.49)	9.52(1.39)	11.43(1.29)	13.66(1.19)
$fSNR$ (dB)	39.69(11.20)	37.43(10.56)	33.29(9.27)	20.36(4.91)
EP (dB)	-21.74(17.10)	-17.89(16.63)	-11.84(15.62)	3.32(12.28)
EF (dB)	10.01(10.94)	10.01(10.94)	10.01(10.94)	10.01(10.94)

Table 5.6: The Statistical Relationship Between $SeLe$ and Other Metrics, Which are in Mean Vale with Standard Deviation in the Bracket. The $CenterFrequency$ fixed at 8KHz; $L = 1024$; $\alpha = 2.6$;

From the experimental results, we note that both $pSNR$ and EP still increase when $SeLe$ is getting larger, this observation is consistent with the previous two experiments. However the noise energy EN will increase, while $fSNR$ will decrease when $SeLe$ is getting larger. This is inconsistent with the previous observation. Through these experiments, we see that both EN and $fSNR$ may increase or decrease when $SeLe$ is getting larger. Therefore, to fulfill the transparency requirement, $SeLe$ can not be too long nor too short.

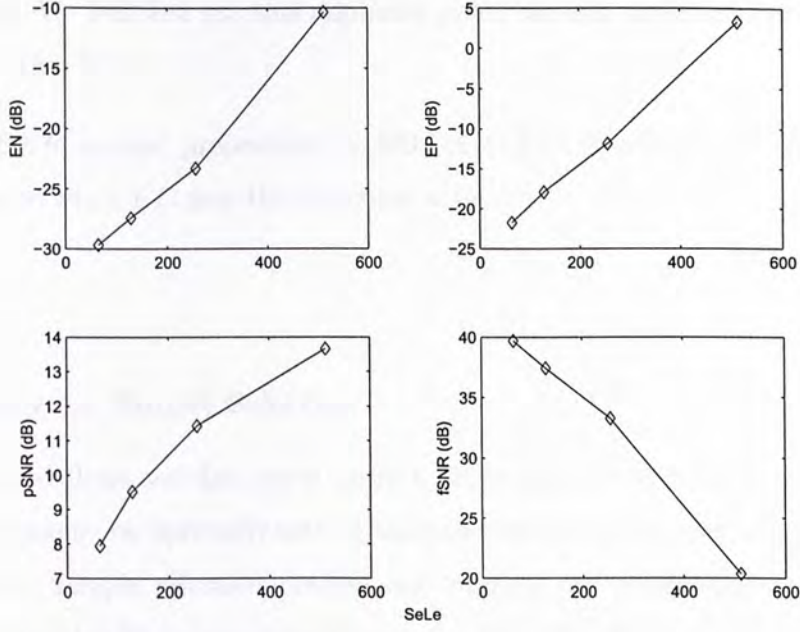


Figure 5.13: The Tendency Relationship Between $SeLe$ and Other Metrics, Which are in Mean Value with Standard Deviation in the Bracket. The $CenterFrequency$ fixed at 8KHz; $L = 1024$; $\alpha = 2.6$;

We have introduced the effect on the transparency property caused by $SeLe$. Combined with the observations in all experiments we can draw the following conclusions about how the performance of the system are affected by $SeLe$:

- Both EP and $pSNR$ of selected sample portion are directly determined by $SeLe$, the longer $SeLe$, the higher $pSNR$ and EP . However $pSNR$ has no obviously relationship with EP .
- ED as well as EN are affected by $SeLe$; the larger the $SeLe$ is, the larger the $r(n)$ is, the cover signal may become more noise-like so the higher $pSNR$ maybe. But with $SeLe$ increases, the total energy of the sample portion increases EP too. Given (5.10) $pSNR_i = EP_i - EN_i(dB)$, thus $EN_i = pSNR_i - EP_i(dB)$. Therefore ED may not decrease nor increase with $SeLe$ getting larger, it is determined by the relationship between the increasing speed of $pSNR$ and EP . There must exits a minimum point for

ED. We will find out this minimum point through theoretical analysis in Section 5.2.4

- *fSNR* inverse proportion to *ED*, thus *fSNR* will got its maximum value when *ED* gets the minimum value.

Summary for Sample Selection

After the analysis and discussion on both *BeginSample* and *SeLe*. We can draw the conclusion: by optimally setting the two selecting parameters *BeginSample* and *SeLe*, *sample selection* process can improve the transparency of watermark embedded Work significantly and maintain the efficiency and robustness of the watermarking system, as shown in Table 5.7 for *BeginFrequency=4kHz, SeLe=256* and in Table 5.8 for *BeginFrequency=4kHz, SeLe=512*. Compared with the result in Table 4.4 Chapter 4, we can see that in both *SeLe=256* and *SeLe=512* cases, the transparency performance have been highly improved after the *sample selection* process. The cost *sample selection* pays is that the bit error rate of the system has slightly increased and the loss of the robustness of watermark signal in certain level. But the error bits can be auto-corrected by EEC and the robustness of the watermark will be verified in Chapter 6.

Compared with other methods [17, 18] by using filter to make the watermark signal imperceptible and transparent, our approach of using sample selection is rather simple yet efficient.

	File1	File2	File3	File4	File5	File6	File7	File8
SNR(dB)	28.52	26.83	46.75	32.07	31.59	25.15	40.66	28.80
BER(%)	0.23	0.47	0	0	0	0	0	4.88
	File9	File10	File11	File12	File13	File14	File15	
SNR(dB)	20.65	42.73	34.95	25.40	24.95	30.37	28.33	
BER(%)	0	0.70	10.38	0.64	0	0	0.81	

Table 5.7: *SNRs* and *BERs* of A Set of Audio File Samples. $L=1024$; $\alpha = 2.6$; $SeLe = 256$; $B = 186$ (BeginFrequency=4kHz). All Frames are Embedded with Information Bit '1'; (These testing files can be found in audiofile1.pdf in the attached CD. To listen these sound files, please click the file names.)

	File1	File2	File3	File4	File5	File6	File7	File8
SNR(dB)	29.35	28.04	48.94	34.31	32.86	26.82	41.76	30.71
BER(%)	0	0.23	0	0	0	0	0	6.74
	File9	File10	File11	File12	File13	File14	File15	
SNR(dB)	22.37	44.35	36.50	27.14	26.30	31.08	30.12	
BER(%)	0	0.70	9.45	0.64	0	0	0.75	

Table 5.8: *SNRs* and *BERs* of A Set of Audio File Samples. $L = 1024$; $\alpha = 2.6$; $SeLe = 512$; $B = 186$ (BeginFrequency=4kHz). All Frames are Embedded with Information Bit '1'; (These testing files can be found in audiofile1.pdf in the attached CD. To listen these sound files, please click the file names.)

5.2.4 Theoretical Analysis of Sample Selection

$B_i(\omega)$ is the frequency representation of the scrambled sample portion $b_i(n)$. Suppose the length of the sample portion is $SeLe$, the total energy of the portion is EP_i

$$\begin{aligned} EP_i &= 10 * \log_{10} \sum_{n=1}^{SeLe} |b_i(n)|^2 \\ &= 10 * \log_{10} \sum_{\omega=1}^{SeLe} |B_i(\omega)|^2 \end{aligned} \quad (5.13)$$

To facilitate the analysis, we assume the ideal case that the random sequence can really randomize the original sample portion into noise-like signal with flat spectrum, in this scenario, each sample in the frequency domain is equal to the others in magnitude, the predefined “largest” sample and “second” largest sample are also equal to each others, thus:

$$B_i(L_Index) = B_i(S_Index) = \dots = B_i(\omega); \quad \omega = 1, \dots, SeLe \quad (5.14)$$

Therefore, the magnitude of each sample in $B_i(\omega)$ can be expressed as:

$$EB_i(\omega) = 10 * \log_{10} |B_i(\omega)|^2 \quad \omega = 1, \dots, SeLe \quad (5.15)$$

given (5.13) and (5.14)

$$\begin{aligned} EB_i(\omega) &= 10 * \log_{10} \frac{10^{EP_i/10}}{SeLe} \\ &= EP_i - 10 * \log_{10}(SeLe) \end{aligned} \quad (5.16)$$

Thus the spectrum of the noise signal D can be derived as follows:

$$\begin{aligned} D_i &= \alpha * B_i(L_Index) - B_i(S_Index) \\ &= (\alpha - 1) * B_i(L_Index) \quad \text{given (5.14);} \\ &= (\alpha - 1) * B_i(\omega) \end{aligned} \quad (5.17)$$

Then we can derive the energy of the noise signal by:

$$\begin{aligned} ED_i &= 10 * \log_{10} |D_i|^2 \\ &= 10 * \log_{10} |(\alpha - 1) * B_i(\omega)|^2 \\ &= 10 * \log_{10} |\alpha - 1|^2 + 10 * \log_{10} |B_i(\omega)|^2 \\ &= 10 * \log_{10} |\alpha - 1|^2 + EP_i - 10 * \log_{10}(SeLe) \end{aligned}$$

From (5.18) we can see that the energy ED_i of the induced noise is determined by α , the energy of sample portion EP_i and the length $SeLe$ of that particular portion. Once α is fixed, ED_i will be determined by the relationship between EP_i and $SeLe$. To simplify the discussion, suppose the $SeLe$ will increase from 2^m to 2^{m+1} , then in order to keep ED_i constant, EP_i should also increase by about $10 * \log_{10} 2 = 3dB$.

Take two of our experiments as examples, in both tests $SeLe$ vary in the pattern as: $\{64, 128, 256, 512\}$. In one of the experiment, the beginning of the sample selection is fixed at 4kHz. In this case, from the experimental result shown in Figure 5.11 and Table 5.5, we can see that the energy of longer sample portion is greater than the one of the previous shorter sample portion by equal to or below 3dB, then the energy of the noise decreases. In another cases, as shown in Figure 5.13 and Table 5.6, when the energy of the longer sample portion is greater than that of the shorter sample portion by above 3dB, then the energy of the noise increases.

Generally speaking, if \overline{SeLe}_1 is the length of the sample portion, the corresponding energy is \overline{EP}_1 , after changing the length of the sample portion, they become \overline{SeLe}_2 and \overline{EP}_2 . Then the energy of the noise induced will increase when:

$$\overline{EP}_2 - \overline{EP}_1 > 10 * \log_{10} \left(\frac{\overline{SeLe}_2}{\overline{SeLe}_1} \right) \quad (5.18)$$

and the energy of the noise will decrease when:

$$\overline{EP}_2 - \overline{EP}_1 < 10 * \log_{10} \left(\frac{\overline{SeLe}_2}{\overline{SeLe}_1} \right) \quad (5.19)$$

ED gets its minimum value when:

$$EP_i = 10 * \log_{10} \sum_{\omega=1}^{SeLe} |B_i(\omega)|^2 = 10 * \log_{10}(SeLe); \quad (5.20)$$

Next, recall the definition of the SNR of the sample portion pSNR in (5.10).

$$pSNR_i = EP_i - EN_i = EP_i - ED_i; \quad (dB)$$

Therefore, we obtain:

$$\begin{aligned}
 pSNR_i &= EP_i - ED_i \\
 &= EP_i - \{10 * \log_{10}|\alpha - 1|^2 + EP_i - 10 * \log_{10}(SeLe)\} \\
 &= 10 * \log_{10}(SeLe) - 10 * \log_{10}|\alpha - 1|^2; \tag{5.21}
 \end{aligned}$$

From (5.21) we can conclude that $pSNR_i$ is determined by the length of the sample portion $SeLe$ and the scale factor α . Once α is fixed, $pSNR_i$ is only related $SeLe$ and has nothing to do with the energy of the sample portion. The larger the $SeLe$, the greater the $pSNR$. This theoretical conclusion has been proven by our observations in Table 5.2, Table 5.4, Table 5.5 and Table 5.6. According to this conclusion, given $SeLe = 512$ and $\alpha = 2.6$ the calculated theoretical $pSNR$ is about $23dB$. However, the experimental value of $pSNR$ under these parameters setting is about $13.5dB$, this is because the assumption our theoretical based, which suppose the PRBS scrambling function can scramble the sample portion into a noise-like signal with evenly distribution, is too critical and too difficult to realize.

Finally, the metrics to be analyzed is the SNR of the whole frame $fSNR$. According to the definition:

$$\begin{aligned}
 fSNR_i &= EF_i - ED_i \\
 \text{given } (5.18) \quad &= EF_i - \{10 * \log_{10}|\alpha - 1|^2 + EP_i - 10 * \log_{10}(SeLe)\} \\
 &= EF_i - EP_i + 10 * \log_{10}(SeLe) - 10 * \log_{10}|\alpha - 1|^2 \tag{5.22}
 \end{aligned}$$

From (5.22), it can be seen that, since the total energy of a frame is a constant, once the $SeLe$ and α is fixed, the frame SNR is determined by the energy of the selected sample, EP_i , recall EP_i at this scenario is determined by the *BeginSample* (*BeginFrequency*), the higher the *BeginFrequency*, the less EP_i , therefore the large $fSNR$, which is confirmed by the experimental result in Table 5.2.

5.3 Frame Selection

Theoretically all the audio frames in the original Work can be used for audio watermarking, because we have seen that by using spread spectrum technique combined with sample selection in each frame, even embedding each frame with one information bit '1', we can achieve our objective of imperceptibly embedding watermark information. In cases when the applications require high signal to noise ratio of the whole audio Work besides subjective transparency, is there any approach that can help to achieve higher **SNR** with *BeginSample* and *SeLe* fixed?

From experimental results in Figure 5.2, 5.3 , 5.5, 5.11 as well as the results in Table 5.3, it can be found that both the noise energy induced in each frame and the *fSNR* of each frame vary with different frames in a large dynamic range. Given $L=1024$, $BeginSample=186$, $SeLe=256$, the statistical mean of *ED* and *fSNR* are $-25dB$ and $35.76dB$ respectively, and their variations are about $16dB$ and $9.9dB$, as shown in Figure 5.14.

It tells us that some frames are good for watermarking, which means given the same parameters setting, including L , $SeLe$, α , these frames can have higher *fSNR*, by suitably selecting part of the frames based on some criteria, we can improve the transparency of the whole watermarked Work.

Frame Selection is feasible also because the number of watermark information bits to be embedded in the audio file is far less than the number of frames available.

Frame Selection will increase the whole audio Work's transparency. However, the compensation for frame selection is the embedding capacity of the system. This is intuitive because we use fewer frames for watermarking embedding, thus the information bits that can be embedded into the whole file decrease as well.

Another problem should be considered is how to identify those frames with watermark information embedded. The properties in which frame selection is based upon may change during the embedding process, storage as well as transmission. Therefore extra care to maintain the important properties is

required.

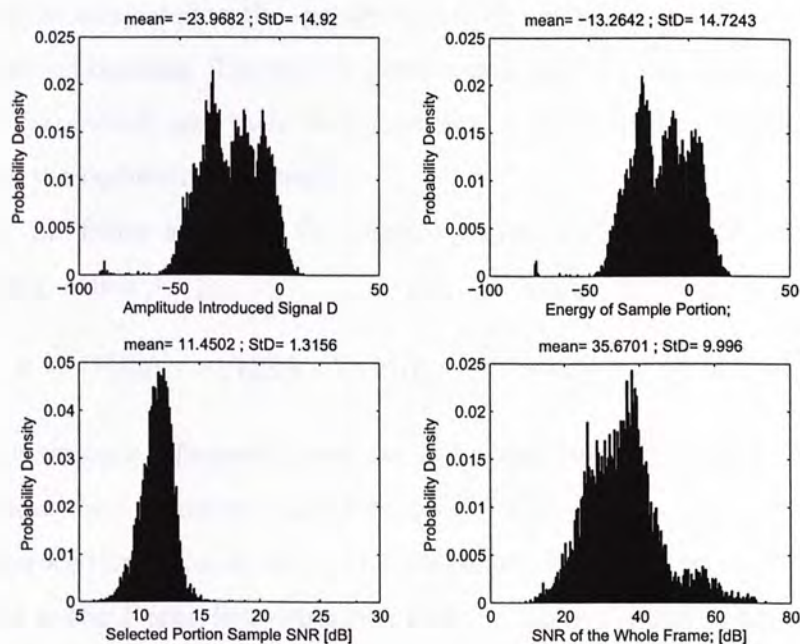


Figure 5.14: PDF of ED, fSNR pSNR and EP, Without Frame Selection.

5.3.1 General Frame Selection

Because we will embed one information bit in each frame, the number of frames selected eventually is determined by the number of bits in the watermark binary stream. Therefore, before we do frame selection, we have to determine the number of required frame . Suppose Q is the number of watermark information bits, Q is the required capacity of the watermarking system, thus we shall select at least Q frames from all available frames.

The procedure of frame selection is shown in the flowchart in Figure 5.15. Frame selection is based on the properties of the selected sample $U_i(\omega)$ portion, which is the output of the *sample selection* process. The frame selection procedure is done off-line and based on the requirements of the applications,

such as the total **SNR** or the capacity **Q**. Before doing the frame selection, a pre-processing is performed and the frame selection criteria determined. All $U_i(\omega)$ will be subjected to the specification to decide whether it can be used for watermark embedding. The sample portion that can be used for watermarking is call $V_i(\omega)$, which will then be transformed into time domain $v_i(n)$ and go thorough the embedding process.

After the frame selection, the selected sample portion form a new sample portion set, called **V**:

$$V = \{V_1(\omega), \dots, V_i(\omega), \dots, V_M(\omega)\}. \quad \text{where} \quad Q \leq N \quad (5.23)$$

The Q watermark information bits are embedded into each sample portion of this set and their sequence is also kept by $\{V_i(\omega)\}$.

The procedure of constructing the watermark embedded audio file Work is described in the *Frame Reconstruction Block* in Figure 5.15. Subsequent to the *content adaptive embedding* process, each sample portion $v_i(n)$ will be modified. Suppose $v_i(n)$ is changed to $\tilde{v}_i(n)$ after the embedding process.

In order to keep the total energy of each sample portion unchanged, which helps to improve the perceptual quality and guarantee the frame identification process while detecting, the watermark embedded frequency sample portion $\tilde{v}_i(n)$ is normalized according to the energy of the original portion $v_i(n)$. The normalized output is called $v'_i(n)$. The normalization process is done according to,

$$v'_i(n) = \sqrt{\frac{E1}{E2}} \tilde{v}_i(n) \quad n = 1, \dots, SeLe; \quad (5.24)$$

where $E1 = \sum_n |v_i(n)|^2$; $E2 = \sum_n |\tilde{v}_i(n)|^2$.

Then $v'_i(n)$ will be transformed into frequency domain $V'_i(\omega)$ by *DCT*. Next $V'_i(\omega)$ is placed in the position where they are originally taken out. Thus the whole frequency frame with watermark embedded $S'_i(\omega)$ is obtained.

By using Inverse-DCT, $S'_i(\omega)$ is transformed back into time domain $s'_i(n)$. The selected frames and the non-selected frames are used together to reconstruct a new audio work $\{x'(n)\}$ with watermark information embedded.

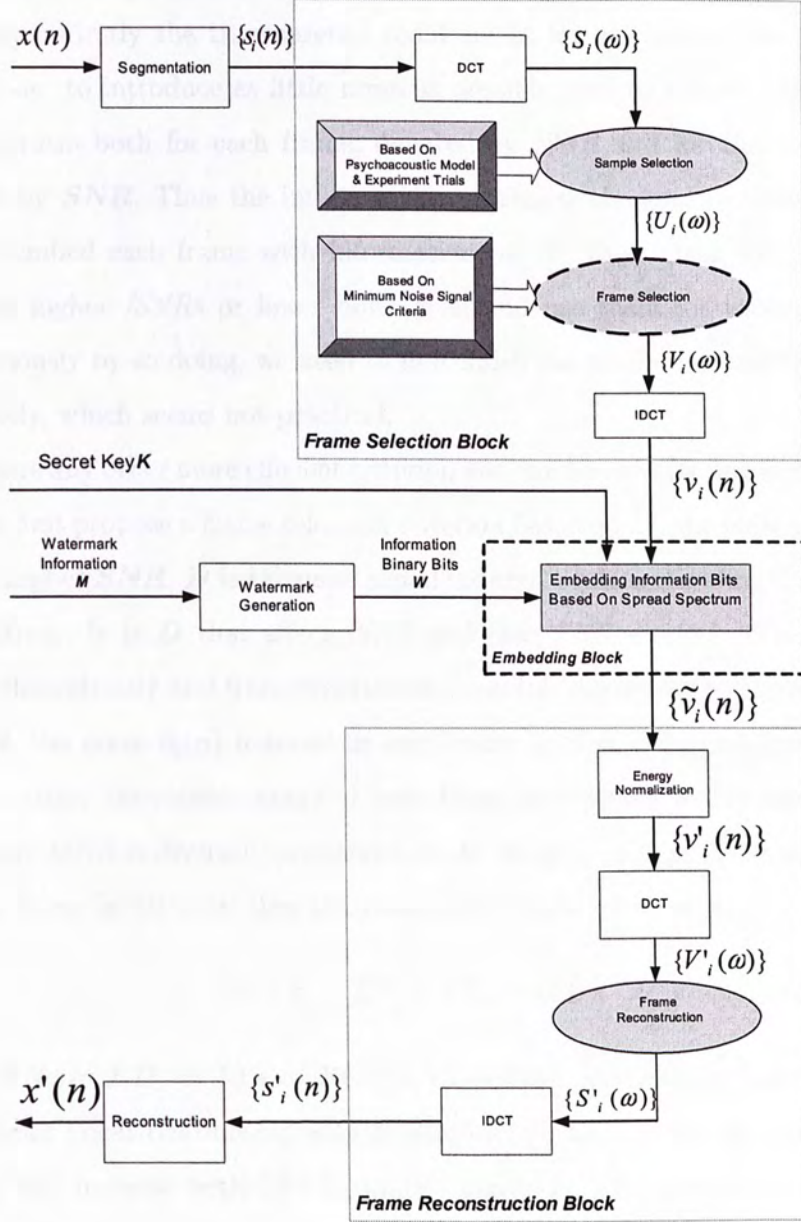


Figure 5.15: The Block Diagram of Watermark Embedding Based on Spread Spectrum Technique Combined With Sample and Frame Selection.

5.3.2 Frame Selection For Transparency Improvement

In our approach, $L=1024$; $BeginSample = 186$ (approximate 4Khz), and $SeLe=256$. How to further improve the transparency of the whole system is a concern. Firstly the transparency requirement for embedding can be summarized as: to introduce as little noise as possible, and to achieve high signal to noise ratio both for each frame, denoted by $fSNR$ and for the whole file, denoted by SNR . Thus the intuitive way to realize the goal of transparency is: first embed each frame with information bit '1', then check which frames have the higher $fSNRs$ or lower noise level and use them for watermarking. But obviously by so doing, we need to first finish the whole embedding process completely, which seems not practical.

Is there any other more efficient criterion and can be used for frame selection? Here we first propose a frame selection criterion based on D , the noise signal, to achieve higher SNR . D is the noise signal induced while embedding watermark information. It is D that affect $fSNR$ and thus SNR directly. This can be proved theoretically and from experimental results. According to the definition of $fSNR$, the noise $n_i(n)$ induced in each frame is D in it is time domain. At the same time the signal energy of each frame is constant within each frame. Therefore, $fSNR$ is inversely proportion to D . Suppose EF_i is the energy of the original frame in dB unit, thus the relationship between them is:

$$fSNR_i = EF_i - EN_i \quad (dB); \quad (5.25)$$

where $EN_i \approx ED_i = 10 * \log_{10}(D_i^2)$, as defined and verified before. Thus $fSNR_i$ has linear relationship with $10 * \log_{10}(D_i^2)$, from which we can see that $fSNR_i$ will increase with $10 * \log_{10}(D_i^2)$ decreases. The analysis is shown in Figure 5.19. It is similar as the previous one for D and EN , the X axis is divided into small sub-section, then average the values fall into that section as the representative for that sub-section. The analysis is denoted by the dot in the figure and a line that best fit the discrete dots is found and represented by the solid line, called as *approximate curve*. The approximate curve satisfies the theoretical relationship governed by equation (5.25).

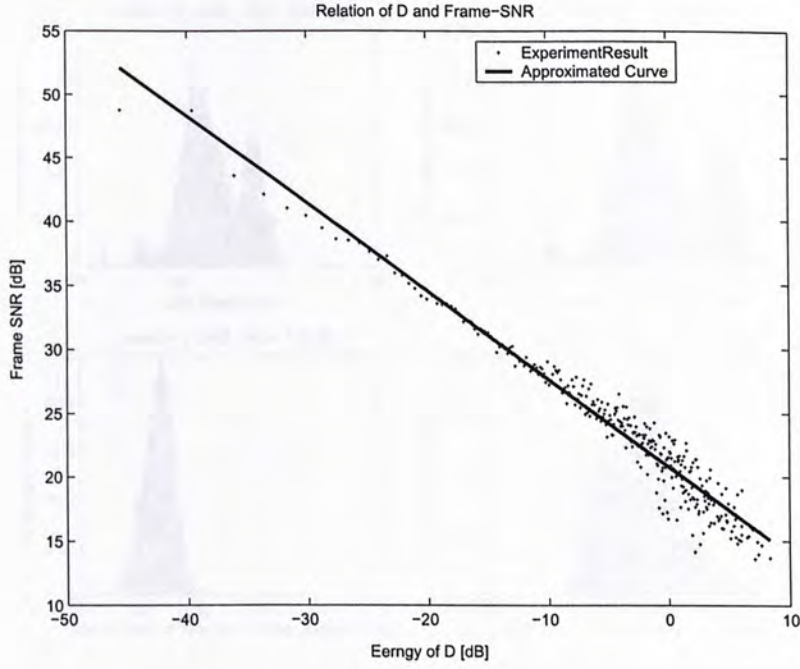


Figure 5.16: Relationship of D and Frame SNR $fSNR$

Therefore we can use D as a criterion for frame selection subject to the requirement of high $fSNR$. Because frame selection criteria based on minimum D will introduce the least noise directly, and it is also consistent with the requirement of high $fSNR$.

The procedure of selected frame based on D is done as follow: in the pre-processing stage, D_i for each scramble sample portion $B_i(L_Index)$ can be found by $D_i = \alpha * B_i(L_Index) - B_i(S_Index)$. Then determined the minimum threshold Th to be set so that there are at least Q sample portions will have D less than this Th and can be used for watermarking.

The experimental result after combined with frame selection based on the D is show in Figure 5.14. From which we can see that the ED has decreased by about $8dB$ as compared with the performance of watermarking without frame selection, and the $fSNR$ increased by $6dB$.

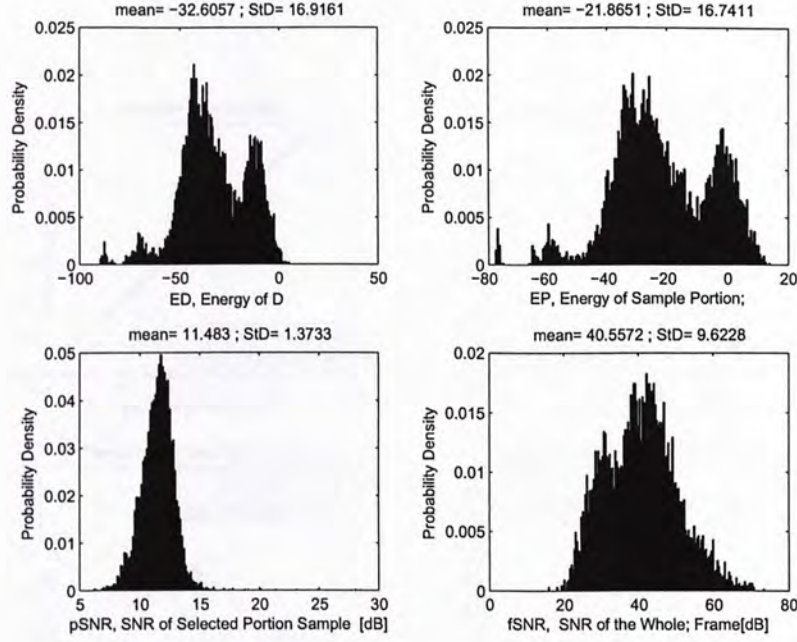


Figure 5.17: *PDF* of *ED*, *fSNR* *pSNR* and *EP*, When Frame Selection Based on *D*

Frame selection based on *D* do improve the performance, both from the metrics of *ED* and *fSNR*. However, to obtain *D*, we have to do the scramble process. Is there any other alternative criterion we can turn to? Let's first further investigate the relationship of *ED* with other parameters, we analysis the experimental data by the same method used to analysis *fSNR* and *ED*. The analysis result is shown in Figure 5.18.

Sub-figure (a) is the relationship between *ED* and the experimental noise energy. Sub-figure(b) is *ED* versus *fSNR*. These two figure have been shown and analyzed already.

Sub-figure (c) is *ED* versus *pSNR*, the dots are the experimental result and the solid line is the curve which best fits the distribution of the analysis result dots. From the result it shows that *pSNR* remains constant as *ED* increases, which means *ED* won't have any effect on *pSNR*. The *ED*-vs-*fSNR* is also plotted in this figure, from which we can see that *pSNR* has very little variation, with *fSNR* decrease, means they don't have inherent relation with each others.

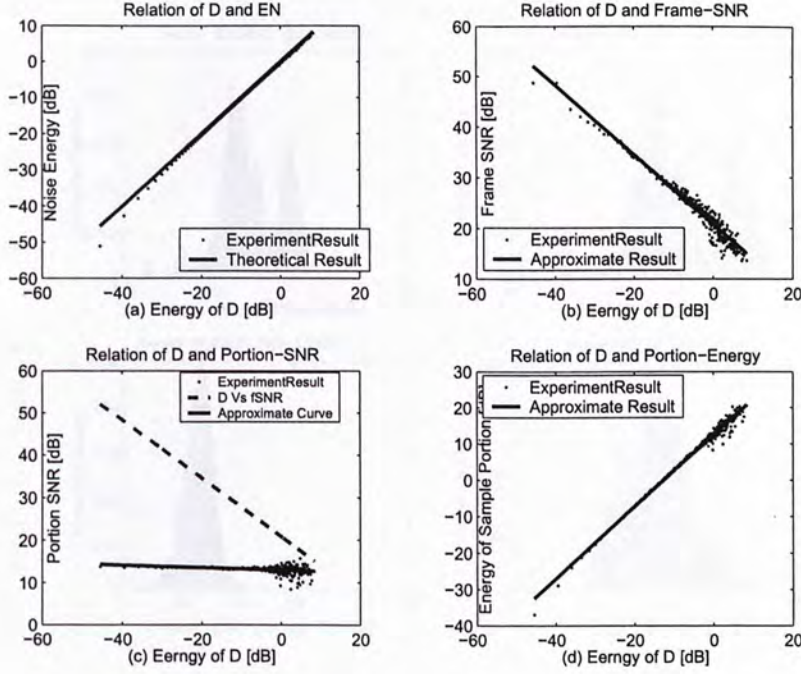


Figure 5.18: Relationship of D with other Metrics

What sub-figure (d) shows is the relation between the energy of ED and the energy of the selected sample portion. From which we can see ED is directly proportional to portion energy. It is consistent with the *content adaptive embedding* scheme, which means the noise signal induced adaptively change with the cover signal, subject to $pSNR = EP - ED$ and $pSNR$ will keep constant when $SeLe$ is fixed. So sample portion energy may be used as an alternative of D .

From the above analysis we propose another frame selection criterion which is based on the sample portion energy. This is because sample portion energy can represent the value of D and it is easy to obtain and analyze during pre-processing stage. The way of frame selection based on sample portion energy is similar to the one based one D . After the sample portions set is obtained, a minimum energy threshold is set so that there are

Q sample portions fall below this threshold which can be used for watermarking.

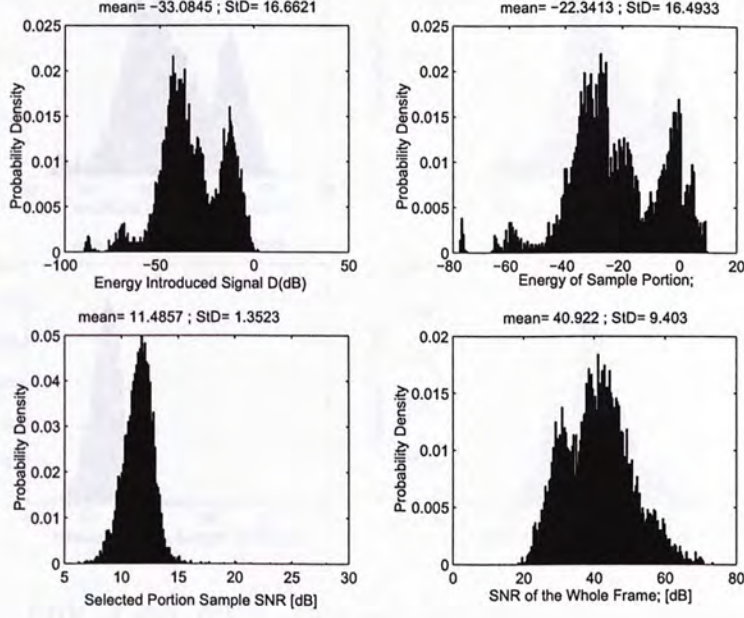


Figure 5.19: PDF of ED , $fSNR$ $pSNR$ and EP , When Frame Selection Based on Sample Portion Energy

In addition to select frame based on the noise signal D and sample portion energy EP , there is another possible criterion for frame selection, which is given in Section 5.2.4 equation (5.22):

$$fSNR_i = EF_i - EP_i + 10 * \log_{10}(SeLe) - 10 * \log_{10}|\alpha - 1|^2$$

From the equation above, once $SeLe$ and α are fixed, $fSNR$ is determined by the value of $EF_i - EP_i$. Therefore in order to select frames with highest SNR, we can select those frames with high $EF_i - EP_i$ value for watermarking. To implement this frame selection method, pre-processing is done to determine the threshold for $EF_i - EP_i$, then those frame with $EF_i - EP_i$ larger than the threshold will be selected. The experimental result of this methods is shown in Figure 5.20.

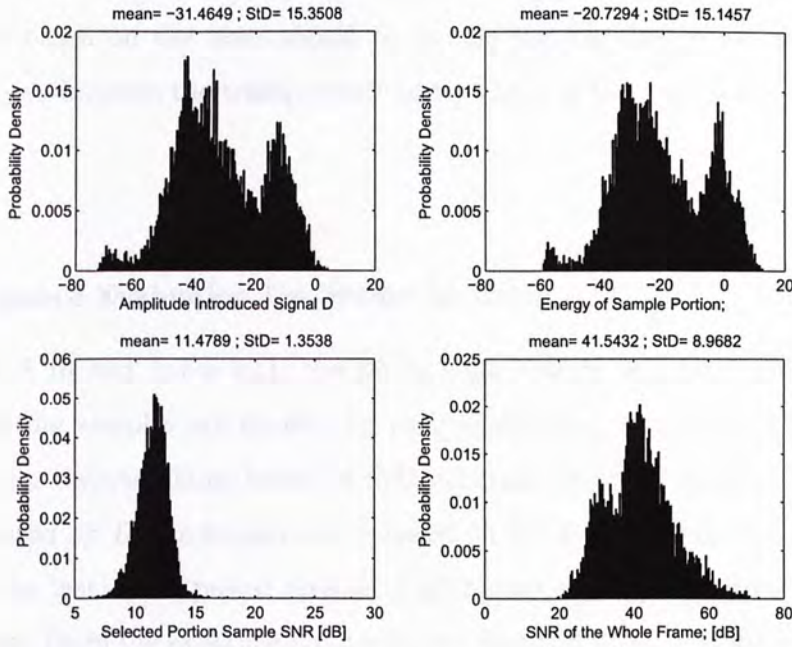


Figure 5.20: PDF of ED , $fSNR$ $pSNR$ and EP , When Frame Selection Based on $EF-EP$

	Based on D	Based on EFEP	Based on EP	No Selection
Mean of ED (dB)	-32.61	-31.46	-33.08	-23.97
Std of ED	16.92	15.35	16.66	14.92
Mean of $pSNR$ (dB)	11.48	11.48	11.49	11.45
Std of $pSNR$	1.37	1.35	1.35	1.32
Mean of $fSNR$ (dB)	40.56	41.54	40.92	35.67
Std of $fSNR$	9.62	8.97	9.40	10.00
Mean of EP (dB)	-21.87	-20.73	-22.34	-13.26
Std of EP	16.74	15.15	16.49	14.72

Table 5.9: Performance of Different Frame Selection Method, with the *BeginFrequency* fixed to 4KHz, in term of *BeginSample*=186; $L = 1024$; $SeLe$ =256. (In each experiment, half of the frames are selected for embedding with information bit '1'.)

From the experiment results shown in Table 5.9, we can conclude that, frame selection based on the noise signal D , sample portion energy EP as well as $EF-EP$, can improve the transparency performance of the system significantly.

Performance Evaluation For Frame Selection

In Table 5.10 and Table 5.11, the $SNRs$, noise energy and bit error rates of 15 audio file samples are shown. In each experiment, half of all frames are selected for watermarking, based on different frame selection criteria. The first one is based on D , the second one is based on $EF-EP$, the third one is based on EP , the last one is picked up half of all frames continuously from the very beginning. From the experimental results, we find that although different frame selection methods will improve the SNR with different levels, all the three frame selection criteria can improve the SNR of the whole file effectively. The noise energy induced during watermarking process also reduced by frame selection.

The performance of these three frame selection methods are close to each others with respect to improvement of SNR and reduction of noise level. Therefore we can claim that frame selection can improve the transparency of our proposed watermarking system effectively.

Since we have applied a process of normalizing the watermarked embedded frequency points portion according to the original one, the frame selection based on sample portion energy EP and $EF-EP$ do not bring on extra error bits significantly, even can reduce the bit error rate. But frame selection based on D brings on extra error bits and result in higher bit error rate than that of no frame selection based process.

Therefore we propose methods which are based on sample portion energy EP and the energy difference between the whole frame and the selected frequency band $EF-EP$ as approaches for frame selection.

	SNR of Audio Files				Noise Level of Audio Files			
Criterion	On D	EFEP	On EP	NoSel	On D	EFEP	On EP	NoSel
File1	40.38	39.51	41.35	31.78	5.96	6.83	4.99	14.55
File2	42.55	42.30	43.24	29.66	4.82	5.07	4.13	17.71
File3	58.65	58.49	60.21	49.77	-18.70	-18.54	-20.26	-9.81
File4	43.51	42.62	44.71	36.74	-5.67	-4.78	-6.87	1.10
File5	45.65	46.04	47.05	35.12	-1.71	-2.10	-3.11	8.82
File6	33.81	34.02	34.82	28.53	14.44	14.22	13.43	19.72
File7	60.01	58.66	61.10	46.32	-16.67	-15.32	-17.76	-2.98
File8	33.68	33.95	35.03	30.81	16.48	16.21	15.14	19.36
File9	27.62	28.08	28.68	22.39	19.59	19.13	18.53	24.81
File10	53.96	51.63	55.27	44.66	-10.46	-8.14	-11.78	-1.16
File11	57.03	48.64	58.94	34.98	-14.64	-6.24	-16.55	7.42
File12	41.57	41.03	43.64	28.05	-2.86	-2.32	-4.94	10.66
File13	32.69	33.16	33.61	26.71	16.67	16.20	15.75	22.65
File14	48.37	48.03	49.15	35.22	-2.26	-1.91	-3.03	10.89
File15	46.55	45.35	48.29	44.01	-9.54	-8.33	-11.28	-6.99

Table 5.10: SNR, Noise Energy of 15 Audio Files, When Subject to Different Frame Selection Criteria; Based on D, EFEP, EP and NoSelection respectively; With $SeLe=256$; $BeginFrequency$ fixed to $4kHz$, in term of $BeginSample=186$; $L = 1024$;

File	Bit Error Rate (%)			
Criterion	On D	On EFEP	On EP	NoSel
File1	3.26	0	0	0.47
File2	1.86	0	0	0.93
File3	0	0	0	0
File4	1.86	0	0	0
File5	0.93	0	0	0
File6	1.86	0	0	0
File7	1.40	0	0	0
File8	6.05	8.37	2.33	6.05
File9	3.26	0	0	0
File10	2.09	0.12	1.39	1.39
File11	21.35	7.19	20.77	0
File12	1.97	0	1.28	1.28
File13	3.02	0	0	0
File14	0.93	0	0	0
File15	2.90	0	1.62	1.62

Table 5.11: Bit Error Rate of 15 Audio Files, When Subject to Different Frame Selection Criteria; Based on D, EFEP, EP and NoSelection respectively; With $SeLe=256$; $BeginFrequency$ fixed to $4kHz$, in term of $BeginSample=186$; $L = 1024$

5.4 Watermark Information Retrieve

The flowchart for watermark information retrieve is shown in Figure 5.21. Suppose at the detection stage, an audio Work $\{\tilde{x}(n)\}$ is waiting for detecting the watermark information embedded with it. It is first segmented into time frames $\tilde{s}_i(n)$ along the time axis, using the same length L as used in embedding process, this procedure is identical to the one at the embedding stage. $\tilde{s}_i(n)$ is transformed into frequency domain $\tilde{S}_i(\omega)$ by DCT. Then using the same parameters *BeginSample* and *SeLe* as used in the embedding procedure to pick out the sample portion $\tilde{U}_i(\omega)$ from $\tilde{S}_i(\omega)$. Followed by using the same criteria as used in frame selection to determine whether the samples portion has been used to embed information. If the sample portion is confirmed to have watermark embedded, it is denoted by $\tilde{V}_i(\omega)$. $\tilde{V}_i(\omega)$ as well as the secret key K are input into the detection block, which has been described in Chapter 4 Section 4.2.2 to check whether a '0' or a '1' information bit has been embedded. The detected bits are used to form the detected binary bit stream W' . Finally the binary stream are process by ECC or/and if applicable then decoded into the watermarking information M' .

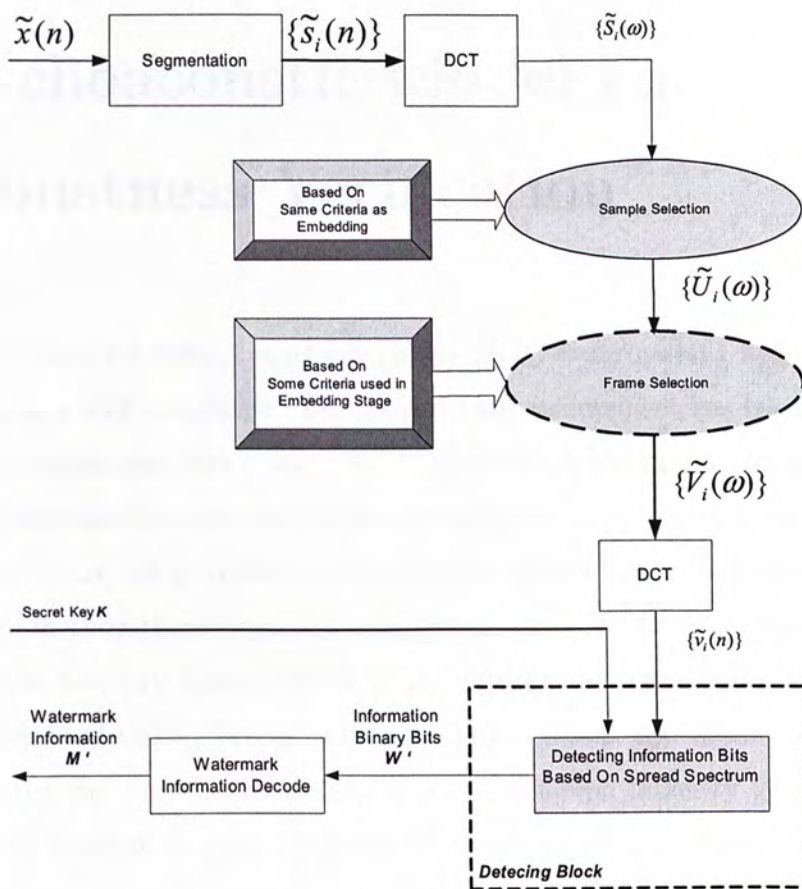


Figure 5.21: The Block Diagram of Watermark Detection Based on Spread Spectrum Technique Combined With Sample and Frame Selection

Chapter 6

Psychoacoustic Model For Robustness Verification

The two basic yet critical requirements for useful watermarking approach are transparency and robustness. The transparency requirement has been fulfilled by using sample and frame selection, as described in Chapter 5. Among all of the possible signal processing, the most severe one is lossy compression. In this Chapter, we are going to verify the robustness against lossy compression of our proposed method theoretically by using psychoacoustic model analysis.

Human auditory system (HAS) is an imperfect detector, which is characterized with *Absolute Hearing threshold*, *Critical Bands* and *Masking* [43, 44]. Human ear can only detect sounds above a minimum intensity level, this is called the *Absolute Hearing Threshold*. HAS has a limited frequency resolution that is characterized by *critical bands*, which are a set of neighboring regions within the human audible frequency range where HAS has uniform audibility and masking properties. Another important phenomenon in HAS is *masking*. Masking refers to the fact that an audible but weak component in a given audio signal becomes imperceptible with the presence of another stronger signal.

Psychoacoustic model exploits the temporal and frequency characteristics of HAS. Therefore many of the data compression techniques nowadays use the characteristics of the HAS implicitly or explicitly. To make sure the watermark can survive these kinds of lossy compression, we shall design our watermarking

system to incorporate with psychoacoustic model.

In this chapter, we will first introduce the three major phenomena of human auditory system. Then we will describe how to use a psychoacoustic model to analyze the audio signals. Finally, we will verify our embedded watermark system can survive lossy compression by using psychoacoustic model to examine the watermark embedded signal.

6.1 Introduction of Human Auditory System

6.1.1 Absolute Hearing Threshold

Before we discuss about absolute hearing threshold, let's first review the definitions of two terms, *dB* and *SPL*. In common audio signal processing, the sound amplitude is measured on a logarithmic scale *decibels* (dB). A decibel scale is a means for comparing the intensity of two sounds:

$$10\log_{10}(I/I_0) \quad (6.1)$$

where I and I_0 are the two intensity levels, with intensity being proportional to the square of the sound pressure P .

Sound pressure level (SPL) is a measure of absolute sound pressure P in dB:

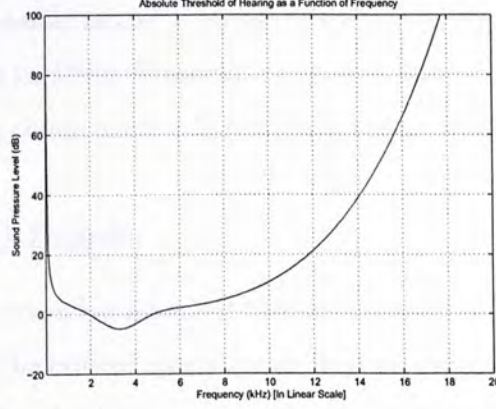
$$SPL(dB) = 20\log_{10}(P/P_0) \quad (6.2)$$

where the reference 0dB corresponds to the threshold of hearing, which is $P_0 = 0.0002\mu\text{bar}$ for a tone of 1kHz.

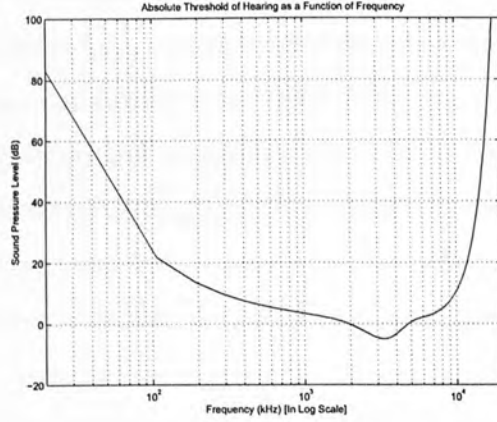
The absolute threshold of hearing is the maximum amount of energy of a pure tone that can't be detected by a listener in a noise free environment. It can also be interpreted as: the minimum amount of energy of a pure tone that can be detected by a listener in a noise free environment. The absolute hearing threshold is a function of frequency. It is typically expressed in terms of dB SPL. The frequency dependence of this threshold was quantified as early as 1940, when Fletcher [42] reported test results for a range of listeners with acute hearing. Base on the subjective experiment on average young man, it can be

approximated using the following equation:

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 (dB SPL). \quad (6.3)$$



(a) Absolute Hearing Threshold with Frequency In Linear Scale



(b) Absolute Hearing Threshold with Frequency In Log Scale

Figure 6.1: Absolute Hearing Threshold as a Function of Frequency

Figure 6.1 indicates the absolute hearing threshold as a function of frequency. In Figure 6.1 (a), the X axis is in linear scale. For better interpretation of this threshold, we take the logarithmic scale of the X axis and get Figure 6.1 (b). From this figure we can see that the sensitivity of human auditory system

varies with the frequency. The curve indicates that the ear is most sensitive to frequency around 3kHz and sensitivity declines at very low (20Hz) and very high(20kHz) frequency. And from the figure we can also know that any sound component under the curve can not be detected.

In the psychoacoustic model analysis, this absolute hearing threshold, complemented with the masking threshold, calculated from the signal itself is used to determine which components will remain in compression coding.

6.1.2 Critical Bands

Human Auditory System has a limited frequency resolution that is characterized by critical bands. The critical bands are defined as: Around a center frequency in which the noise bandwidth is increased until there is a just noticeable difference in the tone at the center frequency [43].

Empirical work by several observers led to the modern notion of critical bands [42]-[47]. Let's consider an example. Suppose the loudness (perceived intensity) remains constant for a narrow band noise source presented at a constant SPL even as the noise bandwidth is increased up to the critical bandwidth. For any increase beyond the critical bandwidth, the loudness then begins to increase.

For an average listener, critical bandwidth tends to remain constant, about 100Hz, up to 500Hz, and increases to approximately 20% of the center frequency above 500Hz. A *Bark scale* can be used to represent the critical bands. The linear frequency in hertz can be converted into Bark scale using function 6.4. The relationship between linear frequency and Bark frequency can be expressed in Figure 6.2.

$$z(f) = 13 \arctan(0.00076f) + 3.5 \arctan[(\frac{f}{7500})^2] \quad (6.4)$$

It is proved to be useful when building practical systems to treat the ear as a discrete set of bandpass filters that conforms to (6.4). In this way, it is supposed that we can achieve a more natural fit with spectral information processing in the ear. One of the possible ways to do this is defined in MPEG-1 standard. Table 6.1 gives an the upper boundaries of each critical bands and the linear

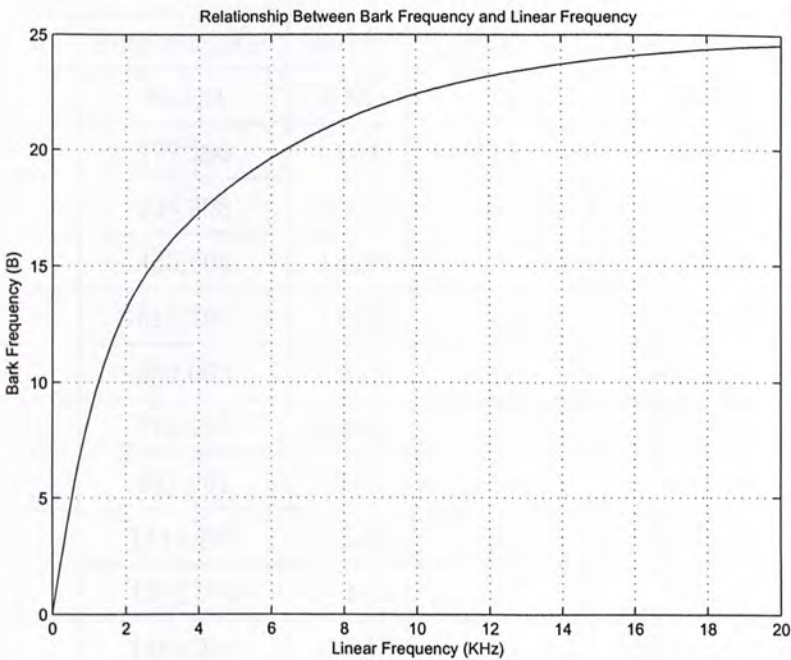


Figure 6.2: Relationship Between Bark Frequency Scale and Linear Frequency Scale

frequency they correspond to. This table can be interpreted in Figure 6.3, in this figure, the X axis represents the linear frequency in hertz. The left Y axis is number of the critical band, from 1 to 25. And the right Y axis is the upper boundary of each critical band. From the curve given in 6.4, the lower and upper frequency of each critical band can be identified. From the figure we can see that the bandwidth of the critical bands increase with the frequency increased.

Critical No.	Frequency[Hz]	Bark[z]	Critical No.	Frequency[Hz]	Bark[z]
1	86.133	0.85	14	2325.586	14.062
2	172.266	1.694	15	2756.25	15.1
3	258.398	2.525	16	3186.914	15.955
4	430.664	4.124	17	3875.977	17.079
5	516.797	4.882	18	4478.906	17.904
6	689.063	6.301	19	5340.234	18.922
7	775.195	6.959	20	6373.828	19.963
8	947.461	8.169	21	7579.688	20.971
9	1119.727	9.244	22	9302.344	22.074
10	1291.992	10.195	23	11369.531	22.984
11	1464.258	11.037	24	15503.906	24.013
12	1722.656	12.125	25	19982.813	24.573
13	1981.055	13.042			

Table 6.1: Critical Band Upper Boundaries

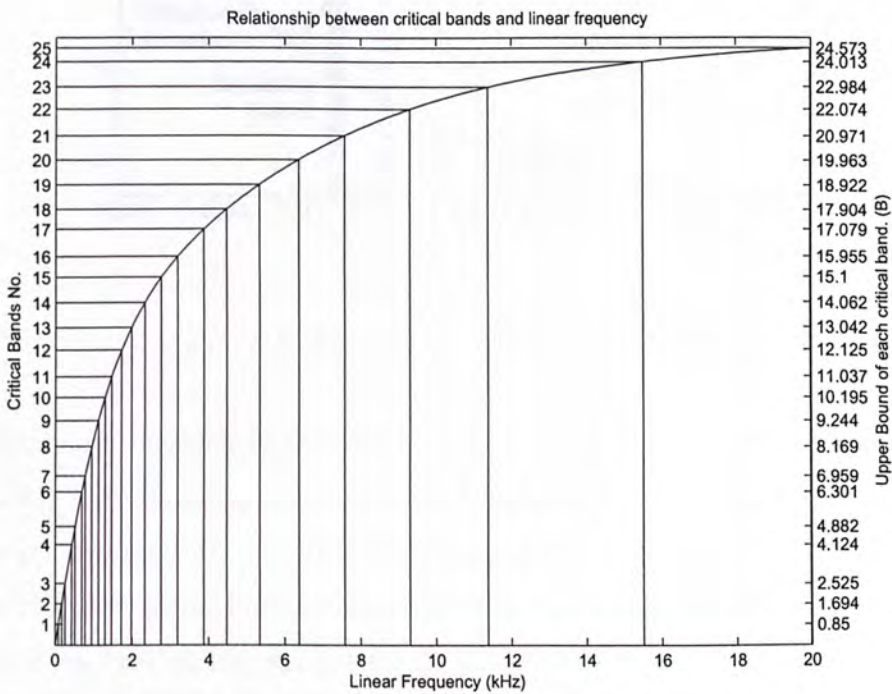


Figure 6.3: Critical Bands Defined in MPEG-1 Psychoacoustic Model 1.

6.1.3 Masking Effect

Audio masking is the effect that a faint audible sound (maskee) becomes inaudible in the presence of another louder audible sound (masker). The masking phenomenon happens both in spectral and temporal domain. called *Frequency masking* and *Temporal Masking*. Frequency Masking refers to the masking phenomenon between frequency components of the audio signal. These two components should occur simultaneously, close together in frequency. Then the stronger masker signal will make the weaker signal inaudible. Both tonal and non-tonal component can be a masker. There is a masking threshold for each masker, as shown in Figure 6.4¹. The masking threshold is determined by the sound pressure level of the masker and the tonal-like or noise-like properties of the masker.

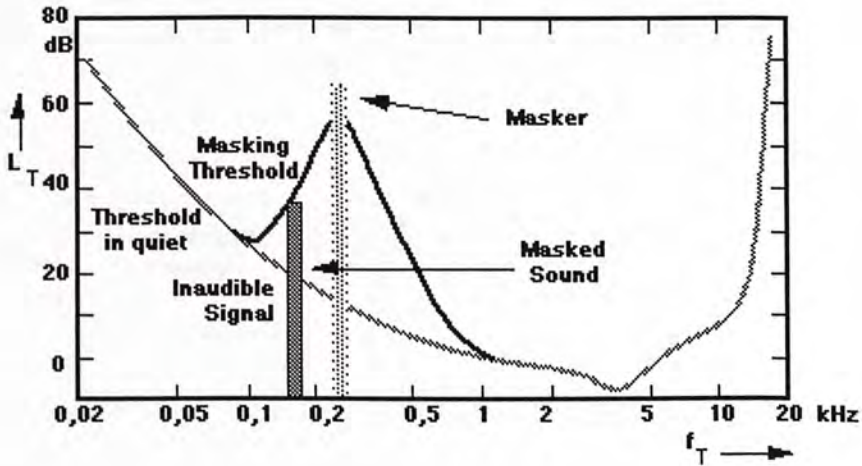


Figure 6.4: An Example of Frequency Masking

Temporal masking happens in time domain and has *pre-masking* and *post masking*. *Pre-masking* refer to weaker signal becomes inaudible before the stronger masker turn on. This masking period can be 5-20ms. *Post-masking* refer to weaker signal becomes inaudible after the stronger masker turn off. This kind of masking period can last for 50-200ms.

¹<http://www.crcg.edu>

6.2 Psychoacoustic Model of Human Auditory System

A psychoacoustic auditory model is an algorithm that tries to imitate the human hearing mechanism. It bases on knowledge from several areas such as biophysics and psychoacoustics. This model is useful for applications like transform coding, compression and audio watermarking.

The psychoacoustic model we are going to introduce here is: “*MPEG 1 psychoacoustic model I Layer 1. ISO/IEC 11172-3. CODING OF MOVING PICTURES AND ASSOCIATED AUDIO FOR DIGITAL STORAGE MEDIA AT UP TO ABOUT 1.5 MBIT/s Part 3 AUDIO.*” [48]

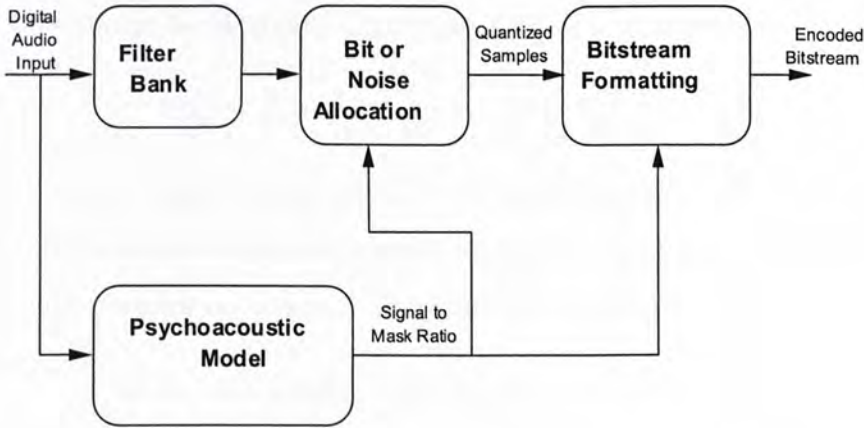


Figure 6.5: The Primary Components of MPEG 1 Encoder

Figure 6.5 is the primary components of MPEG 1 encoder. The filterbank does a time to frequency mapping. It is used to split the broadband signal with sampling frequency f_s into 32 equally spaced subbands with sampling frequencies $f_s/32$. The sound pressure level $L_{sb}(n)$ of subband n is computed.

The procedure of psychoacoustic model 1 layer 1 for analyzing signal with 44.1khz sampling frequency include the following steps:

- Step 1: FFT Analysis;
- Step 2: Finding of the tonal and non-tonal components of the audio signal;
- Step 3: Determination of the threshold in quiet;

- Step 4: Decimation of the maskers, to obtain only the relevant maskers;
- Step 5: Calculation of the individual masking thresholds;
- Step 6: Determination of the global masking threshold.

Step 1: FFT Analysis. The masking threshold is derived from an estimate of the power density spectrum. The parameters of power density spectrum estimation including: FFT-Size $N=512$ samples; which means the window size is 11.6ms; Before doing FFT, a Hann window is applied to the signal frame. The power density spectrum $X(k)$ is normalized to the reference of 96dB SPL.

Step 2: Finding of the tonal and non-tonal components of the audio signal. Since The tonal and non-tonal components have different masking characteristics, they should be identified. First label $X(k)$ as a local maximum if:

$$X(k) > X(k-1) \quad \text{and} \quad X(k) > X(k+1) \quad (6.5)$$

Then listing of tonal components and calculation of the sound pressure level. To determine if a local maximum may be a tonal component, a frequency range df around the local maximum is examined, df is given as:

$$\begin{aligned} df &= 172.266Hz & 0kHz < f \leq 5.512kHz \\ df &= 281.25Hz & 5.512kHz < f \leq 11.024kHz \\ df &= 562.50Hz & 11.024kHz < f \leq 19.982kHz \end{aligned}$$

Therefore a local maximum is labeled as a tonal component if

$$\begin{aligned} X(k) - X(k+j) &> 7dB & (6.6) \\ j &= -2, +2 & \text{for } 2 < k < 63 \\ j &= -3, -2, +2, +3 & \text{for } 63 \leq k < 127 \\ j &= -6, \dots, -2, +2, \dots, +6 & \text{for } 127 \leq k \leq 250 \end{aligned}$$

The sound pressure level of the tonal component is calculated by:

$$X_{tm}(k) = X(k-1) + X(k) + X(k+1) \quad (dB); \quad (6.7)$$

Next step is to find the non-tonal components. First determine the boundaries of critical bands within each critical band, the power of the spectral lines, except the spectral lines examined, are summed to form the sound pressure level of the non-tonal component corresponding to that critical band, denoted by $Xnm(k)$. The index of the new non-tonal component is the value closest to the geometric mean of that critical band.

Step 3: Determination of the threshold in quiet. The threshold in quiet $LTq(k)$, can be derived from the analysis in Section 6.1.1. An offset depending on the overall bit rate is used for the absolute threshold.

$$\begin{aligned} -12dB & \quad \text{for} \quad \text{bit rates} > = 96 \text{ kbit/s per channel} \\ 0dB & \quad \text{for} \quad \text{bit rates} < 96 \text{ kbit/s per channel} \end{aligned}$$

Step 4: Decimation of the maskers, to obtain only the relevant maskers. First reduce the number of maskers which are considered for the calculation of the global masking threshold. Tonal $Xtm(k)$ or non-tonal components $Xnm(k)$ are considered for the calculation of the masking threshold only if:

$$Xtm(k) \geq LTq(k) \quad \text{or} \quad Xnm(k) \geq LTq(k) \quad (6.8)$$

Then decimate of two or more tonal components within a distance of less than 0.5 Bark, keep the component with the highest power.

Step 5: Calculation of the individual masking thresholds. Index j is used to indicate the relevant tonal or non-tonal masking components. Of the original $N/2$ frequency domain samples, indexed by k , only a subset of samples (106), indexed by i , are considered for the global masking threshold calculation. The individual masking thresholds of both tonal and non-tonal components are given by the following expression:

$$\begin{aligned} LTtm[z(j), z(i)] &= Xtm[z(j)] + avtm[z(j)] + vf[z(j), z(i)](dB) \\ LTnm[z(j), z(i)] &= Xnm[z(j)] + avnm[z(j)] + vf[z(j), z(i)](dB) \end{aligned}$$

The term av is called the masking index and vf the masking function of the masking component $Xtm[z(j)]$. The masking index av is different for tonal and

non-tonal masker (*avtm* and *avnm*).

$$\begin{aligned} avtm &= -1.525 - 0.275 * z(j) - 4.5dB, \\ avnm &= -1.525 - 0.175 * z(j) - 0.5dB. \end{aligned}$$

vf is the masking function of a masker. It depends on the distance in Bark $dz = z(i) - z(j)$ to the masker. In this expression *i* is the index of the spectral line at which the masking function is calculated and *j* is that of the masker; The masking function, which is the same for tonal and non-tonal maskers, is given by:

$$\begin{aligned} vf &= 17 * (dz + 1) - (0.4 * X[z(j)] + 6)dB \quad \text{for } -3 \leq dz < -1Bark \\ vf &= (0.4 * X[z(j)] + 6) * dzdB \quad \text{for } -1 \leq dz < 0Bark \\ vf &= -17 * dzdB \quad \text{for } 0 \leq dz < 1Bark \\ vf &= -(dz - 1) * (17 - 0.15 * X[z(j)]) - 17dB \quad \text{for } 1 \leq dz < 8Bark \end{aligned}$$

If $dz < -3Bark$, or $dz \geq 8Bark$, the masking is no longer considered.

Step 6: Determination of the global masking threshold. The global masking threshold $LTg(i)$ at the i^{th} frequency sample is derived from the individual masking threshold of each of the *j* tonal and non-tonal maskers, and in addition from the threshold in quiet $LTq(i)$. The global masking threshold is found by summing the powers corresponding to the individual masking thresholds and the threshold in quiet. The total number of tonal maskers is given by *m*, and the total number of non-tonal maskers is given by *n*.

$$LTg(i) = 10\log_{10}[10^{LTq(i)/10} + \sum_{j=1}^m 10^{LTtm(z(j),z(i))}/10 + \sum_{j=1}^n 10^{LTnm(z(j),z(i))}] \quad (6.9)$$

Figure 6.6 is an example of the psychoacoustic model analysis result.

Finally the minimum masking level $LTmin(n)$ in subband *n* is determined by the following expression:

$$LTmin(n) = MIN[LTg(i)]dB \quad i \text{ is all the index in subband } n \quad (6.10)$$

Applying this psychoacoustic model to an audio signal frame, we can obtain the analysis result as shown in Figure 6.6, in which the power spectrum density (PSD), global masking threshold(GMT) as well as minimum masking threshold (MMT) of each sub-bands are plotted together. From the figure we can see that the PSD curve is above the GMT and MMT in lower frequency part, yet in the higher frequency part, the PSD is below the MMT.

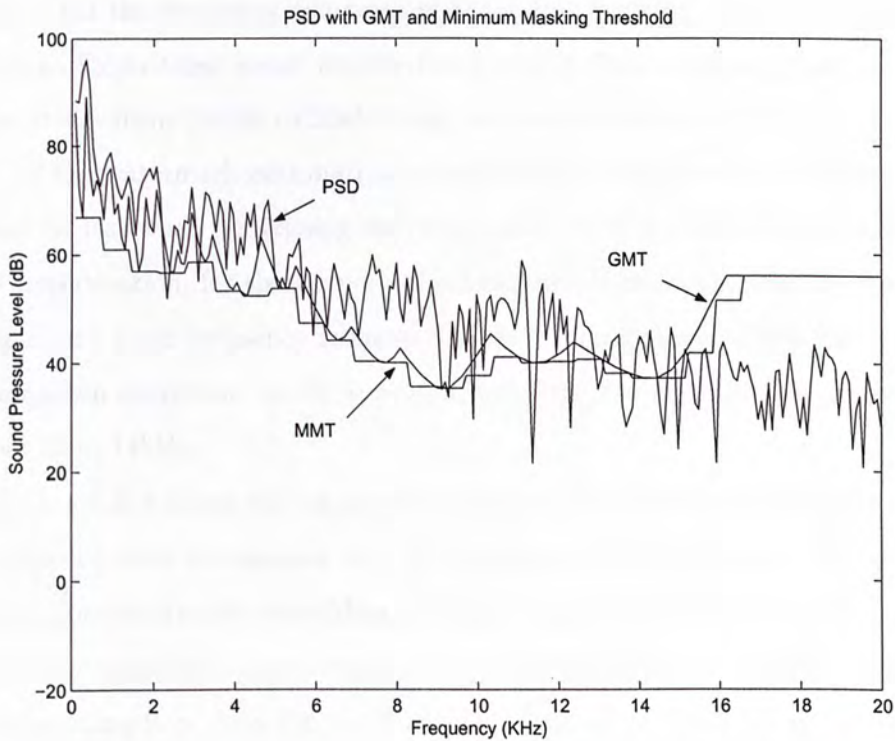


Figure 6.6: An Example of Masking Threshold Calculated by Psychoacoustic Model.

6.3 Robustness Verification by Psychoacoustic Model Analysis

It is known that since the higher frequency components are relatively insignificant, most of the lossy compression may get rid off this part of artifact. As described in the MPEG psychoacoustics model I Layer 1 standard, a masking threshold in frequency domain can be derived from the signal itself, and it is known that the frequency components below this masking curve is likely to be ignored. Experiment result depicted in Figure 6.7 shows that the watermark embedded within 14kHz to 20kHz may be removed or ignored while compression. If the watermark information is embedded in these frequency component, it may be lost when undergoing the compression process. Taken this possibility into consideration, for the purpose of robust sample selection, we should select samples in a lower frequency range so that the watermark embedded can survive compression operation. In our approach, we select frequency sample portion no higher than 14kHz.

Figure 6.7, 6.8 and 6.9 are psychoacoustic model analysis on an audio sample segment with information bits '1' embedded with each frame. The sample portion for watermark embedding is within 4kHz to 10kHz. Figure 6.7 is an example of analysis on one frame. Sample between 4kHz and 10kHz were selected as sample portion for watermarking. The PSD, GMT as well as MMT are shown in the figure. From the result we see that from frequency higher than 14kHz, most of the PSD samples are below the MMT, thus will be ignored during lossy compression. In the frequency range lower than 11kHz, the PSD samples are above the MMT thus promised to survive the compression processing. The sample portion we proposed to select for watermarking lies between 4kHz and 10kHz, and is above the MMT, from this result we can predict the watermark information will still exists even after lossy compression.

In addition to analysis one single frame, we analysis a whole audio file. In Figure 6.8, the PSD and MMT are averaged result of all the frames within the file. This averaged result is also consistent with our conclusion that the PSD

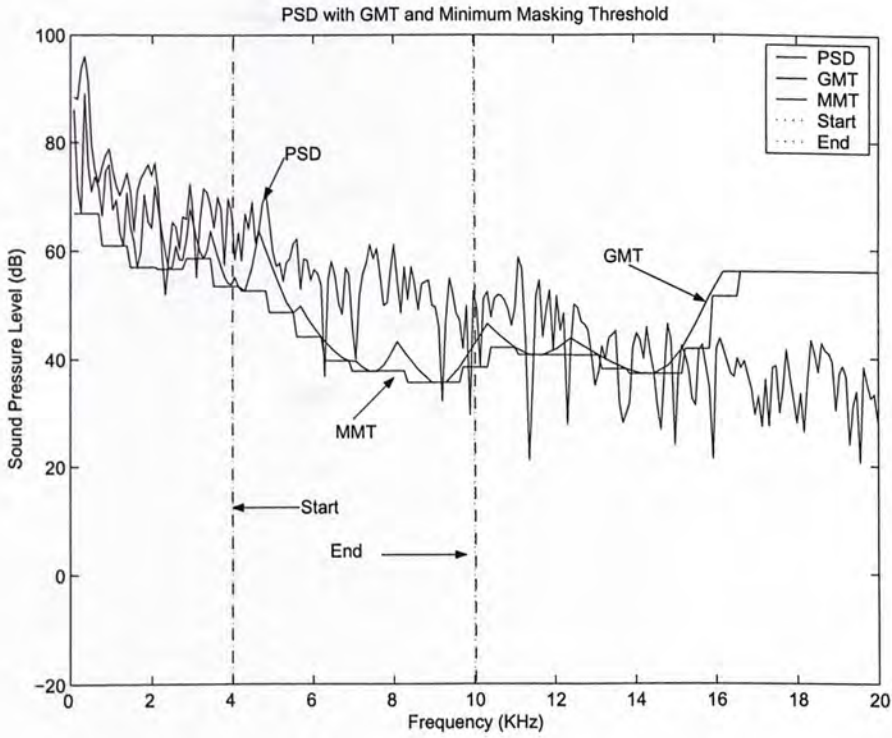


Figure 6.7: Robust Verification On One Frame

sample within 4kHz to 10kHz is above MMT therefore watermark information embedded within this range will not be discarded during lossy compression.

To further prove that PSD within 4kHz to 10kHz is above MMT in each frame, we will analysis each frame separately. There are two experiments being done. One is to select sample portion from 20Hz to 6kHz, thus the comparison of PSD and MMT is within 20Hz to 6kHz. Another is to select sample within the range of 4kHz to 10kHz, thus the comparison of PSD and MMT is between 4kHz and 10kHz. In the experiments, we first define the judgement criterion of “above”. We define as the following: if and only if a certain percentage of samples within the range from 4kHz to 10kHz in PSD is greater than those in MMT, we can claim PSD is above MMT in range from 4kHz to 10kHz. We vary this percentage from 50% to 100% by the step size of 5%. From the experimental results we can see that even the judgement level is set to 70%, all frames are 100% percent above the MMT in both experimental cases. When the judgement level is set to 85%, the experiment select sample between 4kHz and 10kHz has

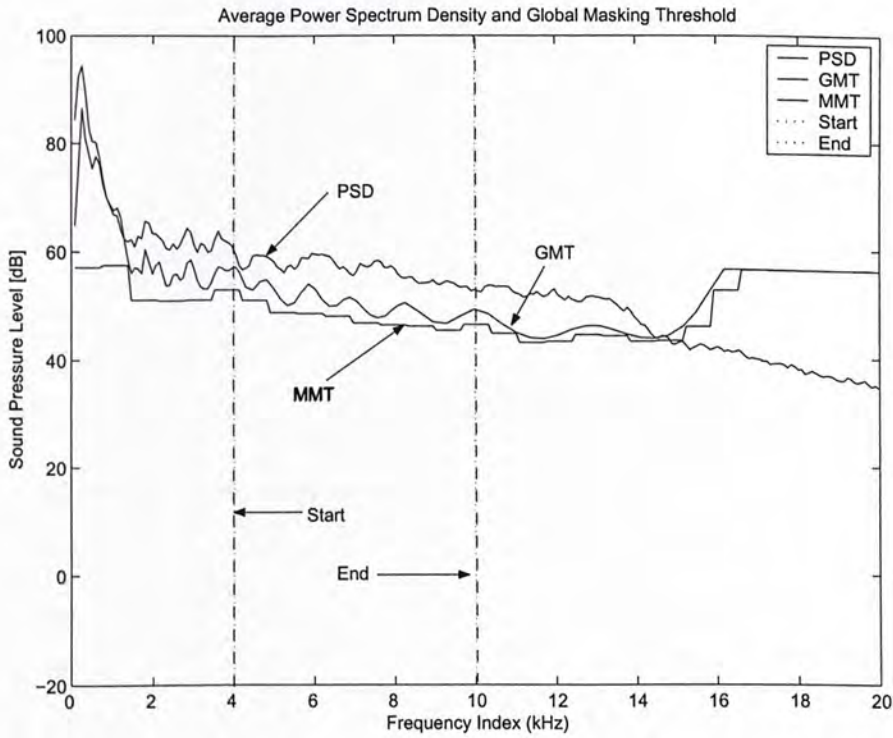


Figure 6.8: Robust Verification On An Whole Audio Sample. Average Result

above 80% frames satisfying the definition of “above” and therefore promised to survive lossy compression. In the case of selecting sample from the lowest frequency, 20Hz to 6kHz, there are about 98% percentage frames satisfying the definition. We can confirm that select samples from lower frequency range will be more robust than those selected from higher frequency components.

From the above experimental result, we perform analysis for one single frame as well as for the entire audio file, it is proved that select sample portion from 4kHz to 10kHz for watermarking is adequate for the watermark to be robust against lossy compression. But verifying the robustness of watermark from psychoacoustic model analysis is not a direct one of lossy compression processing. Therefore real signal processing should be done to validate the effectiveness of the approached watermarking system and tune the parameters including *Begin-Sample*, *SeLe* and α to realize robustness from a more practical way if necessary.

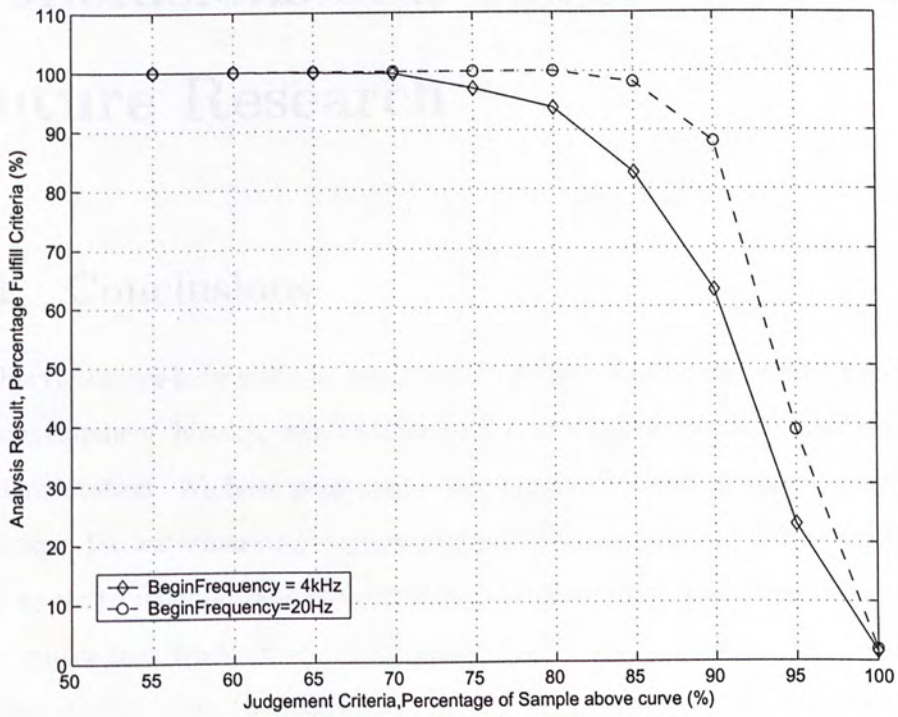


Figure 6.9: Robust Verification On An Whole Audio Sample. Statistical Analysis Result

Chapter 7

Conclusions and Suggestions For Future Research

7.1 Conclusions

In this thesis, we have given an introduction of digital audio watermarking, from its development history, applications and properties to possible methods for implementation. We have presented a new approach based on spread spectrum technique for watermarking implementation. The applications of this approach may apply to include ownership identification and proof, broadcast monitoring and transaction tracking of digital audio Work. Subject to the requirements of these applications, the proposed watermarking system should be transparent and robust.

To fulfill these requirements, we have first defined a series of objective evaluation metrics that can reflect the performance of the system by quantity measurement. Then after describing the general processes for both watermarking embedding and detection based on spread spectrum technique, we propose a more efficient embedding method called *content adaptive embedding*, which can encode the information bits while adapting the energy of the information bits according to the energy of the cover frame. The embedding scheme also allow an ease way of detecting the embedded bits successfully.

However, both objective evaluation metrics and subjective tests confirm that

the above watermarking system can not realize imperceptibly embedding.

To further improve the performance of the watermarking system, especially to ensure the watermarked Work to be as transparency as possible, we propose a new approach by incorporating *Sample Selection* in the frequency domain and *Frame Selection* in the time domain, instead of the traditional method by filtering.

Due to the fact that most of the energy of audio signal is contained in a certain portion of frequency points in the lower frequency band, we can achieve better system performance by embedding watermark information into other frequency bands which composed of less energy but promised to survive after signal processing such as lossy compression.

We then show that optimal sample selection can be achieved by choosing the correct *BeginFrequency* and the bandwidth of the frequency portion to be selected after examining the frequency characteristics of each frame in the frequency domain experimentally.

With optimal sample selection, less distortion noise is induced into the watermark embedded Work and therefore high signal to noise ratio can be obtained. Experiments carried out on a serial of audio file samples proved that both objectively and subjectively, sample selection can improve the transparency of the watermarking system and fulfill the requirement of imperceptibly embedding while maintaining the system's effectiveness.

Large variation exists among all frames within a Work and the fact that information bits are less than available frames provide opportunities for frame selection along the time axis. Followed the description for general frame selection and analysis on the relationships among all the evaluation metrics, we had proposed different frame selection criteria subject to the requirements of high signal to noise ratio and low noise energy level. From observations and analysis on the experimental results, we conclude that *frame selection* can further improve the transparency of the watermarking system while maintaining the effectiveness.

Another important requirement for watermarking system, robustness, had

also been taken into consideration during sample and frame selection. It is also verified by applying psychoacoustic model analysis on the watermark embedded Work.

Currently, there are many organizations putting efforts on audio watermarking research. But there is no standard specific for it.

If optimally setting the parameters according to the experimental discussions, our proposed method can be used for protecting the copyright of digital music. Before selling or distributing the music, the copyright information and transaction information can be embedded as watermark information. Then these information can be extracted when necessary.

7.2 Suggestions For Future Research

Robustness Test

Although we have already proved the robustness of our proposed method, it is necessary to exploit practical lossy compression, as well as other common signal processing for testing the robustness of the system. And adjust the parameters setting during sample and frame selection if necessary.

Discussion and Tuning On the Scale Factor α

Through our whole discussion, we have fixed α for each audio file samples. Although it also work pretty well, we also see that we can improve the performance, specially on bit error rate, by tuning α for each individual audio file.

Realize Realtime embedding and detection

Another promised and challenging future research is to realize frame selection based on realtime process, which means no pre-processing is needed for determining the threshold of frame selection, instead, the embedding module can determine whether to embed watermark information bits with a certain frame once the frame arrive, also the detection module can decide whether the frame

arrived should be check for watermark information bits. This will make digital audio watermarking with a more bright application prospect.

Bibliography

- [1] Pottersma, F.J.P., Zuckerman, R.L., Kahn, M.J., "The 'Watermark Survey' Proceedings of the IEEE", Volume 67, Issue 3, The IEEE Press, 1979, 1062-1078.
- [2] Herodotus, *The Histories*, Translated by E. V. Rieu, 1914, Penguin Books, 1996.
- [3] Ingemar J. Cox, "Robust Audio Watermarking", Ph.D. Thesis, Stanford University, San Francisco, CA, The Morgan Kaufmann, 1995.
- [4] D. Kahn, *The Codebreakers: The Story of Secret Systems*, New York: Scribner, 1997.
- [5] E.P. Hondroch, "Telecommunications Patent", U.S. Patent, Patent 3,504,104, 1970.
- [6] W. Saifanadi, "A Digital Watermarking Method for Copyright Protection of Documents for Authentication", in *Proc. 1998 IEEE Conference on Canadian Computing and Communications Security*, 1998.
- [7] L. Holt, R.L. Green, and A. Orlitzky, "Robust Watermarking of Signals", U.S. Patent 6,326,678, 1999.
- [8] D. Kormatz, and J. Thompson, "Automated Image and Document Image in Document", *Journal of the Electronic Imaging Society*, 1999.
- [9] R. Anderson, *Secure Information Systems*, John Wiley & Sons, 1998.

Bibliography

- [1] Petitcolas, F.A.P.; Anderson, R.J.; Kuhn, M.G.; "Information Hiding—A Survey" *Proceedings of the IEEE* , Volume: 87 Issue: 7 , Jul 1999 Page(s): 1062 -1078
- [2] Herodotus, *The Histories*. Translated by Aubrey de Selincourt London: Penguin Books, 1996.
- [3] Ingemar J. Cox, Matthew L. Miller and Jeffrey A. Bloom, *Digital Watermarking*, San Francisco, Calif. : Morgan Kaufmann, ©2002.
- [4] D. Kahn, *The Codebreakers: The story of Secret Writting*, New York: Scirbner 1967.
- [5] E.F. Hembrooke, "Identification of Sound and Like Signals," *United States Patent* 3,004,104, 1961.
- [6] W. Szepanski, "A Signal Theoretic Method for Creating Forgery-proof Documents for Automatic Verification," in J.S. Jackson,editor, *1979 Canahan Conference on Crime Countermeasures*, pp. 101-109, 1979.
- [7] L. Holt, B.G. Maufe, and A. Wiener, "Encoded Marking of a Recording Signal," *U.K. Patent* GB 2196167A, 1988.
- [8] N. Komatsu and H. Tominaga, "Authentication System Using Concealed Images in Telematics," *Memoirs of the School of Science and Engineering, Waseda University*, 52:45-60,1998.
- [9] R. Anderson. editor *Information Hiding*, volume 1174 of *Lecture Notes in Computer Science*. Berlin; New York: Springer-Verlag, 1996.

- [10] P.W. Wong and E.J. Delp, editor, *Security and Watermarking of Multimedia Contents, Proceedings of the Society of Photo-optical Instrumentation Engineers*, volume 3657, 1999.
- [11] A.E.Bell. "The Dynamic Digital Disk," *IEEE Spectrum*, 36(10):28-35,1999.
- [12] "SDMI Portable Device Specification," Part 1, Version 1.0, Document number pdwg99070802, July 1999
- [13] G. Depovere, T. Kalker, J. Haitsman, M. Mases, L. de Strycker, P. Termont, J.Vandewege, A. Langell, C. Alm, P. Norman, B. O'Reilly, G. Howes, H. Vaanholt, R. Hintzen, P. Donnelly, and A. Hudson. "The VIVA Project: Digital Watermarking for Broadcast Monitoring," *IEEE International Conference on Image Processing*, 2:202-205,1999.
- [14] F. Hartung and M. Kutter. "Multimedia Watermarking Techniques," *Proceeding of the IEEE*, 87(7):1079-1107,1999.
- [15] J. Abbate. "Inventing the Web," *Proceeding of IEEE*, 87(11):1999-2002,1999.
- [16] B. Schneier. "Applied Cryptography." New York: John Wiley & Sons, 1996.
- [17] Boney, L.; Tewfik, A.H.; Hamdy, K.N.; "Digital watermarks for audio signals", *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems*, 1996., 17-23 Jun 1996 Page(s): 473 -480
- [18] Petar Horvatic, Jian Zhao, Niels J. Thorwirth "Robust Audio Watermarking Based on Secure Spread Spectrum and Auditory Perception Model", *IFIP Conference Proceedings* 175 Kluwer 2000, ISBN 0-7923-7914-4 Contents. SEC 2000, Information Security for Global Information Infrastructures, IFIP TC11 Fifteenth Annual Working Conference on Information Security, August 22-24, 2000, Beijing, China.
- [19] United States Code, Title 17.

- [20] S.Craver, N. Memon, B.L. Yeo, and M.M. Yeung. "Can Ivisible Watermarks Solve Rightful Ownerships?" IBM Technical Report RC 20509, IBM Research, July 1996. IBM CyberJournal: <http://www.research.ibm>
- [21] D.R. Stinson. *Cryptography: Theroy and Practice*. Boca Raton , FL: CRC Press, 1995
- [22] B.M. Macq and J.J. Quisquater. "Cryptology for Digital TV Broadcasting", *Proceeding of the IEEE*, 83(6):944-957,1995.
- [23] F. Hartung and M. Kutter. "Multimedia Watermarking Techniques," *Proceedings of the IEEE*, 87(7):1079-1107,1999.
- [24] R.B. Wolfgang, C.I. Podilchuk and E.J. Delp. "Perceptual Watermarks for Digital Images and Video," *Proceedings of the IEEE*, 87(7):1108-1126,1999.
- [25] I.J. Cox and M.L. Miller. "A Review of Watermarking and the Importance of Perceptual Modeling," in *Proceeding of SPIE, Huamn Vision and Electronic Imaging II*, volume 3016,pp.92-99,1997
- [26] Swanson, M.D.; Bin Zhu; Tewfik, A.H.; "Current State of the Art, Challenges and Future Directions for Audio Watermarking", *IEEE International Conference on Multimedia Computing and Systems, 1999.* , Volume: 1 , Jul 1999 Page(s): 19 -24 vol.1
- [27] Jessop, P.; Acoustics, "The Business Case for Audio Watermarking", 1999 *IEEE International Conference on Speech, and Signal Processing, 1999. ICASSP '99. Proceedings.*, Volume: 4 , 15-19 Mar 1999 Page(s): 2077 -2078 vol.4
- [28] Cvejic, N.; Keskinarkaus, A.; Seppanen, T.; "Audio watermarking using m-sequences and temporal masking", *2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001 Page(s): 227 -230

- [29] Silvestre, G.C.M.; Hurley, N.J.; Hanau, G.S.; Dowling, W.J.; "Informed audio watermarking scheme using digital chaotic signals", *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01)*. 2001 IEEE International Conference on , Volume: 3 , 7-11 May 2001 Page(s): 1361 -1364 vol.3
- [30] Changsheng Xu; Jiankang Wu; Qibin Sun; "Digital Audio Watermarking and Its Application in Multimedia Database", *Proceedings of the Fifth International Symposium on Signal Processing and Its Applications, 1999. ISSPA '99.*, Volume: 1 , 1999 Page(s): 95 -98 vol.1
- [31] Changsheng Xu; Jiankang Wu; Qibin Sun; "A Robust Digital Audio Watermarking Technique", *Proceedings of the Fifth International Symposium on Signal Processing and Its Applications, 1999. ISSPA '99.*, Volume: 1 , 1999 Page(s): 95 -98 vol.1
- [32] Bender, W. Gruhl, D., Morimoto, N., & Lu, A. (1996) Techniques for data hiding. *IBM Systems Journal*, 35(5).
- [33] Johnson, N.F., & Katzenbeisser, S.C. (2000). *A survey of steganographic techniques*. In F.A.P. Petitcolas & S. Katzenbeisser (Eds.) *Information hiding: Techniques for steganography and digital watermarking* (1.ed. pp 43-78). Boston: Artech House.
- [34] Dugelay, J.-L., & Roche, S. (2000). *A survey of current watermarking techniques*. In F.A.P. Petitcolas & S. Katzenbeisser (Eds.) *Information hiding: Techniques for steganography and digital watermarking* (1.ed. pp 43-78). Boston: Artech House.
- [35] Hyen O Oh; Jong Won Seok; Jin Woo Hong; Dae Hee Youn; "New echo embedding technique for robust and imperceptible audio watermarking", *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01)*. , Volume: 3 , 2001 Page(s): 1341 -1344 vol.3

- [36] Czerwinski, S., Fromm, R., & Hodes, T. (1999). *Digital music distribution and audio watermarking* (IS 219): University of California - Berkeley.
- [37] M.K. Simon, J. K. Omura, R. A. Scholtz and B.K. Levitt, *Spread-Spectrum Communications Handbook* (McGraw-Hill, New York, 1994)
- [38] R.L. Pickholtz, D.L. Schilling, and L.B. Milstein, "Theory of Spread-Spectrum Communications - A Tutorial," *IEEE Transactions and Communications*, vol. COM-30, pp.855-884(1982May).
- [39] I.J. Cox, J. Kilian, F.T. Leighton, and T. Shamoon. "Secure Spread Spectrum Watermarking for MultiMedia", *IEEE Transactions on Image Processing*, 6(12):1673-1687, 1997
- [40] Kirovski, D., Malvar, H. (2001, April 2001). *Robust cover communication over a public audio channel using spread spectrum*. Paper presented at the Information Hiding Workshop, Pittsburgh, PA.
- [41] Gordy, J.D.; Bruton, L.T.; "Performance Evaluation of Digital Audio Watermarking Algorithms", *Proceedings of the 43rd IEEE Midwest Symposium on Circuits and Systems, 2000*. Volume: 1 , 2000 Page(s): 456 -459 vol.1
- [42] H. Fletcher, "Auditory patterns", *Rev. Mod. Phys.*, pp. 47-65, Jan. 1940
- [43] Painter, T.; Spanias, A.; "Perceptual Coding of Digital Audio", *Proceedings of the IEEE* , Volume: 88 Issue: 4 , Apr 2000 Page(s): 451 -515
- [44] Davis Pan (1996). "A Tutorial on MPEG/Audio Compression". Motorola, Inc. October 7, 1997.
- [45] Saito, S.; Furukawa, T.; Konishi, K.; "A Data Hiding for Audio using Band Division Based on QMF Bank" *IEEE International Symposium on Circuits and Systems, 2002. ISCAS 2002.*, Volume: 3 , 2002 Page(s): 635 -638

- [46] D. D. Greenwood, "Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane," *J. Acoust. Soc. Amer.*, pp. 1344-1356, Oct. 1961.
- [47] B. Scharf, "Critical Bands," in *Foundations, of Modern Auditory Theory*. New York: Academic, 1970.
- [48] Information technology – "Coding of moving pictures and associated audio for digital storage media at up to 1,5 Mbits/s – Part3: audio". British standard. BSI, London. October 1993. Implementation of ISO/IEC 11172-3:1993. BSI, London. First edition 1993-08-01.

CUHK Libraries



004077098