PERCEPTUAL DATA EMBEDDING IN AUDIO AND SPEECH SIGNALS

by

Libo Zhang
Bachelor of Science (B.Sc.), York University,
Toronto, Canada, 2002
Bachelor of Engineering (B.Eng.), Xi'an Jiaotong University,
Xi'an, China, 1990

A thesis
presented to Ryerson University
in partial fulfillment of the
requirement for the degree of
Master of Applied Science
in the Program of
Electrical and Computer Engineering.

Toronto, Ontario, Canada, 2004

Author's Declaration

I hereby declare that I am the sole author of this thesis.
I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.
Signature

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.
Signature

Instructions on Borrowers

Ryerson University requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Perceptual Data Embedding in Audio and Speech Signals

Libo Zhang
Master of Applied Science,
Electrical and Computer Engineering Department, Ryerson University, Toronto, 2004

Abstract

Perceptual embedding is a technique to embed extra information into multimedia signals without fidelity degradation, which is the core of many applications including watermarking and data hiding. Perceptual embedding can be viewed as a telecommunication to transmit the embedded information over the medium consisting of the host signal. This observation divides the current embedding techniques into two categories, i.e. the host-suppressing ones like the quantization-based Quantization Index Modulation (QIM) and Scalar Costa Scheme (SCS), and the non host-suppressing ones like the conventional Spread Spectrum (SS) technique. The former class has significant advantages over the latter in robustness and data rate due to significantly reduced noise levels.

In this research, the conventional SS embedding technique is modified such that it can suppress the host impact mostly. Both the theoretical analysis and simulations show that the modification significantly improve the performance of the conventional scheme and further, outperform the QIM and SCS under the case of watermarking where the attacks can be expected to be very strong. To further increase the robustness and embedding rate, measures like frequency masking effects of the Human Masking Auditory system and Forward Error Correction schemes are employed, such as Turbo code. The second part of this research explores the possibility of high-capacity embedding in telephony speech signals. Another modification to improve the embedding rate is proposed for the conventional SS scheme under weak attacks, which are expected for the case of data embedding.

Acknowledgments

I would like to acknowledge my supervisor, Prof. Sridhar Krishnan at Ryerson University, Toronto, Canada, for his excellent guidance. I would like to acknowledge my supervisor, Dr. Heping Ding at the Institute of Microstructural Sciences of National Research Council (NRC), Ottawa, Canada, for his valuable mentor.

I am grateful to the Multimedia Information and Signal Analysis Research Group at Ryerson University and the Acoustics and Signal Processing Group at the Institute of Microstructural Sciences. They have all contributed to this work by insightful discussions, volunteering for subjective testing or simply creating a pleasant work environment.

I would like to thank Ryerson University and NRC for providing financial assistance to complete this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid development of personal electronics (e.g. MP3 players, CD/DVD recorders, PDAs) has made it possible to create and edit multimedia data much easier than before. Further, an almost errorless transmission of the broadband Internet makes it possible to distribute the large media files without any quality degradation. This has promoted the protection of intellectual copyright and the prevention of the unauthorized tampering to become an important industrial and academic issue.

According to [22], globally annual losses due to piracy (not including Internet piracy) of copyrighted materials are estimated to be as high as US $22 billion. The "2003 Special 301 Report on Global Copyright Protection and Enforcement" from the International Intellectual Property Alliance (IIPA) states that in 2002, deficiencies in the copyright regimes of 56 countries caused US music industries to lose more than US$2.1 billion in trade due to piracy.

To cope with this, a mechanism to embed the copyright information into the media signals seems promising. A watermark, or a digital copyright signature, is hidden in the media transparently and exists permanently no matter what types of processing the media experiences. However, simple mechanisms of embedding the information into header segments of digital files are useless because the headers can be easily removed or changed without fidelity degradation. Instead, to be robust permanently, the copyright information should be fused with the content data seamlessly.

It is worth mentioning that encryption could protect neither the copyright nor the tam-

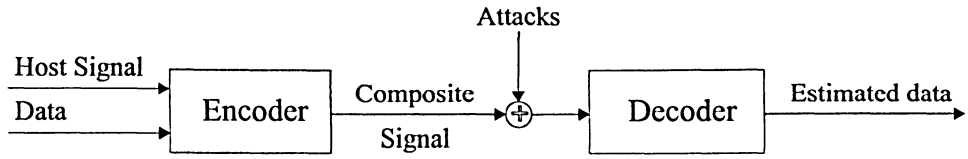

Figure 1.1: Communication model of data embedding system

pering of digital media. Encryption prevents the access to the multimedia content with a decryption key. The client must pay the royalties to get the key. But once the media has been decrypted, it can be repeatedly copied and distributed.

1.1 Research Statements

The general viewpoint thorough the research is to model the data embedding as a telecommunication system. The model is shown in Figure 1.1. With this model some general assumptions can be clarified.

Perceptual embedding means the composite signal should not be perceptually different from the host signal. This requirement constrains the power of the embedded data lower than the hearing threshold, and thus characterizes such embedding as a power-limited communication.

The decoder has no access to the host signal for extraction and this is called *blind extraction.* In most cases an extra channel for the host signal is not possible.

Robustness represents the ability of the decoder to extract the embedded data correctly after attacks of signal processing manipulations. It is desired that the robustness is as high as possible for *all* kinds of attacks.

The data rate in this communication is an important factor that limits the feasibility of different techniques. The difference on the data rates are mainly due to the different noise levels in different scenarios. For example, due to low noise levels in data hiding, it is possible to achieve a data rate higher than that in the watermarking case.

These requirements listed above normally conflict with each other. For example, a very

robust watermark can be obtained by considerably modifying the host signal to increase the embedding strength. However, this large modification will be perceptible. Therefore, an optimal trade-off, which depends on the specific application, is always needed for practical use.

In this research, both robust watermarking for audio signals and high-capacity embedding for telephony speech signals are studied. The first objective is to develop a novel audio watermarking scheme which can provide the state-of-art performance. The second objective is to incorporate the available techniques including the masking effects of Human Auditory System (HAS) and various Forward Error Correction (FEC) schemes to improve the performance (The superior Turbo code is paid special attention to).

This research involves the multidisciplinary areas such as digital signal processing, Human Auditory System (HAS) modeling, telecommunication, information theory.

1.2 Application Areas

The embedded information will be distributed transparently and robustly along with the host signal. This goal enables the technique to be applicable to many potential cases. Some typical scenarios are as follows.

Ownership Protection: This is the most popular application scenario and the main drive behind the research area. The watermark, or the copyright information in its digital form, is embedded in the media data imperceptibly. The watermark, only known to the copyright holder, is expected to be very robust, enabling the owner to demonstrate the presence of this watermark in case of dispute [15] [44]

Additional Services: A high-capacity embedding technique virtually creates a subliminal channel that can be of different usages. For example, due to the transparency of the new channel, a hierarchical service system can be setup on top of the traditional sets with extra encode/decode devices. Many potentially commercial applications can be envisaged with this technique [18].

Authentication: In some applications it is highly desired that the media content should

not be changed any. This can be accomplished by a so-called *fragile* watermark. A fragile watermark is one that has only very limited robustness and usually become invalid after the slightest modification. If a very fragile mark is detected intact, we can infer that the signal has probably not been altered since the watermark was embedded. In this research, the fragile watermarking is not considered.

1.3 Main Contributions and Thesis Outline

The organizations of this thesis are shown in Figure 1.2 and the main contributions include

- Proposed a novel SS embedding scheme that is more robust than the current acclaimed quantization schemes of QIM/SCS;
- Incorporated different FEC schemes, especially Turbo code, to improve the robustness;
- Proposed a watermarking system consisting of multiple watermarks.
- Improved the conventional SS scheme to increase the embedding rate.
- Explored the data hiding in digital telephony signals.

Chapter 2 introduces some background knowledge including time-frequency transformation, HAS models, as well as relevant telecommunication techniques. Chapter 3 presents the proposed scheme in details. The scheme is analyzed theoretically and compared with the quantization-based schemes like SCS/QIM. In Chapter 4 the scheme is applied to audio watermarking case. Different measures to increase the SNR are presented and an extensive simulation is conducted. In Chapter 5 an approach to embed data into the ITU-G.711 μ-law coded speech signals is proposed. The final chapter discusses the future research expected in this area.

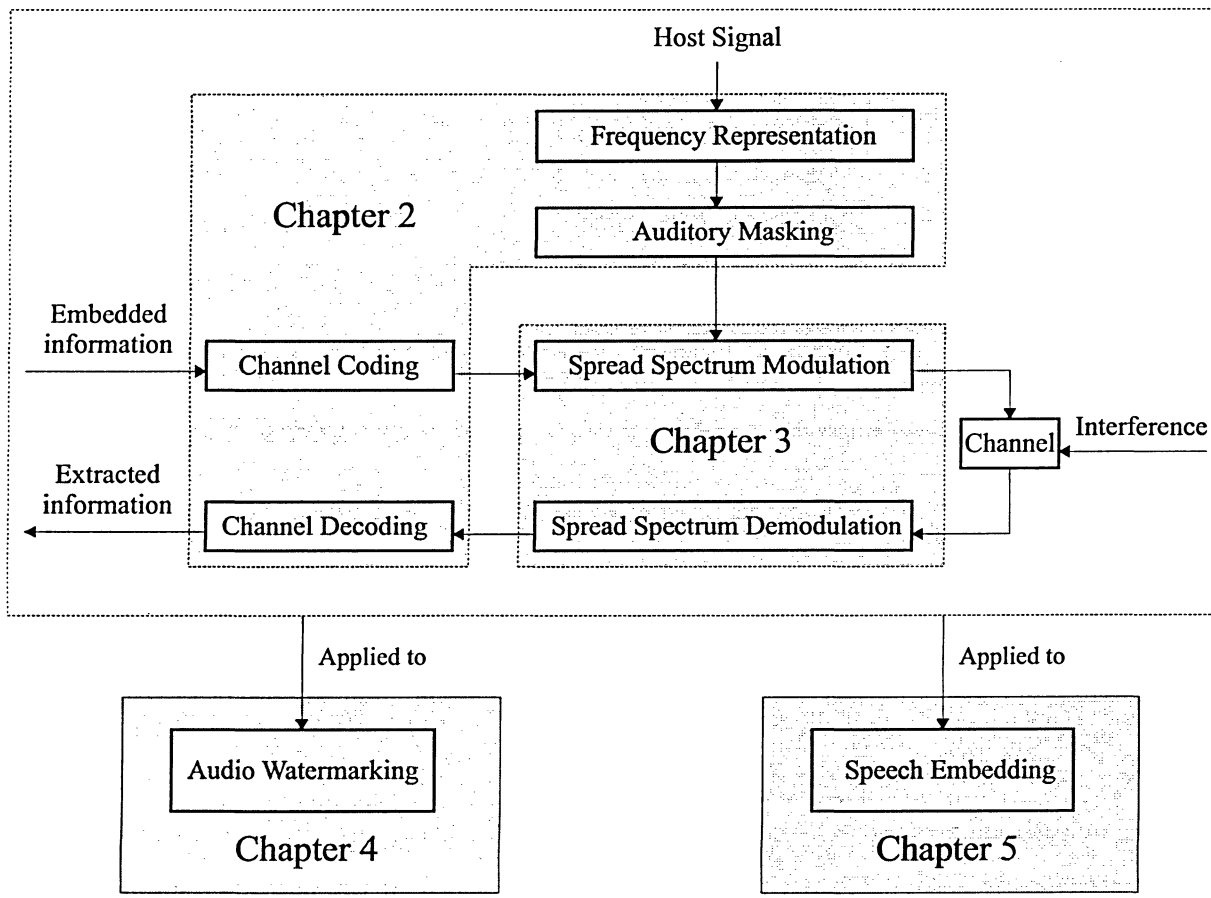

Figure 1.2: Block diagram of the thesis

Chapter 2

Preliminaries

This chapter reviews the background knowledge for the research of data embedding. As the transmission medium of the embedded data, the host signal and its different representation domains are discussed first. As an effective technique to increase the embedding power, the frequency masking effects of HAS are discussed in the second part. The relevant communication modulation technique, i.e. Spread Spectrum, and the FEC schemes are reviewed in the third part. At the last section, a separate part is devoted to Turbo code, where its powerful error-correction capabilities are investigated for data hiding applications.

2.1 Digital Sound Signals Representations

Information can be embedded into the different representation coefficients of the host signal. These different representations correspond to the different channels of the equivalent communication of data embedding.

2.1.1 Time Domain

According to the sampling theory, a continuous, band-limited signal, $x(t)$, can be adequately represented as a discrete-time signal provided that the discrete signal is uniformly sampled at a rate at least twice of its bandwidth. A discrete-time signal is written in the form, $x(nT) = \boldsymbol{x}_n$, where T is the sampling period and n is the integer index into the sequence. A digital signal is a discretely sampled signal where each sample assumes a value from a

discrete range. Quantization is the process where a discrete amplitude value is given to each discrete sample in a digital signal.

Some common sampling rates for audio signals are 16, 20, 32, 44.1, 48, and 96 kHz [36]. The so called 'CD quality' format refers to the linear Pulse Code Modulation (PCM) format with a resolution of 16 bits and a sampling frequency of 44.1 kHz. This is used as the reference format for watermarking in this research.

2.1.2 Fourier Transform Domain

The Fourier transform is an integral transform of a continuous signal, $x(t)$, with the complex exponentials of radian frequencies, $e^{-j\omega}$, as the kernel sequences and defined as

$$F(j\omega) = \int_{-\infty}^{+\infty} x(t)e^{-j\omega t}\, dt \tag{2.1}$$

It can be seen from the definition that the Fourier transform measures the amount of energy at each radian frequency by integral, so it is commonly called the *spectrum* of the signal.

For a discrete signal, x_n, the Fourier transform becomes an infinite sum as

$$F(j\omega) = \sum_{n=-\infty}^{+\infty} x_n e^{-j\omega nT} \tag{2.2}$$

Thus for a finite duration time sequence, the Discrete Fourier Transform (DFT) is obtained from,

$$f_k = F(j\frac{2\pi}{N}k) = \sum_{n=0}^{N-1} x_n e^{-j\frac{2\pi}{N}nk}, \qquad k \in \{0, 1, 2, ..., N-1\} \tag{2.3}$$

Each index of the discrete frequency spectrum is referred to as a *frequency bin.* The magnitude of each bin is the amount of energy at the equivalent discrete frequency range.

The inverse DFT is written

$$x_n = \frac{1}{N}\sum_{k=0}^{N-1} f_k e^{j\frac{2\pi}{N}nk}, \qquad n \in \{0, 1, 2, ..., N-1\} \tag{2.4}$$

The Fast Fourier transform (FFT) can reduce computation of DFT from $O(N^2)$ complex multiplications to $O(N \log N)$ complex multiplications [32].

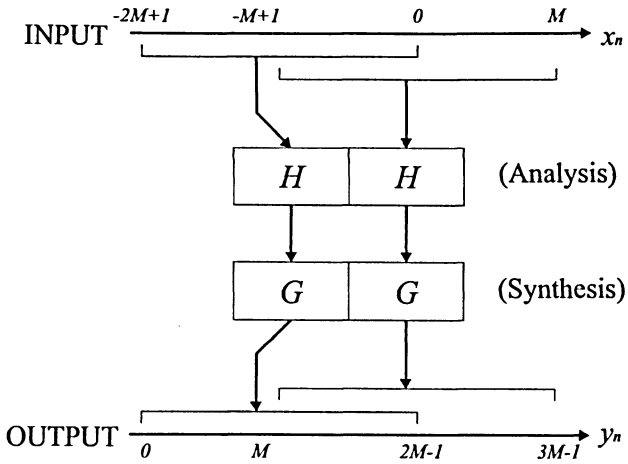

Figure 2.1: MDCT as a transform with 50% overlap between adjacent frames

2.1.3 Modified Discrete Cosine Transform (MDCT) Domain

Signal representation in the MDCT domain has emerged as a dominant tool in high quality audio coding because of its properties including energy compaction, critical sampling, reduction of block effect and flexible window switching [36].

The MDCT of a signal block $\boldsymbol{x}_n$ of length N is a block $\boldsymbol{f}_j$ of length $M = N/2$ and is defined by

$$f_j = \sqrt{\frac{2}{M}} \sum_{k=0}^{N-1} w_k \cos\left[\frac{\pi}{M}(j+\frac{1}{2})(k+\frac{M+1}{2})\right] x_k, \qquad j = 0, 1, ..., M-1 \tag{2.5}$$

where w_k is some analysis window to smooth the boundary effects between successive blocks.

The inverse MDCT is defined as,

$$y_k = w_k \sqrt{\frac{2}{M}} \sum_{j=0}^{M-1} \cos\left[\frac{\pi}{M}(j+\frac{1}{2})(k+\frac{M+1}{2})\right] f_j, \qquad k = 0, 1, ..., N-1 \tag{2.6}$$

Different than a block transform like the above DFT, MDCT is a lapped transform. As a lapped transform, the MDCT works with a 50% overlap between successive blocks of the input signal, this procedure can be described in Figure 2.1. The analysis and synthesis correspond to the MDCT and the inverse MDCT in implementation, respectively.

The following *Princen-Bradley* [38] [39] conditions guarantee that the original signal can be perfectly reconstructed (PR) by adding the inverse MDCT of subsequent overlapping blocks, causing the errors to cancel and the original data to be reconstructed.

$$w_k = w_{N-1-k} \tag{2.7a}$$

$$w_k^2 + w_{k+M}^2 = 1 \tag{2.7b}$$

The following *sine* window is widely used in audio coding, because it provides a good attenuation of the block boundary effect and allows PR [36].

$$w_k = \sin \frac{\pi}{2N}(k + \frac{1}{2}) \tag{2.8}$$

The direct computation of the MDCT formula would require $O(N^2)$ multiplications, as for FFT, it is possible to compute the MDCT with only $O(N \log N)$ complexity by recursively factorizing the computation. Actually the MDCT and the inverse MDCT can be calculated using only one $n/4$ point FFT and some pre- and post-rotation of the sample points [10].

In [1], a relationship between the MDCT and the FFT was shown as follows,

$$f_j^{MDCT} = \sqrt{\frac{2}{M}} |f_{j+\frac{1}{2}}^{FFT}| \cos [\frac{(M+1)(j+\frac{1}{2})\pi}{N} - \angle f_{j+\frac{1}{2}}^{FFT}] \tag{2.9}$$

It can be seen that MDCT coefficients are approximately the corresponding DFT ones with a rapid modulation. This similarity indicates that the MDCT-based psychoacoustic model can be borrowed directly from the corresponding one based on Fourier coefficients. For this reason, as well as for reduced complexity (MDCT is a real transform, while DFT is a complex one), the MDCT is used as the frequency embedding domain in this research. Some other possible representations include the DCT, Wavelet, Wavelet Packet.

2.2 Human Auditory System (HAS)

HAS has a dynamic frequency range of $20Hz \sim 20kHz$ and an intensity range from $20\mu Pa$ to $20Pa$ [48]. Sounds are commonly characterized by their logarithmic level, i.e. Sound Pressure Level (SPL),

$$L = 20 \log_{10} \frac{p}{p_0} \qquad (dB) \tag{2.10}$$

where the reference pressure, p_0, has a value of $20\mu Pa$.

It is generally convenient to evaluate the level of a sound from its frequency domain representation. For discrete spectra of periodic signals, the overall level is calculated by summing the levels of individual spectral components. Individual component levels are directly related to the squared magnitude of the Fourier series coefficients of the signal.

Perceptual embedding is challenging due to the wide dynamic range and high sensitivity of HAS to Additive White Gaussian Noises (AWGN); such noises as low as 70 dB below the ambient can be perceived.

2.2.1 Critical Band

Sounds are not perceived equally well at all frequencies, the concept of *critical band* is proposed to explain this phenomenon [48]. A critical band is a bandwidth around a center frequency within which sounds of different frequencies are blurred as perceived by us. The critical band itself is a function of frequency and two adjacent bands have the difference of 1 in *Bark* unit.

Thus the HAS is usually modeled as a bandpass filterbank consisting of strongly overlapped bandpass filters. Within the spectrum high up to 22.05 kHz, 26 critical bands have to be taken into account [37], and this subdivision is used in this research.

2.2.2 Auditory Masking

Auditory masking is the process by which the perception of one sound, i.e. the *maskee*, is suppressed by another one, i.e. the *masker*. Masking is characterized by an increase in the audibility threshold of the maskee in the presence of the masker. The amount of masking corresponds to the quantity by which the threshold is augmented above the Threshold in Quiet (TiQ) curve [36].

Frequency masking, or *simultaneous* masking, occurs when the masker and maskee are presented to the ear concurrently. Actually, this accounts for the most masking effects of HAS. Another kind, i.e. *temporal* masking, occurs when the masker and the maskee have

a temporal offset with each other. In this research only the frequency masking effects are considered for implementation reasons.

2.2.3 Masking Model

Auditory masking effects enable the noises up to some extent while still inaudible. Different masking models are proposed to compute this masking ability. As an example, the popular Model 1 used in MPEG-1 Layer 3 [36] [37] is described below and used in this research.

Step 1. The power spectrum is obtained from the FFT of the input signal and normalized to an anticipated playback level of $96dB$ SPL. The whole spectrum is then subdivided into 26 non-uniform subbands.

Step 2. Each subband may have several tonal maskers and one equivalent noise masker. Tonal components are identified through the detection of local maxima within the predefined neighbors. The SPLs of a tonal masker is computed from the sum of its neighbors. The energy left in one critical band are summed together to yield an equivalent noise masker.

Step 3. The number of maskers considered for threshold computation needs to be reduced. At first, only maskers having an SPL above the TiQ are retained. A decimation process then occurs between multiple tonal maskers that lie within half of a critical band. The tonal masker having the highest level is maintained while the other elements are removed from the maskers.

Step 4. The masking abilities of a masker is represented by a spreading function. The shapes and parameters of the spreading function are determined by the masker's type.

Step 5. A global masking threshold is computed by summing the individual masking contributions from each masker along with the TiQ. A Signal-to-Mask-Ratio (SMR) is calculated by subtracting the global masking threshold from the signal power in each sub-band.

The initial intention of computing the SMR is to remove those perceptually irrelevant components or quantize in such a way that the quantization noise level remains below the masking threshold [27]. This can be shown clearly in the Figure 2.2. For the case of perceptual embedding, the embedded information is virtually the noises injected into the host

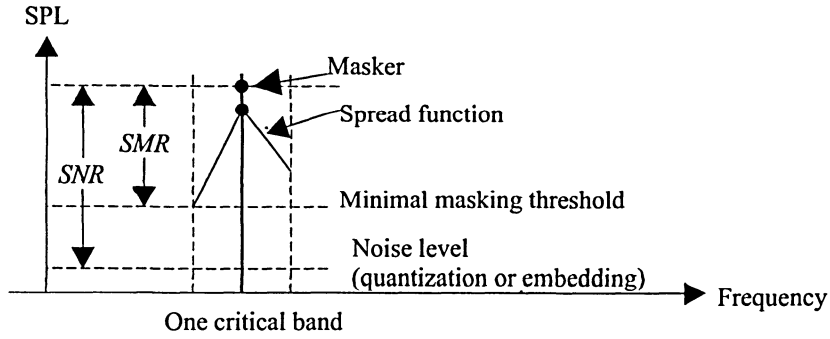

Figure 2.2: Masking threshold

signal. Similarly, to be inaudible, its Signal-Noise-Ratio (SNR) should be bigger than SMR. i.e. $SNR > SMR$. In other words, the noise level should be below the global masking threshold.

2.3 Telecommunication Model

A general telecommunication, including the case of data embedding, highly depends on the aspects like channel models, modulation techniques and channel coding schemes.

2.3.1 Channel Model

A memoryless channel means that the noise affecting one received bit is independent from the noises affecting other received bits. The theoretical definition of channel capacity of a memoryless channel is maximally mutual information between the input and output over all possible input distributions; the operational definition of channel capacity is the highest rate in bits per channel use at which information can be sent with arbitrarily low probability of errors. Shannon's channel coding theorem establishes that these two capacities are equal to each other, so the operational channel capacity serves as a good measure of the potential for transmission [12].

In an Additive White Gaussian Noising (AWGN) channel with the pulse energy of $E > 0$, the received bit is $r = \sqrt{E}b + n$, where n is Gaussian random variable with zero-mean,

$E[n_j] = 0$, and variance $E[n_j^2] = \sigma^2$, i.e.

$$p(n) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{n^2}{2\sigma^2}} \tag{2.11}$$

In this case, the optimal demodulation is a simple level detector: $b = sign(r)$. The operational capacity of an AWGN channel with power constraint E is given by [12],

$$C = \frac{1}{2}\log_2\left(1 + \frac{E}{\sigma^2}\right) \tag{2.12}$$

The Bit Error Rate (BER) of the system is determined by the SNR ratio E/σ^2 as follows,

$$p = Q\left(\frac{\sqrt{E}}{\sigma}\right) = Q\left(\sqrt{SNR}\right) \tag{2.13}$$

where the function $Q(\cdot)$ is the complementary error function defined as follows,

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} e^{-\frac{u^2}{2}}\, du = \frac{1}{2} erfc(\frac{x}{\sqrt{2}}) \tag{2.14}$$

where the function $erfc(\cdot)$ is the corresponding one defined in MATLAB.

The simplest way to prevent errors is to repeat the transmitted bits, this corresponds to virtually increasing the SNR. For example, for one information bit, a sequence of N pulses is transmitted, that is, if $b = i$, then $\{i, i, ..., i\}$ is transmitted, where $i \in \pm 1$, and the pulse energy is still $\sqrt{E}$. The received sequence is, $\boldsymbol{r} = \sqrt{E}b + \boldsymbol{n}$. The optimal receiver employs a correlation as,

$$y = \frac{1}{N}\sum_{j=0}^{N-1} r_j = \sqrt{E}b + \frac{1}{N}\sum_{j=0}^{N-1} n_j \tag{2.15}$$

Thus the decision variable y has mean $\sqrt{E}$ and variance σ^2/N. Now the performance is determined by the increased SNR of $N\sqrt{E}/\sigma$, which is N times of that in non-repetition case.

For some noisy channels repetition can not be avoided since the SNR is too low. It is the last resort to increase the SNR up to some threshold so that some FEC scheme could take effective.

2.3.2 Spread Spectrum

The spread-spectrum is a repetition of the same information bit according to some deterministic pattern, rather than a simply identical repetition [23]. The pattern is known as the *spreading sequence* $\boldsymbol{c}$, which satisfies the following three conditions,

$$Periodity: \sum_{n=0}^{N-1} c_n c_{n+N} = N \tag{2.16}$$

$$Zero-mean: \frac{1}{N}\sum_{n=0}^{N-1} c_n = 0 \tag{2.17}$$

$$Orthogonality: \frac{1}{N}\sum_{i=0}^{N-1} c_i c_{m+i} = \begin{cases} 1, & \text{if } m = 0 \\ 0, & \text{if } 0 < |m| < N \end{cases} \tag{2.18}$$

The receiver for the spread-spectrum signal performs the following correlation:

$$y = \frac{1}{N}\sum_{i=0}^{N-1} r_i c_i = \sqrt{E}b + \frac{1}{N}\sum_{i=0}^{N-1} n_i c_i \tag{2.19}$$

Hence, the decision variable is again Gaussian with mean $\sqrt{E}$ and variance σ^2/N. This shows that the spreading yields no improvement in the ideal AWGN channel. But if the channel contains an interferer: an unknown constant I is added to the received signal. It is easy to show that the decision variable for the non-spread system would have a mean of $N(Eb + I)$, which will render the system unusable for $|I|$ sufficiently large. On the other hand, for the spread system, the received sequence is

$$r_j = \sqrt{E}b + I + n_j, \qquad j = 0, ..., N-1 \tag{2.20}$$

Then the correlation receiver produces the decision variable

$$y = \frac{1}{N}\sum_{j=0}^{N-1} r_j c_j = \sqrt{E}b + \frac{I}{N}\sum_{j=0}^{N-1} c_j + \frac{1}{N}\sum_{j=0}^{N-1} n_j c_j = \sqrt{E}b + 0 + \frac{1}{N}\sum_{j=0}^{N-1} n_j c_j \tag{2.21}$$

So the interference is suppressed by the de-spreading (correlation) operation. It can be seen that Spread Spectrum is superior to the simple repetition technique since, in addition to increasing the SNR, the former also suppresses the constant interference. So in this research, when repetition is needed, spread spectrum techniques is always used instead.

2.4 FEC Schemes

FEC, or channel coding schemes, are used in digital communication systems to protect the channel from noise and interference. They are mostly accomplished by selectively introducing redundant bits into the transmitted information stream. These additional bits will allow detection and correction of bit errors in the received data stream. The cost of using channel coding is a reduction in data rate or an expansion in bandwidth.

There are three basic types of channel codes, namely block codes, convolutional codes and concatenated codes.

Block code accepts a block of k information bits and produce a block of n coded bits. By predetermined rules, $n - k$ redundant bits are added to the k information bits to form the n coded bits. Some commonly used block codes include Hamming codes, Bose-Chaudhuri-Hochquenghem (BCH) codes.

Convolutional code converts the entire data stream into one single codeword. The encoded bits depend not only on the current k input bits but also on past input bits. The main decoding strategy for convolutional codes is Viterbi algorithm.

Concatenated code consists of two separate codes that are combined to form a larger code. The concatenation could be used to develop a powerful code using relatively weaker codes since the minimum distance of the concatenated code is typically larger than the minimum distances of the inner and outer codes. Turbo code is a concatenation of two convolutional codes and it has been shown that it can achieve the performance within 1 dB of Shannon's capacity [5] [6].

2.4.1 Error Correction Abilities

The usefulness of channel coding schemes can be represented by the concept of *coding gain.* For a given BER, the coding gain is defined as the reduction in SNR that can be realized through the use of a channel code as shown in Figure 2.3. The coding gain, G can be expressed as,

$$G = SNR_{uncoded} - SNR_{coded} \qquad dB \tag{2.22}$$

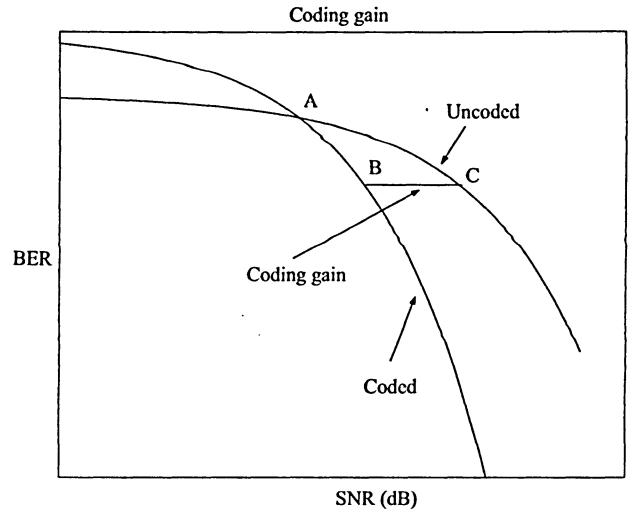

Figure 2.3: Coding gain of channel codes

Every channel code has a specified error correcting threshold. When the channel SNR exceeds some threshold, the code can correct virtually all or most errors. On the other hand, when the channel SNR is below this threshold, the decoder fails catastrophically and the decoded bitstream appears random. This is shown in Figure 2.3 by the crossover between the coded and uncoded curves (point A). Turbo code can provide performance improvements at low SNR.

When the received channel values are used directly or multi-bits quantized, by the channel decoder, this is called *soft-decision* decoding. Alternatively, *hard-decision* decoding uses 1-bit quantization on the received channel values. Soft value could be expressed succinctly by one number, i.e. *Log-Likelihood Ratio* (LLR) defined as follows,

$$L(u_i) = \log \frac{P(u_i = +1)}{P(u_i = -1)} \qquad or \qquad L(u_i|y) = \log \frac{P(u_i = +1|y)}{P(u_i = -1|y)} \tag{2.23}$$

The sign of the number corresponds to the hard decision of the bit, i.e. $x = sign(LLR)$, while the magnitude gives a reliability estimate.

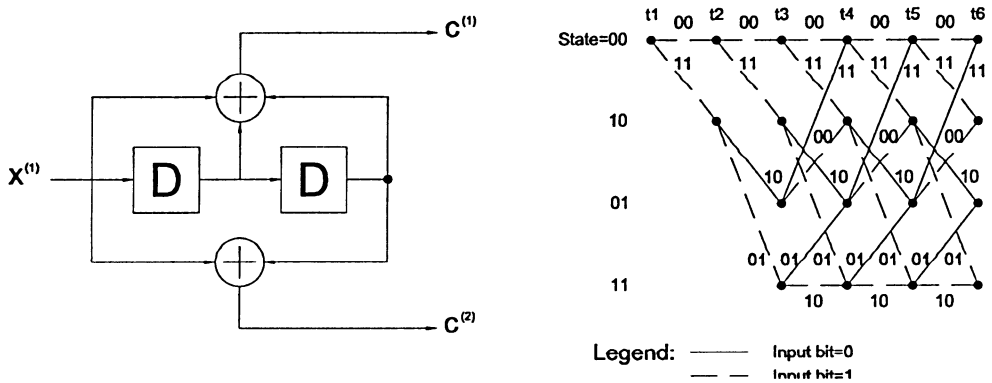

Figure 2.4: Convolutional code and its trellis, generator=[7,5], k=1, n=2, K=3

2.4.2 BCH Code

BCH codes are the most important class of linear block codes including both binary and non-binary alphabets. Binary BCH codes can be constructed with parameters (n, k, t), where $n = 2^m - 1$ is the length of the code word, k is the length of the embedded data and t is the number of bit errors the BCH code can correct. The constraints are $m \geq 3$, $n - k \leq mt$ and t are arbitrary integers, e.g. BCH(7,4,1), BCH(127, 64, 10).

BCH can be decoded very efficiently by the so called syndromes decoding algorithms [31]. According to [43], the coding gain of BCH(127, 64, 10) can reach a gain of about 3.3 dB at the BER of $p = 10^{-5}$ over the Gaussian channel.

2.4.3 Convolutional Code

A convolutional code introduces redundant bits into the data stream through the use of Linear Shift Registers (LSR). The input parallel information bits are input into LSR and the parallel output encoded bits are obtained by modulo-2 addition of the input information bits and the contents of the shift registers. The shift registers store the state information of the encoder. An example convolutional encoder and its trellis is shown in Figure 2.4.

Optimum decoding of a convolutional code involves a search through the trellis for the most probable sequence. For a general convolutional code, the input information sequence contains kL bits, where k is the number of parallel information bits at one time interval

and L is the number of time intervals. There are exactly 2^{kL} distinct paths in the trellis diagram, and as a result, an exhaustive search for the most probable sequence would have a computational complexity on the order of $O(2^{kL})$ [43].

Viterbi algorithm reduces this complexity based on the observation that, if any two partial paths in the trellis merge to a single state, one of them could always be eliminated. A partial path metric is determined from the starting state $s = 0$ at time $t = 0$ to some particular state $s = k$ at time $t \geq 0$. At each state, the minimal partial path metric is chosen from the paths terminated at that state. The selected metric represents the survivor path and are stored, while the remaining metrics represent the non-survivor paths and are discarded. Trace-back of the survivor path would provide the decoded sequence.

The Viterbi algorithm reduces the complexity by performing the most probable search one stage at a time in the trellis. Its complexity is on the order of $O(2^k * L)$. This significantly reduces the number of calculations because the number of time intervals L is now a linear factor in the complexity.

The performance of convolutional codes, in terms of coding gain, ranges from $3 \sim 7.5$ dB, which depends on the specific structures [43].

2.5 Turbo Code

Turbo encoder is built using two identical Recursive Systematic Convolutional (RSC) coders with parallel concatenation, as shown in Figure 2.5. An RSC encoder is obtained from a non-recursive one by feeding one encoded outputs back to the input. An RSC encoder tends to produce codewords with increased weight relative to a non-recursive one. This results in fewer codewords with lower weights and leads to better error performance [25].

The input sequence $\boldsymbol{x}$ produces a low-weight recursive convolutional code sequence $\boldsymbol{c_1}$ from encoder #1. To avoid having encoder #2 produce another low-weight recursive output sequence, an interleaver permutes the input sequence $\boldsymbol{x}$ to obtain a different sequence that hopefully produces a high-weight recursive convolutional coded sequence $\boldsymbol{c_3}$. Thus, the Turbo code's weight is moderate, combined from encoder #1's low-weight code and encoder #2's

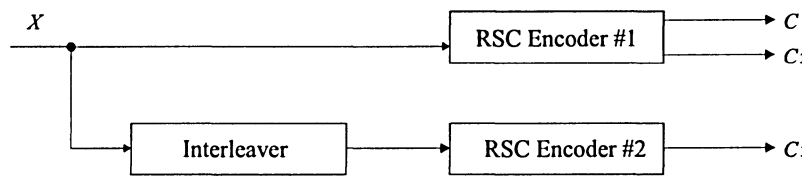

Figure 2.5: Structure of Turbo Encoder

high-weight code.

Optionally, a puncturer could be used to raise the data rate of Turbo code by periodically deleting selected bits. Without the puncturer, $3N$ bits would be transmitted for every N bits which results in a data rate of only $1/3$, however, for the case of puncturers included, the date rate is raised to $1/2$ with a minor loss of performance.

The decoder is a concatenation of two cooperating component convolutional decoders, but in a serial way. The basis for this cooperation is the *soft output* from each component decoder. The soft value represents the certainty of one decoder decodes one bit, and such information could be used by the other decoder as *a priori* and thus improve the decoding reliability. There are mainly two types of component decoders. i.e. Maximum A Priori (MAP) algorithm and Soft Output Viterbi Algorithm (SOVA).

2.5.1 MAP algorithm

MAP is a symbol-by-symbol decoding algorithm [2]. It examines every possible path through the trellis to minimize the BER and outputs *a posteriori* LLR for each decoded bit u_i. The main idea of MAP algorithm is that, as shown in Figure 2.6, the trellis state transition, as a Markov process, can be factored into three independent parts as follows,

$$P(s_i \rightarrow s_{i+1}, y) = \alpha(s_i)\gamma(s_i \rightarrow s_{i+1})\beta(s_{i+1}) \tag{2.24}$$

In the above, $\alpha(s_i)$ represents the probability that the current state is s_i and is called *forward* state metric. $\gamma(s_i \rightarrow s_{i+1})$ represents the probability of the state transition $s_i \rightarrow s_{i+1}$ given the current state is s_i and is called *branch* metric. $\beta(s_{i+1})$ represents the probability

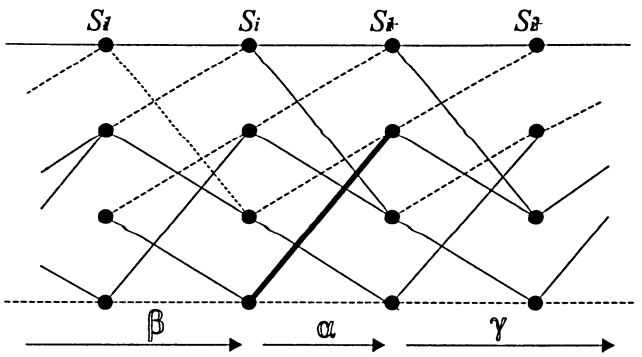

Figure 2.6: MAP decoder trellis

that the next state is s_{i+1} and is called *backward* state metric. According to the definitions, the metrics α, β can be computed recursively. The branch metric γ is given by the product of *a priori* probability of the current input bit and the probability of receiving y_i given the codeword x_i was transmitted. Generally, it is a function of the modulation and channel characteristics.

The above MAP algorithm can be simplified greatly if transformed into log domain, which is called log-MAP, based on the following relation.

$$\log(e^x + e^y) = \max(x, y) + f_c(|x - y|) \tag{2.25}$$

The so called Max-Log-MAP algorithm just ignores the correction item f_c and yields a slight degradation of about 0.35dB in performance compared to the log-MAP algorithm [40].

2.5.2 SOVA algorithm

SOVA [24] minimizes the probability of an incorrect path through the trellis. In this case, the soft output is the Euclidean distance associated with the sequence of the received symbols, as opposed to the individual symbols in the above MAP algorithm. In [24], the authors derived the reliability metric for each decoded symbol from the sequence metrics of the classical Viterbi algorithm.

SOVA is based on the observation that the probability that a hard decision on a given

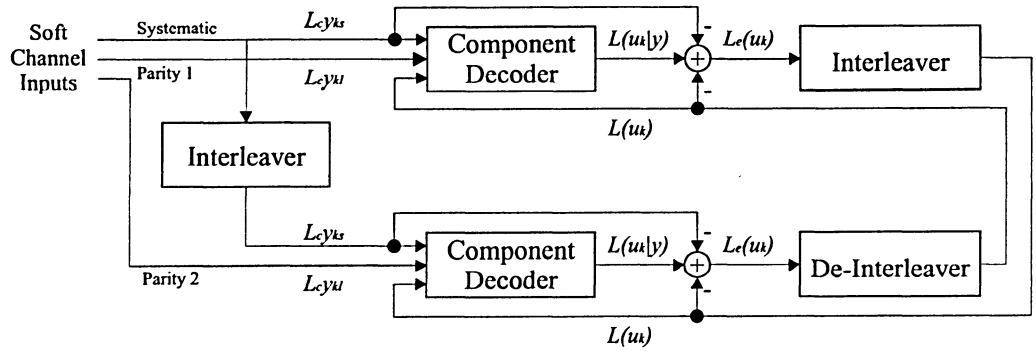

Figure 2.7: Iterative decoding of Turbo code

symbol in the Viterbi algorithm is correct is proportional to the difference in path metrics between the survivor sequence and the associated non-survivor sequences. This observation forms an estimate of the error probability, or the probability of a correct decision for each symbol by comparing the path metrics of the surviving path with the path metrics of non-surviving paths.

It has been shown that SOVA algorithm is about half as complex as the Max-Log-MAP algorithm. However, the SOVA algorithm is also the least accurate one. When used in an iterative Turbo decoder, it performs about 0.6dB worse than MAP [40].

2.5.3 Iterations

It is shown in [25] that the soft output can be expressed as three additive terms as

$$L(u_k|y) = L_c y_{ks} + L(u_k) + L_e(u_k) \tag{2.26}$$

where $L_e(u_k)$ is derived from the *a-posteriori* information $L(u_k|y)$ sequence and the received channel information sequence y, excluding the received systematic bits y_{ks} and the *a-priori* information $L(u_k)$ for the bit u_k. Hence it is called the extrinsic LLR for the bit u_k. This extrinsic information is actually supplied by the constraints imposed on the transmitted sequence by the code. So this extrinsic information can be used as the *a-priori* information for another component decoder after interleaved.

For each bit u_i, decoder #1 receives soft information from decoder #2 and uses it as *a-priori* information. Similarly, decoder #2 receives soft information from decoder #1 and the decoding iteration proceeds as #1 $\rightarrow$ #2 $\rightarrow$ #1 $\rightarrow$ #2 $\rightarrow$ ... with the previous decoder passing soft information along to the next decoder at each half iteration. The idea behind extrinsic information is that decoder #2 provides soft information to decoder #1 for each u_k, using only information not available to decoder #1, i.e., encoder #2 parity; decoder #1 does likewise for decoder #2.

As the number of iterations increases, the decoders become more certain about the values of the bits and hence the magnitudes of the LLRs gradually becomes larger, thus the improvement in performance for each additional iteration carried out falls as the number of iterations increases. Normally around six to eight iterations are carried out, as no significant improvement in performance is obtained with higher number of iterations.

2.5.4 Complexity

The MAP algorithm is extremely complex due to the multiplications needed for the recursive calculation and the multiplication and natural logarithm operations required to calculate LLR. However, the Log-MAP algorithm gives the same performance as the MAP algorithm at a significantly reduced complexity and without the numerical problems described above. Viterbi states [46] that the complexity of the Log-MAP-Max algorithm is no greater than three times that of a Viterbi decoder. According to [40], the SOVA algorithm is about half as complex as the Max-Log-MAP algorithm.

Chapter 3

Spread Spectrum Embedding

3.1 Conventional Spread Spectrum

Perceptual embedding can be viewed as a spread spectrum communication problem due to its low-energy property and robustness against interference [13]. This embedding scheme in its basic form can be described as follows.

A bipolar information bit, $b \in \{\pm 1\}$, is spread by a spreading sequence, $\boldsymbol{w}$, of the length N, which is called the *spreading factor*, to generate the spread information, $b\boldsymbol{w}$; the host signal, $\boldsymbol{x}$, of the same length is embedded with this spread information in an additive way resulting in the composite signal, $\boldsymbol{y} = \boldsymbol{x} + \alpha b\boldsymbol{w}$, where the *perceptual factor*, α, controls the perceptibility of the embedded information. After distorted by the transmission noises, $\boldsymbol{n}$, the received signal can be expressed as,

$$\boldsymbol{r} = \boldsymbol{x} + \alpha b\boldsymbol{w} + \boldsymbol{n} \tag{3.1}$$

To extract the embedded information, the normalized correlation between the received signal, $\boldsymbol{r}$, and the spreading sequence, $\boldsymbol{w}$, is computed as follows,

$$c = \boldsymbol{r} \cdot \boldsymbol{w} = (\boldsymbol{x} + \alpha b\boldsymbol{w} + \boldsymbol{n}) \cdot \boldsymbol{w} = \alpha b + (\boldsymbol{x} + \boldsymbol{n}) \cdot \boldsymbol{w} \tag{3.2}$$

where the normalized correlation of two length-N vectors $\boldsymbol{u}$, $\boldsymbol{v}$ is defined as follows,

$$\boldsymbol{u} \cdot \boldsymbol{v} = \frac{1}{N} \sum_{i=1}^{i=N} u_i v_i \tag{3.3}$$

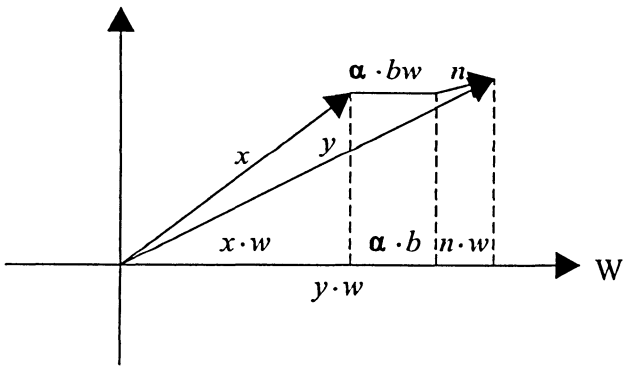

Figure 3.1: Data embedding with conventional Spread Spectrum

Assuming the Gaussian distributions of $\boldsymbol{x} \sim \boldsymbol{N}(0, \sigma_x^2)$ and $\boldsymbol{n} \sim \boldsymbol{N}(0, \sigma_n^2)$, the distribution of the correlation is also Gaussian as $c \sim \boldsymbol{N}(\alpha b, \frac{\sigma_x^2+\sigma_n^2}{N})$, the deduction can be referred to Appendix B. When the spreading factor, N, is big enough, the variance $\frac{\sigma_x^2+\sigma_n^2}{N}$ will decrease to zero, thus αb dominates the correlation c and the embedded information bit can be extracted by $b' = sign(c)$. This strategy is shown in Figure 3.1 geometrically. It can be seen that the idea of this scheme is to make the embedded information dominate the projection of the composite signal on the specific direction. This requires some considerable embedding energy or, equivalently, a considerably large spreading factor.

The BER of this Gaussian channel can be easily derived as follows,

$$p = Q\left(\frac{m_c}{\sqrt{\sigma_c}}\right) = Q\left(\sqrt{\frac{N\alpha^2}{\sigma_x^2 + \sigma_n^2}}\right) \tag{3.4}$$

For simplicity, the following ratios are defined

$$SNR = \frac{Signal}{Noise} \Rightarrow 10\log_{10}\frac{\sigma_x^2}{\sigma_n^2} \qquad (dB) \tag{3.5a}$$

$$WNR = \frac{Data}{Noise} \Rightarrow 10\log_{10}\frac{\alpha^2}{\sigma_n^2} \qquad (dB) \tag{3.5b}$$

$$SWR = \frac{Signal}{Data} \Rightarrow 10\log_{10}\frac{\sigma_x^2}{\alpha^2} \qquad (dB) \tag{3.5c}$$

Thus Equ. (3.4) can be re-written as follows,

$$p = Q\left(\sqrt{(N \cdot WNR) \cdot \frac{1}{1+SNR}}\right) \tag{3.6}$$

From Equ. (3.4), it can be seen that both the host signal $\boldsymbol{x}$ and the attacks $\boldsymbol{n}$ contribute to the noises of the embedding channel. Comparing Equ. (3.6) with the bound performance in Equ. (2.13), it can be seen that the gap is $\frac{1}{1+SNR}$ that depends on the value of SNR.

Since the perceptual factor α is constrained by the perceptual transparency requirement, a large N is always needed to decrease the variance of the correlation in Equ. (3.2). But this reduces the embedding bit rate correspondingly as well. For this reason, some efforts are made to suppress the host impact by different "whitening" procedures.

One of them utilizes the Linear Predictive Coding (LPC) [42] on the audio signals. In speech coding, LPC models a sound as an autoregressive random process, thus decomposes it as the sum of a parametric model and a white noise process. The watermark can be thought of as a white noise due to the randomness of the spreading sequence and should only exist in the white noise part after LPC decomposition. So the received signal can be written as $\boldsymbol{r} = LPC(\boldsymbol{r}) + \boldsymbol{n} + \boldsymbol{w}$, only the residual $\boldsymbol{r} - LPC(\boldsymbol{r})$ may contain the watermark and the channel attacks. Using this residual in the above correlation will result in a smaller noise term and could increase the extraction. Another estimation is based on polynomial curve fitting [16]. The received is thought as the sum of a spline curve and the noises. Similar to the above, only the residual $\boldsymbol{r} - fitting(\boldsymbol{r})$ is used in the correlation.

However, these schemes are theoretically ambiguous, because the estimation of the host signal is based on the corrupted and watermarked signal. Such estimations will definitely introduce new noises in addition to the existing attacks. In other words, there is no guarantee that $LPC(\boldsymbol{x}) = LPC(\boldsymbol{r})$ with LPC estimation, so the residual $\boldsymbol{r} - LPC(\boldsymbol{r})$ is actually $\boldsymbol{n} + \boldsymbol{w} + LPC(\boldsymbol{r}) - LPC(\boldsymbol{x})$, i.e. a new noise $LPC(\boldsymbol{r}) - LPC(\boldsymbol{x})$ is introduced into the correlation.

3.2 Spread Spectrum of Improved Robustness

It is desirable to develop an embedding method so that the host impact can be suppressed more accurately at the extraction. The first thought is to introduce another auxiliary sequence $\boldsymbol{v}$ that is orthogonal to the spreading sequence $\boldsymbol{w}$, i.e. $\boldsymbol{w} \cdot \boldsymbol{v} = 0$. Embedding this auxiliary sequence with some amplitude β into the host signal as,

$$\boldsymbol{r} = \boldsymbol{x} + \alpha_M b\boldsymbol{w} + \beta\boldsymbol{v} + \boldsymbol{n} \tag{3.7}$$

In order to suppress the host effect, the factor β can be determined by computing the following two correlations,

$$\boldsymbol{r} \cdot \boldsymbol{w} = (\boldsymbol{x} + \alpha_M b\boldsymbol{w} + \beta\boldsymbol{v} + \boldsymbol{n}) \cdot \boldsymbol{w} = \alpha_M b + (\boldsymbol{x} + \boldsymbol{n}) \cdot \boldsymbol{w} \tag{3.8}$$

$$\boldsymbol{r} \cdot \boldsymbol{v} = (\boldsymbol{x} + \alpha_M b\boldsymbol{w} + \beta\boldsymbol{v} + \boldsymbol{n}) \cdot \boldsymbol{v} = \beta + (\boldsymbol{x} + \boldsymbol{n}) \cdot \boldsymbol{v} \tag{3.9}$$

If we choose $\beta = \boldsymbol{x} \cdot (\boldsymbol{w} - \boldsymbol{v})$, the correlation difference becomes

$$c = \boldsymbol{r} \cdot \boldsymbol{w} - \boldsymbol{r} \cdot \boldsymbol{v} = \alpha_M b + \boldsymbol{n} \cdot (\boldsymbol{w} - \boldsymbol{v}) \tag{3.10}$$

Comparing with the conventional scheme in Equ. (3.1), the equation $b' = sign(c)$ still determines the embedded information bit, but now the host impact is suppressed with the help of the auxiliary sequence $\boldsymbol{v}$. That is, the distribution of the correlation $c \sim \boldsymbol{N}(\alpha b, \frac{2\sigma_n^2}{N})$. The BER of this Gaussian channel is as,

$$p = Q\left(\frac{m_c}{\sqrt{\sigma_c}}\right) = Q\left(\sqrt{\frac{N\alpha_M^2}{2\sigma_n^2}}\right) \tag{3.11}$$

Clearly, the embedded information power is $\alpha_M^2 + \frac{2\sigma_x^2}{N}$. For a fair comparison between this scheme and the conventional one, the embedding power must be same. That is,

$$\alpha_M^2 + \frac{2}{N}\sigma_x^2 = \alpha^2 \Rightarrow \alpha_M^2 = \alpha^2 - \frac{2}{N}\sigma_x^2 \tag{3.12}$$

Thus the BER in Equ. (3.11) can be re-written as follows,

$$p = Q\left(\sqrt{\frac{N\alpha^2 - 2\sigma_x^2}{2\sigma_n^2}}\right) = Q\left(\sqrt{(N \cdot WNR) \cdot \left[1 - \frac{2SWR}{N}\right] \cdot \frac{1}{2}}\right) \tag{3.13}$$

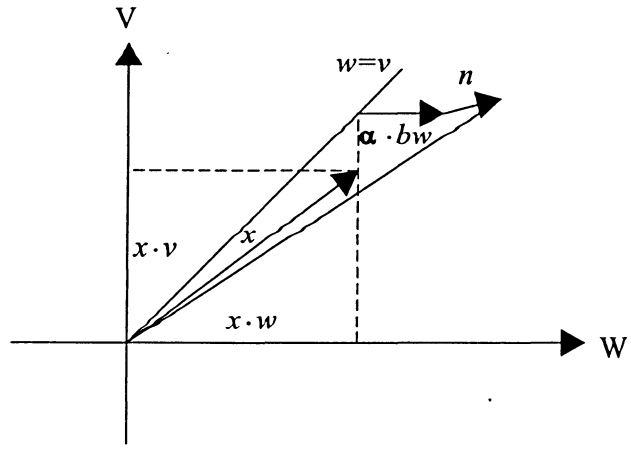

Figure 3.2: Initial Improvement on the conventional SS scheme

Though this scheme is far away from optimum, this idea of suppressing the host impact does improve the performance in terms of BER. For example, when SNR=5 dB and SWR=25 dB, with the spreading factor N=2000, the BER of the conventional technique is $p = 0.0142$; the BER of the proposed scheme is $p = 0.00446$ and $p = 3.872 \times 10^{-6}$ in the ideal bound case.

This algorithm can be explained geometrically in Figure 3.2. It can be seen that the proposed scheme actually maps the original point $(\boldsymbol{x} \cdot \boldsymbol{w}, \boldsymbol{x} \cdot \boldsymbol{v})$ to the point $(\boldsymbol{x} \cdot \boldsymbol{w}, \boldsymbol{x} \cdot \boldsymbol{w})$ on the line $\boldsymbol{w} = \boldsymbol{v}$ in the $\boldsymbol{w} - \boldsymbol{v}$ plane. This mapping results in a distortion of $\frac{2\sigma_x^2}{N}$; the information bit is then embedded into one direction as $\alpha b \cdot \boldsymbol{w}$. So the difference of the projections denotes the embedded information bit αb and the possible difference of the attacks in these two directions.

The perceptual degradation from the embedding algorithm occurs when it replaces the projection of $\boldsymbol{x} \cdot \boldsymbol{v}$ with its orthogonal projection $\boldsymbol{x} \cdot \boldsymbol{w}$. When the spreading factor N is large enough, in other words, in a very high dimensional space, the projection of a signal on any direction must be small enough such that the replace should not be audible.

However, if the original point $(\boldsymbol{x} \cdot \boldsymbol{w}, \boldsymbol{x} \cdot \boldsymbol{v})$ is mapped to the point $(\frac{\boldsymbol{x} \cdot \boldsymbol{w} + \boldsymbol{x} \cdot \boldsymbol{v}}{2}, \frac{\boldsymbol{x} \cdot \boldsymbol{w} + \boldsymbol{x} \cdot \boldsymbol{v}}{2})$, the projection on the line $\boldsymbol{w} = \boldsymbol{v}$, the worst distortion is only $\frac{\sigma_x^2}{N}$. Figure 3.3 shows this

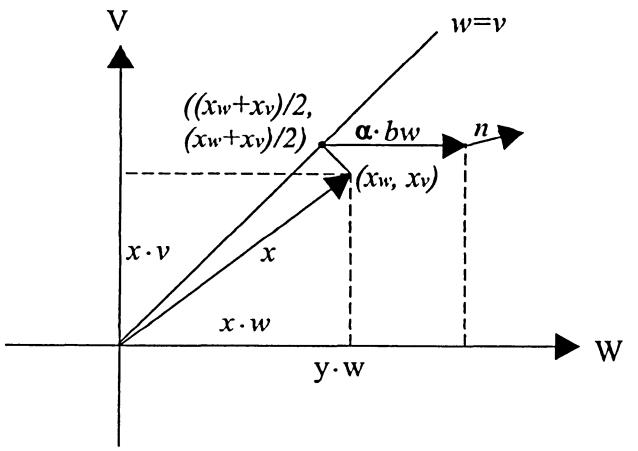

Figure 3.3: Further Improvement on the conventional SS scheme

procedure geometrically. Accordingly, the embedding is changed as follows,

$$\boldsymbol{r} = \boldsymbol{x} + \alpha_M b\boldsymbol{w} + \frac{\boldsymbol{x}\cdot\boldsymbol{v} - \boldsymbol{x}\cdot\boldsymbol{w}}{2}\boldsymbol{w} + \frac{\boldsymbol{x}\cdot\boldsymbol{w} - \boldsymbol{x}\cdot\boldsymbol{v}}{2}\boldsymbol{v} + \boldsymbol{n} \tag{3.14}$$

The BER of this Gaussian channel is as,

$$p = Q\left(\frac{m}{\sqrt{\sigma}}\right) = Q\left(\sqrt{\frac{N\alpha_M^2}{2\sigma_n^2}}\right) \tag{3.15}$$

The embedding power is now $\alpha_M^2 + \frac{\sigma_x^2}{N}$. To see the improvement of this scheme over the conventional scheme, the embedding powers must be same. That is,

$$\alpha_M^2 + \frac{\sigma_x^2}{N} = \alpha^2 \Rightarrow \alpha_M^2 = \alpha^2 - \frac{\sigma_x^2}{N} \tag{3.16}$$

Thus the corresponding BER is as follows,

$$p = Q\left(\sqrt{\frac{N\alpha^2 - \sigma_x^2}{2\sigma_n^2}}\right) = Q\left(\sqrt{(N\cdot WNR)\cdot\left(1 - \frac{SWR}{N}\right)\cdot\frac{1}{2}}\right) \tag{3.17}$$

In order to eliminate the factor $\frac{1}{2}$, we consider the case where the two directions are not limited to be orthogonal. In the following deduction, it is shown that two opposite vectors can actually cancel the factor and thus further increase the performance by $2 \simeq 3$dB.

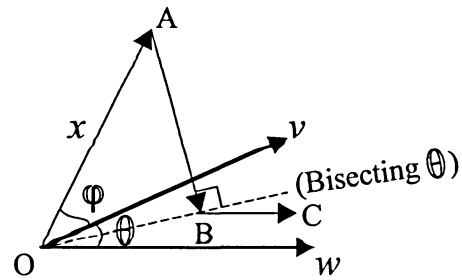

Figure 3.4: Suppressing the host impact with two random vectors

As shown in the Figure 3.4, the angle between the two vectors $\boldsymbol{w}$ and $\boldsymbol{v}$ is θ. The received signal and its projections are as follows,

$$\boldsymbol{r} = \boldsymbol{x} + \overrightarrow{AB} + \overrightarrow{BC} + \boldsymbol{n} = \overrightarrow{OB} + \alpha_M b\boldsymbol{w} + \boldsymbol{n} \tag{3.18}$$

$$\boldsymbol{r} \cdot \boldsymbol{w} = \overrightarrow{OB} \cdot \boldsymbol{w} + \alpha_M b + \boldsymbol{n} \cdot \boldsymbol{w} \tag{3.19}$$

$$\boldsymbol{r} \cdot \boldsymbol{v} = \overrightarrow{OB} \cdot \boldsymbol{v} + \alpha_M b(\boldsymbol{w} \cdot \boldsymbol{v}) + \boldsymbol{n} \cdot \boldsymbol{v} \tag{3.20}$$

The projections' difference is also Gaussian,

$$c = \boldsymbol{r}\cdot\boldsymbol{w} - \boldsymbol{r}\cdot\boldsymbol{v} = \alpha_M b(1-\boldsymbol{w}\cdot\boldsymbol{v}) + \boldsymbol{n}\cdot\boldsymbol{w} - \boldsymbol{n}\cdot\boldsymbol{v} \sim \boldsymbol{N}\left(\alpha_M b \cdot 2\sin^2\frac{\theta}{2},\ \frac{\sigma_n^2}{N}\cdot 4\sin^2\frac{\theta}{2}\right) \tag{3.21}$$

Clearly, when $\theta = \pi$, the SNR is optimized. The distortion still follows Equ. (3.16). So the performance is as follows,

$$p = Q\left(\sqrt{\frac{N\alpha_M^2 \cdot \sin^2\frac{\pi}{2}}{\sigma_n^2}}\right) = Q\left(\sqrt{\frac{N\alpha_M^2}{\sigma_n^2}}\right) = Q\left(\sqrt{(N \cdot WNR) \cdot \left(1 - \frac{SWR}{N}\right)}\right) \tag{3.22}$$

Using the same numerical example as above, the BER is $p = 2.036 \times 10^{-5}$ by this scheme, which corresponds to an improvement by an order of about 2 over the conventional one. Figure 3.5 compares the performance of the proposed scheme and the theoretical bound. It can be seen the proposed scheme suppresses the host impact significantly.

In [33], the authors proposed an Improved Spread Spectrum (ISS) as follows. The embedding considered as a slight perturbation as follows,

$$\boldsymbol{r} = \boldsymbol{x} + (\alpha b - \lambda(\boldsymbol{x} \cdot \boldsymbol{w}))\boldsymbol{w} + \boldsymbol{n} \tag{3.23}$$

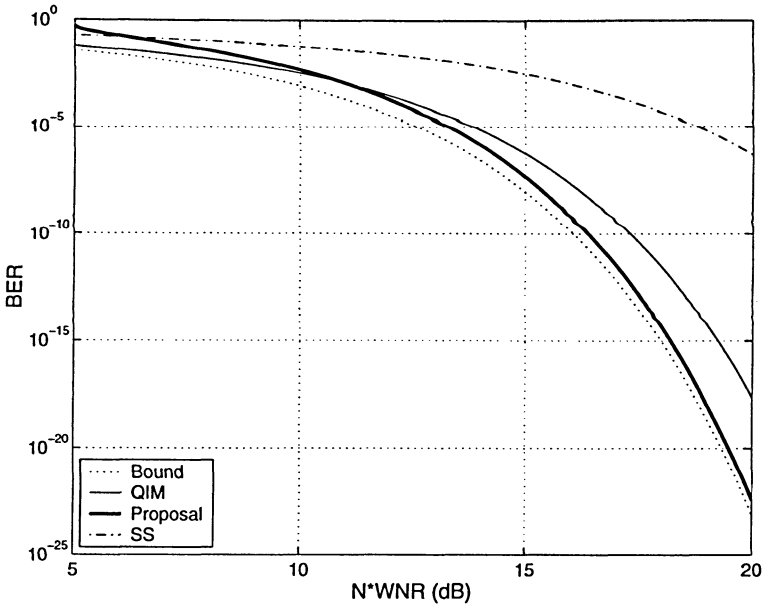

Figure 3.5: Theoretical comparison of embedding schemes for Gaussian signal and Gaussian attack (SWR=25dB, SNR=5dB)

The new item $\lambda(\boldsymbol{x} \cdot \boldsymbol{w})$ controls the suppression of the host signal at extraction. The traditional SS is a special case of ISS when $\lambda = 0$.

The decoding also uses the normalized correlation,

$$\boldsymbol{r} \cdot \boldsymbol{w} = \alpha b + (1 - \lambda)(\boldsymbol{x} \cdot \boldsymbol{w}) + \boldsymbol{n} \cdot \boldsymbol{w} \tag{3.24}$$

As λ tends to 1, the host impact is removed from the correlation. An analysis shows that the scheme is equivalent to the proposed scheme in Equ. (3.18).

Though equivalent in performance, our deduction shows more clearly that the improvement of the modified SS over the conventional one roots from the explicit use of the host-suppressing strategy. Without this deduction, the following further improvement to improve the embedding capacity would be impossible.

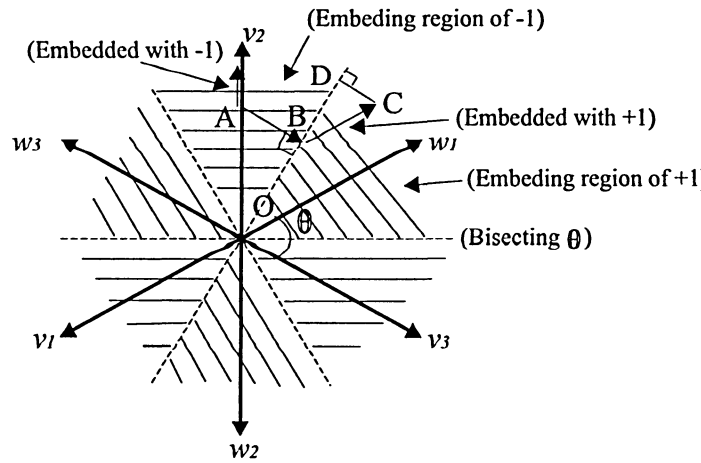

Figure 3.6: Dividing the signal plane with more vectors

3.3 Spread Spectrum of Improved Capacity

According to Equ. (3.22), the minimal spreading factor must satisfy $N > SWR$. This is because the embedding power, as shown in Figure 3.4, is at least $|\overrightarrow{AB}|^2 = \frac{\sigma_x^2}{N}$. This means that, even under weak attacks, the spreading factor N has to be large enough to make the embedding inaudible.

The embedding power of $\frac{\sigma_x^2}{N}$ is necessary when only one pair of vectors is used. Suppose now there are n (an odd number) pairs of vectors distributed uniformly, as shown in Figure 3.6 for n=3. By comparing the projections of the received signal onto these n directions, the extraction can be conducted by *the sign of the projection with maximum magnitude.* Thus the whole signal plane are divided into 6 interleaved embedding regions for a bipolar information bit when n=3.

The maximal embedding power, which occurs when the signal point locates on the center of the region reverse to the embedded information bit, is as follows,

$$W = |\overrightarrow{AB}|^2 + |\overrightarrow{BC}|^2 = \alpha_M^2 + \sin^2\frac{\theta}{2} \cdot \frac{\sigma_x^2}{N} \tag{3.25}$$

The deduction of the accurate performance is very complex and the result can not be

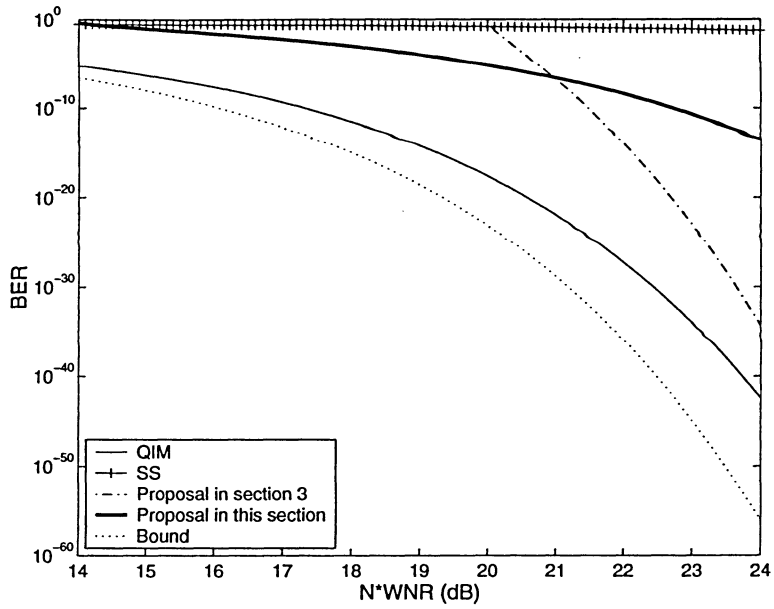

Figure 3.7: Theoretical comparison of embedding schemes for Gaussian signal and Gaussian attack (SWR=30dB, SNR=20dB)

expressed in a closed form. We simply consider the following worst case.

$$p = Q\left(\sqrt{\frac{|\overrightarrow{BC}|^2 \cdot \sin^2\frac{\theta}{2}}{\frac{\sigma_n^2}{N}}}\right) = Q\left(\sqrt{\frac{\alpha_M^2 \cdot \sin^2\frac{\theta}{2}}{\frac{\sigma_n^2}{N}}}\right) \tag{3.26}$$

Similarly, to embed with the same power as before,

$$\alpha_M^2 + \sin^2\frac{\theta}{2} \cdot \frac{\sigma_x^2}{N} = \alpha^2 \Rightarrow \alpha_M^2 = \alpha^2 - \sin^2\frac{\theta}{2} \cdot \frac{\sigma_x^2}{N} \tag{3.27}$$

The above performance can be re-written as,

$$p = Q\left(\sqrt{\left(N \cdot WNR - SNR \cdot \sin^2\frac{\theta}{2}\right) \cdot \sin^2\frac{\theta}{2}}\right) \tag{3.28}$$

Specifically, when $n = 3$,

$$p = Q\left(\sqrt{\left(N \cdot WNR - \frac{SNR}{4}\right) \cdot \frac{1}{4}}\right) \tag{3.29}$$

According to Equ. (3.29), the minimal spreading factor now is $SWR/4$. When the noise is weak, this proposed scheme outperforms the conventional and modified SS schemes

in capacity. Specifically, by equating Equ. (3.29) and Equ. (3.22), the capacity of the proposed scheme will be higher than that of the previous if $N < 1.25 \cdot SWR$. The BER at the crossing point is $p = Q\left(\sqrt{\frac{1}{4} \cdot SNR}\right)$. For example, if the noise is weak so that $SNR = 20dB$, this crossing BER is 2.866×10^{-7}. A simple deduction can also show that the proposed scheme outperforms the conventional SS scheme under some similar condition. This means that, under weak attacks, the proposed scheme can improve the capacity and achieve an acceptable BER. Figure 3.7 compares the proposed scheme with other embedding algorithms.

It should be noted that the actual performance is better that the worst case shown by Equ. (3.28) because it is the average of the cases of b=+1 and b=-1. For example, at the point A in Figure 3.6, the performance of b=-1 will be much better than that of b=+1 because the embedding power is reduced to α_M^2 and the host signal itself, $|\overrightarrow{OA}|^2 = \frac{\sigma_x^2}{N}$, along with the embedding power of α_M^2, contributes to the performance concurrently.

The interleaved embedding regions can only be constructed when n is an odd number. However, the cases of $n > 3$ are not worthwhile due to the fast reducing factor, $\sin^2 \frac{\theta}{2}$, as shown in Equ. (3.28).

Chapter 4

Digital Audio Watermarking

This chapter is organized as follows. First, the specific requirements of audio watermarking are checked closely and defined clearly. The early techniques are quickly reviewed in the second part. The quantization-based schemes including QIM and SCS are reviewed next. At last, the proposed SS scheme, incorporated with HAS masking and FEC schemes, is implemented for audio watermarking. A comparison between the proposed scheme and QIM/SCS schemes is also conducted.

4.1 General Requirements

An equivalent communication model of audio watermarking is shown in Figure 4.1. The embedding process consists of two steps. First, the watermark message is mapped into a pattern (the watermark) $\boldsymbol{w}$ of the same type and dimension as the host signal, $\boldsymbol{x}$. This mapping may be done with a key, $\boldsymbol{k}$, for security reasons, as well as the host signal $\boldsymbol{x}$. Next, the watermark, $\boldsymbol{w}$, is embedded into the host to produce the watermarked signal, $\boldsymbol{y}$.

4.1.1 Perceptual Transparency

A watermarking algorithm must embed the data without affecting the perceptual quality of the host signal. The objective measures such as the Signal-to-Watermark-Ratio (SWR) has not been shown to be reliably related to the perceived audio quality, because it can not distinguish inaudible artifacts from audible noise. SWR is measured in decibels and defined

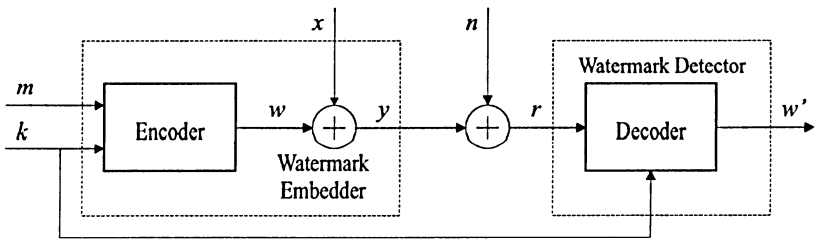

Figure 4.1: Watermarking as Communication

Rating	Impairment	Quality
5	Imperceptible	Excellent
4	Perceptible, not annoying	Good
3	Slightly annoying	Fair
2	Annoying	Poor
1	Very annoying	Bad

Table 4.1: ITU-R Rec. 500 Quality Rating

by the formula,

$$SWR = \frac{\sum_{i=1}^{i=N} x_i^2}{\sum_{i=1}^{i=N} (y_i - x_i)^2} \tag{4.1}$$

where x_i corresponds to the ith sample of the original audio signal, and y_i to the ith sample of the watermarked signal. Normally, one could expect the noise distortion of SWR less than 30 dB to be imperceptible.

Formal listening tests like the rating scale in the Table 4.1 have been considered to be the only relevant method for judging audio quality.

4.1.2 Blind Extraction

From Figure 4.1 it can be seen that this communication is actually one with side information (i.e. the host signal, $\boldsymbol{x}$) as the channel state at the the encoder. This idea is called *informed-embedding* [3]. Though the decoder has no access to the original host signal, the encoder knows the host signal in advance and hence it can take some measures to reduce the impact of the decoder's blindness.

According to Costa [11], the capacity of this channel is totally determined by the watermark signal and channel interference.

$$C = \frac{1}{2}\log_2(1 + \frac{\sigma_w^2}{\sigma_n^2}) \qquad (bits/sample) \tag{4.2}$$

The performance depends solely on the Watermark-to-Noise-Ratio (WNR), which is $10\log_{10}\frac{\sigma_w^2}{\sigma_n^2}$ (in decibels). The noise effect of the host signal does not affect the capacity theoretically.

A natural performance bound for the informed embedding is the result that could be achieved if the decoder has the knowledge of the original host signal. With this knowledge, the decoder could remove all host impact from the decision variable. The BER bound is, as shown in Chapter 2, repeated as follows

$$p = Q(\sqrt{WNR}) \tag{4.3}$$

4.1.3 Robustness and Attacks

Manipulations that modify the host signal also modify the embedded watermark. Furthermore, third parties may attempt to modify the host signal to thwart detection of the embedded watermark. An algorithm should guarantee a robust extraction of the embedded watermark even if the watermarked signal is distorted by these unintentional and intentional attacks.

Early literature considered the extraction process successful only if the whole watermark message was extracted. This was in fact a binary robustness metric. The Bit Error Rate (BER) has become common recently, as it allows for a more detailed scale to evaluate the extraction. The BER is defined as the ratio of incorrectly extracted bits to the total number of embedded bits.

To claim about a watermarking scheme being robust is difficult to prove due to the lack of testing standards. It becomes necessary to create a detailed and thorough test for measuring the ability that a watermark has to withstand a set of clearly defined signal operations. Given this, a set of the most common signal operations must be specified, and watermark

resistance to these must be evaluated. The common signal processing operations can be classified into six different groups as follows. [45]

Dynamics change the loudness of the audio signals. The most basic way of performing this consists of increasing or decreasing the loudness directly. More complicated operations include limiting, expansion and compression, which constitute non-linear operations that are dependant on the host signal.

Filters cut off or increase a selected part of the audio spectrum. Specialized filters include low-pass, high-pass, all-pass, equalizers.

Conversions convert from digital to analog representation and back, and might induce significant quantization noise.

Lossy compressions reduce the amount of data needed to represent an audio signal, which save bandwidth and storage. These compression algorithms are normally based on psychoacoustic models and delete regions where information is not perceived by the listener. If a watermarking algorithm embeds in these regions, the lossy compression could remove the watermark totally.

Modulation like vibrato, chorus, amplitude modulation and flanging are not common post-production operations. However, they are included in most audio editing software packages and thus can be used against watermarking.

Time stretch and pitch shift either change the duration of an audio signal without changing its pitch or change the pitch without changing the duration. They are used for fine tuning or fitting audio parts into time windows.

It is not always clear how much processing a watermark should be able to withstand. That is, the specific parameters of the diverse filtering operations that can be performed on the host signal are not easy to determine. Guidelines and minimum requirements for audio watermarking schemes have been proposed by different organizations such as the Secure Digital Music Initiative (SDMI).

One popular benchmark is Stirmark for audio [45] , and is adopted in this research. Table 4.2 summarizes common attacks defined in Stirmark benchmark.

Table 4.2: Attacks on Audio Watermarking

Name	Definition & Parameters	Implementation
AddDynNoise	Adds dynamic noise. *Dynnoise:* strength.	Stirmark
AddFFTNoise	Adds white noise in FFT domain. *FFTNoise:* strength.	Stirmark
AddNoise	Adds white noise. *Noisep:* strength.	Stirmark
AddSinus	Adds a sinus signal. *AddSinusFreq:* frequency; *AddSinusAmp:* strength.	Stirmark
Compressor	Increase/decrease the loudness. *Threshold:* threshold for compressor; *CompressValue:* ratio.	Stirmark
FFT-HighPass	High-pass filter in FFT domain. *HighPassFreq:* high pass frequency.	Stirmark
FFT-Invert	Inverts all FFT coefficients.	Stirmark
FFT-RealReverse	Reverses the real part of FFT coefficients.	Stirmark
Invert	Inverts all samples.	Stirmark
LSBZero	Sets all LSB to zero.	Stirmark
RC-HighPass	RC high-pass filter. *HighPassFreq:* high pass frequency.	Stirmark
RC-LowPass	RC low-pass filter. *LowPassFreq:* low pass frequency.	Stirmark
Smooth	Smoothes the samples.	Stirmark
ZeroCross	Set samples less than the threshold to zero.	Stirmark
Echo	Adds a constant echo. *Delay:* delay time; *Decay:* Percentage of the original.	Sound Forge.
Chorus	Adds an echo with vary delay time, strength modulated to the original.	Sound Forge.
A/D Conversion	Consequently change the resolution to 8-bit and back to 16-bit.	Sound Forge.
Re-sample	Consequently change the sampling rate to half and back to the original.	Sound Forge.
MPEG	Compress raw audio with MPEG algorithm at different rates.	Sound Forge.

4.2 Early Works

Many early audio watermarking efforts were based on some special features of HAS [4], and often characterized by small embedding capacity. Most of them were proved weak when attacked with some specified manipulations. Some typical ones are exemplified as follows.

Echo watermarking is a blind technique to embed information in the original audio signal $x(t)$ by introducing a repeated version of a component of the audio signal with a small delay (usually around $1ms$) and decay rate of $\alpha x(t - \Delta t)$ to make it imperceptible. In the most basic scheme, the information is encoded in the signal by modifying the delay such that two different values Δt_1 and Δt_2 are used in order to encode either a 0 or a 1. For watermark recovery, a technique known as *cepstrum* autocorrelation is used. This technique produces a signal with two pronounced amplitude spikes. By measuring the distance between these two spikes, one can determine if a 1 or a 0 was initially encoded in the signal.

Phase Coding works by substituting the phase of the original audio signal with one of two reference phases representing the information bit 1 and 0, respectively. That is, the watermark bit is represented by a phase shift in the phase of the audio signal. The phase shifts between consecutive signal segments must be preserved in the watermarked signal. This is necessary because HAS is very sensitive to relative phase differences, but not to absolute phase changes. In other words, the phase coding method works by substituting the phase of the initial audio segment with a reference phase that represents the data. After this, the phase of subsequent segments is adjusted in order to preserve the relative phases between them.

Least Significant Bit (LSB) Modulation takes advantages of the quantization errors. The lower order bits of the digital sample can be fully substituted with a pseudo-random (PN) sequence that contains the watermark message. This scheme is desired by its remarkable embedding capacity, the major disadvantage is its poor immunity to attacks. This technique can be implemented in a transform space rather than in the time domain.

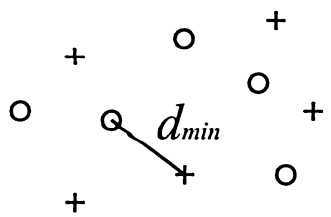

Figure 4.2: Quantization Index Modulation

4.3 Quantization Embedding Schemes

In [8] and [9] the authors proposed a new class of data embedding system, Quantization Index Modulation (QIM). In its implementable form, the scheme is a *subtractive dithered quantization*, in which the host signal is dithered by a watermark-modulated sequence and then quantized by a uniform quantization.

In QIM, an embedding function $\boldsymbol{y} = s(\boldsymbol{x}, \boldsymbol{m})$ maps the host signal, $\boldsymbol{x}$, and the watermark, $\boldsymbol{m}$, to a watermarked signal, $\boldsymbol{y}$, subject to the perceptual distortion constraint. QIM views the embedding function, $s(\boldsymbol{x}, \boldsymbol{m})$, as a collection or ensemble of functions of $\boldsymbol{x}$, indexed by $\boldsymbol{m}$. The rate R determines the number of possible values for $\boldsymbol{m}$, and hence, the number of functions in the ensemble. The distortion constraint suggests that each function in the ensemble is close to an identity function so that $s(\boldsymbol{x}, \boldsymbol{m}) \approx \boldsymbol{x}$ for all $\boldsymbol{m}$. That the system needs to be robust to noise suggests that the points in the range of one function in the ensemble should be "far away" from the points in the range of any other function. At the very least, the ranges should be non-intersecting. These properties suggest that the functions $s(\cdot)$ be discontinuous. Quantizers are just such a class of discontinuous, approximate-identity functions. QIM refers to modulating an index or sequence of indices with the embedded information and quantizing the host signal with the associated quantizer or sequence of quantizers.

Figure 4.2 illustrates QIM information embedding for the $N = 2$ and $R = 1/2$ case. In this example, one bit $m \in \{0, 1\}$ is embedded. The reconstruction points of the two required quantizers are represented with $+$'s and $\circ$'s. If $m = 0$, for example, $\boldsymbol{x}$ is quantized with the

+-quantizer, i.e., $\boldsymbol{y}$ is chosen to be the + closest to $\boldsymbol{x}$. If $m = 1$, $\boldsymbol{y}$ is quantized with the o-quantizer.

The number of quantizers in the ensemble determines the data embedding rate. The size and shape of the quantization cells determine the embedding-induced distortion. Finally, the minimum distance between the sets of reconstruction points of different quantizers in the ensemble determines the robustness of the embedding.

Intuitively, the minimum distance measures the noise vectors that can be tolerated by the system. In case of AWGN channel of the noise variance σ_n^2, if SNR is high enough, the minimal distance characterizes the BER of the minimal distance decoder as follows,

$$p = Q\left(\sqrt{\frac{d_{min}^2}{4\sigma_n^2}}\right) \tag{4.4}$$

For implementation reasons, the quantizers take the form of dithered quantizers, which are quantizer ensembles where the quantization cells and reconstruction points of every quantizer are shifted versions of some base quantizer $q(\cdot)$. For data embedding, the dither vector $d(\cdot)$ is modulated by the embedded data $\boldsymbol{m}$. Thus the embedding function is

$$s(x, m) = q_\Delta(x + d(m)) - d(m) \tag{4.5}$$

Consider the binary case, $q_\Delta(\cdot)$ is a uniform, scalar quantizer with step size Δ, i.e. $q(\cdot) = round(\frac{\cdot}{\Delta})\Delta$, and the dither vector is constructed by $d(0) = \Delta/4$ and $d(1) = -\Delta/4$ such that the two quantizers are maximally far away from each other. The embedded data can be extracted blindly by the following function,

$$m_i = mod\left[round\left(\frac{y_i - d(0)}{\Delta/2}\right), 2\right] \tag{4.6}$$

When the quantization cells are small enough such that the host signal can be modeled as uniformly distributed within each cell, the expected squared-error distortion, i.e. the watermark signal power is as follows,

$$\sigma_w^2 = \frac{1}{\Delta}\int_{-\Delta}^{+\Delta} x^2\, dx = \frac{\Delta^2}{12} \tag{4.7}$$

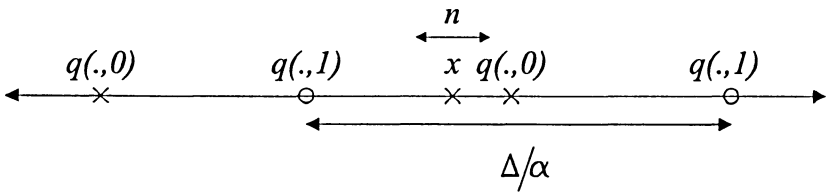

Figure 4.3: Distortion-compensated Quantization Index Modulation

Since,

$$d_{min}^2 = |d(0) - d(1)|^2 = \frac{\Delta^2}{4} \tag{4.8}$$

The BER can be written as,

$$p = Q\left(\sqrt{WNR \cdot \frac{3}{4}}\right) = Q\left(\sqrt{WNR - 1.25dB}\right) \tag{4.9}$$

Compared with (4.3), it can be seen that QIM is within the gap of 1.25 dB to the bound performance *asymptotically.* A theoretical comparison can be made between the QIM and the proposed scheme using the above equation, as shown in Figure 3.5. It can be seen that for this specific case, the proposed scheme outperforms the QIM as early as from p=0.001. Thus it can be concluded that the modified SS scheme is superior to the quantization-based schemes for the case of watermarking were accurate extraction is desired.

Since in dithered modulation the minimal distance is just the quantization step Δ, increasing this step by $\Delta/\alpha, 0 < \alpha < 1$ can increase the robustness, as well as the QIM embedding-induced distortion. One way to compensate for this distortion is to add back some of the quantization error to the reconstruction point to form the composite signal. Specifically, the embedding function is,

$$s(x, m) = q_\Delta\left(x + d(m), \Delta/\alpha\right) + (1 - \alpha)\left(x - q_\Delta(x + d(m), \Delta/\alpha)\right) \tag{4.10}$$

The deflection is a source of interference, along with the channel distortion during decoding. The above algorithm is called Distortion-Compensation QIM (DC-QIM). This is shown in the Figure 4.3.

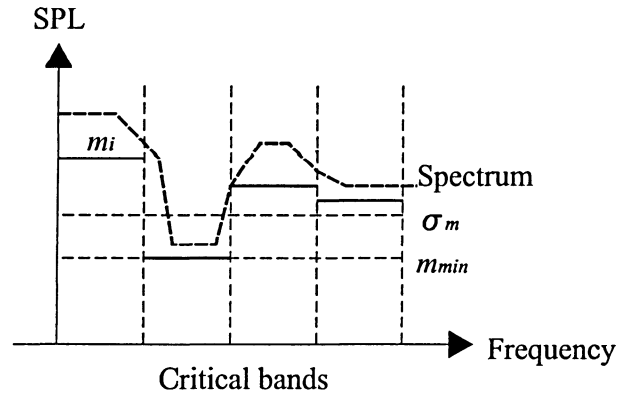

Figure 4.4: Increase SNR by masking effects

In [20] and [21] the authors proposed the so called Scalar Costa Scheme (SCS), which is a sample-wise embedding algorithm based on the method proposed by Costa [11]. Costa presents a theoretic scheme that involves a random codebook, which is very huge and not practical. The codebook is needed to be available at both the encoder and the decoder for decoding. Instead, SCS uses a structured codebook that can be expressed as subtractive dither quantization. In fact, SCS is equivalent to DC-QIM in performance but with a simpler implementation.

4.4 Proposed Audio Watermarking Scheme

The proposed scheme is already described in the last chapter. This scheme can be further improved by the following measures.

4.4.1 Use masking effects to increase the robustness

When an embedding occurs in frequency domain, a uniform perceptual factor α is actually confined by the weakest region of the whole spectrum in terms of masking ability. It is possible to maximize the watermark magnitude with the local masking abilities, thus increase the power of the transmitted signal, as well as conforming to the perceptual constraints. This idea is shown in Figure 4.4.

The global masking threshold curve $\boldsymbol{m}$, which is the output from some HAS masking model, represents the maximal changes a frequency component can tolerate without causing perceptual distortion. It actually denotes the local masking abilities of each frequency range and can be used to implement this maximization. The signal power when no making explored is constrained by the minimal value m_{min} and that when masking explored is, roughly saying, the mean value σ_m As a non negative function, there is always, $m_i > m_{min}$, which means the signal power is increased by exploring the masking effects.

4.4.2 Use attacks characterization to increase the robustness

Generally saying, the attacks falls into two categories, i.e. the wideband noises that corrupt the whole spectrum, and the narrowband noises that corrupt some regions only and leave other regions (almost) unchanged.

For wideband attacks, choosing different domains does not effect much differently. For example, the watermarked signal is corrupted by an additive white Gaussian noise, the noise effect is almost identical in terms of SNR, either in time domain or frequency domain. Figure 4.5 and 4.6 show such an attack in time domain and MDCT domain, respectively.

On the other hand, for those narrowband attacks, a watermark residing on the unchanged region should be more robust because of less noise in extraction. A good example is the attacks of low-pass filtering, where the noise effect in high frequency region is much higher than that in low frequency region. If a watermark occupies the low frequency region only, the performance can be expected to be much better than that of a watermark that occupies the whole spectrum. Figure 4.7 and 4.8 show such an attack in time domain and MDCT domain, respectively.

Of the two watermark schemes, no one can take on each other. The wideband watermark is of longer length and will perform better for the strong wideband noise where the narrowband watermark performs poorer due to its shorter length. On the other hand, for the narrowband noise, a longer wideband watermark will introduce more noise and thus is not desired.

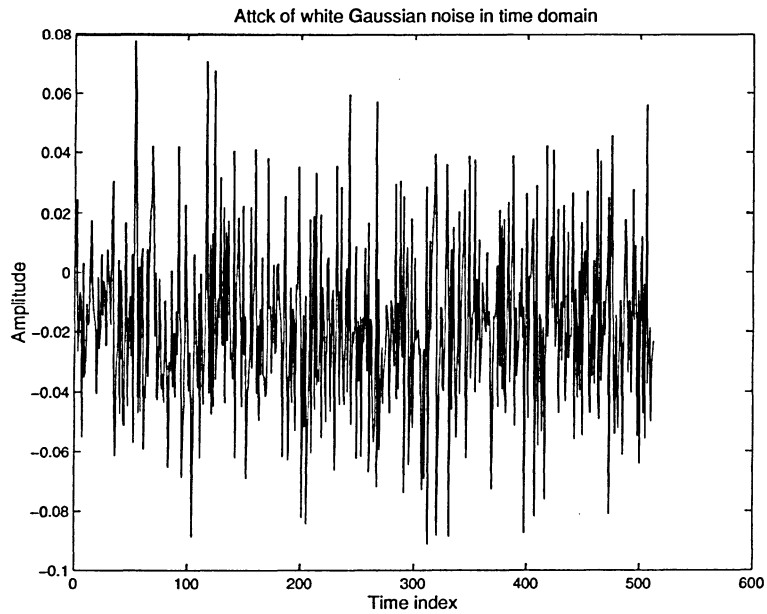

Figure 4.5: Attack of additive white Gaussian noise in time domain

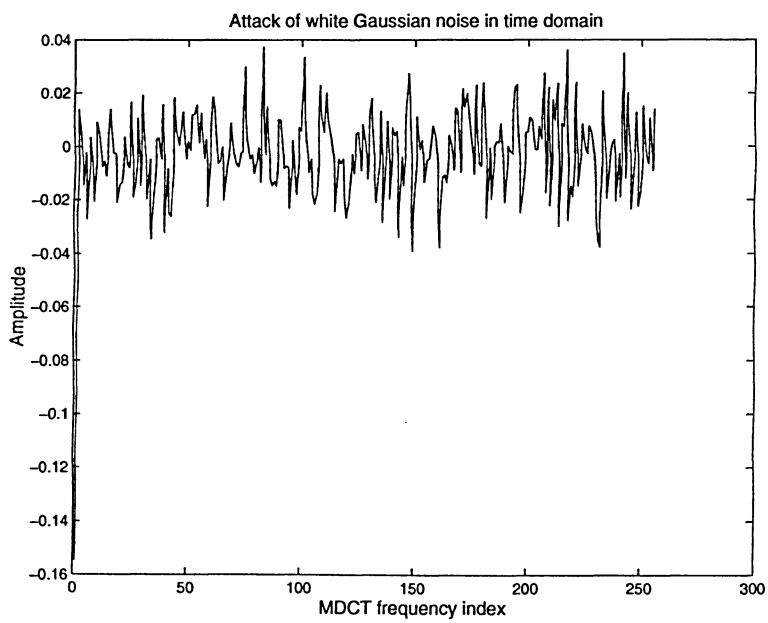

Figure 4.6: Attack of additive white Gaussian noise in MDCT domain

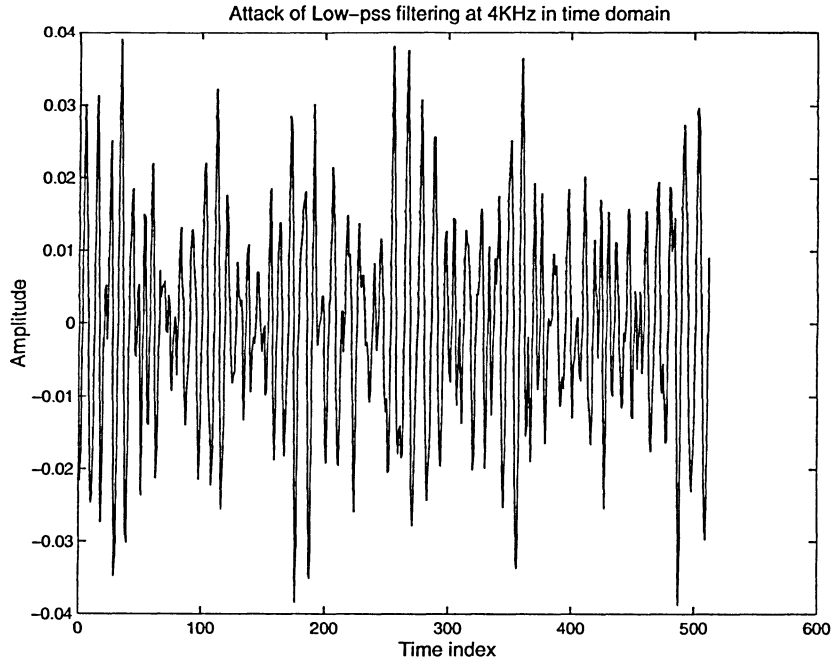

Figure 4.7: Attack of low-pass filtering at 4 kHz in time domain

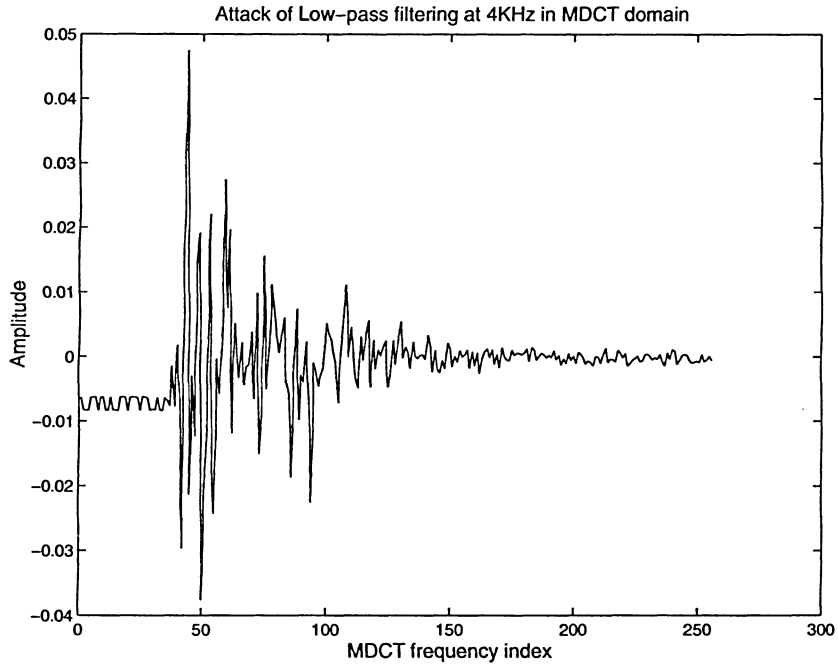

Figure 4.8: Attack of low-pass filtering at 4 kHz in MDCT domain

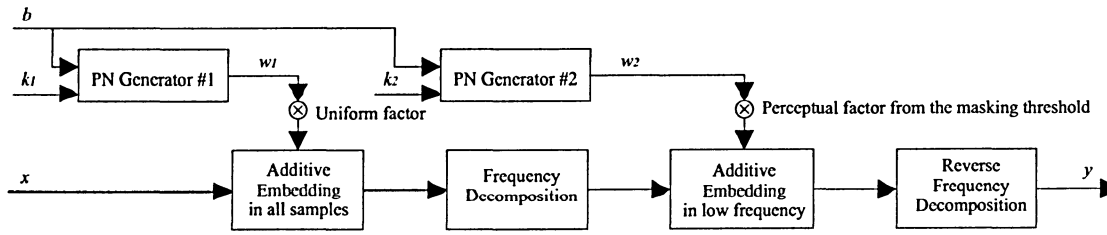

Figure 4.9: Proposed watermark embedding system

Based on the above duality analysis, a watermarking system with two different watermarks is proposed and the embedding block is shown in Figure 4.9.

The first watermark is implemented in time domain and occupies the whole frequency spectrum, this is a wideband watermark and expected to be against wideband attacks effectively. Another watermark is implemented in frequency domain and occupies only the low frequency region, this is a narrowband watermark and expected to be against narrowband noise effectively. On the receiving side, the extraction performs in both time and frequency domains independently, any extraction with high confidence (low BER) can prove the ownership. The serial order of embedding is not important in the scheme. This is because one watermark is taken as a part of the host signal, to the extraction of another watermark, and thus can be suppressed.

The trade-off for this scheme is that the total available energy has to be distributed between the two watermarks. Thus, the SNR of each watermark transmission has to be decreased accordingly. This signal power loss is compensated for by the greater reduction of noise power in the scheme.

4.4.3 Use FEC to increase the robustness

An FEC scheme can only take effect when the BER of the uncoded channel is below some threshold, this threshold is determined by the correction ability of the specific scheme and relates to the minimal distance of the code directly.

For a watermarking channel, the noise due to attacks normally makes the SNR very low.

Considering this, only Turbo code is expected to be superior to others.

4.5 Simulation Results

To evaluate the proposed watermarking system, which is called Modified Spread Spectrum (MSS) in the following, a comparison between SS, QIM and MSS is conducted to illustrate the above analysis.

The simulation steps are designed to illustrate the host-suppression property, the improvements due to attacks characterization and the improvements due to FEC schemes. At last, a final complete scheme is tested thoroughly on different real audio signals against the attacks defined previously.

In the following, all watermark sequences are of 128-bit length and all results are the average from 5 or 10 repeated tests for smoothness. Some common structures are listed as follows.

- BCH code is with the structure of BCH(127, 64, 10).
- Convolutional code is of generator [6, 7] and the constraint of 3.
- Turbo code is unpunctured, with generator of [7, 5] and the random interleaver; decoded with log-MAP by 4 iterations.

4.5.1 Host Suppression

In this simulation, a time-domain watermark sequences with the same embedding powers are embedded into host signals using different schemes respectively. The watermarked signals are then attacked by the same typical distortion, specifically AWGN attacks. Comparing their BERs, we are able to illustrate the performance improvements due to host-suppression.

The cases of the Gaussian host signal under AWGN of SNR=0dB and SNR=5dB are shown in Figure 4.10 and Figure 4.11 respectively. The cases of real non-Gaussian audio signals are shown in Figure 4.12 and Figure 4.13, respectively.

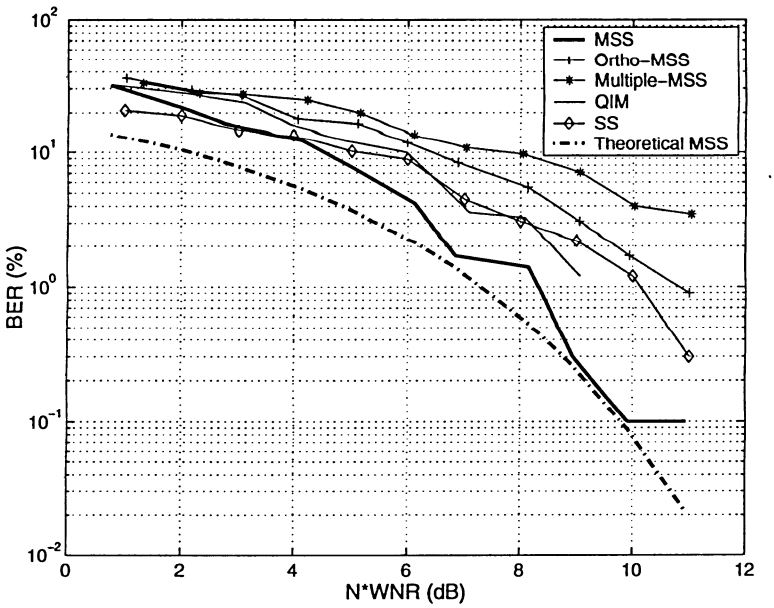

Figure 4.10: Performances of different embedding schemes for Gaussian signal and Gaussian noise attack (SWR=25dB, SNR=0dB). **Note:** Because the embedded sequence must be long enough to guarantee the precision of the measured BER, and the tested audio clips normally do not last long enough to hold such a long embedded sequence, that is why there is some discrepancy between the measured and theoretical performances.

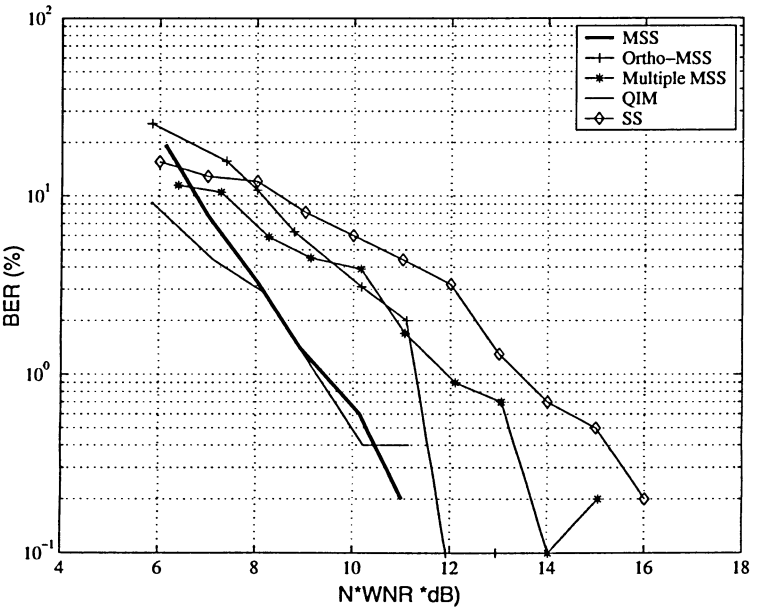

Figure 4.11: Performances of different embedding schemes for Gaussian signal and Gaussian noise attack (SWR=25dB, SNR=5dB). **Note:** At high N*WNR values (> 11 dB), the measured BERs of MSS and QIM are both zeros and could not be drawn gracefully by MATLAB, but it can be seen clearly that MSS begins to outperform QIM from about N*WNR=10.5 dB.

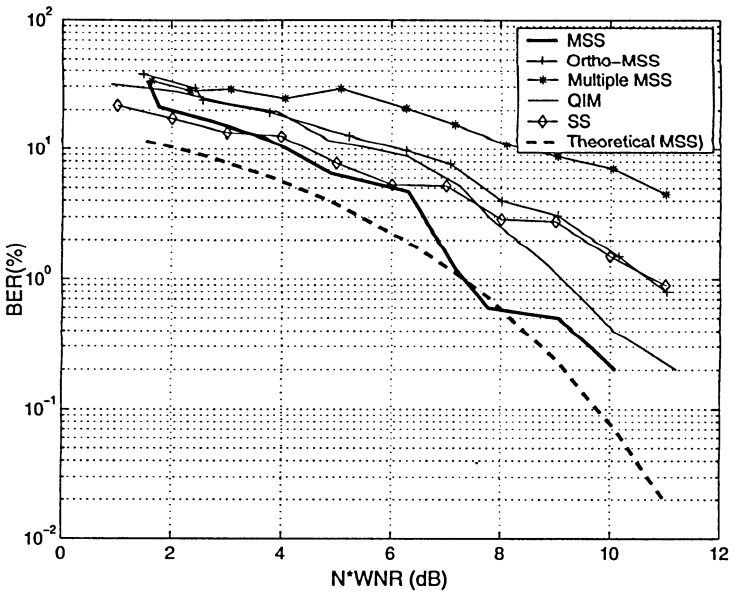

Figure 4.12: Performances of different embedding schemes for non-Gaussian signal and Gaussian noise attack (SWR=25dB, SNR=0dB). **Note:** Because the embedded sequence must be long enough to guarantee the precision of the measured BER, and the tested audio clips normally do not last long enough to hold such a long embedded sequence, that is why there is some discrepancy between the measured and theoretical performances.

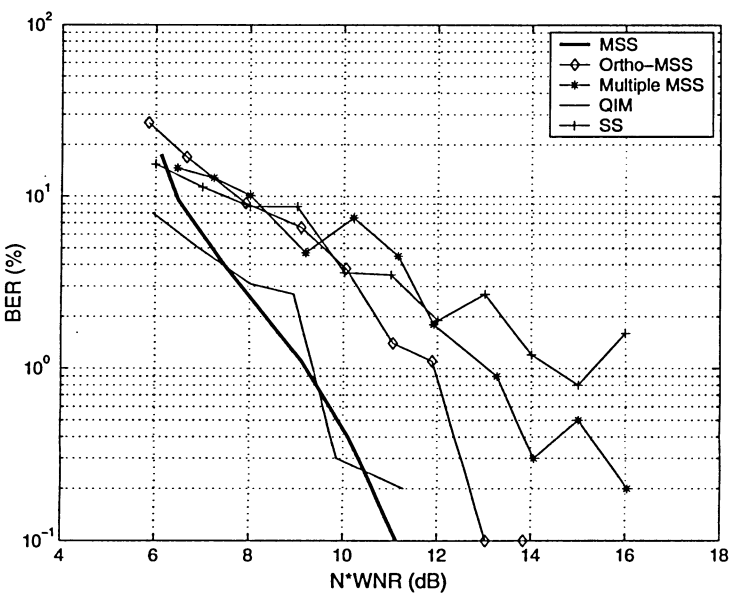

Figure 4.13: Performances of different embedding schemes for non-Gaussian signal and Gaussian noise attack (SWR=25dB, SNR=5dB). **Note:** At high N*WNR values (> 11 dB), the measured BERs of MSS and QIM are both zeros and could not be drawn gracefully by MATLAB, but it can be seen clearly that MSS begins to outperform QIM from about N*WNR=10.5 dB.

In Figure 4.10 the close coincidence of the measured MSS BER curve and the theoretical one suffices to show that the deduction is correct.

In all these figures the proposed scheme outperforms the host-suppressing schemes of QIM when the spreading factor N is big enough, especially when under strong attacks (e.g. SNR=0dB). Its significant improvements over the conventional SS is clearly seen in these figures as well.

It is also worth noting that, when the spreading factor is not large enough, the induced distortion is comparatively large and results in poor performance in these regions. These regions are characterized by high BERs also for other schemes in the cases of strong attacks. Thus these would not be reasonable operational regions for all schemes too. When the spreading factor is increased, after some threshold determined by Equ. (3.17), the performance of MSS begins to be improved significantly, which is characterized by the steep slopes in the figures, and outperforms all three other schemes.

4.5.2 FEC Schemes

To illustrate the effects of different FEC schemes, the same time-domain watermark sequences are coded with BCH, convolutional code and Turbo code respectively, and then embedded into signals with same embedding power. The results are shown in the Figures 4.14, 4.15.

It can be seen that each FEC scheme can take effect only when the effective SNR is above some threshold. Comparing Figure 4.14 and Figure 4.15, it can be seen that in terms of WNR, Turbo code does take effect earlier with MSS scheme than with SS scheme. This is because that MSS suppresses the host impact and thus equivalently increases the SNR of the channel.

Figure 4.15 shows that Turbo code generates a coding gain more than 2dB starting from BER of 1%, and that is about 1dB with BCH code. The effect of convolutional code is not apparent since its SNR threshold to take effect is larger than that of the others.

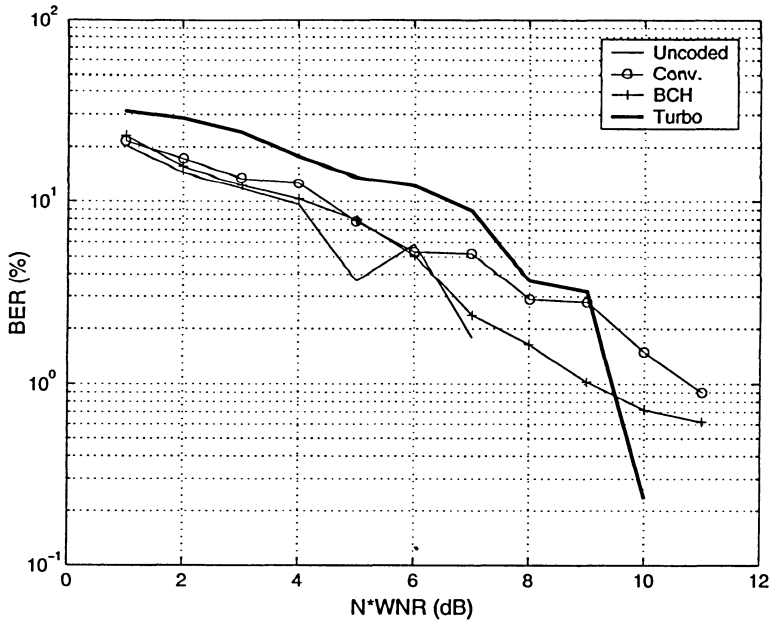

Figure 4.14: Performance of SS scheme with different FEC schemes for non-Gaussian signal and Gaussian attack (SWR=25dB, SNR=0dB)

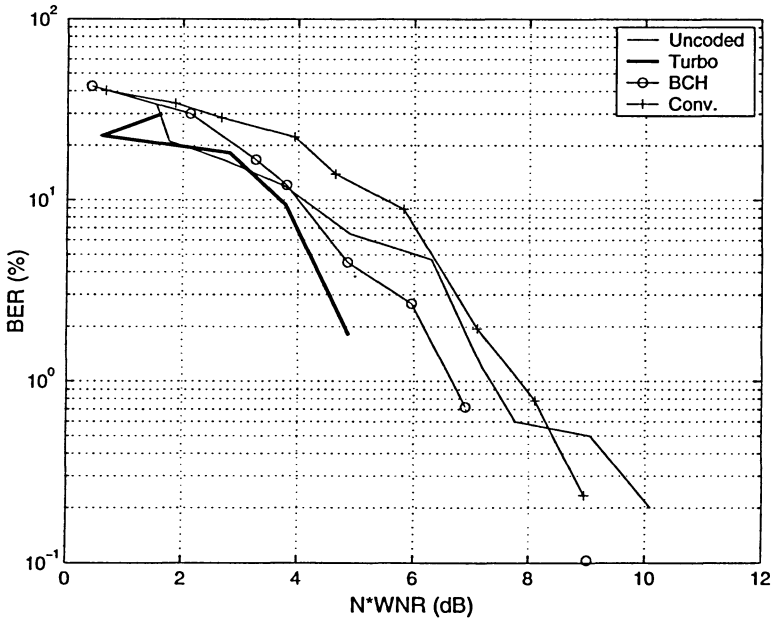

Figure 4.15: Performance of MSS scheme with different FEC schemes for non-Gaussian signal and Gaussian attack (SWR=25dB, SNR=0dB). **Note:** Because the measured BER values of Turbo code case are zeros at high N*SNR values, MATLAB could not express them gracefully in logarithmal scale.

4.5.3 Multiple Watermarks

To illustrate the effects of characterization, the same watermark sequences are embedded with the same embedding rate into time-domain and MDCT-domain, respectively. The MDCT embedding domain spans $1 \sim 4$ kHz only. Their performances under different attacks are shown in Figure 4.16 and Figure 4.17.

Under the Gaussian attacks, the time-domain watermark can give a much lower BER than the MDCT-domain watermark. On the other hand, for the Low-Pass filtering attack, the MDCT one is superior to the time domain one. The huge difference between their performance shows that multiple watermarks can be complementary under different attacks.

4.5.4 Final Scheme and Parameters

To show the potentials of the proposed watermarking system, it is tested against audio clips of different genres and strengths. All testing audio clips are sampled at 44.1 kHz and of 16-bit resolution. All of them are of durations between $25s \sim 30s$ so that a watermark sequence with 128 bits can be embedded in both time and MDCT frequency domain.

To evaluate the subjective fidelity, 10 listeners were asked to report dissimilarities between the original and the watermarked using the 5-point impairment scale of MOS, as defined in Table 4.1.

After many testing on all audio clips, the following structures and parameters are regarded as optimized. The MDCT domain watermark sequence is first Turbo encoded, then embedded into $1 \sim 4kHz$ MDCT frequency range. The amplitude of the watermark sequence is computed by the masking threshold. The embedding rate is about 34.8 bps. The time domain watermark sequence is also Turbo encoded first, and then embedded into the result signal with a spreading factor of 500. The amplitude of the watermark sequence is controlled by SWR=32dB. This correspond to an embedding rate of about 29.4 bps.

At the receiving end, the decoding is conducted in both time-domain and MDCT-domain. The one with the lowest BER represents the performance of the whole system. The averaged results are shown in the Table 4.3.

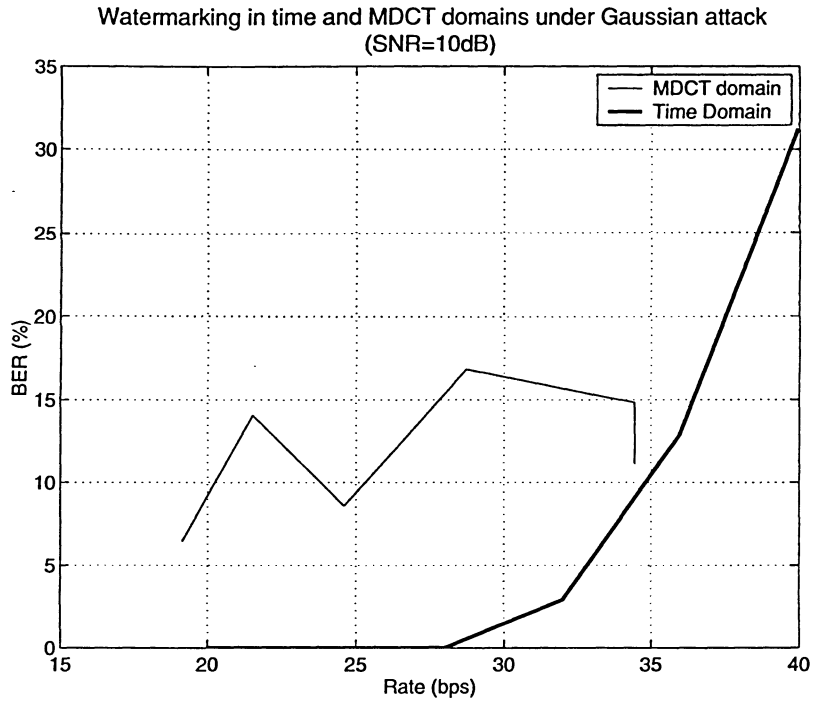

Figure 4.16: Performance of MSS scheme with Turbo code schemes attacked by Gaussian noise (SWR=25dB, SNR=0dB)

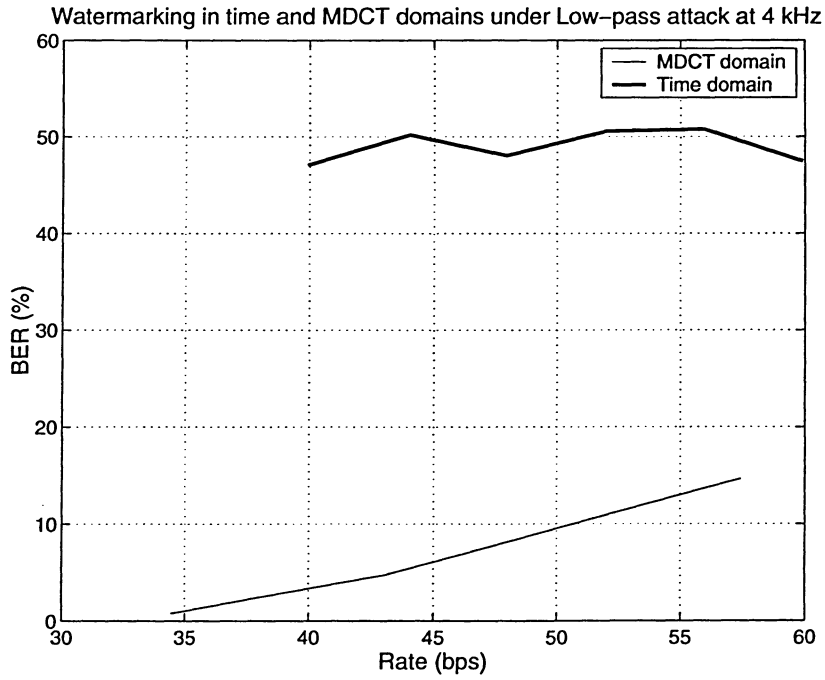

Figure 4.17: Performance of MSS scheme with Turbo code schemes attacked by Low-Pass filtering at 4 kHz attack (SWR=25dB)

Under this configuration, the averaged MOS out of five different clips from 10 listeners is 4.8 with variance of 0.146.

It is worth noting that, under LP, re-sampling and MP3 attacks the MDCT watermarks outperform the time watermarks significantly. Comparing the BERs of the two watermarks, it can be concluded that this is a robust watermarking system with high embedding rate.

Compared the results shown in [17], under almost the same MOS evaluation and BER limit, our proposed scheme increases the embedding data rate by at least 50%.

4.6 Chapter Summary

Both the theoretical analysis and the simulations show that, in the cases of watermarking, the proposed SS scheme is superior to the quantization-based techniques in terms of robustness, especially when facing strong attacks. The quantization-based techniques are superior to the SS techniques when facing weaker noises. Actually the SCS/QIM can be used as high-capacity embedding techniques for these cases, while the SS techniques can not due to the strong distortions introduced under these cases.

Both SS and SCS/QIM need a large spreading factor to virtually increase the SNR as desired. This means that a perfect synchronization is needed for them. Thus those de-synchronization attacking methods, e.g. swapping or dropping samples, are the most serious ones. In [29], some measures were proposed to protect SS against de-synchronization attacks.

Another disadvantage of quantization-based techniques is that a simple scaling attack could easily destroy the watermark embedded by SCS/QIM. Some efforts has been made to alleviate this attacks [21].

FEC schemes, especially the Turbo code, introduce heavy complexities to trade with robustness. The simulation shows that Turbo code is the best one among them, in terms of coding gain. which makes it desirable when very high robustness is required, like the case of DVD watermarking.

Attacks characterization is a very effective measure to increase the robustness, as shown in this chapter. In fact, instead of "one fits all", there is a trend to design specific watermarks

Table 4.3: Averaged BER under different attacks

Name	timeBER; freqBER
AddDynNoise (*Dynnoise*=)2%.	0.000%; 0.000%
AddFFTNoise (*FFTNoise*=).	0.000%; 0.000%
AddNoise (*Noisep*=2%).	0.000%; 0.000%
AddSinus (*AddSinusFreq*=900 Hz; *AddSinusAmp*=)4%.	0.000%; 0.000%
Compressor.	0.000%; 0.000%
FFT-HighPass (*HighPassFreq*=2000 Hz).	0.000%; 0.000%
FFT-LowPass (*LowPassFreq*=8000 Hz).	0.000%; 0.000%
FFT-Invert.	0.000%; 0.000%
FFT-RealReverse.	0.000%; 0.000%
Invert.	0.000%; 0.000%
LSBZero .	0.000%; 0.000%
RC-HighPass (*HighPassFreq*=2000 Hz).	0.000%; 0.000%
RC-LowPass (*LowPassFreq*=8000 Hz) .	0.000%; 0.000%
Smooth.	1.120%; 0.050%
ZeroCross.	0.000%; 0.100%
Normalize.	0.000%; 0.000%
Echo (*Delay*=400 ms; *Decay*=10%).	0.000%; 0.000%
Chorus.	0.000%; 0.000%
A/D Conversion.	0.000%; 0.000%
Re-sample.	3.750%; 0.000%
MPEG (96 kb).	0.000%; 0.000%
MPEG (80 kb).	0.000%; 0.000%
MPEG (64 kb).	0.950%; 0.050%
MPEG (56 kb).	2.500%; 0.050%
MPEG (48 kb).	5.450%; 0.090%
MPEG (40 kb).	7.750%; 0.120%
AddNoise (SNR=10 dB).	0.00%; 5.350%
LowPass filtering (4 kHz).	12.350%; 0.000%

for specific attacks, as discussed in [30]. This can be expected as an active branch of research in the future.

Taking advantage of the masking effects of HAS leads to a really effective way to increase SNR of the watermark transmission. The disadvantage is that it introduces heavy complexities when summing up the energies of critical bands. Also when incorporated with SCS/QIM schemes, extra errors are introduced due to the blind estimation of the quantization step at the receiving end. It is worth noting that MPEG encoding transmits the quantization steps directly.

Chapter 5

Data Hiding in Digital Speech Signals

Data hiding also denotes the techniques to embed extra data imperceptibly in the multimedia of any kind such as speech, audio, image, and video. Data hiding is intended to hide larger amounts of data into the host signal to provide additional functionalities, rather than just to check for authenticity and copyright information. While most of current research of data embedding concentrates on watermarking, the high-capacity data hiding is at present receiving considerable attention.

The relevant research in this thesis focuses on embedding data as much as possible into the ITU G.711 μ-law encoded digital telephony signal. An application of this research is to embed some wide-band information into the narrow-band signals transmitted on the Public Switched Telephone Network (PSTN), thus to improve the voice quality and intelligibility. Other potential applications includes providing additional services and features using the embedded information [18].

In the following sections, after a quick review of the different speech coders, the μ-law encoded telephony signal coding is described in details. Then the proposed spread spectrum technique in the pervious chapter is applied to this scenario.

5.1 Speech Coders Review

Speech coders are usually divided into two main classes: waveform coders and voice coders (*vocoders*). In addition, there are hybrid coders that combine the characteristics of the two

main types [35].

Waveform coding means that the amplitudes of the analog signal are described by a number of quantized values. These values are then pulse-coded and sent to the receiving end. The signal's analog appearance is reproduced in the receiving end by means of the received values. The method makes it possible to obtain a very high level of voice quality, since the received voice curve is a true copy of the one transmitted. There are techniques that operate on the waveform in the time domain such as Pulse Code Modulation (PCM), Adaptive Pulse Code Modulation (ADPCM), and delta modulation. Other techniques, such as Sub-Band Coding (SBC), operate on the signal in the frequency domain.

The voice coder is a parametric coder. Instead of transmitting a direct description of the voice curve, a number of transmitted parameters describe how the curve has been generated. Parametric coding requires a defined model of how the voice curve is created. The quality will be average but, on the other hand, signals can be transmitted with a very low bit rate. The most popular scheme is Linear Predictive Coding (LPC). The process is called linear prediction since the next output value of the system is determined from a weighted sum of past output values and plus an input value. This is characteristic of a finite impulse response filter. The excitation signal is also characterized and sent along with the other parameters for synthesis at the receiver. The many different forms of LPC vary in the way the excitation signal and the other parameters are represented and transmitted.

A hybrid coder sends a number of parameters as well as a certain amount of waveform-coded information. This type of voice coder, which provides a reasonable compromise between voice quality and coding efficiency, is used in digital mobile telephone systems.

Table 5.1 summarizes some common speech standards and their MOS ratings [35].

5.2 Telephony Speech Signals

Speech sound can be broken into three distinct classes of phonemes i.e. voiced, unvoiced, and plosive. In general, the amplitude of voiced phonemes is approximately ten times that of unvoiced and plosive phonemes. Thus the telephone system must provide a large range

Coder	MOS
64 kbps PCM (ITU G.711)	4.3
32 kbps ADPCM (ITU G.721)	4.1
16 kbps LD-CELP (ITU G.728)	4.0
8 kbps CS-ACELP (ITU G.729)	4.0
4.8 kbps CELP (FS 1016)	3.2
2.4 kbps LPC-10e (FS 1015)	2.3

Table 5.1: Speech coders performance

of signal dynamics. Although lower in amplitude, unvoiced and plosive phonemes contain more information and the telephone system must also provide higher resolution for lower amplitude signals.

In addition, the telephone network is also subject to bandwidth restrictions with respect to the human speech and auditory ranges. The telephone network restricts transmission to a 3.1 kHz portion, from 0.3 $\sim$ 3.4 kHz. This frequency range coincides with the region of greatest intelligible speech, retaining only the first three frequency formant of the sampled speech signal. Surrounded by two guard bands of 0 $\sim$ 0.3 KHz and 3.4 $\sim$ 4 KHz to provide a buffer against conversation interference, the telephone network has a total bandwidth of 4 kHz. For accurate reproduction, according to Nyquist, a speech signal must be sampled at a rate of at least 8 kHz.

5.2.1 Companding

For digital transmission, the analog speech signal is converted to a digital signal with a fixed precision. A uniform quantization transforms the discrete signal into digital signals. Coding of the signal is performed by converting the midpoint of each quantization level to a codeword.

In general, speech signals are composed of relatively fewer voiced phonemes than unvoiced phonemes. Unfortunately, the uniform quantizer provides unneeded quality for large signals which are least likely to occur, and pronounced truncation effects for the more frequent small amplitude signals. As a result, uniform quantization does not perform as well as a quantizer

with wider zones at high amplitudes and narrower zones at lower amplitudes.

Conversion to a logarithmic scale coincides with the processing of Human Auditory System and allows quantization intervals to increase with amplitude, and it ensures that low-amplitude signals are digitized with a minimal loss of fidelity. This specific quantization may be achieved by first passing the signal through a compressor, a nonlinear device which compresses the peak amplitudes. This is followed by a uniform quantizer, such that uniform zones at the output correspond to non-uniform zones at the input. At the receiving end, the compressed signal is passed through an expander, another nonlinear device used to cancel the nonlinear effect of the compressor. The combined process is known as *companding.*

In addition to reducing quantization error, companding decreases the required bandwidth of the system. Systems solely employing uniform quantization require 13-bit codewords for equivalent performance requirements of the telephone system. However, systems using nonlinear companding may reduce the required codeword length to 8-bits or less. Fewer bits per sample are necessary to provide a specified SNR for small signals and an adequate dynamic range for large signals.

Two international companding standards that retain up to 5 bits of precision by encoding signal data into 8 bits are μ-law and A-law, as defined in ITU G.711. μ-law is the accepted standard of North America and Japan, while A-law is accepted in Europe.

5.2.2 $\mu-$law Companding

[7]

The μ-law compression is defined mathematically by the continuous equation:

$$y = sgn(x)\frac{\ln(1+\mu|x|)}{\ln(1+\mu)}, -1 \leq x \leq 1 \tag{5.1}$$

where μ is the compression parameter (μ=255 for the U.S. and Japan), and x is the normalized signal sample to be compressed. The actual compression algorithm is a piece-wise linear approximation of this mathematical definition. An 8-bit μ-255 codeword is composed of 1 sign bit concatenated with a 3-bit chord and a 4-bit step.

Table 5.2: μ-law Encoding

Linear Input Data													μ-law Encoded Output							
0	0	0	0	0	0	0	1	A	B	C	D	X	S	0	0	0	A	B	C	D
0	0	0	0	0	0	1	A	B	C	D	X	X	S	0	0	1	A	B	C	D
0	0	0	0	0	1	A	B	C	D	X	X	X	S	0	1	0	A	B	C	D
0	0	0	0	1	A	B	C	D	X	X	X	X	S	0	1	1	A	B	C	D
0	0	0	1	A	B	C	D	X	X	X	X	X	S	1	0	0	A	B	C	D
0	0	1	A	B	C	D	X	X	X	X	X	X	S	1	0	1	A	B	C	D
0	1	A	B	C	D	X	X	X	X	X	X	X	S	1	1	0	A	B	C	D
1	A	B	C	D	X	X	X	X	X	X	X	X	S	1	1	1	A	B	C	D

During compression the sample magnitudes are limited to 13 bits. The least significant bits of large amplitudes are discarded. The number of insignificant bits deleted is encoded into a field called the *chord.* Before chord determination, the sign of the original integer is removed and a bias of 33 is added to the absolute value of the integer. Due to the bias, the magnitude of the largest valid sample is reduced to 8159 and the minimum step size is reduced to 2/8159.

Each chord of the piece-wise linear approximation is divided into equally sized quantization intervals called *steps.* The step size between adjacent codewords is doubled in each succeeding chord. Chord determination may be reduced to finding the most significant 1 bit of the binary representation of the biased integer value, while the step equals the four bits following the most significant 1. Also encoded is the sign of the original integer. The polarity bit is set to 1 for positive integer values. The Table 5.2 better illustrates the translation from linear to μ-law compressed data. Of the compressed codeword, bits $4-6$ represent the chord and bits $0-3$ represent the step.

Finally, before transmission, the entire μ-law code is inverted. The codeword is inverted since low amplitude signals tend to be more numerous than large amplitude signals. Consequently, inverting the bits increases the density of positive pulses on the transmission line, which improves the hardware performance.

μ-law expansion is defined by the continuous inverse equation:

$$x = sgn(y)\frac{(1+\mu)^{|y|}-1}{\mu}, -1 \leq y \leq 1 \tag{5.2}$$

Table 5.3: μ-law Decoding

μ-law Encoded Input								Linear Output Data												
S	0	0	0	A	B	C	D	0	0	0	0	0	0	0	1	A	B	C	D	1
S	0	0	1	A	B	C	D	0	0	0	0	0	0	1	A	B	C	D	1	0
S	0	1	0	A	B	C	D	0	0	0	0	0	1	A	B	C	D	1	0	0
S	0	1	1	A	B	C	D	0	0	0	0	1	A	B	C	D	1	0	0	0
S	1	0	0	A	B	C	D	0	0	0	1	A	B	C	D	1	0	0	0	0
S	1	0	1	A	B	C	D	0	0	1	A	B	C	D	1	0	0	0	0	0
S	1	1	0	A	B	C	D	0	1	A	B	C	D	1	0	0	0	0	0	0
S	1	1	1	A	B	C	D	1	A	B	C	D	1	0	0	0	0	0	0	0

The implementation of the μ-law expansion is as follows. Before expansion, the μ-law code is inverted again to restore the original code. During expansion, the discarded least significant bits are approximated by the median of the interval, to reduce the loss in accuracy. That is, if six of the least significant bits of the original binary integer were discarded during compression, these six least significant bits will be approximated by 100000_2 during expansion. After decoding the μ-law code, the bias is removed and the sign of the binary integer is restored according to the polarity bit. This procedure is shown in the Table 5.3.

From the paradigm of communication, the quantization noise from μ-law companding can be thought of as an attack that corrupts the extraction of the embedded information. It can be seen that this noise is highly correlated with the original signal, because the companding is a sample-wise non-linear procedure in which the quantization step is dependent on the corresponding sample amplitude. Clearly this noise is not white and according to the implementation, the maximal error is about 3% which occurs at the highest amplitude.

5.3 Problem Statement

The procedure of data hiding in μ-law speech signals is illustrated in the Figure 5.1. As shown in the figure, the companding part is determined by the telephony network. This research work assumes that the bitstream is not changed.

Just like the case of watermarking, the speech data hiding problem also has four fundamental constraints: *imperceptibility*, *capacity*, *robustness*, and *security*.

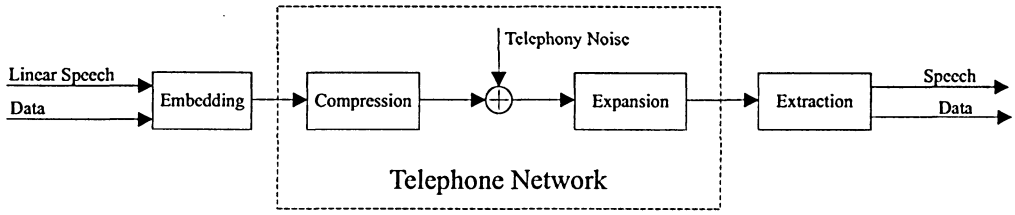

Figure 5.1: Data Embedding in μ-law Speech Signal

Imperceptibility is referred to the perceptual difference between the composite and original signal and can be tested subjectively. Especially in this application, imperceptibility must not be viewed as a binary condition. Different levels of perceptibility means different allowable energy of the embedded data, which determines the robustness directly.

Capacity is the most important consideration in this hiding system. The main objective is to embed a large amount of data into the host media. However, increasing the amount of embedded data causes the hidden information perceptible and degrades the perceptual quality of the composite signal.

Robustness is desired when the composite signal passes through some distortions. If the hidden data is still detectable after these distortions, then the system is regarded as robust. Due to the limitation of telephony channel, the embedded information should occupy the frequency range of $0.3 \sim 3.4kHz$ only. The attacks in this scenario mainly consist of the quantization noise from μ-law companding, as well as AWGN. A complete simulation of the telephony channel is described in [26].

Security is referred to as the resistance to the hostile attacks. The security criteria is application-dependent and is not considered in this research.

5.4 Possible Techniques

To be robust for high capacity embedding, a possible technique must be of host-suppressing at least. This requirement excludes most of currently available embedding techniques. The proposed algorithm is already analyzed in Chapter 3 and the steps for speech embedding are

summarized as follows.

5.4.1 Proposed Scheme and Simulation

The main steps of embedding consist of the following,

Turbo encoding: The binary information sequence is encoded with a Turbo encoder.

Time frequency analysis: Each frame of the speech signals is analyzed to output the frequency coefficients $\boldsymbol{x}$ and the masking threshold $\boldsymbol{m}$.

Embedding: Aided by three PN vectors $\boldsymbol{w_1}$, $\boldsymbol{w_2}$ and $\boldsymbol{w_3}$, each encoded information bit b is embedded into $\boldsymbol{x}$, a group of N MDCT coefficients.

The three vectors are all of length N and divide a plane uniformly with the angle of $\frac{2\pi}{3}$ from each other. Assuming that $\boldsymbol{w_1} = \{x_1, y_1, x_2, y_2, ...\}$, the two other vectors can be constructed as follows,

$$\{x_1 \cos\theta - y_1 \sin\theta, x_1 \sin\theta + y_1 \cos\theta, x_2 \cos\theta - y_2 \sin\theta, x_2 \sin\theta + y_2 \cos\theta, ...\} \tag{5.3}$$

where $\theta = \frac{2\pi}{3}$ for $\boldsymbol{w_2}$ and $\theta = \frac{4\pi}{3}$ for $\boldsymbol{w_3}$.

The embedding strategy is as follows,

1. $sort(||\boldsymbol{x} \cdot \boldsymbol{w_1}||, ||\boldsymbol{x} \cdot \boldsymbol{w_2}||, ||\boldsymbol{x} \cdot \boldsymbol{w_3}||)$, suppose the maximum result is from $\boldsymbol{w_1}$;

2. If $sign(b) = sign(\boldsymbol{x} \cdot \boldsymbol{w_1})$, embed the bit in $\boldsymbol{w_1}$ by

$$\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{x} \cdot \boldsymbol{w_1^{\perp}} + \alpha b \boldsymbol{w_1} \tag{5.4}$$

   where $\boldsymbol{w_1^{\perp}}$ is orthogonal to $\boldsymbol{w_1}$ and can be constructed by Equ. (5.3) of $\theta = \frac{\pi}{2}$.

3. Else choose the direction that yields $max(\boldsymbol{x} \cdot \boldsymbol{w_2}, \boldsymbol{x} \cdot \boldsymbol{w_3})$, suppose the direction is $\boldsymbol{w_2}$.

4. Embed the bit in the direction $\boldsymbol{w_2}$ as follows,

$$\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{x} \cdot \boldsymbol{w_2^{\perp}} + \alpha b \boldsymbol{w_2} \tag{5.5}$$

   where $\boldsymbol{w_2^{\perp}}$ is orthogonal to $\frac{\boldsymbol{w_1}+\boldsymbol{w_2}}{2}$ and can be constructed similarly as above.

The perceptual factor α in the above is determined by the following constraints to guarantee imperceptibility.

$$\alpha = \sqrt{m^2 - \sin\frac{\theta}{2} \cdot \frac{\sigma_x^2}{N}} \tag{5.6}$$

Reconstruction: use inverse MDCT transform to reconstruct the embedded signal in time domain. This signal may then be corrupted by companding and Gaussian noise during the transmission.

The main steps of extraction consist of the following,

Time frequency analysis: Each frame of the received speech signals is analyzed to output the frequency coefficients $\boldsymbol{y}$. Only the frequency coefficients of the range 0.3 $\sim$ 3.4 kHz are used for the following extraction.

Extraction: Compute and sort the projection of $\boldsymbol{y}$ onto the three vectors, $sort(||\boldsymbol{y} \cdot \boldsymbol{w_1}||, ||\boldsymbol{y} \cdot \boldsymbol{w_2}||, ||\boldsymbol{y} \cdot \boldsymbol{w_3}||)$. Choose the one (say, $\boldsymbol{w_1}$)with maximal absolute value. This direction is regarded as the one carrying the embedded information. The embedded bit is then decoded by $b' = sign(\boldsymbol{y} \cdot \boldsymbol{w_1})$

Turbo decoding: The extraction information is decoded with the corresponding Turbo decoder and the results is the information embedded.

In the simulation, the AWGN noise level is assumed, in terms of SNR, to be 30dB and 40dB to simulate different channels. A sequence of 128-bits are Turbo-encoded first. The embedding frequency range is chosen as 1 $\sim$ 3.4 kHz. According to author's testing, any embedding under $1kHz$ will introduce audible noises and therefore, that frequency range is left untouched. The embedded signals are rated with MOS of 4.3 on the average from the same subjects in the watermarking case.

The results are summarized in Figure 5.2. It can be seen that, under the attack of $SNR = 40dB$ in addition to the companding noise, the embedding rate could reach about 40 bps with a BER less than 1%.

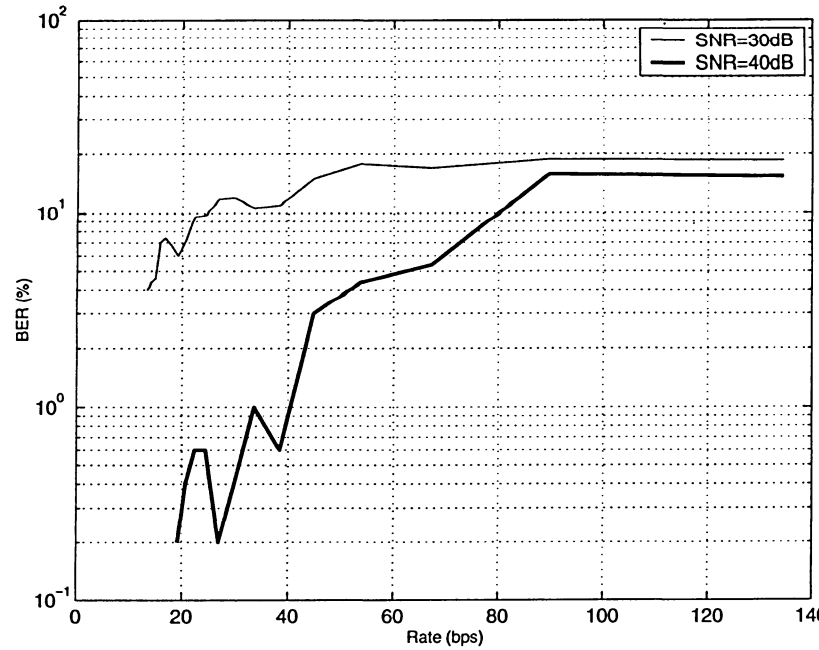

Figure 5.2: Embedding with the proposed scheme under different attacks (Turbo coded, HAS masking)

5.4.2 SCS scheme

In [41], the authors use SCS, incorporated with perceptual masking, to implement a perceptual embedding algorithm at the rate of 300 bps with BER as low as 10^{-4}, though the noises of channel is not stated clearly. To the author's knowledge, this result is the best one in the literature.

5.5 Chapter Summary

High-capacity embedding is a relatively new research area and little has been reported in the literature. The theoretical capacity bound can be computed by Equ. (4.2), though that of the implementable schemes is far below.

Different than the case of watermarking, the attacks in data embedding are supposed much less so that higher capacity can be expected. On the other hand, the sampling rate of speech signals is much less than that of audio signals, thus there is much less embedding

space available.

To be an effective embedding technique, host-suppression is highly desired. Among the current techniques only quantization-based ones possess this property. The disadvantage is the quantization steps should be known also at the receiving end, or have to be estimated blindly at the receiving end. The latter case will introduces errors in addition to that caused by channel interference.

Chapter 6

Conclusions and Future Research

In this thesis, we studied the blind, perceptual audio/speech embedding techniques. These techniques are the basis of many applications including digital watermarking and data hiding. The theoretical research on the capacity of such embedding shows that, many proprietary schemes are trivial either in robustness or embedding rate. Current research focuses on the two competing public techniques, i.e. SS and quantization-based schemes.

6.1 Audio Watermarking

6.1.1 Main Results

- Our proposals show that the modified SS embedding scheme can actually approach capacity when the spreading factor is large enough. It can outperform the current state-of-the-art quantization-based schemes including QIM and SCS. In fact, which of the SS or quantization-based schemes is superior is always an open question in the literature. Our research is aimed at answering this. The biggest disadvantage of the conventional SS scheme is the significant host impact at extraction. Our research uses an explicit strategy such that it can suppress the host impact better than the quantization-based schemes. Especially under strong attacks as in the watermarking cases, the modified SS scheme is superior to the quantization-based scheme.

- The proposals are general. All advantages of the conventional SS schemes including

low-energy and interference-suppressing are kept. And further, all efforts to increase the SNR of conventional SS schemes can still be used, which include, at least, the exploration of HAS, incorporation of FEC schemes, choice of the specific domains.

- In this research, we compare different typical FEC schemes when incorporated with various embedding techniques. Just as expected, the Turbo code is superior because it can actually decrease the BER to a very low level, whereas, other schemes like BCH or convolutional can only decrease the BER moderately. This property enables the Turbo code to be a promising scheme for the case of high accurate embedding, such as DVD watermarking.

- Exploring the masking effects of HAS is shown to be an effective measure to increase the SNR of the perceptual embedding. Actually, this is just the exact reason why MPEG coding can compress the audio signals effectively. In the cases with no such exploration, the permitted embedding power is actually the lowest masking threshold value in the whole spectrum. Depending on the specific audio clips, this power can be very small if the music happens to be very smooth in the duration.

- The quantization-based schemes are susceptible to dynamic scaling attacks. This is an important advantage of SS over quantization-based schemes. Though some efforts to cope with uniform scaling attacks for QIM/SCS had been made, quantizing the watermarked samples with different quantization steps can destruct the QIM/SCS watermarks easily. SS and its variants are immune to such attacks.

6.1.2 Discussions

- Both SS and quantization-based schemes need to be synchronized, i.e. the embedded samples must be synchronized with the spreading sequence at extraction. Simple attacks like dropping only one sample can be catastrophic.

- Ownership deadlock is not considered in the current research. Several parties can embed their own watermarks into the media by different spreading sequences or dif-

ferent information respectively. Then they can claim the ownership with the same algorithm. This deadlock could happen with the SS and its modifications, as well as quantization-based scheme.

- The frequency masking effects of HAS are much stronger in high frequency regions than that in low frequency regions. To cope with the attacks like low-pass filtering, the regions above the cutoff frequency have to be left unused by embedding. A cutoff frequency as low as 4 kHz can greatly reduce the embedding rate. Moreover, this also greatly reduces the ability to increase the SNR by exploring the masking effects of HAS.

- Also, a frame-by-frame real-time estimation of the masking threshold can be a big burden when embedding. An option is to use the TiQ as the masking threshold for every frame uniformly. This can be implemented by storing the TiQ curve initially and thus save significant computations. The trade-off is that dynamic masking effects are not utilized. Turbo code is also very heavy in complexity due to its iterations in decoding.

6.1.3 Future Research

- Attacks characterization can be expected as an active research area in watermarking. Its effectiveness has already been shown naively by our proposed scheme of the system consisting of multiple watermarks. An universally robust watermark against all kinds of attacks seems impractical, if not impossible. Designing specific watermarks against specific attacks seems very effective. Carefully choosing embedding domain seems to be a good implementation of this strategy.

- Exploring the masking effects of HAS can continue to be an effective measure for such embedding. Actually, the current MPEG masking models are too conservative. As shown in some literatures, the juxtaposition of the masking effects from different components are actually not linear. A non-linear approach can be expected to increase

the robustness and embedding rate significantly.

- An important problem needs to be cared is the synchronization of the current high-capacity embedding techniques including the modified SS and quantization based schemes. Some efforts have been made as shown in the literature. Generally saying, these ideas use some special spreading sequences to find the beginning point by projections, but they do not solve the attacks like dropping some samples in the middle. This could happen during transmission by intentional attacks or unintentional manipulations.

6.2 Speech Data Hiding

Data hiding in telephony speech signals can be of many potential applications in the industry. The research on this area is still in its early stage and few results are published.

Comparing with the embedding in audio signals, the usable bandwidth is greatly reduced, from 22.05 kHz to 4 kHz. According to Costa's result, the corresponding capacity will be reduced to 1/5th accordingly. Further, due to the characteristics of telephony network and inherent noise from the non-linear companding to compress speech signals, this expected capacity has to be further reduced. Another important factor limiting this embedding is that, because the sampling frequency is bounded by 4 kHz, exploring the HAS will not be so effective as in audio embedding since much of the masking effects occur in high frequency regions.

The embedding rate reached in our current research is low, and it seems to be far from the expected, and even from the publish results (though the channel noises is unknown in the publications). The future research can focus on two aspects as follows,

- A theoretical research on the capacity will be of great value. With the additional constraints of data hiding in speech signals, the embedding capacity should be different from Costa's result. This research can reveal the potentials of this technique and evaluate future schemes.

- Exploring the masking effects will definitely help to increase the embedding rate in speech signals. The effects of the narrow-band speech signals should be different from that of wide-band audio signals. Careful simulations are needed to setup the effective masking models of the speech signals. This research can also benefit other research branches in speech processing.

Bibliography

[1] H. N. Azghandi and P. Kabal, *Improving Perceptual Coding of Narrowband Audio Signals at Low Rates,* IEEE International Conference of Acoustics, Speech, Signal Processing, Vol. 2, pp. 913-916, March 1999.

[2] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, *Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate,* IEEE Transactions on Information Theory, Vol. 20, No. 2, pp. 284-287, March 1974.

[3] M. Barni, C. I. Podilchuk, F. Bartolini, E. J. Delp, *Watermark Embedding: Hiding a Signal Within a Cover Image,* IEEE Communications Magazine, Vol. 39, No. 8, pp. 102-108, August 2001.

[4] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. *Techniques for Data Hiding,* IBM Systems Journal, Vol. 35, No. 3-4, pp. 313-336, 1996.

[5] C. Berrou, A. Glavieux, and P. Thitimajshima, *Near Shannon Limit Error-Correcting Coding and Decoding: Turbo Codes,* Proceedings of International Communications Conference, Vol. 2, pp. 23-26, May 1993.

[6] C. Berrou and A. Glavieux, *Near Optimum Error Correcting Coding and Decoding,* IEEE Transactions on Communications, Vol. 44, No. 10, pp. 1261-1271, October 1996.

[7] C. W. Brokish, M. Lewis, *A-law and μ-aw companding implementations using the TMS320C54x–Application Note: SPRA163A,* Digital Signal Processing Solutions, Texas Instruments, December 1997.

[8] B. Chen and G. W. Wornell, *Preprocessed and Postprocessed Quantization Index Modulation Methods for Digital Watermarking,* Proc. of SPIE: Security and Watermarking of Multimedia Contents II, Vol. 3971, pp. 48-59, January 2000.

[9] B. Chen and G. W. Wornell, *Quantization Index Modulation: A Class of Provably Good Methods for Digital Watermarking and Information Embedding,* IEEE Transactions on Information Theory, Vol. 47, No. 4, pp. 1423-1433, May 2001.

[10] M. Cheng and Y. Hsu, *Fast IMDCT and MDCT Algorithms–A Matrix Approach,* IEEE Transactions on Signal Processing, Vol. 51, No. 1, pp. 221-229, January 2003.

[11] M. Costa, *Writing on Dirty Paper*, IEEE Transactions on Information Theory, Vol. 29, No. 3, pp. 439-441, May 1983.

[12] T. Cover and J. Thomas, *Elements of Information Theory,* John Wiley and Sons Inc., 1991.

[13] I. J. Cox, J. Killian, T. Leighton, and T. Shamoon, *Secure Spread Spectrum Watermarking for Multimedia,* IEEE Transactions on Image Processing, Vol. 6, No. 12, pp. 1673-1687, December 1997.

[14] I. J. Cox, M. L. Miller, and A. L. McKellips, *Watermarking as Communications with Side Information,* Proceedings of the IEEE, Special Issue on Identification and Protection of Multimedia Information, Vol. 87, No. 7, pp. 1127-1141, July 1999.

[15] I. J. Cox, M. L. Miller, *Electronic Watermarking: the first 50 years,* IEEE 4th Workshop on Multimedia Signal Processing, No. 3-5, pp. 225-230, October 2001.

[16] N. Cvejic and T. Seppnen, *Audio Prewhitening based on Polynomial Filtering for Optimal Watermark Detection,* Proceedings of XI European Signal Processing Conference, Vol. 3, pp. 69-72, July 2002.

[17] N. Cvejic and T. Seppnen, *Increasing Robustness of an Audio Watermark using Turbo Codes,* IEEE International Conference on Multimedia and Expo, Vol. 1, No. 6-9, pp. 17-20, July, 2003.

[18] H. Ding, *Sub-channel below the Perceptual Threshold in Audio,* Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 49-52, April 2003.

[19] J. J. Eggers, J. K. Su and B. Girod, *A Blind Watermarking Scheme Based on Structured Codebooks,* IEE Seminar on Secure Images and Image Authentication, Vol. 4, pp. 1-21, April 2000.

[20] J. J. Eggers, J. K. Su and B. Girod, *Performance of a Practical Blind Watermarking Scheme,* Proceedings of SPIE: Electronic Imaging 2001, Security and Watermarking of Multimedia Contents III, Vol. 4314, pp. 1-12, January 2001.

[21] J. J. Eggers, R. Buml, R. Tzschoppe and B. Girod, *Scalar Costa Scheme for Information Embedding,* IEEE Transactions on Signal Processing, Vol. 51, No. 4, pp. 1003-1019, April 2003.

[22] A. M. Eskicioglu, *Protecting Intellectual Property in Digital Multimedia Networks,* IEEE Computer Society, Special Issue on Piracy and Privacy, Vol 36, No. 7, pp. 39-45, July 2003.

[23] P. G. Flikkema, *Spread-Spectrum Techniques for Wireless Communication,* IEEE Signal Processing Magazine, Vol. 14, No. 3, pp. 26-36, May 1997.

[24] J. Hagenauer and P. Hoeher, *A Viterbi algorithm with Soft-Decision outputs and its applications,* Proceedings of GLOBECOM, Vol.3, pp. 1680-1686, November 1989.

[25] J. Hagenauer, E. Offer and L. Papke, *Iterative Decoding of Binary Block and Convolutional codes,* IEEE Transcations on Information Theories, Vol. 42, pp. 429-445, March 1996.

[26] ITU-T. *Network Transmission Model for Evaluating Modern Performance over 2-wire Voice Grade Connections.* Technical Report V.56 bis, August 1995.

[27] J. Johnston, *Transform Coding of Audio Signals using Perceptual Noise Criteria,* IEEE Journal on Selected Areas of Communication, Vol. 6, No. 2, pp. 314-323, February 1988.

[28] D. Kirovski and H. S. Malvar, *Robust Spread Spectrum Audio Watermarking,* IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3, No. 7-11, pp. 1345-1348, May 2001.

[29] D. Kirovski and H. S. Malvar, *Spread Spectrum Watermarking of Audio Signals,* IEEE Transactions on Signal Processing, Vol. 51, No. 4, pp. 1020-1033, April 2003.

[30] D. Kundur and D. Hatzinakos, *Diversity and Attack Characterization for Improved Robust Watermarking,* IEEE Transactions on Signal Processing, Vol. 49, No. 10, pp. 2383-2396, October 2001.

[31] S. Lin and D. J. Costello, *Error Control Coding: Fundamentals and Applications*, Prentice Hall Inc., Englewood Cliffs, 1983.

[32] V. K. Madisetti and D. B. Williams, *The Digital Signal Processing Handbook,* CRC Press and IEEE Press, 1998.

[33] H. S. Malvar and D. A. Florencio, *Improved Spread Spectrum: A New Modulation Technique for Robust Watermarking,* IEEE Transactions on Signal Processing, Vol. 51, No. 4, pp. 898-905, April 2003.

[34] P. Noll, *MPEG Digital Audio Coding,* IEEE Signal Processing Magazine, Vol. 14, No. 5, pp. 59-81, September 1997.

[35] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, 2000.

[36] T. Painter and A. Spanias, *Perceptual Coding of Digital Audio,* IEEE Proceedings, Vol. 88, No. 4, pp. 451-515, April 2000.

[37] D. Pan, *A Tutorial on MPEG/Audio Compression,* IEEE Multimedia, Vol. 2, No. 2, pp. 60-74, June 1995.

[38] J. P. Princen and A. B. Bradley, *Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation,* IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 34, No. 5, pp. 1153-1161, October 1986.

[39] J. P. Princen, A. W. Johnson and A. B. Bradley, *Subband/Transform Coding using Filter Bank Designs Based on Time Domain Aliasing Cancellation,* IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 12, pp. 2161-2164, April 1987.

[40] P. Robertson, E. Villebrun, and P. Hoeher, *A Comparison of Optimal and Sub-Optimal MAP Decoding Algorithms Operating in the Log Domain,* Proceedings of International Communications Conference, Vol. 2, No. 18-22, pp. 1009-1013, June 1995.

[41] A. Sagi and D. Malah, *Data Embedding in Speech Signals using Perceptual Masking*, to be published in European Signal Processing Conference 2004.

[42] J. Seok and J. Hong, *Audio Watermarking for Copyright Protection of Digital Audio Data,* Electronics Letters, Vol. 37, No. 1, pp. 60-61, August 2001

[43] B. Sklar, *Digital Telecommunications, Fundamentals and Applications*, Prentice Hall, 2001.

[44] M. Swanson, B. Zhu and A. Tewfik, *Current state-of-art, Challenges and Future directions for Audio Watermarking,* Proceedings of IEEE Internaltional Conference on Mutimedia Computing and Systems, Vol. 1, pp. 7-11, June 1999 .

[45] M. Steinebach, F. A. P. Petitcolas, F. Raynal, J. Dittmann, C. Fontaine, C. Seibel, N. Fates and L. C. Ferri, *StirMark benchmark: Audio Watermarking Attacks,* International Conference on Information Technology: Coding and Computing, Vol. 2, No. 4, pp. 49-54, April 2001.

[46] A. J. Viterbi, *Approaching the Shannon limit: Theorists' dream and practitioners' challenge,* Proceedings of International Conference on Millimeter Wave and Far Infrared Science and Technology, Vol. 2, No. 4, pp. 111-114, August 1996

[47] J. P. Woodward and L. Hanzo, *Comparative Study of Turbo Decoding Techniques: An Overview,* IEEE Transactions on Vehicular Technology, Vol. 49, No. 6, pp. 2208-2233, November 2000.

[48] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer, 1999.

Appendix A

List of Publications

In this section, we list the publications resulted from our research work for the thesis.

- L. Zhang, S. Krishnan and H. Ding, "Modified Spread Spectrum Audio Watermarking Algorithm," 2004 Annual Conference of Canadian Acoustic Association, Ottawa, October 2004.
- L. Zhang, S. Krishnan and H. Ding, "Improved Spread Spectrum Audio Watermarking Algorithm," submitted to 2005 IEEE International Conference on Acoustics, Speech and Signal Proceedings (ICASSP).

Appendix B

Important Mathematical Deductions

Given the sequences of $\boldsymbol{x} = \{x_1, x_2, ..., x_N\}$, $\boldsymbol{n} = \{n_1, n_2, ..., n_N\}$, x_i , $i = 1, 2, ..., N$, are statistically *independent and identically distributed* (i.i.d.) random variables, each having a finite mean $m_x = 0$ and a finite variance σ_x. Similarly, n_i, $i = 1, 2, ..., N$, are i.i.d. random variables with a finite mean $m_n = 0$ and a finite variance σ_n. The zero-mean spreading sequence, $\boldsymbol{w} = \{w_1, w_2, ..., w_N\}$, is *statistically independent* of both $\boldsymbol{x}$ and $\boldsymbol{n}$.

B.1 Conventional Spread Spectrum Embedding Scheme

As shown in Equ. (3.2), the normalized correlation is,

$$c = \boldsymbol{r} \cdot \boldsymbol{w} = (\boldsymbol{x} + \alpha b \boldsymbol{w} + \boldsymbol{n}) \cdot \boldsymbol{w} = \alpha b + (\boldsymbol{x} + \boldsymbol{n}) \cdot \boldsymbol{w} = \alpha b + \frac{1}{N} \sum_{i=1}^{i=N} (x_i + n_i) w_i \tag{B.1}$$

By the *central limit theorem*, the correlation c, being the sum of the i.i.d. random variables with finite mean and variance, approaches a Gaussian distribution as $N \to \infty$, i.e. $c \sim \boldsymbol{N}(m_c, \sigma_c)$.

The mean of c is,

$$m_c = E\left(\alpha b + \frac{1}{N} \sum_{i=1}^{i=N} (x_i + n_i) w_i\right) = \alpha b + \frac{1}{N} \sum_{i=1}^{i=N} E[(x_i + n_i) w_i] \tag{B.2}$$

$$= \alpha b + \frac{1}{N} \sum_{i=1}^{i=N} E[x_i + n_i] E[w_i] = \alpha b \tag{B.3}$$

The variance of c is,

$$\sigma_c^2 = E[(c-m_c)^2] = E\left(\left[\frac{1}{N}\sum_{i=1}^{i=N}(x_i+n_i)w_i\right]^2\right) = \frac{1}{N^2}\sum_{i=1}^{i=N}E\left([(x_i+n_i)w_i]^2\right) \tag{B.4}$$

$$= \frac{1}{N^2}\sum_{i=1}^{i=N}E\left[(x_i+n_i)^2\right] = \frac{1}{N^2}\sum_{i=1}^{i=N}[E(x_i^2)+E(n_i^2)] = \frac{\sigma_x^2+\sigma_n^2}{N} \tag{B.5}$$

Thus it is proved that $c \sim \boldsymbol{N}(\alpha b, \frac{\sigma_x^2+\sigma_n^2}{N})$, which results in the BER performance shown in Equ. (3.4).

B.2 Improved Spread Spectrum Embedding Scheme with Two Orthogonal Spreading Sequences

Given the spreading sequence $\boldsymbol{w}$, the auxiliary orthogonal sequence $\boldsymbol{v}$ can be constructed by the following procedure,

$$\boldsymbol{w} = \{w_1, w_2, w_3, w_4, ...\} \Rightarrow \boldsymbol{v} = \{-w_2, w_1, -w_4, w_3, ...\} \tag{B.6}$$

As shown in Equ. (3.10), the normalized correlation is,

$$c = \boldsymbol{r}\cdot\boldsymbol{w} - \boldsymbol{r}\cdot\boldsymbol{v} = \alpha_M b + \boldsymbol{n}\cdot(\boldsymbol{w}-\boldsymbol{v}) = \alpha_M b + \frac{1}{N}\sum_{i=1}^{i=N}(w_i - v_i)n_i \tag{B.7}$$

Similar to the above, by the *central limit theorem*, the correlation c, being the sum of the i.i.d. random variables with finite mean and variance, approaches a Gaussian distribution as $N \to \infty$, i.e. $c \sim \boldsymbol{N}(m_c, \sigma_c)$.

The mean of c is,

$$m_c = E\left(\alpha_M b + \frac{1}{N}\sum_{i=1}^{i=N}(w_i - v_i)n_i\right) = \alpha_M b + \frac{1}{N}\sum_{i=1}^{i=N}E[(w_i - v_i)n_i] \tag{B.8}$$

$$= \alpha_M b + \frac{1}{N}\sum_{i=1}^{i=N}E[w_i - v_i]E[w_i] = \alpha_M b \tag{B.9}$$

The variance of c is,

$$\sigma_c^2 = E[(c-m_c)^2] = E\left(\left[\frac{1}{N}\sum_{i=1}^{i=N}(w_i - v_i)n_i\right]^2\right) = \frac{1}{N^2}\sum_{i=1}^{i=N} E\left([(w_i - v_i)n_i]^2\right) \tag{B.10}$$

$$= \frac{1}{N^2}\sum_{i=1}^{i=N} E\left[(w_i^2 + v_i^2)n_i^2\right] = \frac{2\sigma_n^2}{N} \tag{B.11}$$

Thus it is proved that $c \sim \boldsymbol{N}(\alpha b, \frac{2\sigma_n^2}{N})$.

The embedding power is as follows,

$$W = E[||\boldsymbol{y} - \boldsymbol{x}||] = E[(\boldsymbol{y} - \boldsymbol{x}) \cdot (\boldsymbol{y} - \boldsymbol{x})] = E[(\alpha_M b\boldsymbol{w} + \beta\boldsymbol{v}) \cdot (\alpha_M b\boldsymbol{w} + \beta\boldsymbol{v})] \tag{B.12}$$

$$= \alpha_M^2 + E(\beta^2) = \alpha_M^2 + E\left(\left[\frac{1}{N}\sum_{i=1}^{i=N} x_i(w_i - v_i)\right]^2\right) \tag{B.13}$$

$$= \alpha_M^2 + E\left(\left[\frac{\sqrt{2}}{N}\sum_{i=1}^{i=N} x_i w_i\right]^2\right) \tag{B.14}$$

$$= \alpha_M^2 + \frac{2\sigma_x^2}{N} \tag{B.15}$$

This results in the improved performance shown in Equ. (3.13).

B.3 Improved Spread Spectrum Embedding Scheme with Two Non-orthogonal Spreading Sequences

Given one spreading sequence $\boldsymbol{w}$, the auxiliary non-orthogonal sequence $\boldsymbol{v}$ can be constructed by the following procedure,

$$\boldsymbol{w} = \{w_1, w_2, w_3, w_4, ...\}$$

$$\boldsymbol{v} = \{w_1\cos\theta - w_2\sin\theta, w_1\sin\theta + w_2\cos\theta, w_3\cos\theta - w_4\sin\theta, w_3\sin\theta + w_4\cos\theta, ...\} \tag{B.16}$$

such that,

$$\boldsymbol{w} \cdot \boldsymbol{v} = \frac{1}{N}\sum_{i=1}^{i=N} w_i v_i = \frac{1}{N} \cdot \sum_{i=1}^{i=N} w_i^2 \cdot \cos\theta = \cos\theta \tag{B.17}$$

As shown in Equ. (3.10), the normalized correlation is,

$$c = \boldsymbol{r}\cdot\boldsymbol{w} - \boldsymbol{r}\cdot\boldsymbol{v} = \alpha_M b(1-\boldsymbol{w}\cdot\boldsymbol{v}) + \boldsymbol{n}\cdot(\boldsymbol{w}-\boldsymbol{v}) = \alpha_M b\cdot 2\sin^2\frac{\theta}{2} + \frac{1}{N}\sum_{i=1}^{i=N}(w_i - v_i)n_i \quad \text{(B.18)}$$

Similar to the above, by the *central limit theorem*, the correlation c, being the sum of the i.i.d. random variables with finite mean and variance, approaches a Gaussian distribution as $N \to \infty$, i.e. $c \sim \boldsymbol{N}(m_c, \sigma_c)$.

The mean of c is,

$$m_c = E\left(\alpha_M b\cdot 2\sin^2\frac{\theta}{2} + \frac{1}{N}\sum_{i=1}^{i=N}(w_i - v_i)n_i\right) = \alpha_M b\cdot 2\sin^2\frac{\theta}{2} + \frac{1}{N}\sum_{i=1}^{i=N}E[(w_i - v_i)n_i] \quad \text{(B.19)}$$

$$= \alpha_M b\cdot 2\sin^2\frac{\theta}{2} + \frac{1}{N}\sum_{i=1}^{i=N}E[w_i - v_i]E[w_i] = \alpha_M b\cdot 2\sin^2\frac{\theta}{2} \quad \text{(B.20)}$$

The variance of c is,

$$\sigma_c^2 = E[(c-m_c)^2] = E\left(\left[\frac{1}{N}\sum_{i=1}^{i=N}(w_i - v_i)n_i\right]^2\right) = \frac{1}{N^2}\sum_{i=1}^{i=N}E\left([(w_i - v_i)n_i]^2\right) \quad \text{(B.21)}$$

$$= \frac{1}{N^2}\sum_{i=1}^{i=N}E\left((w_i^2 + v_i^2 - 2w_iv_i)n_i^2\right) = 2(1-\cos\theta)\cdot\frac{\sigma_n^2}{N} = \frac{\sigma_n^2}{N}\cdot 4\sin^2\frac{\theta}{2} \quad \text{(B.22)}$$

Thus it is proved that $c \sim \boldsymbol{N}\left(\alpha_M b\cdot 2\sin^2\frac{\theta}{2},\ \frac{\sigma_n^2}{N}\cdot 4\sin^2\frac{\theta}{2}\right)$, which results in the maximized performance when $\theta = \pi$ as shown in Equ. (3.22).