A HIGH CAPACITY WATERMARKING TECHNIQUE FOR STEREO AUDIO

Xing He, Alexander I. Iliev, Michael S. Scordilis
Department of Electrical and Computer Engineering, University of Miami
Coral Gables, Florida 33124, U.S.A.
x.he@umiami.edu, {ailiev, m.scordilis}@miami.edu

ABSTRACT

A novel audio watermarking technique is introduced, which is based on the processing of the difference phase spectrum of stereo audio. Psychoacoustical principles of binaural hearing are employed to compute a frequency-dependent threshold of the interaural phase difference, which is used to determine locations suitable for watermark insertion. Spread spectrum is used to mark watermark locations and to provide channel synchronization. Listening tests have indicated that the using this method to insert watermarks in stereo audio results in a signal that indistinguishable from the original audio. This technique is computationally efficient and it provides payload of up to 20 kb/s making it suitable for audio information enhancement. Its fragility necessitates a friendly environment, such as in authentication applications.

1. INTRODUCTION

Successful watermarking is a compromise between several performance measures, such as the computational complexity, level of introduced distortion and robustness. Various watermarking techniques have been proposed over the past decade. Depending on the application environment, a particular method may provide good performance on a specific measure at the expense of the remaining measures. For example, in audio copyright management, survivability of the watermark whenever the host signal undergoes a variety of processing such as compression, transmission or conversion to analog, is of paramount importance. On the other hand, in content enhancement where a non-hostile environment is assumed, payload is the primary goal. Regardless of the method, however, the watermarking process should not introduce audible artifacts and in fact the virgin and processed versions of the host audio should be indistinguishable.

The primary concern in audio watermarking is to make new the information inserted in the host audio inaudible and yet to manage to extract it at the detector. This is a challenge since the human auditory system is extremely sensitive to the presence of information unrelated to a host signal. Abiding by psychoacoustical principles is important in designing procedures that ensure inaudibility.

Audio watermarking techniques can be distinguished into those that use the spectral properties of the auditory system and those that capitalize on the temporal properties of human audition. Among the former, spread spectrum is one of the most successful methods that estimate the masking properties of audio by analyzing its short-time spectral features [1,2]. Here, the frequency domain representation of the host signal is viewed as a communication channel to be used for the transmission of the watermark. The watermark is spread over a wide frequency region and furthermore it is shaped so as to be restricted below the masking threshold thus making it inaudible. Signal transformations are considered as adding noise to the 'channel' that the host signal should be able to resist.

Methods that utilize temporal properties of the human auditory system or operate on the temporal representation of the host audio signal include the Least Significant Bit (LSB) replacement [1], noise quantization [11], phase coding [7], echo hiding [3, 4, 5], and others.

The use of the LSB for inserting meaningful information is one of the simplest techniques. Inaudibility is achieved by altering the LSB in of samples represented with high precision (e.g., 16 bit quantization) in order to carry the watermark information. Although this method is simple and it results in an inaudible watermark it suffers from low robustness since adding even minuscule levels of noise would destroy the watermark.

In [3] and [5], Bender et al. proposed the echo hiding method where a single positive echo is used to carry one bit of information. While the system achieved 16 b/s data payload its robustness was weak making it easily prone to detections and malicious attacks. Xu et al. [4] proposed a multi-echo embedding technique, where instead of embedding one large echo into an audio segment, four smaller echoes with different time offsets were chosen. The presence of multiple echoes contributes to the reduction in the echo magnitude thus reducing the chances of detection by third parties. On the other hand, this technique does not increase robustness because even small audio timbre changes can alter echo amplitudes [6].

In [7], Tilki and Beex proposed a watermarking method which is based on altering the phase of a single channel audio signal. Encoding is accomplished by imposing slight and controlled changes on the phase spectrum of short time signal segments. Modest phase alterations remain inaudible, while offering a substantial payload, which can reach 5 kb/s. Methods such as this are well-suited for content enhancement, despite their fragility.

In this paper we present a novel technique that carves a new channel in the phase of stereo digital audio. This is achieved by determining the minimum audible phase difference between the two channels. A large payload is achieved while maintaining inaudibility. The remainder of the paper outlines the psychoacoustical principle used for the development of this watermarking technique, describes the algorithm for it implementation, reports on the robustness tests, followed by conclusions.

2. THE PROPOSED SYSTEM

A functional representation of the watermarking system is shown in Figure 1. The encoder accepts digital audio and it determines the amount of information that may be inserted in the stereo signal by computing the short-time phase relationship between the two channels. Binary data provided by the auxiliary channel is imperceptibly encoded into the phase spectrum of the host audio. The watermarked auxiliary channel is reconstructed at the decoder after the watermark is detected and extracted.

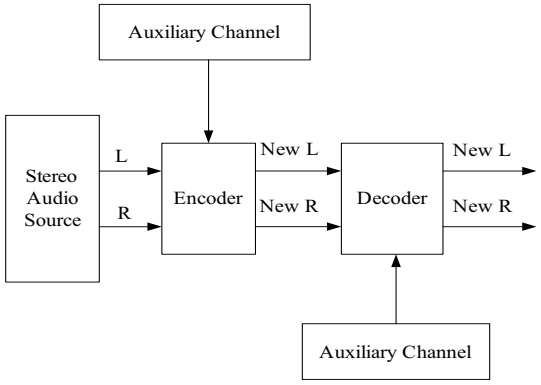

Figure 1. Functional diagram of the stereo watermarking system

2.1 Interaural Phase Difference (IPD) Relationships

The estimation of direction of a sound source by a listener requires the utilization of binaural audible information. Psycho-acoustical studies have established that this is achieved by the processing of binaural differential information in the brain [8, 9] and it includes the Interaural Phase or Time Difference (IPD/ITD), Interaural Intensity or Loudness Difference (IID/ILD), and Spectral notches whose locations depend on the elevation angle of the source. Utilizing IPD and IID alone can localize sources restricted on the azimuth plane. For all other locations processing of spectral notches is necessary as well.

The geometric relationship of the parameters related to a sound source located on the azimuth plane is shown on Figure 2.

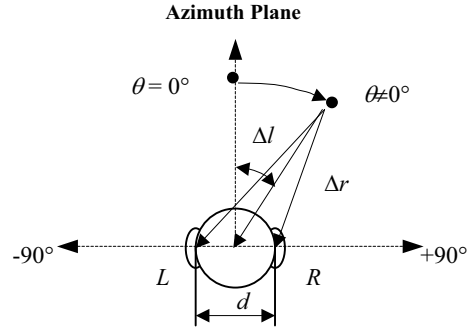

Figure 2. Sound source movement on the azimuth plane

Let θ be the azimuth angle, r the distance from the sound source to the center of the listener's head, and d the interaural distance (17 cm for a typical male head). The distance of the source from the right and the left ear are Δr and Δl, respectively. Δd is their difference and it is given by:

$\Delta r^2 = (r*\cos\theta)^2 + (r*\sin\theta - d/2)^2$, and
$\Delta l^2 = (r*\cos\theta)^2 + (r*\sin\theta + d/2)^2$,so
$\Delta d = \Delta r - \Delta l$.

It can be observed that Δd, is independent of the source distance. The interaural phase difference (IPD), Φ, is a function of frequency and it is given by the expression:

$\Phi = \Delta d * (f / c) * 360°$ or $\Phi = \Delta d * (f / c) * 2*\pi$ radians,

2.2 Watermarking Encoder

Audio and speech being stochastic signals must be processed on a short-time basis (10 to 40 ms). Here, processing is performed on adjacent, not overlapping pairs of N samples of the input signals (N Left and N Right samples). The temporal relationship of the two members of this pair must be maintained throughout the watermarking process. As there are no special needs for particular windowing in extracting the required data frames simple rectangular window is sufficient and add no overhead to the computation.

The short-time Fourier analysis of the Left and Right segments provides spectral magnitude and phase in modulo 2π. The actual phase can be estimated by unwrapping this principal phase sequence. However, this is unnecessary since this method requires the difference of Interaural Phase (IPD) rather than the actual IP values. This is efficiently obtained by adapting the IPD test on the cosine value of the Left and Right principal phase values.

The psychoacoustical phase threshold is set to correspond to a MAA of 1°, which is approximately the smallest value detectable in the audible range binaural hearing as reported in [10]. In other words, changes to the IPD can take place if the following relationship is satisfied.

$$\cos\{(\text{phase}[X_L(f_i)] - \text{phase}[X_R(f_i)]\} < \cos(-0.003104 * f_i)$$

where f_i are the frequency components provided by the Fourier analysis.

An energy threshold is used to avoid frequency components whose real and imaginary parts are too small to provide a meaningful phase value. For the remaining components, provided the stereo data at a particular frequency satisfies the threshold conditions one bit of watermarking data may be inserted according to the following scheme:

$$\text{phase}[X_L(f_i)] = \text{phase}[X_R(f_i)] \rightarrow \text{logical } 0$$
$$\text{phase}[X_L(f_i)] = kIPD_{max}(f_i) \rightarrow \text{logical } 1$$
$$\text{phase}[X_L(f_i)] >= IPD_{max}(f_i) \rightarrow \text{no encoding}$$

The approach taken in this process is to use the right channel as reference and to alter the phase of the left channel. Constant k in the above equation specifies the amount of phase difference within the IPD threshold, which would denote a logical one. Typically, $k = 1/2$ may be used. A block diagram of the encoder is shown in Figure 3.

All the frequency components of the left channel, both those altered as well as those left unchanged by the application of the psychoacoustical threshold, are collected and the new frequency representation of the left channel is constructed which now contains the masked data. The new time waveform may now be constructed.

The quantization process involved in converting the high precision processed signal into the standard 16-bit PCM format results in distortion that occasionally causes the encoded data to be corrupted. An iterative procedure is included in the encoder, which corrects this problem, whereby the altered frequency data of the Left channel are converted to the time domain via an Inverse DTF, quantized and converted back to the frequency domain. Decoding is performed and if the extracted data are identical to the inserted data set then encoding is considered successful, the data is stored and the next N-data set is processed. Otherwise, the frequency components corresponding to erroneous inserted data are marked so that they may be avoided in the future. This marking is achieved by altering their phase so that it may fall outside the IPD threshold. The process is repeated until the inserted and the extracted data sets are identical.

Marking the beginning of watermarked regions effectively is of paramount importance. This will also provide the necessary inter-channel synchronization. Timing errors of even a single sample may have detrimental effects in the decoding process. Spread-spectrum was used to provide robust and inaudible sample marking. A header is inserted in each channel just before the stereo watermark that provided effective synchronization. This way, single sample marking was achieved and the location watermarks in digital signals can be robustly recovered.

2.3 Watermarking Decoder

At the decoder, shown in Figure 4, the presence of a watermarked region is detected and the channels are checked to remove any introduced temporal shifts. Short-time Fourier analysis of the new Left and Right audio channels is performed by computing the DFT of signal frames obtained sequentially by N-point rectangular windows. The resulting phase information for every frequency component is compared against the IPD psycho-acoustic threshold, expressed in the equation above. Detection of the presence of encoded data masked into the audio signal is achieved according to the following process:

$$| phase[X_L(f)] - phase[X_R(f)] | =< r_1 IPD_{max}(f) \rightarrow \textbf{logical 0}$$

$$r_1 IPD_{max}(f) < | phase[X_L(f)] - phase[X_R(f)] | =< r_2 IPD_{max}(f) \rightarrow \textbf{logical 1}$$

$$| phase[X_L(f)] - phase[X_R(f)] | > r_2 IPD_{max}(f) \rightarrow \textbf{no data}$$

Constants r_1 and r_2 specify the ranges of phase differences used in the decoding process to extract logical 0, logical 1 or to indicate that no encoding was included in the particular frequency component under examination. Typically, $r_1 = ¼$, and $r_2 = ¾$ may be used.

It has already been noted that the energy level of the data in the frequency domain is important in the encoding process. Signal components below a certain threshold are avoided because their phase is meaningless. Similarly, the energy profile of the audio signal is computed prior to the encoding process. Signal regions with energy below a certain value are avoided as having been unsuitable for encoding. Such small signals suffer substantially from quantization effects and the iterative correction in the encoder may never end, resulting in an infinitely looping process.

3. PERFORMANCE EVALUATION

CD quality audio (stereo, sampling rate of 44100 Hz, 16 bits per sample) was used for the evaluation. The achieved peak watermarked payload was of the order of 20 kb/s.

3.1 Audibility

Systematic tests with 20 listeners with normal hearing abilities were conducted and they showed that (a) when the phase was randomly disturbed (no watermark inserted) the original and the processed audio were indistinguishable; (b) when meaningful watermarks were inserted no difference or preference between the two audio versions (virgin and processed) was noticed. The insertion of spread spectrum sequences used for synchronization did not add audible artifacts either, and the two audio versions remained overall indistinguishable.

3.2 System Robustness

The survival and integrity of the watermark was tested under various conditions. This is a fragile watermark that does not survive compression and D/A-A/D transformation. Nevertheless, it survives the following processes:

3.2.1. Cutting and splicing

A typical type of attack on digital audio is cuts and splices of various watermarked and unwatermarked content. Fortunately, this type of distortion can alter the audio dramatically, often making it useless. We have tested this type of attack with random cuts and splices of watermarked audio. Without prior knowledge of the approximate location of the watermarks these attacks largely failed because watermarks were repeated in many different audio locations and therefore we were still able to extract the information in all cases tested.

3.2.2. Temporal shifting

Another type of attack on digital stereo audio is the temporal shifting of the two channels thus corrupting the inter-channel phase relationship. Spread spectrum marking of both channels provided an effective countermeasure and ensures correct information recovery.

3.2.3. Downsampling

The watermarking scheme can successfully be applied to audio of varying bandwidths and sampling rate, with the smaller bandwidth permitting less information to be inserted. However, while downsampling of watermarked audio resulted in distortions, watermarks inserted after downsampling were correctly extracted.

4. CONCLUSION AND FUTURE WORK

In this paper we introduced a novel high capacity digital audio watermarking technique. Its operation is based on psychoacoustical principles of binaural hearing. Experimental results show that this watermarking system can successfully embed large amounts of information into stereo audio without introducing perceptual distortion. Its large payload makes it suitable for audio information enhancement. Its fragility necessitates a friendly environment, such as in authentication applications. Future work will address survivability under compression and methods for payload increase.

REFERENCES

[1] I.J. Cox, M.L. Miller, J.A. Bloom, "Digital Watermarking", pp.153-154, Morgan Kaufmann Publishers, 2002

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

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

[4] C. Xu, J. Wu, Q. Sun, K. Xin, "Applications of Digital Watermarking Technology in Audio Signals*", Journal Audio Engineering Society,* Vol. 47, No. 10, pp. 805-812, 1999.

[5] D. Gruhl, A. Lu, W. Bender, "Echo Hiding", *Proceedings of the 1996 Information Hiding Workshop,* pp. 295-315, 1996.

[6] H.O Oh, J.W. Seok, J.W. Hong, D. H. Youn, "New Echo Embedding Technique for Robust and Imperceptible Audio Watermarking", *Proceedings of the 2001 IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 1341-1344, 2001

[7] J. F. Tilki and A. A. Beex, "Encoding A Hidden Auxiliary Channel Onto A Digital Audio Signal Using Psychoacoustic Masking", IEEE Southestcon 1997, pp. 331-333, 1997

[8] W.A Yost, *Fundamentals of Hearing*, Academic Press, 1993.

[9] A.S. Bregman, *Auditory Scene Analysis,* The M.I.T. Press, 1990.

[10] A.W. Mills, "Auditory Localization", in Tobias, (Ed.) *Foundations of Auditory Theory,* Academic Press, 1972.

[11] Hyong Joong Kim, "Audio watermarking techniques", Invited Lecture, Pacific Rim Workshop on Digital Steganography 2003, July 3-4, Japan.

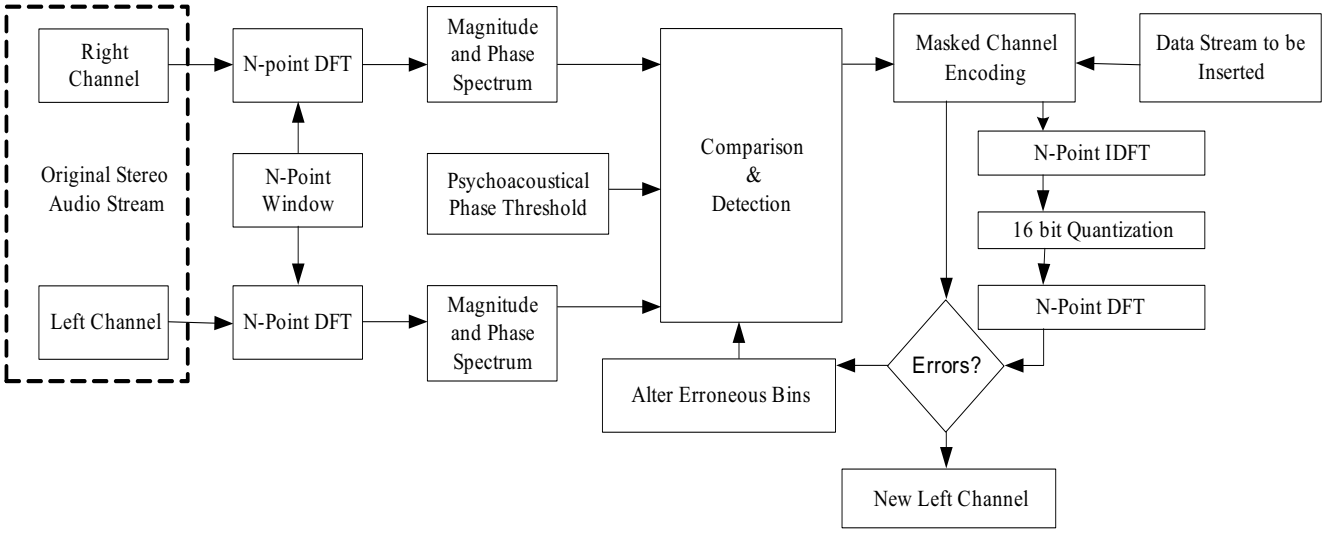

Figure 3. Functional diagram of the perceptual encoder

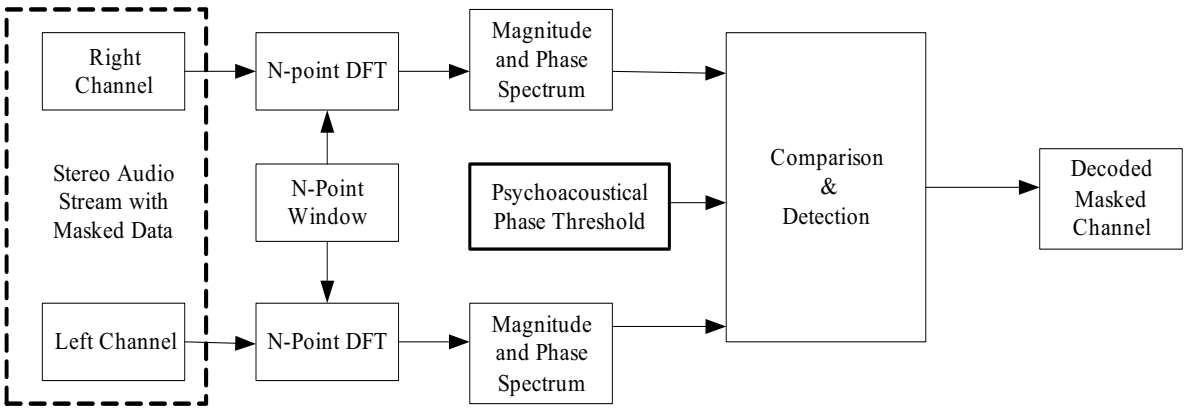

Figure 4. Functional Diagram of the Perceptual Decoder