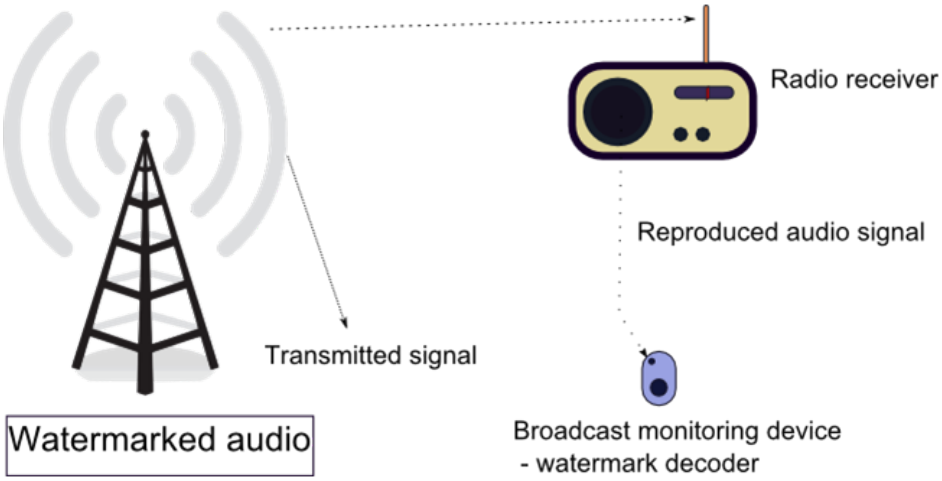
CHALMERS



Exploiting side information for audio watermarking: A sample system case study

Master of science thesis

Chandramouli Soorian

Department of Signals and Systems, Signal Processing Group
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden, 2010
Report No. Ex071/2010

Report No. Ex071/2010

Exploiting side information for audio watermarking: a sample system case study

Chandramouli Soorian
Personal number: 840126 - 9470

Examiner: Prof. Irene Yu-Hua Gu (Chalmers Univ. of Technology, Sweden)

Supervisor: Dr.-Ing. Giovanni Del Galdo (Fraunhofer IIS, Germany)

Chalmers University of Technology, Dept. of Signals and Systems
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr.-Ing. Giovanni Del Galdo for all the hours he had spent, not only in brainstorming or discussing the implementation details, but also in helping me write good code and testing it. He has been more than a good supervisor, he has been a good teacher and from him I have learnt the power of simplifying a problem and the power of graphical visualization.

I would also like to thank Ing. Tobias Bleim for helping me navigate through the large amounts of code and documentation. His in depth knowledge of the system, the code and his eye for details have made this journey a lot easier. I would also like to thank Ing. Alexandra Craciun and Ing. Florian Kolbeck, for reviewing my thesis document through its many iterations, and suggesting ways to improve the material.

I would also like to thank Ing. Stefan Kraegeloh for providing me the opportunity to work on my thesis here at Fraunhofer IIS. I would also like to thank the other members of the audio watermarking team: Ing. Juliane Borsum and Ing. Bert Greevenbosch, for their support during my time here at Fraunhofer, IIS.

DECLARATION

I declare to have written this thesis independently. The use of sources (paraphrasing, ideas, quotations) has been properly documented.

Erlangen, the 16. August 2010

ABSTRACT

Audio watermarking is the process of embedding information into an audio signal so that the embedded information is inseparable from it and imperceptible to the listener. There are various applications for such embedded information, one such commercial application is broadcast monitoring. In this application, a device is used to extract the information embedded into broadcast signals. The extracted information allows the device to track the media consumed by the user.

In this thesis, the various subsystems of the low complexity, real-time system currently designed for broadcast monitoring, the watermarking system under study, are studied. The initial goal of the thesis was to find ways to improve or adapt the current system in order to make it suitable for other applications. During the study, it was found that the system, modeled as a Communication System does not use the available side information (the original audio) completely.

The original audio in the current system acts as an interference. According to the paper *Writing on Dirty Paper* [3], it is theoretically possible to exploit the side information at the transmitter with the aim of removing the interference from the audio. In this thesis we will explore two such methods the *Quantization Index Modulation (QIM)* and *Dirty Paper Trellis Codes (DPTC)*.

QIM was found to perform very well at high SNRs but poorly at low SNRs. The system was tested with the intention of increasing the rates, it was found to perform very well compared to the original. Further, the QIM under the effects of compression and codec conversion, was found to perform better than the original system.

CONTENTS

1. INTRODUCTION TO AUDIO WATERMARKING

Watermarking is the process of embedding information into an object (video, audio or image) such that it is imperceptible and inseparable from it [1]. Among the many possible applications of such a watermark, the following are included: ownership/source authentication, copyright protection, fingerprinting, copy/modification control, covert communication (steganography) and broadcast monitoring [5, p. 20-22].

In the current context, watermarking can be interpreted as embedding information into a signal, with the aim of later being able to identify the signal again. Depending on the application, it can also have other interpretations, such as embedding information that cannot be easily removed/altered without distorting the original signal significantly.

Figure 1.1 shows a basic block description of a watermarking system, where the message is added into the object by the encoder. The decoder in turn, retrieves the message from the watermarked object.

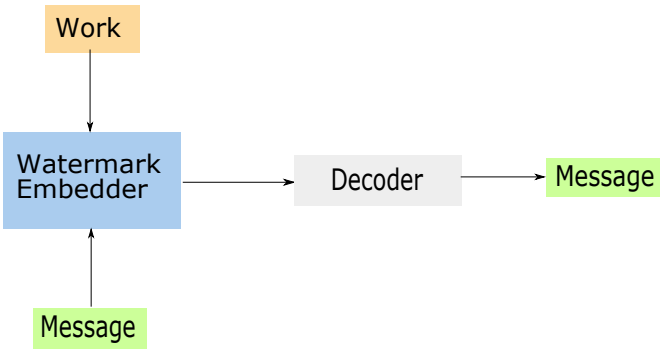


Fig. 1.1: Basic Audio watermarking description

The addition of a watermark to an audio signal should be done without causing any *perceptible* distortion to the original audio signal. By exploiting the properties of the Human Auditory System (HAS), we can make sure that the distortion due to the watermark is not perceptible. In Chapter 2.1 we introduce psychoacoustics as the

study of how sounds are perceived by the human brain and we define the psychoacoustic model, which can be used to exploit the properties of the HAS.

New challenges are introduced when adding digital information to the analog signal, which is analogous to transmitting binary information in a noisy communication channel. Techniques used to build a communication system through a noisy channel can be adapted to an audio watermarking system (see [5, p. 24]).

An audio watermarking system has been developed at Fraunhofer Institute for Integrated Circuits (FIIS) for broadcast monitoring. Some of its inherent properties include low complexity and real time performance. In this thesis, we seek to identify areas that can be improved without sacrificing significantly on the properties or design of the original watermarking system under study. As stated earlier, the system is modeled as a communication system with a noisy channel. In addition to this, our model tries to minimize the perceptible distortion caused by the transmitted information. The watermarking system under study consists of a number of blocks present in any communication system, such as: synchronization, channel coding, modulation/demodulation etc. Apart from these blocks, it also includes a *psychoacoustic module* that helps us minimize the perceptible distortions.

The modules that make up the system were investigated for possible improvements and their performance was studied under various conditions including compression and codec conversion. The current system does not use the information about the interference from the audio known to it at the transmitter (*side information*). However, there are various ways to make use of the side information in order to improve the system. Two such methods are Quantization Index Modulation (QIM) and Dirty Paper Trellis Codes (DPTC). The QIM was implemented and showed improved bit-error rate (BER) at high signal-to-noise ratio (SNR), while at low SNR it showed poor performance due to its susceptibility to noise. The DPTC is intuitively identified as a better option than the QIM because it works at a higher layer (channel coding layer) than the QIM (modulation layer). This provides more flexibility by moving the operating point at a lower SNR, which is not the case with QIM. Nevertheless, this is achieved at the cost of increased complexity both at the encoder and at the decoder side.

This thesis is organized in six chapters. Chapter 2 provides a brief overview of the background material needed to understand the psychoacoustic model. It also includes a list of the audio signal properties, which could be exploited for data hiding. Possible application areas and the required properties of the underlying watermarking system are also covered. The communication channel model is briefly described at the end of this chapter.

Chapter 3 provides an introduction to the watermarking system under study, the broadcast monitoring application and the various subsystems included in our model. Chapter 4 elaborates on the goals of the thesis, the approach used to study the system and the reasoning behind the intermediate decisions.

Chapter 5 presents Costa's *Writing on Dirty Paper*, on which we base our attempts of achieving better system performance. The next part of the thesis discusses QIM, its use in image and audio watermarking, its adaptation to the current system and the results obtained. This is followed by a brief overview of DPTC, its advantages over QIM and the challenges it poses. We will also discuss possible methods of using it in the current system. Finally, we sum up the results of the simulations and present possible future work in this direction.

Finally in Chapter 6, we conclude with the results and possible future work/extensions that could be made to this thesis.

2. BACKGROUND INFORMATION FOR AUDIO WATERMARKING

This chapter introduces the field of *psychoacoustics*, based on which we try to perceptually hide the embedded data. Then we give an overview of the psychoacoustic properties being exploited by various watermarking techniques. This is followed by a discussion of the trade-offs between the desired properties. The next section presents various applications of watermarking and their requirements. Finally, we give a brief introduction on how to model a watermarking system by using the communication model.

2.1 Perceptual hiding

An important property of a watermarking system is to make sure that the presence of embedded information is transparent to the end user (certain applications may also require an audible/visible watermark). To perceptually hide information one would have to study the perceptual properties of the host signal. Audio watermarking poses greater challenges than image/video watermarking, mainly because of the wider dynamic range and higher sensitivity to AWGN of the HAS, compared to the Human Visual System (HVS). But in contrast to its large dynamic range, its differential range is limited. This means that louder sounds can *mask* out weaker ones [10]. The HAS also has a number of other properties which can be exploited to perceptually hide the distortions caused by embedding information.

2.2 Audio watermarking techniques

The watermarking techniques depend heavily on the study of the HAS and psychoacoustics. The properties studied can be used to make the added watermark imperceptible to the human ear, while still retaining sufficient strength for robustness.

2.2.1 Psychoacoustic properties

Phase

In the technique exploiting phase properties, the watermark is embedded by modifying the phase of the original audio signal. Thus, the embedding is done in frequency subbands which are sufficiently high such that a change in phase does not result in significant shift of energy in time. According to [5, p. 41], the introduction of a phase change is imperceptible to the HAS if the change remains within limits:

$$\|\Delta\phi(Z)\|/\Delta Z < 30\,\mathrm{deg},$$

where $\phi(Z)$ is the change in phase, and Z is the bark scale. Since the change in energy has to be slow in time, the phase change has to be implemented over long blocks($N = 2^{14}$). Alternatively if smaller block sizes are desired, the lower frequencies can be shifted only slightly in phase whereas the higher frequencies can freely choose their phase without introducing perceptual artifacts to the audio signal. One of the drawbacks of this method is that it is suceptible to phase rotation introduced by the channel which varies in frequency and time (slowly).

Least significant bit

In this method, the least significant bit in the original audio is used to carry the watermark information. The additional distortion in the least significant bits can be seen as noise and since the power contributed by this is very low, it can be considered perceptually insignificant. The samples chosen for embedding the data is chosen using a secret key. The HAS is suceptible to noise and is quite sensitive to even low additive noise, this would mean that not all the chosen bits would be perceptually transparent. This system provides good data rates (44.1 kbps using one LSB) because the perceptual distortion in the original audio can be kept low. This is also a very low complexity embedding/decoding and has very little algorithmic delay since there are no demanding transformations. On the flip side, addition of a small noise or re-quantization can easily destroy this watermark and hence, it is not a very robust system. It is also highly unlikely that the embedded information would survive an analog to digital or digital to analog conversion (see [5, p. 40]).

Echo

Echoes are imperceivable if the delay is not too large, the acceptable limit being usually assumed to be one ms. This property can therefore be used to embed information bits in the audio [5, p. 42]. However, the audio signal is often reproduced in a room, this can introduce echos of its own. With extraneous echos added to the signal, the embedded message can become distorted beyond recoverability.

2.2.2 Masking

The ability of an audio signal to make other audio signals in its neighborhood (both in time and frequency) imperceptible to the human ear is called *masking*. A loud sound distorts the absolute threshold of hearing and thus renders inaudible a weaker sound that may otherwise be audible [10]. We can observe in Figure 2.1, which depicts the masking effect, that the *masker* hides/masks any signal falling within its *masking threshold* (like the *masked sound*). *The threshold in quiet* is the minimum sound pressure level of a signal that can be heard by the HAS in the absence of other sounds.

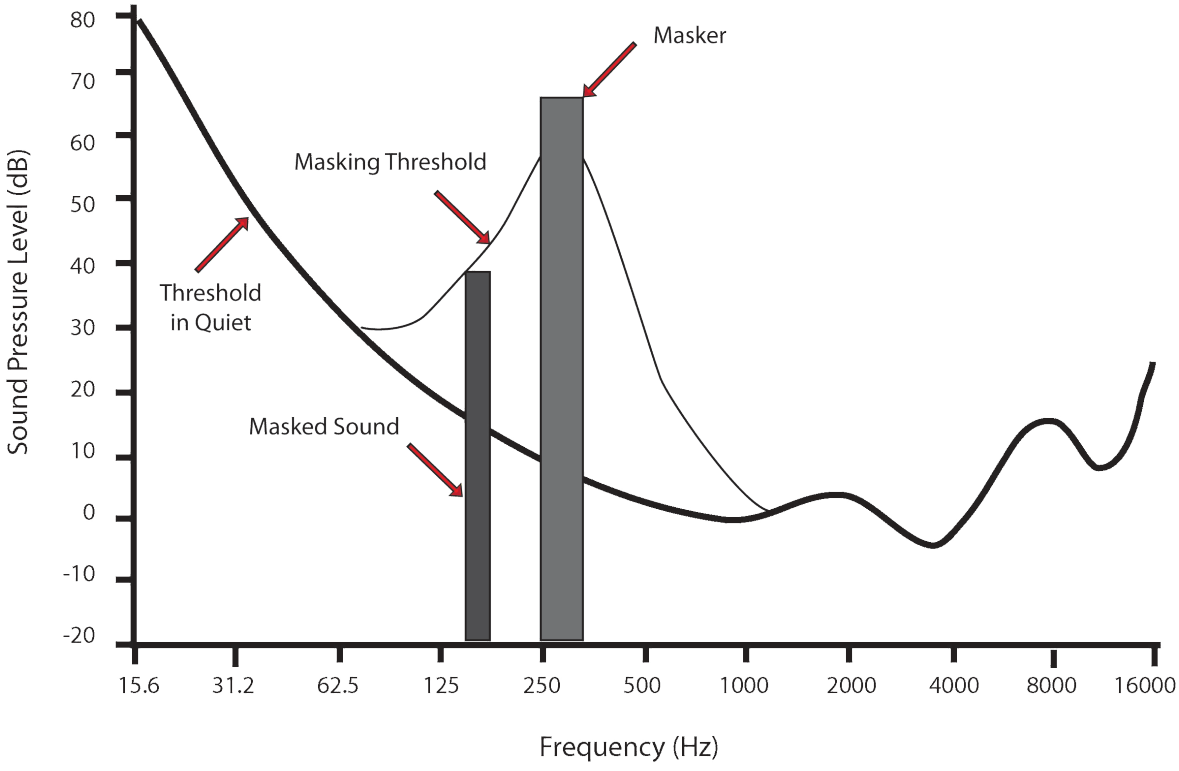


Fig. 2.1: Masking effect ([2])

2.3 Masking mechanism

The current system makes use of the masking effects of the audio signal to perceptually hide the embedded data. The dynamic range of frequencies that can be perceived by the HAS are divided nonlinearly into *critical bands*. The critical bands are themselves modeled as strongly overlapping bandpass filters, being a close analogue of the *basilar membrane*. The basilar membrane consists of elongated fibers located inside the *cochlea* (the inner ear), which vibrate in response to an incoming sine wave (sound). Any signal, which falls within a range (critical band) of frequencies from the center frequency, is picked up by this membrane (filter). The range of frequencies within a critical band represent the frequencies, which can be masked by other signals within this range.

2.3.1 Temporal and spectral masking

Based on studies of the HAS, it is possible to add an information signal to an audio signal such that it is imperceptible to the human ear. The information signal is masked by the original audio signal. The masking can be either temporal or spread across frequency.

Temporal masking

An audio signal at any point in time can make signals in the neighborhood imperceptible to hearing. The maximum strength of the signal that can be masked by the masker decreases with distance from the masker. As a general observation, the effect of temporal masking precedes the masker by 20 ms and the effects last for about 50 to 200 ms after the masker [10].

Frequency/Simultaneous masking

A tone or a noiselike signal in a particular frequency band can mask signals at other nearby frequencies that occur simultaneously. Noise like maskers have better masking properties than tonal maskers. The net masking effect of the temporal and frequency masking is not additive. Instead, the more dominant masking effect prevails [10].

2.4 Evaluation - Qualitative and quantitative study of watermarks

Depending on the application, a watermarking system is expected to exhibit certain properties. It is desirable to study these qualitatively as well as quantitatively. There

are no absolute or unambiguous ways to quantify the effectiveness of a watermarking system. It is nevertheless valuable for evaluating one watermarking mechanism over another for a particular application. In the following sections, we introduce the properties of watermarking systems. These can be used to compare the different systems listed in [5, p. 22-23].

Perceptual transparency

Addition of a watermark to the host signal causes a degradation in the quality of the original signal. However, the watermarked signal can be made perceptually similar using psychoacoustics. The fidelity of a watermark is often defined as the perceptual closeness of the watermarked signal to the original signal. However, in many applications such as broadcast monitoring the signal undergoes significant degradation before reaching the consumer. As a result, it would be sufficient to define the fidelity of the signal at the point where it reaches the consumer.

Perceptual transparency is a subjective measure and can only be measured by subjective tests called *listening tests*. The listening tests usually involve human test subjects chosen from a sample population who volunteer to listen to sample audio material. This material may or may not contain distortions, which the listeners need to rate against a reference with no distortion.

Bit rate

The bit rate requirements depend heavily on the application. Copy control for instance might require a maximum of 0.5 bit/s, whereas broadcast monitoring, where the advertisements may be very short, might require average bit rates of 15 bit/s [5, p. 22]. In addition, applications holding metadata such as lyrics might require much higher data rates.

Robustness

Robustness of an audio watermarking is defined as its immunity to distortions. These can be caused by signal processing manipulations such as compression, filtering or by the channel through which the signal is transmitted. While robustness of varying degrees might be desirable in certain applications, in others it might be completely undesirable, such as in *fragile audio watermarking*. In fragile watermarking algorithms any change in the signal should result in the failure of the extraction algorithm.

Blind vs. informed detection

Depending on whether the decoder has access to the original host signal and makes use of it or not, the detection is termed either *informed detection* or *blind detection*.

Security

The adversary must not be able to detect the presence of a watermark, remove or modify the watermark data. This can be done by giving the watermark a statistically random distribution and by making it invisible to various signal analysis techniques. *Steganalysis* is the analysis of a cover material to detect the presence of hidden information meant for covert communication [4, p. 49].

Computational complexity and cost

Depending on the application, the computational complexity can vary a lot. Though the data rates are low compared to general digital communication systems, the audio watermarking system operates in an adverse environment at very low bandwidths. Therefore, complex channel coding becomes a necessity. Sometimes, real-time constraints apply such as in the case of broadcast monitoring applications. At other times the computational complexity and latency are not such pressing issues. Nevertheless, power and device complexity become limiting constraints in the case of portable embedded devices.

2.5 Trade-offs

As seen in the previous section, a watermarking system exhibits a number of qualitative and quantitative properties. Those among them, which characterize the immediate goals of a watermarking system include *robustness*, *perceptual transparency* and *bit rate*. These also form the *magic triangle*, which is representative of the fact that a gain in one of these properties also results in a loss in the other. The key is to find and reach an optimal trade-off suitable for our application.

The flexibility to choose a suitable trade-off however often comes at the cost of increased complexity [5, p. 51].

2.6 Applications of audio watermarking

As explained before, a given watermarking system is designed to operate at a particular region of the *bit rate-transparency-robustness triangle*. It gains in one of these at the cost of the other two and therefore watermarking systems are in general designed to be application specific.

2.6.1 Application areas

In this section, the various possible applications of watermarking and their requirements will be described (see [5, p. 20 - 22]).

Ownership protection

A watermark containing information that can identify the owner is embedded into the audio signal. This allows the owner to prove that the work belongs to him in case of a dispute of ownership. He can do that by showing that the work contains the watermark known only to him. The watermark needs to be robust against intentional attacks as well as common signal processing operations. The watermark detection should also have a very low probability of false alarm. At the same time, the capacity does not need to be very high since only an identifying signature needs to be embedded.

In some cases, the adversary could potentially tamper or replace the original watermark with one of his own and claim it to be his. In such a case, the detector may be made unavailable to the adversary. Even more effective would be to show that the tampered work was probably derived from the original.

Authentication and tampering detection

In the content authentication application the goal is to determine, if the original material has been tampered with and if possible, to identify the parts that have been tampered. There is no motivation for the attacker to remove the watermark, and there is no need for the watermark to be hidden from him. However, it should not be possible for the attacker to forge a valid signature or watermark into a work that is not original. In general it is desirable to have a high capacity and it should be possible to detect the watermark without the need of the original signal (blind detection).

Fingerprinting

Fingerprinting in audio watermarking is used to identify the sender or the recipient of a multimedia file. The watermark in this application needs to be robust against intentional and unintentional attacks. It also requires good anti-collusion properties, i.e. it should not be possible to embed more than one ID number to the host file, since the detector would not be able to distinguish between the IDs. The bit rates required for these applications are not very high, with only few bits required per second of media.

Broadcast monitoring

There are a number of possible applications of audio watermarking in broadcast monitoring. The watermarking system is compatible with a wide range of broadcast and reception equipment and does not depend so much on the transmission/reception technology. More details can be found in Chapter 3.1.1.

Copy control and access control

Copy control or access control applications include for instance the case of a DVD player, which may refuse to play/copy material not owned by the user. Since the identifying license is embedded into the media, attempts to remove it would result in substantial distortion in the media and hence discourage illegal copying.

Information carrier

The watermark could be used to carry an identity tag or metadata independent of the file format, so that it is preserved during file format conversions or when it is used along with some other material. It can also be used to track duplicated or reused material on the Internet. There is generally no compelling reason to make it robust against intentional attacks, but it must be robust against unintentional attacks like file compression.

2.7 Modeling watermarking systems

An audio watermarking system is in general modeled around the *communication channel model*, where the audio itself is considered an interference. Since the audio is known at least at the transmitter side, the system can be modeled as a communication system

in a noisy channel with known side information. In the case of non-blind decoding, where the audio is known at the receiver, the interference due to the audio is completely known as well [5, p. 23,65].

3. DESCRIPTION OF THE WATERMARKING SYSTEM UNDER STUDY

This chapter starts with a brief introduction of the broadcast monitoring system, followed by the requirements for such a system. We then provide a high level view of the watermarking system under study and a description of the important subsystems of the watermarking system under study.

3.1 The broadcast monitoring system under development

3.1.1 Broadcast monitoring system

Advertisers, content providers and media networks are interested in the viewership of their material. The data collected on viewership is of interest to estimate the effectiveness of advertisements, popularity of the program and in general affects the amount of money or time slots alloted to promote the broadcast material. This data is in general collected from a sample population, who volunteer to have their viewing habits monitored. At present, one of the possible ways to capture this information is for the user to manually set his/her profile on a device that tracks and transmits his/her program viewing habits. This is quite a cumbersome solution and there is need for a portable/wearable device that can automatically monitor and keep track of the material viewed by the sample population.

3.1.2 Description of the watermarking system and design goals

The audio watermarking system being developed is targeted to be used in broadcast monitoring devices. In this scheme, a small portable device at the viewer's location (viewer pertains to the chosen sample population) picks up the watermark sent along with the TV or Radio broadcast and thus monitors the viewership.

The watermark is designed to be robust against distortions/losses in the channel. The system is built to be robust against reverberations of the room, additive noise from the microphone and channel, distortions in the sounds reproduced by the speakers and to some extent Doppler effects. The decoder or monitoring scheme has been designed to run in real-time on a low-power portable device. Also considering that there are few users in the sample population, the number of false positives are not of much concern. Nevertheless, it is not designed to be robust against intentional attacks or unintentional changes like low bit rate compression. This is because the broadcast service is expected to have good control over the transmitted watermark, but not over the channel. Adapting the techniques to other applications would thus require considering other constraints on the system, while relaxing some.

3.2 Overview of the system

The basic system is shown in Figure 3.1.

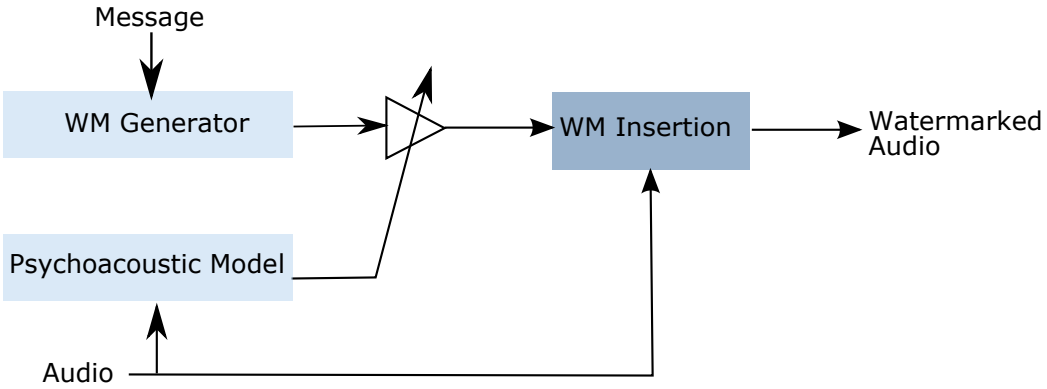


Fig. 3.1: Basic high level functional diagram

At high level, the watermarking system under study producing the watermarked signal consists of a watermark generator, a psychoacoustic model and the actual watermark insertion process. The purpose of the watermark generator is to transform the input message into a more robust form. This process modifies the original message through a number of steps which might involve channel coding, interleaving, time and frequency spreading and synchronization.

In the psychoacoustic analysis, the original audio is analyzed in order to estimate the amount of energy that can be added to the original audio in a given time-frequency slot, so as to make the watermark signal imperceptible to the human ear. This is done by taking into account the temporal and frequency masking effects of the original signal.

In the final process of actual watermark embedding, the threshold calculated by the psychoacoustic model is used to modulate the coded bits from the watermark generator. These are then embedded into the original audio signal additively to obtain the final watermarked audio signal (AWM).

The block diagram of the encoder and decoder are shown in Figures 3.2 and 3.3.

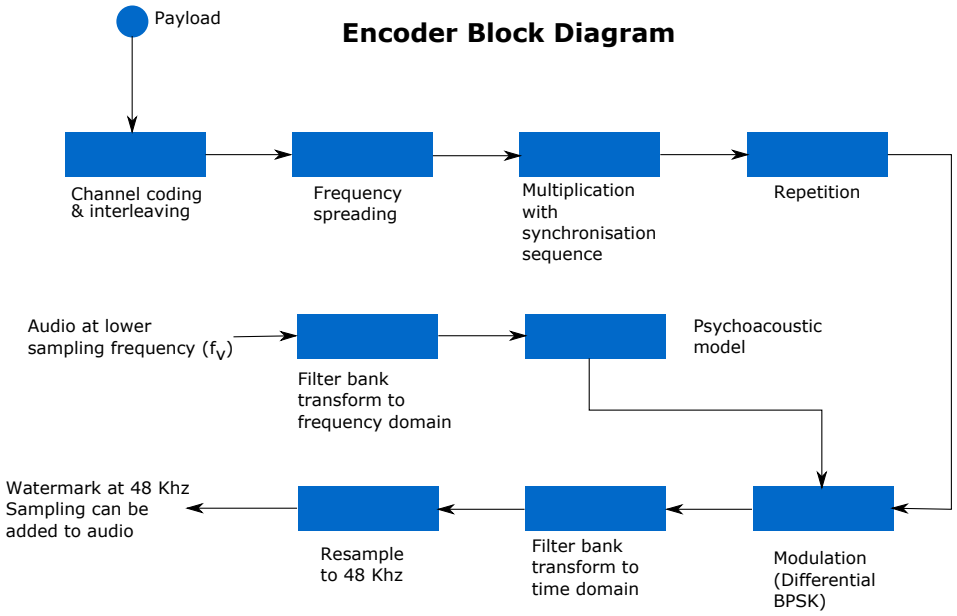


Fig. 3.2: Encoder

The decoder works in the reverse order, first moving the watermarked signal to the frequency domain and then demodulating after normalizing the subbands with it neighbors. It then sums up the spreaded data, following this the decoder tries to find the synchronization point and when this happens, it triggers the Viterbi decoding block and retrieves the message from the coded soft bits.

3.3 Subsystems and design decisions

3.3.1 Convolutional encoder and Viterbi decoder

The basic idea of channel coding is to allow the information bits to be spread as much as possible within the transmitted bits. This results in maximizing the likelihood of retrieving the correct information at the receiver end. We can do this by introducing redundancy within the transmitted code and as a general principle, by mapping a short sequence of information bits onto a longer sequence of transmitted bits.

By such a mapping procedure, a smaller dimension is transformed into a larger dimension, allowing for a certain error correction capability. The design of this type

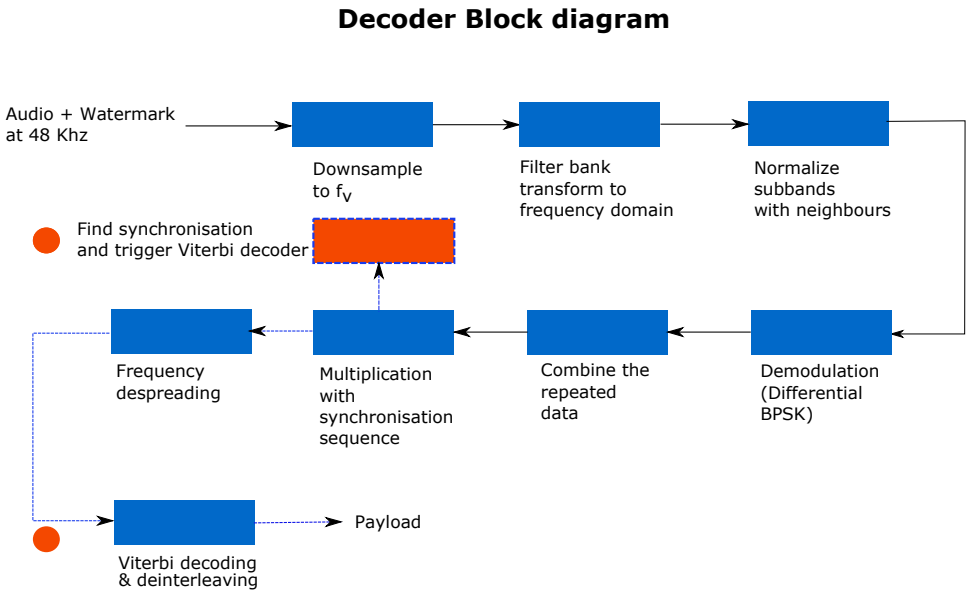


Fig. 3.3: Decoder

of system is generally carried out by using simulations with an appropriate channel model. Robust systems usually require longer sequences and more complex decoding (often iterative) [7].

In terms of hardware implementation, convolutional codes are an efficient implementation of such a mapping procedure. The *code rate* is defined as the ratio between the number of information bits transmitted and the corresponding coded bits. It can also be interpreted as the change in data rate introduced by encoding it. The code rate is an important parameter, which tells us how much we lose in terms of data rate in order to achieve a particular robustness of the transmitted information. The encoded signal can be decoded in a number of ways, one such method that is considered efficient is the Viterbi decoder.

The output of the convolutional encoder is transmitted over the channel and fed into the Viterbi decoder. The decoder attempts to recover the most probable sequence of information bits that were input into the convolutional encoder. It is based on the Hidden Markov Model (HMM) and it is done using the soft decoding method. In this method, the soft information from the bits is preserved during the decoding, as opposed to the hard decoding method, where all bits are treated equally. This results in a much better performance because we preserve the knowledge about how likely a bit is either a one or a zero. More precisely, the bits that have a high probability of being either a one or a zero have greater influence over the output than the ones that are ambiguous.

The Viterbi algorithm builds a state transition matrix which keeps track of state changes, the most probable paths and the probabilities associated with them. A much

more detailed description of the Viterbi can be found in Chapter 5.4.

3.3.2 Time and frequency spreading

The encoded bits are spread both in time and frequency before which they are also interleaved in order to avoid burst errors. This spreading in frequency and time increases the robustness of the message. The spreading used here is not the most efficient way of utilizing the available resources. However, considering the minimal computational complexity introduced and the ability to easily implement real-time decoding makes this an optimal scheme for the current application. This becomes increasingly relevant, where the despreading process has to be applied at numerous points of the incoming stream until synchronization has been reached.

3.3.3 Synchronization

In the current system, synchronization is not carried out by introducing known pilot/sync symbols and instead, an orthogonal sequence of codes is multiplied with the bits to be transmitted. On the receiver side the same orthogonal sequence is multiplied with the incoming stream to recover the bits. Since the codes are orthogonal, any misalignment would result in very poor recovery and a low energy signature. However, a good alignment would result in an almost complete recovery of the bits and a high energy signature. This is illustrated in Figures 3.4 and 3.5, the first diagram shows how the synchronization symbols are added to the datastream. And the second illustrates how proper alignment/synchronization results in strong recovery of the original datastream.

To obtain a set of orthogonal bit sequences we use the *Walsh-Hadamard sequences*. A matrix of size eight used to generate the orthogonal bit stream in our system is shown here.

$$\mathcal{H} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

Since the sequences are orthogonal to each other, they have the property that

$$\mathcal{H} \cdot (\mathcal{H}^T) = n * \mathcal{I},$$

where $\mathcal{H}$ is of the size (n x n) This allows us to multiply the datastream with the $\mathcal{H}$ at the encoder and the at the receiver side with its transpose to retrieve the datastream. In case we have not chosen the right synchronization point because of the orthogonality we would not be able to retrieve zero energy from the data bits.

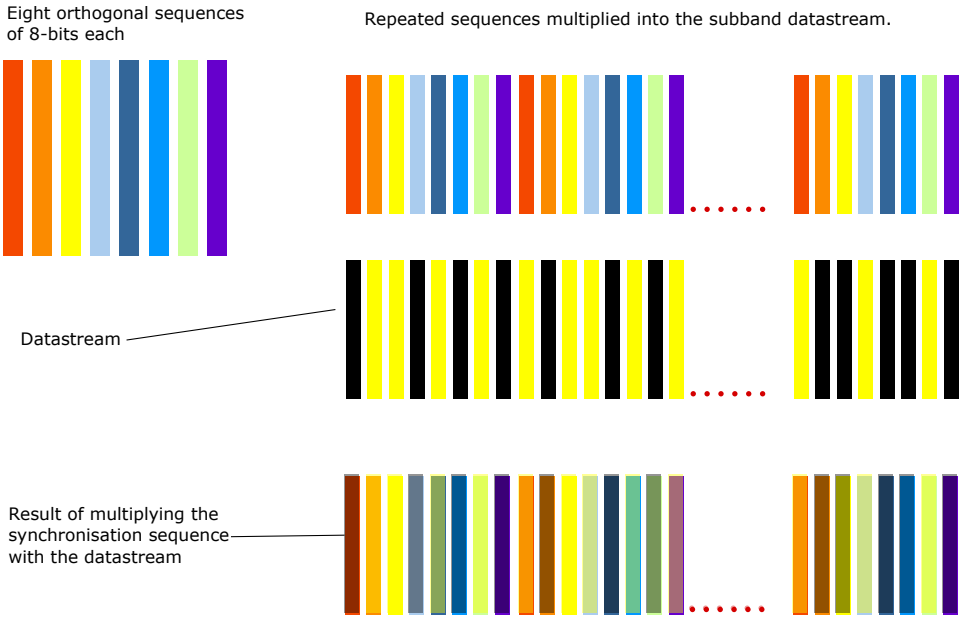


Fig. 3.4: Synchronisation

3.3.4 DBPSK - Differential Binary Phase Shift Keying

The channel introduces a phase rotation of the AWM; this phase rotation depends on the channel response, which varies with frequency and time (can however be assumed to vary slowly). This makes it difficult to retrieve the bits by embedding only along fixed directions. The bits can instead be embedded differentially i.e. the bit information can be contained in the phase difference between two signals rather than in their absolute value. Since the change in response with time is assumed to be slow, the phase difference is preserved.

3.3.5 Filter bank

The purpose of the filter bank is to convert the signal in the time domain to time-frequency blocks and back again to the time domain. A filter bank decomposes signal

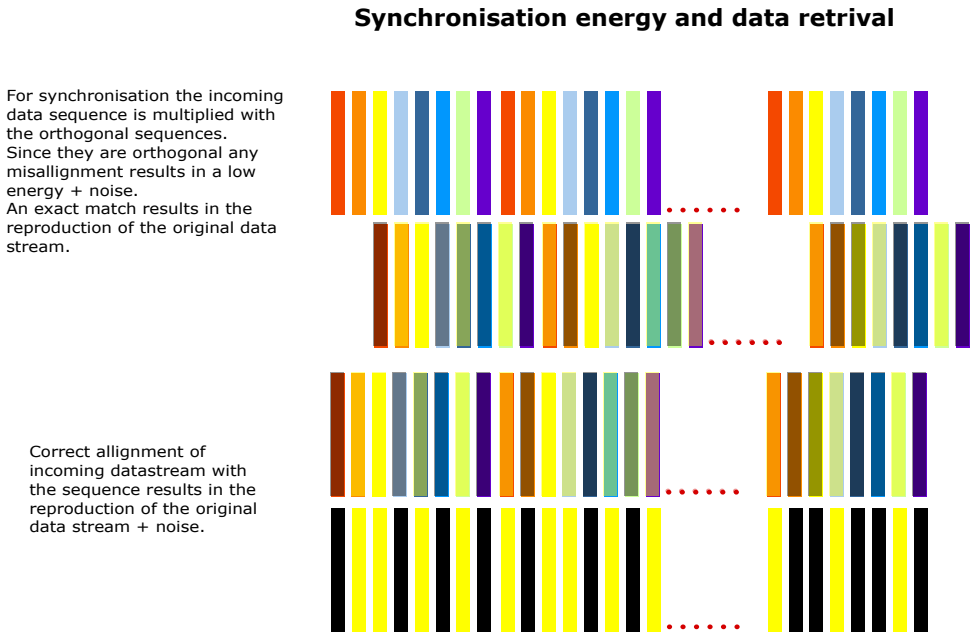


Fig. 3.5: Synchronisation energy retrieval

into multiple bands, each band is a time sequence at specific frequency band. This functionality finds use in many places. For example in psychoacoustic analysis, windows of different sizes are used to analyze the masking effect. The information bits are embedded in the frequency domain and thus require conversion of the time domain signal to the frequency domain and vice versa.

3.3.6 Psychoacoustics

The psychoacoustic analysis model approximates the energy that can be added within each critical band and which can effectively be masked by the temporal and frequency masking effects of the original audio signal. However, these two effects are not additive, the more dominant one determines the masking effect.

The model used in the current system is a simplified one and has been found to be sufficiently effective. At a high level, the system works as follows: the psychoacoustic module estimates the energy within each critical band in a given time window (two kinds of time windows are used - long and short - in order to obtain good temporal and frequency resolution). Then the estimated energy, is used to calculate the approximate energy of a signal within that critical band that can be masked. This is currently done using a fixed set of multiplying factors for each band. The masking effects due to both temporal and frequency masking are combined to get the final thresholds of the watermark. These can afterwards be added to the original audio without introducing large perceptual distortion.

3.4 Challenges, limitations and relaxations

The design of the said watermarking system consists of a number of assumptions about the environment in which the system is to be put to use. This enforces some limitations on the design of the system. The system is itself expected to work in a harsh environment, therefore faces a lot of challenges. However, there are certain conditions that are assumed to be under control and therefore do not influence the design.

Due to the environment to which the watermarked signal is to be exposed, there are some distortions introduced in the signal. The design of the system aims to overcome these. The channel response of the room makes it necessary for the symbol lengths to be long enough so as to avoid destructive interference from the echoes (room reverberation). The phase rotation introduced by the channel to various frequency bands requires *Differential BPSK (DBPSK)* instead of BPSK.

The implementation and deployment in the real world imposes some limitations on the system such as: power, speed , memory and computational capability. The device, on which the decoder should operate, works at a sampling frequency of $\approx f_v$ KHz, thus limiting the highest usable frequency subband to $\approx \frac{f_v}{2}$ KHz due to the Nyquist criterion. In addition, the operations of the device should be performed in real-time and ideally, the complexity should be spread over time, so that peak computational requirements are minimized. Since the decoding device is designed to be small and energy efficient, there are also limitations on the buffer size and other memory use.

There are however few relaxations on the environmental factors compared to other applications of watermarking. There are for example no active signal processing attacks and most of the signal conversion processes such as compression and codec conversion are the anticipated ones.

4. PROBLEM STATEMENT AND DIRECTION OF WORK

The first part of this chapter talks about the initial motivations of the thesis and continues to discuss about the areas that were considered for further investigation. The following sections give a brief overview of the results of studying the subsystems and their performances under different conditions and modifications.

4.1 Thesis topic

There are many applications of audio watermarking, but the watermarking system under study was designed specifically for broadcast monitoring. As seen in the section about trade-offs (Chapter 2.5), a watermarking system built for a specific application can perform poorly for other applications. The initial aim of the thesis was to find ways to use the existing system for applications outside broadcast monitoring without making considerable modifications to the existing system.

4.2 Initial goals

The current system has certain inherent properties and limitations that fit well to broadcast monitoring. Particularly its low complexity and the ability to work in real-time, allow it to be used in embedded portable broadcast monitoring devices. The psychoacoustic model gives a very good approximation of the masking threshold without introducing excessive complexity.

Tracking files on multimedia sharing sites such as *youtube* seemed to be an attractive application of audio watermarking. Depending on the application, the requirements and limitations could be contrasting with the current system. Therefore, the current watermarking system may have to undergo considerable modification before it can be used in this new application. We would however like to preserve as much as possible the inherent properties of the current system.

4.3 File tracking

Multimedia files uploaded to media sharing sites are often duplicated and the material could contain commercial or copyrighted information or advertising material. Embedding watermarks to identify and track the material could help prevent duplication, track advertisements, commercial material and promote selling media. There are several challenges though: the watermark could undergo compression, Codec conversion, both intentional and unintentional modifications like time stretching, time compression, cropping, noise addition, echo addition, filtering etc. However, there are also a few advantages over the broadcast monitoring environment: the potential absence of noise and interference from a reverberant medium. Due to the lack of real-time or low complexity requirements, sacrificing these non-critical functionalities could help build a system suitable for the desired application.

4.3.1 Compression and codec conversion

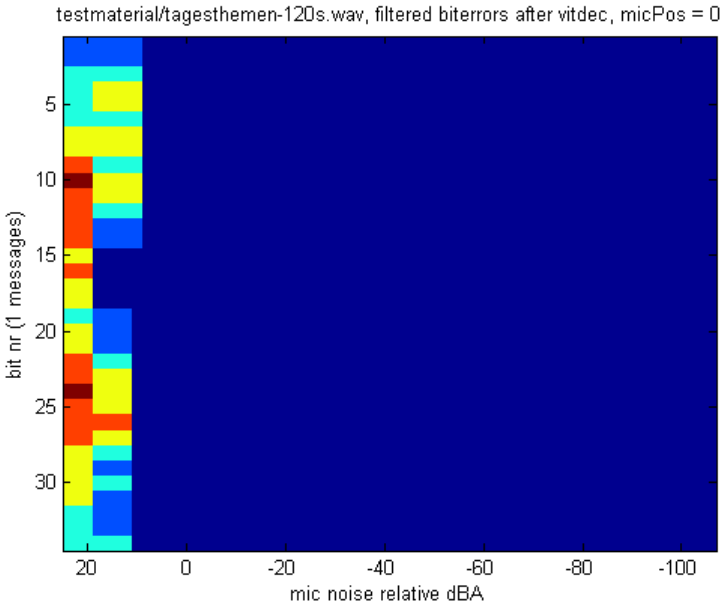


Fig. 4.1: Original system - where 'mic noise relative dBA' stands for the A-weighted sound pressure level ratio of the added noise level to the original audio, the y axis represents the bits of the message recovered. Blue in the background represents no errors and the other colours represent an error in the bit, a deep red shows that there are errors in the neighbouring bits too.

The current system was found to work well with MP3 compression provided the data rates are not very low. Layer 2 and MP3 compression at 64 Kbps and 128 Kbps were tried out. The message bits could still be recovered after compression and codec

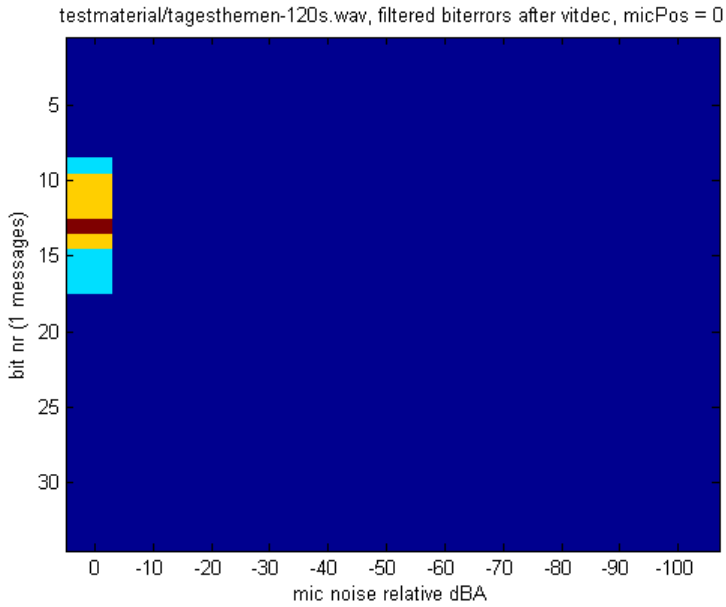


Fig. 4.2: Simulation results: Errors in the message retrived after compression - MP3 at 64 Kbps

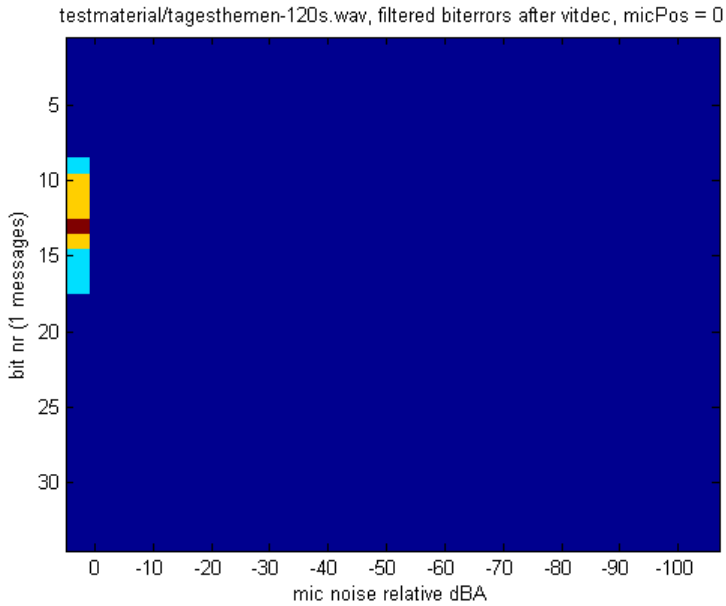


Fig. 4.3: Simulation results: Errors in the message retrived after compression - MP3 at 128 Kbps

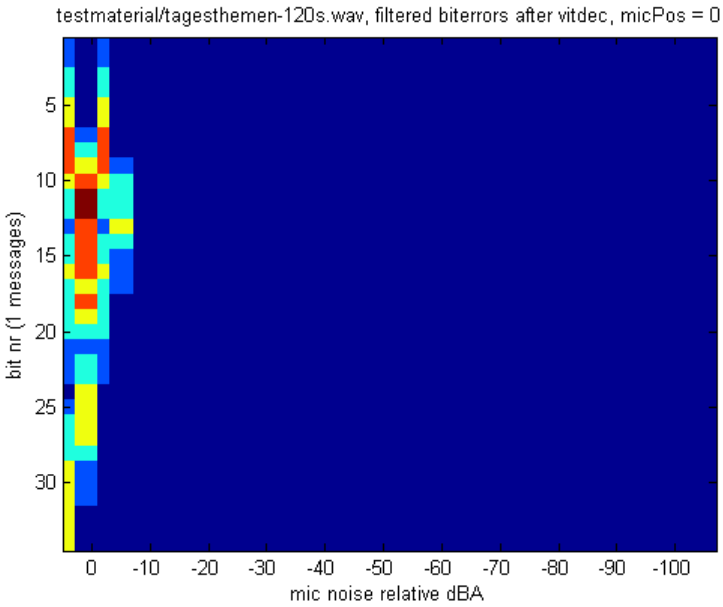


Fig. 4.4: Simulation results: Errors in the message retrived after compression - Layer 2 at 64 Kbps

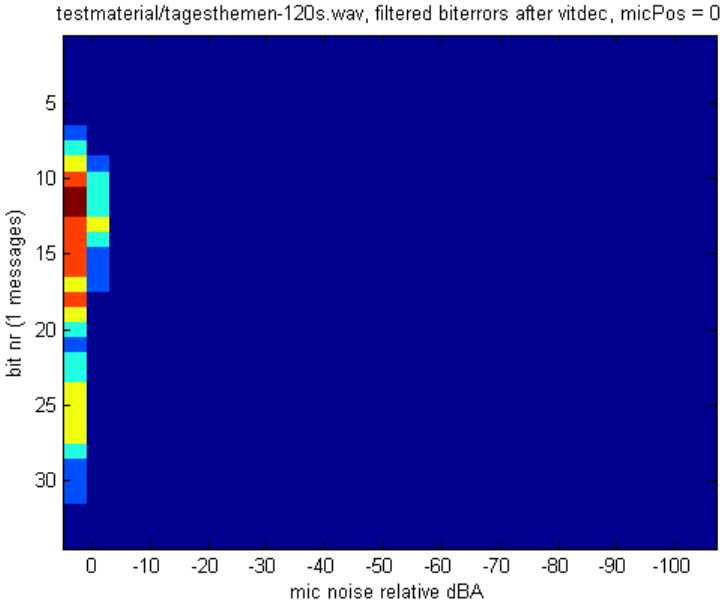


Fig. 4.5: Simulation results: Errors in the message retrived after compression - Layer 2 at 128 Kbps

conversion and the performance in terms of its robustness to additive noise was found acceptable compared to the original system.

Figures 4.2, 4.3, 4.4 and 4.5 show the performance in terms of the bit errors after viterbi of one message. The bit errors in each message are plotted against various levels of additive noise added to the watermarked signal. The colors on the plots are used to distinguish the burst errors from random errors, deep red indicates the presence of burst errors and the lighter shades, the random errors.

We see that by using various codecs and rates the performance degrades significantly compared to the original system which works at 10 dBA (see Figure 4.1). Nevertheless, the system still works with an acceptable performance under such modifications.

4.3.2 Time stretching and time compression

Small unintended time stretching or compression result in small leakages and shift of the energy from the subbands. The system is nevertheless robust to small shifts and oversampling done during synchronization could improve the chances of correct alignment.

4.3.3 Strength of the error correction coding

Since the real-time and complexity constraints can be relaxed, increasing the complexity and/or decreasing the rate could result in a watermark of greater robustness. Vice versa we could increase the bit rates of the watermarking system at the cost of complexity or robustness.

4.4 Studying the subsystems

Without the ability to obtain clear specifications required for a certain application, designing a useful and implementable system is challenging. We decided instead to concentrate on identifying the shortcomings and improvements that could be made to the different subsystems.

4.4.1 Frequency hopping

By introducing *frequency hopping* in our system, the subbands are allowed to switch to pseudorandom center frequencies within their own critical bands. In this way, the

used subbands are not fixed and thus could be hidden from any steganalysis that might be carried out. This adds a level of security against intentional distortion of the watermark.

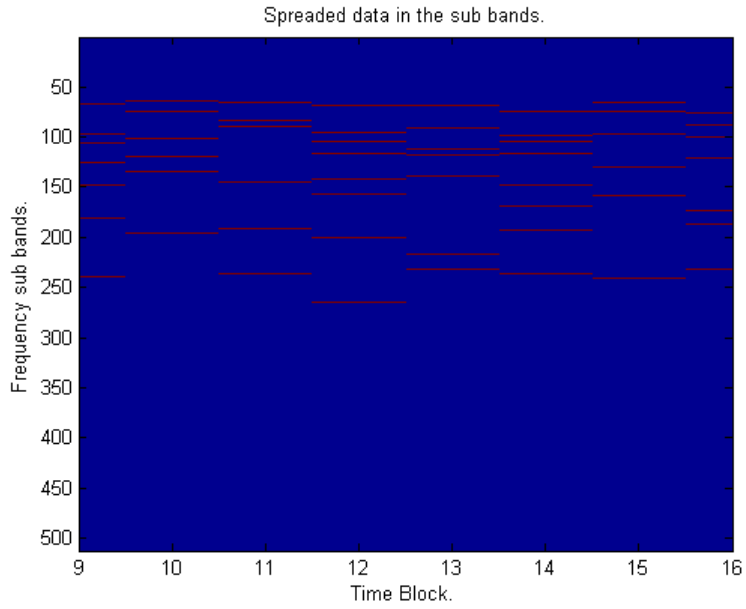


Fig. 4.6: Frequency hoping - Data embedded into different time frequency slots

4.4.2 Partitioning the energy within critical bands

Initially, the system used nine fixed subbands, this was modified to use 12 subbands within the same frequency range. The new implementation was tested for artifacts and was found to have good psychoacoustic properties and performance similar to that of the original. When the number of subbands was increased to 18 and the watermarked audio was tested for artifacts, it seemed not to have introduced any perceptible distortion, though this is debatable due to the lack of extensive listening tests.

4.4.3 Avoid repeating data across the subbands

By removing the frequency spreading block, and utilizing the various subbands to carry different bits we can improve the data rates at the cost of robustness. However, by increasing the strength of the convolutional code, it is possible to increase the data rates without affecting the robustness significantly. Figure 4.7 shows the plot of such a system. We were able to increase the bit rates by seven times by removing the spread in frequency and decreasing the rate of the convolutional encoder to one-ninth from

the current one-seventh. As seen in the figure, this can be reached within reasonable SNR values.

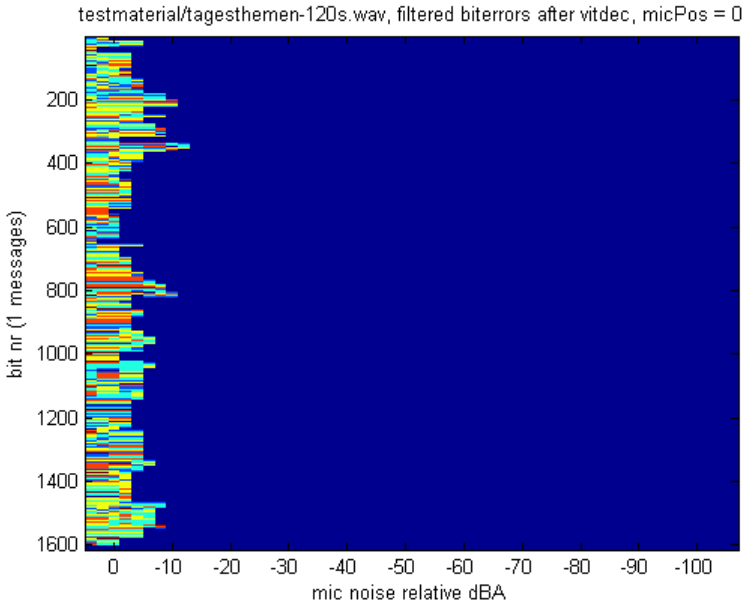


Fig. 4.7: Simulation results: Errors in the message retrived with Frequency spreading removed and using a rate $\frac{1}{9}$ instead of $\frac{1}{7}$, removal of frequency spreading increases the datarate whereas lowering the rate results in increased robustness.

4.4.4 Phase of the audio signal

The threshold information obtained from the psychoacoustic module says nothing about the change in phase of the original audio that would be perceptible to the HAS. Using the limits for modifying the phase (see Chapter 2.2.1), the phase of the original audio was used to embed the information. Using the limits of the phase change for the original audio together with the threshold value that is calculated by the psychoacoustic module it is possible to achieve better robustness. However, this makes it a non linear operation since the theshold calculation after the phase modulation can be different from the original threshold. This becomes particularly challenging with the introduction of differential modulation. As a result, it requires extensive listening tests before it can be accepted as is or with modified psychoacoustic parameters. However, this makes a valuable dirrection for future work.

5. WRITING ON DIRTY PAPER

5.1 Writing on Dirty Paper

In the communication model used to represent the current audio watermarking scheme, the original audio is considered as interference. But if the interference is known at the transmitter, at least in principle it is possible to transmit over the channel as if the interferer were not present [3]. This is analogous to writing on dirty paper. Though difficult, it is not impossible to write on dirty paper as if it were a clean sheet.

5.2 Costa's - Writing on Dirty Paper

Costa showed in his paper [3] that given a channel modeled as

$$Y = X + S + Z,$$

where X is the transmitted signal of power P, Y the received signal, S the interference $\mathcal{N}(0, Q)$ and Z the Gaussian noise $\mathcal{N}(0, M)$, where $\mathcal{N}$, is the normal distribution then the channel capacity (C) is given by

$$C = \frac{1}{2} log\left(1 + \frac{P}{(M+Q)}\right).$$

But if S is known at the transmitter and Z is known neither at the transmitter, nor the receiver, then the capacity of the channel is given by

$$C = \frac{1}{2} log\left(1 + \frac{P}{M}\right).$$

The above relationship is equivalent to the case where the interference S is absent.

Though Costa showed that it is theoretically possible, he did not specify ways to achieve this. Nevertheless, over the years many have proposed methods to make use of the side information.

Considering a communication channel model, where part of the interference is

known, we could choose to ignore the knowledge about this interference. Otherwise, we could try to actively suppress this interference by expending the energy that should have gone into the modulation. Neither of these methods is very efficient, instead we could code/modulate the data stream such that it follows the interference as closely as possible.

The basic idea in these methods, particularly the ones proposed for (image) watermarking, is to associate more than one possible codeword for a given message to be embedded. Among these possible codewords, the one which is closest to the original signal is chosen. In this way, using the side information the message can be embedded with minimum distortion of the original signal. But in our case, we already have a good estimate of the amount of energy that can be added within a critical band without making the watermarked signal perceptually distorted. We can use the side information instead to strengthen our embedded message and make the decoding more robust.

5.3 Practical methods and trade-offs

Based on Costa's paper [3], the gain should theoretically come at no extra cost. In practice though, the proposed methods are in general implemented with at least a marginal trade-off elsewhere. These methods achieve distortion minimization of the original signal at the cost of a more compressed symbol constellation, lowered bit rate and often increased complexity.

The key would be to find an ideal operating point for an application, such that the distortion to the original signal is minimized, without introducing excessive complexity or losing too much elsewhere. Some implementations/methods offer greater flexibility in choosing the operating point, but they also introduce greater complexity.

Among the many different ways in which this can be solved, we will concentrate on two of them: Quantization Index Modulation (QIM) and Dirty Paper Trellis Codes (DPTC).

5.4 Dirty-Paper Trellis codes

One way of reducing the interference due to the host signal is by choosing a sequence of bits to represent the message such that it matches as close as possible the host signal. This can be done with trellis shaping, by selecting a path from the trellis that represents the message and at the same time is as close to the audio as possible.

5.4.1 Current system

The current system uses a convolutional encoder with a code rate of $\frac{1}{7}$ and a soft bit Viterbi decoder. More details about the convolutional coder and decoder can be found here [6]. For each bit input (message bit) into the convolutional encoder, $\frac{1}{rate}$ output bits are generated, representing the code bits. The output bits depend not only on the current input bits, but also on the current state of the convolutional encoder. The current state is by itself determined by the past inputs. In this way, the information of a particular input bit is spread across the output bits.

The convolutional encoder starts at a particular fixed state, generally an all zero state and generates output bits based on the current state and the input bit. The input bit along with the current state of the convolutional encoder determines the next state. The Viterbi decoder on the receiver side uses the bits generated by the convolutional encoder to determine the most probable original message sequence. This is achieved by maintaining the state transition information for each input message bit and its associated probabilities. These probabilities depend on the closeness of the received (soft) bits to the output bits associated with this transition and also on the past transitions leading to the current state. It would at present suffice to remember that in a conventional convolutional encoder, given an input message bit and a current state, there is only one possible next state and a corresponding unique set of output bits.

5.4.2 Rethinking the Viterbi decoder and convolutional coding

The design of the conventional convolutional encoder is particularly efficient when implemented in hardware. It requires only few memory buffers to hold the state and few XOR gates for the output bits. However, this should not restrict what coding can be used. The Viterbi decoder itself requires only a look-up table to associate a transition from one state to another with a corresponding message bit and a set of output bits.

This allows us to associate more than one possible transition from the current state with a given message bit. Given this, the flexibility to choose the next state and the corresponding output bits allows us to make the output bits match the underlying audio as closely as possible, reducing the interference from the audio. This comes however at the cost of the trellis code strength since there is an extra uncertainty introduced in the state transition.

Figure 5.1 shows a conventional trellis, where at any given current state there are

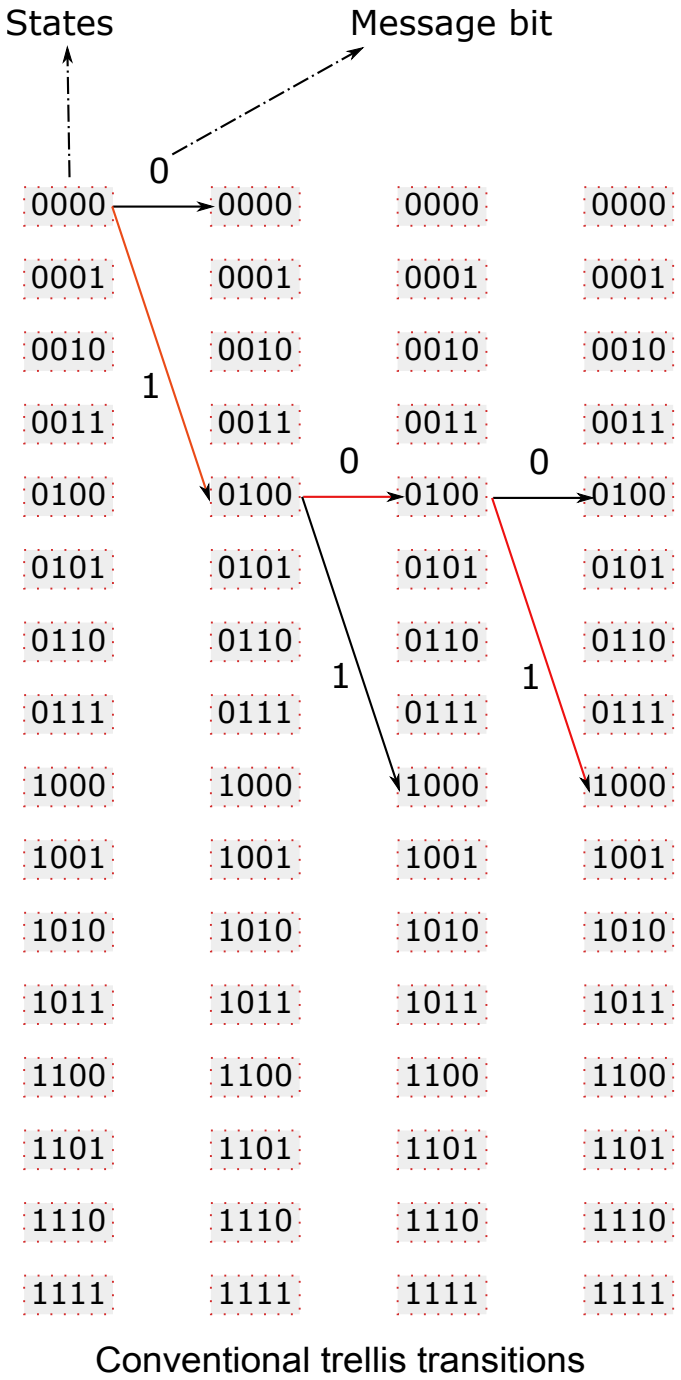


Fig. 5.1: Convolutional encoder - There are only two paths leaving from one state to the next state, one representing input message 'one' and the other representing 'zero'.

exactly two possible next states. The trellis also contains the corresponding output bit sequences and the encoded bits. The modified trellis (see Figure 5.2) can however have a number of possible transitions (here four, two each for message bit 0 and bit 1). To send a sequence of message bits (0s and 1s in red in the figures), there is only *one unique path* in the conventional trellis. In the modified trellis however due to the fact that at each step there are two possible transitions for the same message bit, there are *a number of paths* associated with a given message sequence.

Instead of just two output bit sequences as in the conventional trellis (see Figure 5.1), there are now a number of them (see Figure 5.2). The constellation is thus crowded, increasing the probability of error. In the absence of a strong known interference, this would equal the error correcting capabilities of a weaker or higher rate convolutional code. However, if the audio signal (interference) is significant, we can gain substantially in using a modified trellis.

5.4.3 Current techniques in image watermarking trying to exploit DPTC

In the current watermarking techniques where attempts have been made to use the side information, DPTC though computationally complex and harder to implement compared to the QIM is seen as an interference reduction method holding great potential. This is not only because of the increased potential for fine tuning but also because the QIM suffers from inherent problems like being sensitive to scaling and re-quantization.

Both QIM and DPTC have largely been studied as a means of minimizing the distortion of the original material, usually considering the case of image watermarking. That is as ways to reduce the distance between the original host signal and the final watermarked signal. The variables that affect the performance of DPTC are listed in the paper introducing Trellis Coded Modulation (TCM) [8] as a suitable technique for DPTC. A brief description of the various parameters are given below.

Number of states in the trellis The number of states in the trellis affects the robustness of the code. The larger the number of states, the stronger the code is expected to be. However, this comes at the cost of increased complexity.

Number of arcs entering/leaving each state When a large number of arcs (the transition path from one state to the next) can be used to represent a message bit at any state, we have a diverse choice of bit sequences that can be embedded in the original signal. This ensures that we are likely to get a sequence that follows

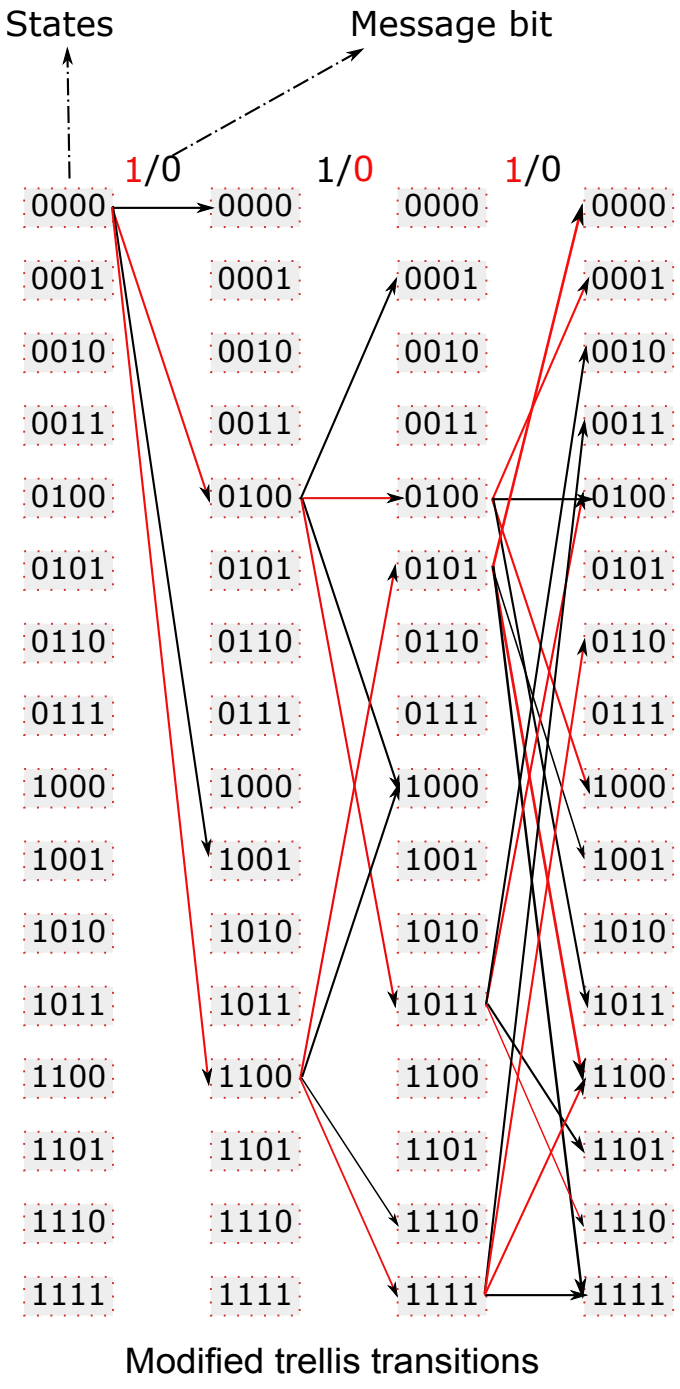


Fig. 5.2: Modified Trellis: There are a number of possible paths to encode the same message sequence.

the original signal very closely. This could decrease the distortion or increase the robustness depending on how we use it. However, it comes at the cost of closeness of the sequences to each other in the constellation, which results in a reduced error correcting capability of the trellis.

Reference pattern used to represent the arcs The reference patterns are the output bits of the encoder that are embedded into the original signal. Using long sequences of reference bits allows for increased robustness due to larger minimum distance between the sequences. This would however also result in lower data rates.

Connectivity between the states Due to the possibility of a number of arcs to enter/leave a state in the modified trellis, there is a need to find an optimal connectivity between the states.

Mapping between the reference patterns and the arcs The larger the distance between the reference patterns representing arcs leaving the same state, the greater the likelihood that we can find the correct next state. However, this would force other arcs of other states to have reference patterns closer to the given arcs or just closer to each other, increasing the probability of errors in those states. As a result, we need to find an optimal distribution of the reference patterns on the arcs.

Mapping between the message bits and the arcs This would affect the way the reference patterns are mapped to represent ones and zeros of the message bits, as well as the choice of next states.

5.4.4 Adapting the modified trellis to the watermarking system under study

The conventional convolutional encoder - Viterbi decoder pair can be replaced by a modified trellis encoder - Viterbi decoder pair. There are many accompanying challenges in adapting a DPTC to audio watermarking and a few that are specific to the watermarking system under study.

Computational complexity

The current system uses a convolutional code requiring a trellis of 64 states. Given a message of size 40, in the current system the number of possible paths considering a

fully connected trellis are 32^{40} or 2^{200}. This is calculated as follows: to transmit the first message bit starting from a all zero state, we have 32 possible paths to the next state. The other 32 path are for transmitting the inverse of the bit. From this new state we again have a choice of 32 different paths and so on. If for each message bit to be transmitted we have 32 different paths to choose from, then for n message bits we have 32^n paths. This problem is very complex and cannot be solved in a reasonable amount of time. We would instead have to use a *heuristic search* (see Chapter 5.4.5).

Co-decoding

The current system uses nine subbands, so in order to build a comparable system would require nine modified Viterbi decoders working in parallel and exchanging their decoded message information. An optimal way to use this exchanged information is not very straightforward.

5.4.5 Possible solutions

Ant colony optimization (ACO)

Ant colony optimization is a probabilistic technique used to solve computational problems that can be reformulated to a problem of finding a good path through a graph. Our problem of finding a good path through the trellis is similar. This falls within the class of heuristic search methods. We know the metric for different state transitions at any given time and we need to find the path that follows the audio the closest.

In the following, we will explain how the ACO algorithm can be applied to our system. Initially, we could maintain a matrix that represents the trellis and let $\mathcal{U}$ number of *ants* leave the first node of the trellis (the all-zero state). We then allow them to take random paths available to them from any node of the trellis depending on the message bit. The closeness of the output bits to the audio can be regarded as the metric in this case.

Once the ants reach a particular node, they can compute the metric of their path so far. Then they compare this metric with the best metric registered at this node by any other ant that might have reached it earlier. If the new metric is better, then the best metric of the node gets updated to the new value. By doing so, the ants walk the trellis randomly identifying the suboptimal paths to a number of nodes.

In the second iteration we send a set of $\mathcal{V}$ ants, but starting from some random nodes that have already been visited and using the best path up to that node found by

ants preceding it. Likewise, several iterations can be carried out to reach a suboptimal solution to the whole problem. Alternatively, it is also possible to choose only the best node at any message point for spawning new ants.

No spreading in the subbands

If we were to spread the data in the subbands as in the current system, but use DPTC along each subband, then the current synchronization system can no longer be used. However, better use of the subbands could be achieved by with a Viterbi decoder of lower rates. We can do this either by increasing the constraint length and the number of output bits or by just increasing the number of output bits. In the first case the complexity increases, but the distance between the output bits remains the same. In the second case the distance between the output bits increases, but the diversity provided by the output bits decreases.

Suboptimal solutions

The ACO does not guarantee an optimal solution and neither do the trellis configurations. The resulting watermark is therefore expected to be suboptimal. In [9] the authors suggest studying the effects of the different parameters such as trellis structure and reference pattern on the performance in terms of robustness (BER), fidelity (distortion to the original signal) and complexity. It was found that they had a significant impact and particularly in image watermarking it was found that a configuration of 64 states and 64 arcs leaving and entering each node gave the best trade-off among the performance metrics. It was also discovered that the choice of a reference with a distribution similar to the original work gave the best performance in terms of BER.

5.5 Quantization Index Modulation (QIM)

In the previous section (DPTC), we saw a way of reducing the interference from the audio working at the channel coding level, by modifying the trellis used for coding. One other way to reduce the interference is by introducing more points in the constellation space such that the distance between the host signal and a feasible point is reduced.

In the current scheme we try to embed the watermark either in phase or out of phase (180 °) with the previous AWM. Since the audio itself is random, it acts as an interference. However, instead of embedding the data at 0 ° or 180 ° phase difference, we can split it into smaller segments. A binary zero could be embedded using a 0 °

or a 180° phase difference, while a binary one using a 90° or 270° phase difference. By applying it this way, we achieve greater freedom to choose where to embed the watermark.

If for example the original audio were to be in the fourth quadrant (with respect to the previous AWM) and we would like to embed a binary one, the threshold should have been strong enough to allow the AWM to move from the fourth quadrant to the third. We could instead choose to embed the watermark along the 270° phase difference using the modified constellation. However, this comes at a cost: the symbols in the constellation are now closer and thus more susceptible to noise. It performs poorly compared to the original system at low SNR. Nevertheless, at higher SNR it has a much better raw BER and also a better lower error floor. This can be useful in case the application that it is targeted for requires high bit rates and operates at high SNR.

5.5.1 QIM in image watermarking

Unlike audio, in images noise is not easily perceived. Therefore it is quite effective to introduce a watermark as a noise that causes very little distortion. In QIM, the components of the image such as color and intensity are split into regions and each region is associated with a code. If the same code is associated with multiple regions, then to encode a message would mean to change the properties of the image at that location in order to move it to one of the possible regions representing that code. In order to minimize the distortion we would preferably move it to the region nearest to the original region.

However, there are some drawbacks to this method. Modifying the color space or intensity slightly such that the difference is not perceptible to the Human Visual System (HVS) would result in distorting or completely destroying the message. Scaling the image would have a similar effect. Nonetheless, both problems could be solved with a more exhaustive search.

5.5.2 QIM in audio watermarking

The proposed methods for applying QIM in audio watermarking often involve quantizing the phase of the audio signal and shifting the phase of the original audio to the nearest quantized code. But there are some drawbacks to this, since the HAS is quite sensitive to changes in phase, though not as sensitive to additive noise.

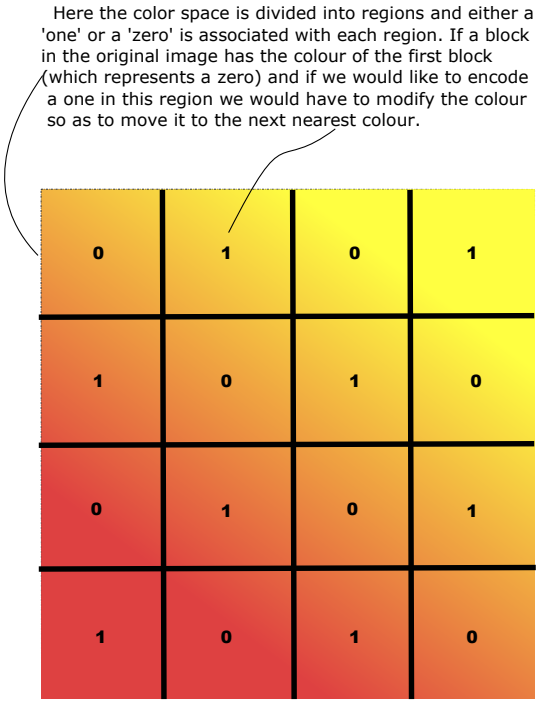


Fig. 5.3: QIM in image watermarking - Quantizing the color space

In order to keep it within the perceptual limits, we would have to use a QIM of high granularity. This would mean that the regions in the quantization space are very small and the constellation space is compressed. This would in turn make it highly susceptible to noise and would require coding of great strength to compensate.

5.5.3 Applying QIM to the current watermarking system

The psychoacoustic model in the current system quantifies the amount of energy that could be added to each critical band at any time slot, without introducing perceptible distortions in the original audio. Nonetheless, it does not give an estimate of the phase change in the original audio that would be imperceptible to the HAS.

In our attempts to introduce phase change to the original audio in encoding data (see [5, pg. 41]), we found that modifying the phase of the original audio introduces perceptible artifacts. We decided against modifying the original audio to any extent, since it requires subjective listening tests to quantify the transparency of the system. We opted to base possible improvements to the system only on the threshold information from the psychoacoustic model.

5.5.4 Current modulation technique

The current system uses Differential Binary Phase Shift Keying (DBPSK), where the message is encoded in the phase difference between two vectors in the time-frequency blocks obtained from the original watermarked audio signal. The *decoder* estimates the projection of one vector over the other. If they are in phase, then it outputs a binary one, otherwise a binary zero.

Mathematically, if $\mathcal{A}$ and $\mathcal{B}$ are the two vectors, then the encoded soft bit (s) is represented as

$$s = \Re\{\mathcal{A} * conj(\mathcal{B})\}.$$

So the encoder adds the watermark in phase (0°) with the previous watermarked audio vector or out of phase (180°) depending on if the encoder is to encode a one or a zero respectively.

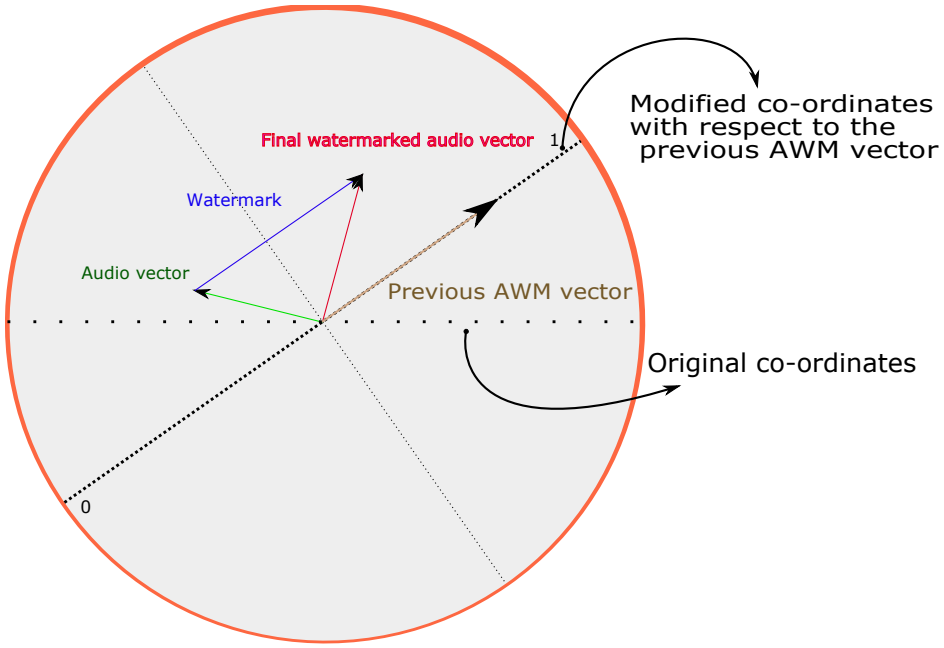


Fig. 5.4: Modulation using DBPSK

5.5.5 Deriving the optimal encoding for DBPSK

Let S, W and A be the complex coefficients for the signal (audio+watermark), watermark, and audio respectively, after the analysis filterbank for a specific time and frequency slot. Due to the linearity of the filterbank, we have $S = W + A$.

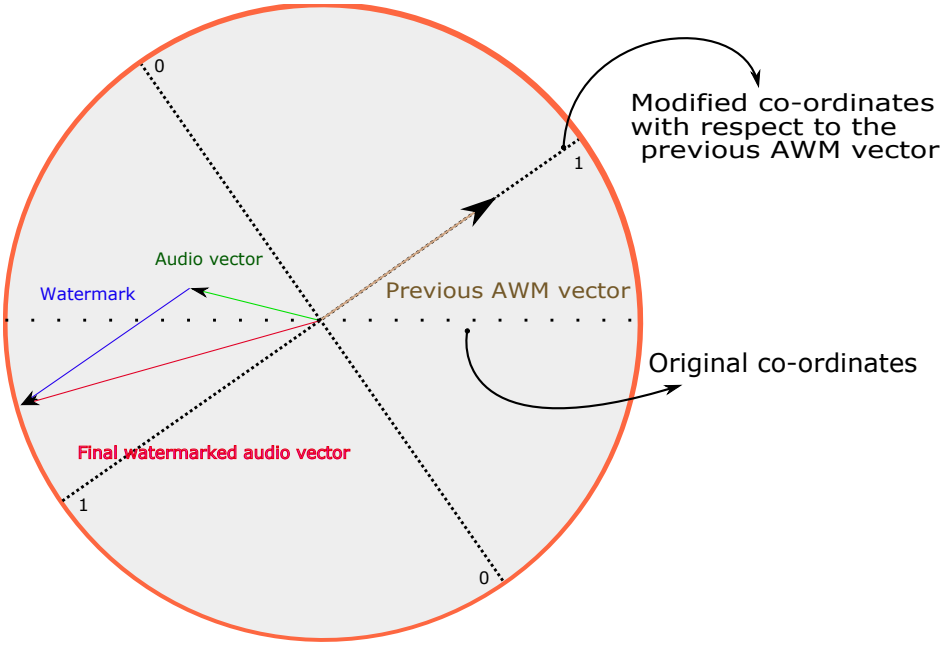


Fig. 5.5: Modulation using QIM

The differential decoding for the n-th soft bit $b(n)$ computes the following

$$b(n) = \Re\left\{S(n) \cdot S^*(n-1)\right\}, \tag{5.1}$$

where * denotes conjugation. Assuming that

$$S(n) = \rho_S(n)e^{j\theta_S(n)}, \tag{5.2}$$

where $\rho_S(n)$ is the magnitude and $\theta_S(n)$ is the phase information of the vector $S(n)$, we have:

$$b(n) = \rho_S(n)\rho_S(n-1)\cos(\theta_S(n) - \theta_S(n-1)). \tag{5.3}$$

Ideally, the cosine is either 1 or -1 and the product of the magnitudes gives the amplitude of the soft bit.

The audio + watermark can be written as:

$$S(n) = \rho_S(n)e^{j\theta_S(n)} = \rho_A(n)e^{j\theta_A(n)} + \rho_W(n)e^{j\theta_W(n)}, \tag{5.4}$$

where $\rho_A(n), \rho_W(n), \theta_A(n)$ and $\theta_W(n)$ are the maginitude and phase of the original audio and the watermark respectively.

In the system before introducing DBPSK, the audio carrier (i.e. A) is implicitly seen as noise during the embedding, meaning that the modulation scheme assumes

that it is unknown to both transmit and receive side. Since the original audio itself is considered irrelevant to the embedding process, then $\theta_W(n)$ is chosen simply to be either 0 or π depending on the bit to be embedded. That is, we embed the information by adding or subtracting from the real component (can also be done in the imaginary part); one drawback of this is that we do not have a constant phase reference since the watermarked material can undergo phase rotations.

In the new method, we design $\theta_W(n)$ assuming that A and thus S are known. Substituting (5.2) in (5.1), we obtain:

$$b(n) = \Re\left\{\rho_S(n-1)e^{-j\theta_S(n-1)} \cdot \rho_S(n)e^{j\theta_S(n)}\right\}. \tag{5.5}$$

We define the optimal watermark W^{opt} as the one which maximizes the SNR. This corresponds to maximizing $b(n)$ when the bit sent is $+1$ and maximizing $-b(n)$ when it is -1. From (5.5) and (5.4) we get that:

$$\begin{aligned}
W^{\text{opt}}(n) &= \arg \max_{\rho_W(n),\theta_W(n)} \pm\Re\left\{\rho_S(n-1)e^{-j\theta_S(n-1)} \cdot \left(\rho_A(n)e^{j\theta_A(n)} + \rho_W(n)e^{j\theta_W(n)}\right)\right\} = \\
&= \arg \max_{\rho_W(n),\theta_W(n)} \pm\rho_S(n-1)\rho_A(n)\cos(\theta_A(n) - \theta_S(n-1)) \\
&\qquad + \rho_S(n-1)\rho_W(n)\cos(\theta_W(n) - \theta_S(n-1)) = \\
&= \arg \max_{\rho_W(n),\theta_W(n)} \pm\rho_S(n-1)\rho_W(n)\cos(\theta_W(n) - \theta_S(n-1)).
\end{aligned}$$

The maximization of this expression leads to

$$W^{\text{opt}}(n) \Rightarrow \begin{cases} \rho_W^{\text{opt}}(n) = \gamma \\ \theta_W^{\text{opt}}(n) = \theta_S(n-1) + \kappa\pi, \end{cases} \tag{5.6}$$

where γ is the threshold computed by the psychoacoustic model and the term κ is either 0 or 1 depending on the information bit (i.e. $+1$ or -1).

The derivation for the optimal watermark embedding procedure using QIM will be explained in the next section.

5.5.6 The QIM method

In the phase rotation method, the bits are embedded in the direction of the previous AWM vector or in the opposite direction, depending on whether the bit is a 1 or a 0. However, in the case of the QIM-4, the bits are embedded along the previous AWM or in the direction perpendicular to it. We now have two ways of embedding the

watermark for any given bit and intuitively we choose the direction that is closest to the current audio.

In the case of QIM, the differential decoding for the n-th soft bit $b(n)$ computes the following: Let S, W, and A be defined as in Chapter 5.5.5.

$$\widehat{b}(n) = \|\Re\{S(n) \cdot S^*(n-1)\}\| - \|\Im\{S(n) \cdot S^*(n-1)\}\|,$$

where * denotes conjugation. We know that

$$S(n) = \rho_S(n)e^{j\theta_S(n)} = \rho_S(n)(\cos\theta_S(n) + i\sin\theta_S(n)). \tag{5.7}$$

Since the resulting AWM can be written as a sum of the audio and the watermark, then:

$$S(n) = \rho_S(n)e^{j\theta_S(n)} = \rho_A(n)e^{j\theta_A(n)} + \rho_W(n)e^{j\theta_W(n)}. \tag{5.8}$$

We now need to maximize $\widehat{b}(n)$ for 1 and minimize it for -1.

Let $D = \{S(n) \cdot S^*(n-1)\}$, using this and equation (5.7) and (5.8):

$$\begin{aligned}
D &= \rho_S(n-1)e^{-j\theta_S(n-1)} \cdot \{\rho_A(n)e^{j\theta_A(n)}\rho_W(n)e^{j\theta_W(n)}\} = \\
&= \rho_A(n)(\cos\theta_A(n) + j\sin\theta_A(n))\rho_S(n-1)(\cos\theta_S(n-1) - j\sin\theta_S(n-1)) \\
&\quad +\rho_W(n)(\cos\theta_W(n) + j\sin\theta_W(n))\rho_S(n-1)(\cos\theta_S(n-1) - j\sin\theta_S(n-1)) = \\
&= \rho_A(n)\rho_S(n-1)(\cos\theta_A(n)\cos\theta_S(n-1) + \sin\theta_A(n)\sin\theta_S(n-1) \\
&\quad +j\sin\theta_A(n)\cos\theta_S(n-1) - j\cos\theta_A(n)\sin\theta_S(n-1) \\
&\quad +\rho_W(n)\rho_S(n-1)(\cos\theta_W(n)\cos\theta_S(n-1) + \sin\theta_W(n)\sin\theta_S(n-1) \\
&\quad +j\sin\theta_W(n)\cos\theta_S(n-1) - j\cos\theta_W(n)\sin\theta_S(n-1)) = \\
&= \rho_A(n)\rho_S(n-1)(\cos(\theta_A(n) - \theta_S(s-1)) + j\sin(\theta_A(n) - \theta_S(n-1)))) \\
&\quad +\rho_W(n)\rho_S(n-1)(\cos(\theta_W(n) - \theta_S(n-1)) + j\sin(\theta_W(n) - \theta_S(n-1))) = \\
&= \rho_S(n-1)(\rho_A(n)\cos(\theta_A(n) - \theta_S(n-1)) + \rho_W(n)\cos(\theta_W(n) - \theta_S(n-1))) \\
&\quad +j\rho_S(n-1)(\rho_A(n)\sin(\theta_A(n) - \theta_S(n-1)) + \rho_W(n)\sin(\theta_W(n) - \theta_S(n-1))).
\end{aligned}$$

For a 1 to be transmitted, $\widehat{b}(n) = \|\Re\{D\}\| - \|\Im\{D\}\|$ needs to be maximized. At the same time, for a 0 to be transmitted, $\widehat{b}(n)$ needs to be minimized.

$$\begin{aligned}\widehat{b}(n) \;=\; & \arg \max_{\rho_W(n),\theta_W(n)} \\ & \{\|\rho_S(n-1)\{\rho_A(n)(\cos(\theta_A(n)-\theta_S(n-1))) + \rho_W(n)(\cos(\theta_W(n)-\theta_S(n-1)))\}\| \\ & - \|\rho_S(n)\{\rho_A(n)(\sin(\theta_A(n)-\theta_S(n-1))) + \rho_W(n)(\sin(\theta_W(n)-\theta_S(n-1)))\}\|\}\end{aligned} \tag{5.9}$$

Since all the angles are with respect to $\theta_S(n-1)$, we can write

$$\begin{cases}(\theta_W(n)-\theta_S(n-1)) = \phi_W(n) \\ (\theta_A(n)-\theta_S(n-1)) = \phi_A(n).\end{cases} \tag{5.10}$$

Then the equation (5.7) can be rewritten as

$$\widehat{b}(n) = \arg \max_{\rho_W(n),\phi_W(n)} \quad \{\|\rho_S(n-1)\{\rho_A(n)(\cos(\phi_A(n))) + \rho_W(n)(\cos(\phi_W(n)))\}\| \tag{5.11}$$

$$- \|\rho_S(n-1)\{\rho_A(n)(\sin(\phi_A(n))) + \rho_W(n)(\sin(\phi_W(n)))\}\|\}. \tag{5.12}$$

We can now use the previous equation to find the optimal values of $\rho_W(n)$ and $\theta_W(n)$.
The results of the numerical simulation for optimal values is shown in Figure 5.6. It can be clearly seen that the maxima are at 0 ° and 180 ° and the minima at 90 ° and 270 ° .

To transmit a 1 the optimal angle is chosen such that

$$\theta_W(n) - \theta_S(n-1) \;= \begin{cases}0, & \theta_A(n)-\theta_S(n-1) = [-\frac{\pi}{2}\ \frac{\pi}{2}] \\ \pi, & \theta_A(n)-\theta_S(n-1) = (\frac{\pi}{2}\ \frac{3\pi}{2})\ .\end{cases} \tag{5.13}$$

To transmit a 0 the optimal angle is chosen such that

$$\theta_W(n) - \theta_S(n-1) \;= \begin{cases}+\frac{\pi}{2}, & \theta_A(n)-\theta_S(n-1) = [0\ \pi] \\ -\frac{\pi}{2}, & \theta_A(n)-\theta_S(n-1) = (-\pi\ 0)\ .\end{cases} \tag{5.14}$$

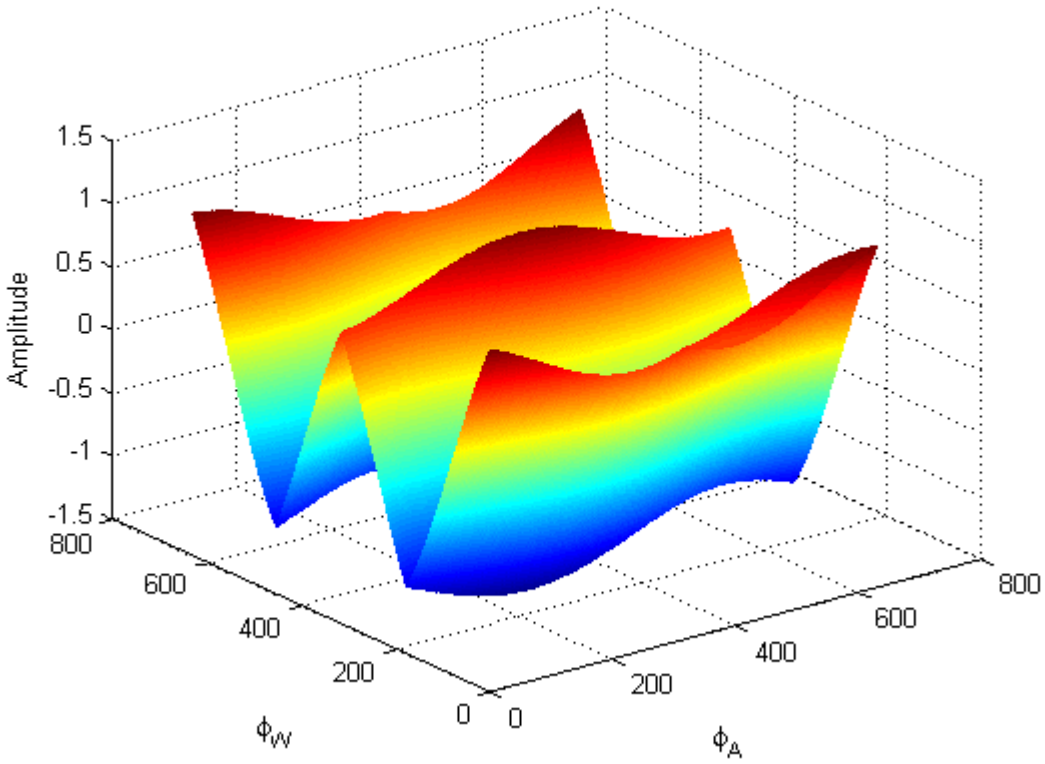


Fig. 5.6: Numerical simulation for optimal angle to increase the robustness of the watermark. This is a plot of $\widehat{b}(n)$ with $\rho_A(n)$ and $\rho_W(n)$ fixed and ϕ_A and ϕ_W taking values from 0 ° to 360 ° .

5.6 QIM results

Figure 5.7 shows the results from simulating QIM using 12 subbands. It can be seen that at sufficiently high SNR, the QIM performs better than DBPSK.

For applications such as *fragile watermarking* (see Chapter 2.4), tampering with the original media should result in the destruction of the message, whereas in the absence of any modification or external distortion the message could be retrieved reliably. Therefore, QIM can be very useful in such cases. In addition, the method works well in applications operating at high SNRs, but requiring high data rates. In such applications we are forced to increase the number of channels used, save on the frequency spreading and increase the strength of the channel coding. By increasing the number of channels within a critical band, we are spreading the energy of the usable threshold into a larger number of time-frequency bins. The usable threshold in each bin is now smaller and thus the interference from the original audio is greater. In this case, by using QIM we reduce the interference from the audio and increase the gain.

Figure 5.8 shows the BER plots for DBPSK and QIM-4. In this case the nine

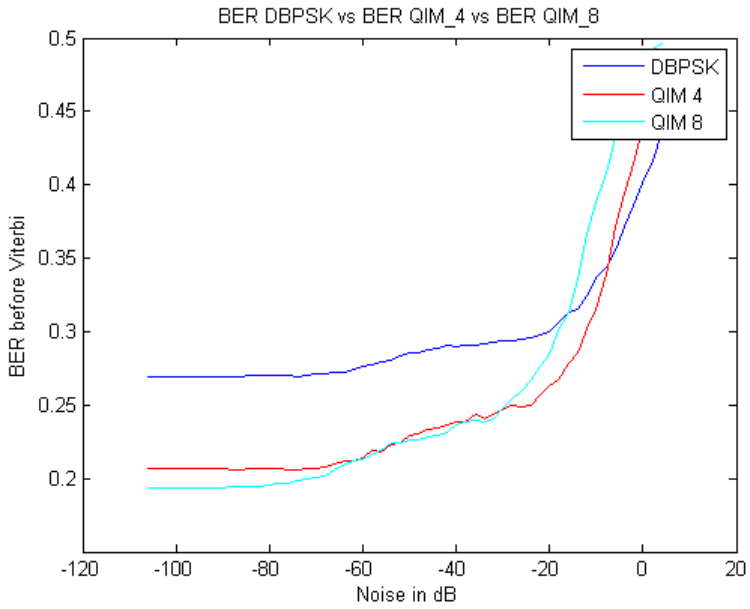


Fig. 5.7: QIM simulation of 12 subbands

subbands used have been expanded to 18 equally spaced subbands, but they are still restricted to the same frequency range. Frequency spreading has been removed, increasing the rate 18 times. However, the rate of the Viterbi decoder has been reduced from $\frac{1}{7}$ to $\frac{1}{9}$ to increase the strength of the convolutional coder. We have thus increased the overall data rate $\frac{18*7}{9} = 14$ times.

The two bit error plots cross over at about 20 dB. This is also reflected in Figures 5.9 and 5.10.

5.6.1 QIM with codec

Since QIM follows the audio as closely as possible, on compression/codec conversion it is expected to be less affected than DBPSK. However, the process also introduces noise, to which QIM is more susceptible than DBPSK.

In the simulation, the watermarked audio is compressed with MP3 at 128 Kbps. Figures 5.11 and 5.12 show the errors in the message at different noise levels and we observe that the QIM performs better, with no error at noise levels lower that -20 dBA, whereas in the case of DBPSK the errors remain down at about -30 dBA.

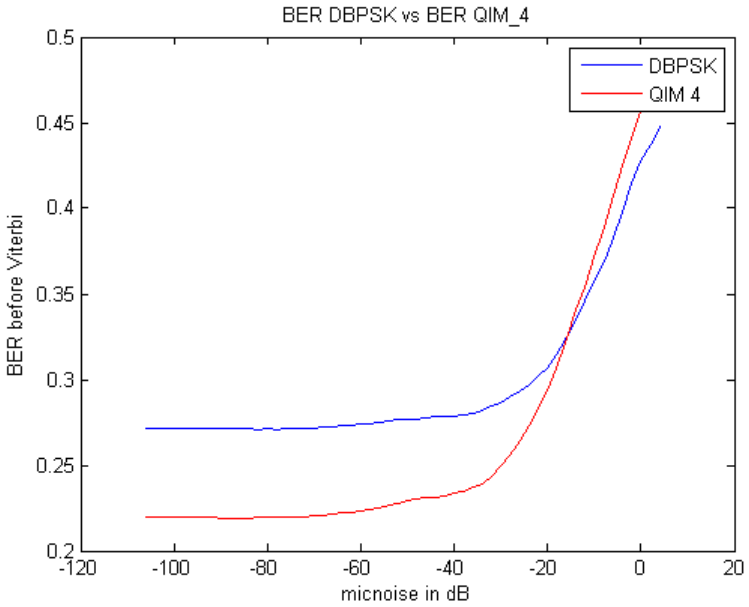


Fig. 5.8: Simulation using 18 subbands, a code rate of 1/9 and no frequency spreading.

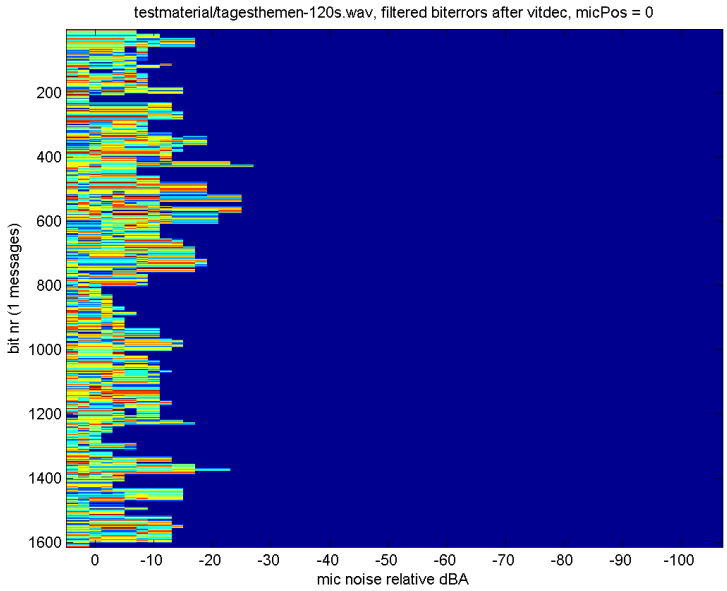


Fig. 5.9: Simulation results: Errors in the decoded message at different noise levels, embedding data in 18 subbands using DBPSK - error remains until -27 dBA

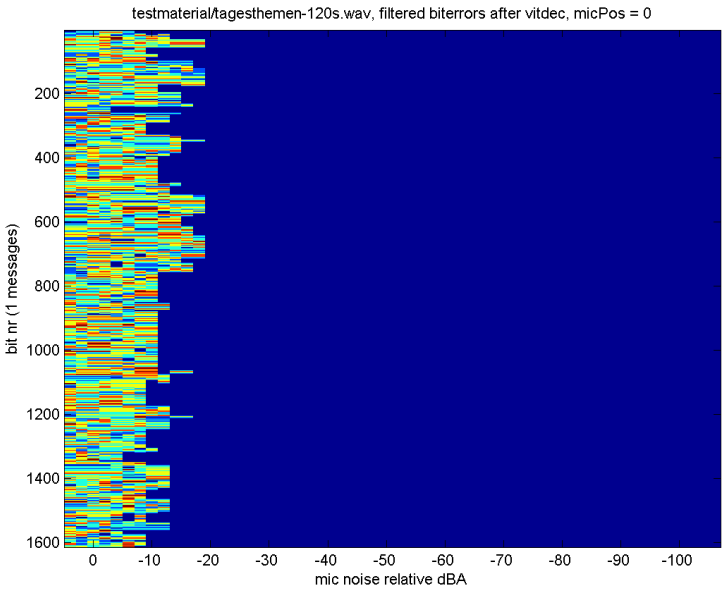


Fig. 5.10: Simulation results: Errors in the decoded message at different noise levels, embedding data in 18 subbands using QIM 4 - error remains until -19 dBA, which is better than the original

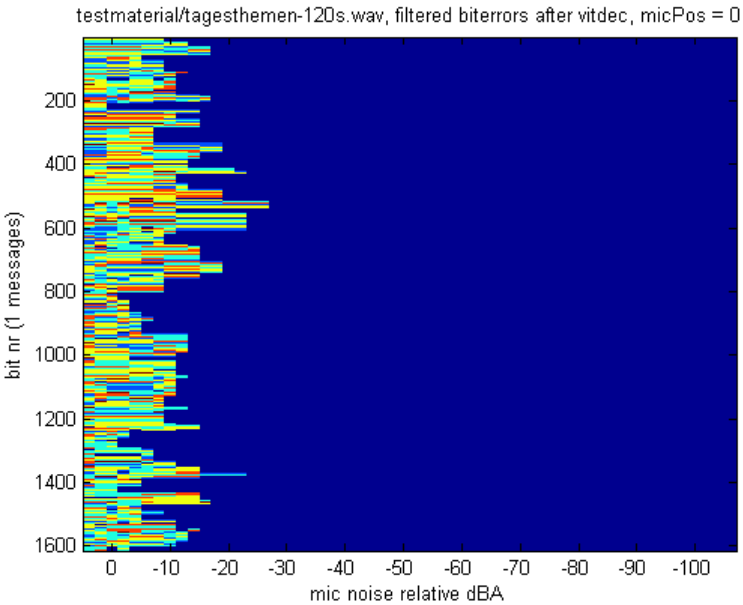


Fig. 5.11: Simulation results: Errors in the decoded message at different noise levels, embedding data in 18 subbands using DBPSK and after compression (MP3 at 128 Kbps)- error remains until -28 dBA

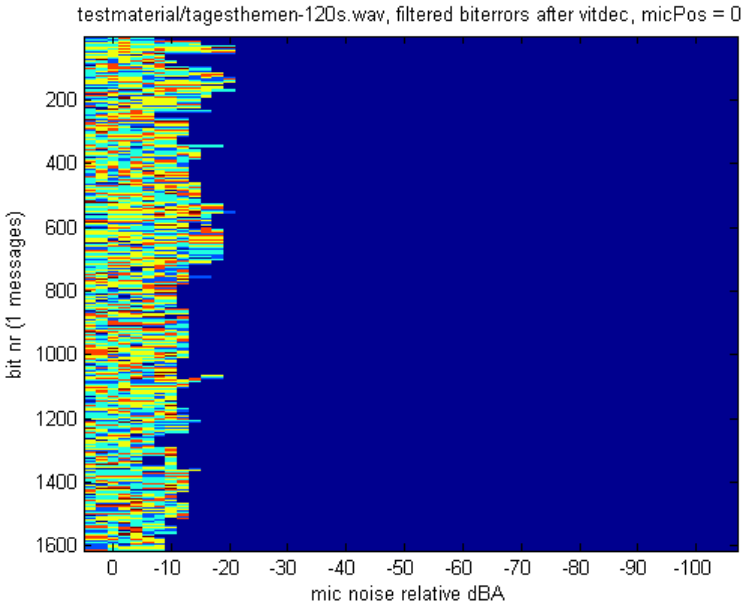


Fig. 5.12: Simulation results: Errors in the decoded message at different noise levels, embedding data in 18 subbands using QIM-4 and after compression (MP3 at 128 Kbps)- error remains until -21 dBA, which is also an improvement over the original system

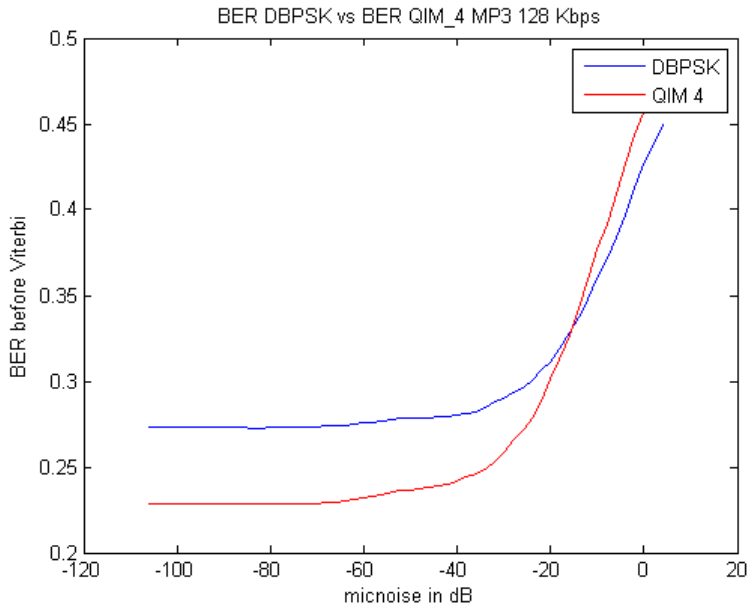


Fig. 5.13: BER of DBPSK vs QIM-4 with MP3 at 128 Kbps

6. CONCLUSIONS

A brief survey of the different digital audio watermarking techniques, based on the exploitation of the various psychoacoustic properties was given. Followed by a comprehensive overview of the watermarking system under study, which was being developed for the broadcast monitoring application. The thesis contributes to the current IIS project of modifying the current watermarking system to employ it to applications other than the broadcast monitoring.

In our attempt to identify applications for which the current system could be adapted with the least changes, we identified the minimum characteristic requirements of potential applications and carried out a performance study. Among these requirements were robustness against codec compression, high data rates, and robustness against steganalysis. In these studies, the various relaxations on the requirements such as lowering the constraints on the computational complexity and lowering the need for real-time performance were also considered. The system was able to recover the watermark after codec conversion to MP3 and Layer 2 at rates of 64 Kbps and 128 Kbps, but this could be achieved only at higher SNRs regions than the ones for which the original system was designed. Higher data rates were achieved by limiting the repetition in time and frequency. With a rate seven times higher and using stronger convolutional coding, it was possible to achieve zero error rate at about 15 dB higher signal-to-noise ratio (SNR) than for the original system. The system was also found to work well after introducing a known frequency hopping sequence (but working without the differential modulation). This pseudo random hopping prevents or impedes an adversary from identifying or attacking the subbands carrying data. The rates could be further increased by increasing the number of subbands within the same frequency range used by the original system, the watermarked audio did not exhibit any perceptible distortion.

Particular emphasis was laid on the exploitation of the side information, namely the fact that the information about the host audio signal is known at the transmitter. Dirty Paper Trellis Codes (DPTC) and Quantization Index Modulation (QIM) are two methods that have been studied to make use of the side information. DPTC works at the channel coding level, this is implemented with convolutional coding using a

modified trellis. In this modiffied trellis a number of code sequence can represent the same message sequence, thereby allowing the encoder to choose between a number of output sequences. Using a sequence that is closest to the host audio we can reduce the effect of interference. DPTC is as yet not completely understood and there is no guarenteed method for effective exploitation. We propose a number of algorithms to simplify the implementation of DPTC for our specific system but we carry out no further studies as extensive analysis needs to be carried out before obtaining an effective system.

QIM on the other hand works at the modulation level and is easier to implement, however it is less flexible than DPTC. In this method we introduce a number of points in the constellation, instead of just two as in the current system. This ensures that the host audio is more likely to be closer to one of the feasible points in the constellation. This was implemented in the system as an extention of the current Differential Binary Phase Shift Keying (DBPSK) system and an optimal modulation method was derived. On simulations it was found that the bit-error rate (BER) is good at high SNRs but is worse than the original system at low SNRs. This is in keeping with the intutive idea that by introducing more points in the constellation we reduce the distance between the host signal and a feasible point. However, we also get a compressed constellation which makes it more succeptible to noise. On introducing more subbands (to increase the bit rate), it was found that QIM showed a significant improvement over DBPSK alone. Also under compression and codec conversion, QIM performed better than DBPSK.

GLOSSARY OF ACRONYMS, SYMBOLS AND NOTATION

ACO Ant colony optimization

AWM watermarked audio signal

BER bit-error rate

DBPSK Differential Binary Phase Shift Keying

DPTC Dirty Paper Trellis Codes

fis watermarking system under study

FIIS Fraunhofer Institute for Integrated Circuits

HAS Human Auditory System

HMM Hidden Markov Model

HVS Human Visual System

QIM Quantization Index Modulation

SNR signal-to-noise ratio

TCM Trellis Coded Modulation

BIBLIOGRAPHY

[1] Wikipedia entry - digital watermarking.

[2] Wikipedia entry - psychoacoustics: Masking effects.

[3] M. Costa. Writing on dirty paper (corresp.). *Information Theory, IEEE Transactions on*, 29(3):439 – 441, may 1983.

[4] Ingemar Cox. *Digital Watermarking and Steganography*. Morgan Kaufmann Publishers, San Francisco, 2008.

[5] Nedeljko Cvejic. *Algorithms for audio watermarking and steganography*. PhD thesis, Faculty of Technology, University of Oulu,, University of Oulu, P.O.Box 4500, FIN-90014 University of Oulu, Finland, 2004.

[6] Andrea Goldsmith. *Wireless Communications*. Cambridge University Press, Cambridge, 2005.

[7] David Mackay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, 2004.

[8] Chin Kiong Wang, Gwenaël Doërr, and Ingemar Cox. Trellis coded modulation to improve dirty paper trellis watermarking. SPIE Conf. on Security, Steganography and Watermarking of Multimedia Contents IX 2007, 2007.

[9] Jun Xiao and Ying Wang. Study on the performances of trellis dirty paper watermarking. In *MultiMedia and Information Technology, 2008. MMIT '08. International Conference on*, pages 829 –832, 30-31 2008.

[10] Eberhard Zwicker and Hugo Fastl. Psychoacoustics: Facts and models. 2006.