A Perceptually Adaptive QIM Scheme for Efficient Watermark Synchronization

Hwai-Tsu HU† and Chu YU†, Members

1. Introduction

In recent years, the rapid advance of computer and networking technology has accelerated the speed of spreading digital media over the Internet. Such convenience creates the problem of unauthorized replication and distribution of multimedia data. While copyright protection of digital data has increasingly become an important issue, digital watermarking technology is receiving much attention for its ability to resolve security concerns.

In general, an audio watermarking scheme takes into consideration four aspects, namely, imperceptibility, security, robustness, and capacity. Different approaches have been attempted on a variety of domains (time, spectrum, cepstrum, discrete cosine, wavelet, etc.) to achieve effective watermarking. The main weaknesses of many existing algorithms consist in the vulnerability to time-shifting and cropping [1], apart from low payload and insufficient robustness.

One of the promising remedies against time-shifting is the use of a synchronization code to track the watermark. A previous method [2] performed well but suffered from the disadvantage that the range of the embedded sequence has to be long to render an unambiguous indication during the watermark extraction. On the other hand, the method developed in [1], [3] was suitable for a shorter segment and proved to be efficient in many attack situations. The adopted parameters were nonetheless tentative and did not exploit the human auditory properties. Hence the main aim in this study is to develop a robust synchronization scheme to combat the shifting and/or cropping attacks without the compromise of perceptual quality.

SUMMARY This study presents an adaptive quantization index modulation scheme applicable to a small audio segment, which in turn allows the watermarking technique to withstand time-shifting and cropping attacks. The exploitation of auditory masking further ensures the robustness and imperceptibility of the embedded watermark. Experimental results confirmed the efficacy of this scheme against common signal processing attacks.

key words: audio watermarking, adaptive quantization index modulation, frame synchronization

2. Adaptive Quantization Index Modulation

The chosen embedding scheme belongs to the category of quantization index modulation (QIM) [1], [3], [4]. The audio signal is first segmented into frames of length $L_f$, and each frame is further divided into $N_s$ subsections. This study assigns $L_f = 8820$ and $N_s = 70$. Hence the length of each subsection equals $L_s = L_f / N_s = 126$. Let $x(k_f L_f + k_s L_s + k_t)$ correspond to the $k_t$th sample in the $k_s$th subsection of the $k_f$th frame. To embed a single bit, the proposed scheme modifies the audio samples in a specific subsection using

$$\hat{x}(k_f L_f + k_s L_s + k_t) = x(k_f L_f + k_s L_s + k_t) + (\hat{m}(k_f, k_s) - m(k_f, k_s))M(k_t)$$

for $k_t = 0, 1, \ldots, L_s - 1$ (1)

where $m(k_f, k_s)$ and $\hat{m}(k_f, k_s)$ respectively denote the original and modified means of the samples in the $k_s$th subsection of the $k_f$th frame. $M(k_t)$ is a function designed to have a flat level in the middle but descend to zero on both sides, i.e.

$$M(k_t) = \lambda \times \left\{ \begin{array}{ll} 0.5 - 0.5 \cos(2\pi k/35), & k_t = 0, 1, \ldots, 17; \\ 1, & k_t = 18, \ldots, 107; \\ 0.5 - 0.5 \cos(2\pi (k - 90)/35), & k_t = L_s - 18, \ldots, L_s - 1. \end{array} \right. \tag{2}$$

The variable $\lambda$ in Eq. (2) is a scaling factor used to attain a mean of unity for $M(k_t)$, i.e. $\frac{1}{L_s} \sum_{k_t=0}^{L_s-1} M(k_t) = 1$. The declination on both ends of $M(k_t)$ can be considered as the fade-in and fade-out effects of the embedding process for each bit. The elimination of sharp transition across subsections helps reduce the magnitude spectrum in the middle and high frequencies.

Depending on the watermark bit $S_b(i)$, the QIM adjusts the mean of the subsection, i.e. $m(k_f, k_s)$, by

$$\hat{m}(k_f, k_s) = \left\{ \begin{array}{ll} \lfloor m(k_f, k_s) / D \rfloor D + D/4 & \text{if } S_b(i) = -1 \\ \lfloor m(k_f, k_s) / D \rfloor D + 3D/4 & \text{if } S_b(i) = 1 \end{array} \right.$$

for $i = 0, 1, \ldots, L_{sync} - 1$ (3)

where $\lfloor \cdot \rfloor$ represents the floor function. The synchronization code $\lfloor S_b(i) \rfloor$ is a bipolar random sequence with the length $L_{sync}$ equal to 70. The quantization step $D$ is of paramount importance because it is related to the robustness.

Manuscript revised September 2, 2012.
†The authors are with the Department of Electronic Engineering, National I-Lan University, Taiwan, ROC.
E-mail: hthu@mail.niu.edu.tw
DOI: 10.1587/transinf.E95.D.3097

Copyright © 2012 The Institute of Electronics, Information and Communication Engineers
and perceptibility of the watermark. Enlarging \( D \) will intensify the watermark robustness but at the cost of increasing quantization noise. In this study, the \( j \)th \( D \) is acquired by referring to the root-mean-square of \( N \) recently past lowpass filtered samples using

\[
D_j = k_L D_f + k_s D_s + k_i
= \left( \frac{1}{N} \sum_{n=1}^{N} x_{lp}^2(k_f L_f + k_s L_s + k_i - n) \right)^{\frac{1}{2}} \times 10^{-10/20} \tag{4}
\]

where \( x_{lp}(i) \) denotes the output of feeding \( x(i) \) through a 4th order Butterworth lowpass filter with the cutoff frequency set at \( f_s/126 \) (\( 350 \) Hz when \( f_s = 44100 \)). \( N \) is chosen as 1512 (\( \approx 34.3 \) ms). In the above equation, the scaling factor \( 10^{-10/20} \) aims at attenuating the signal power by 10 dB. For watermark extraction, the synchronization bit \( \hat{b}(i) \) is recovered by first obtaining the mean value for the next \( L_s \) samples, termed \( \bar{m}_i \), and the \(-10 \) dB RMS of previous \( N \) lowpass filtered samples, termed \( \bar{D}_i \), from the watermarked signals, and then polarizing the value obtained via the validity test of the inequality shown below:

\[
\hat{b}(i) = 2 \left( \bar{m}_i - \left[ \bar{m}_i/\bar{D}_i \right] \bar{D}_i \right) > 0.5 \bar{D}_i - 1. \tag{5}
\]

Notice that the computation required in the derivation of \( \hat{b}(i) \) and \( \bar{D}_i \) is low as the entire process can be formulated in a recursive manner, i.e., taking a new sample from the upcoming signal and eliminating the old at the front end of the memory buffer. Following the acquisition of the bit sequence, the existence of a synchronization code is detected by assessing the cross correlation between the code \( S \beta(i) \) and a decimated version of \( \hat{b}(i) \):

\[
r_\beta(i) = \frac{1}{L_{\text{sync}}} \sum_{k=0}^{L_{\text{sync}}-1} S \beta(L_{\text{sync}} - 1 - k) \hat{b}(i - kL_s). \tag{6}
\]

Because Eq. (6) places the synchronization code in a backward direction, the presence of a value near 1 indicates the starting of the watermark bit sequence.

3. Consideration of Human Auditory Perception

According to the signal to noise ratio (SNR) analysis given in [1], the QIM specified by Eq. (3) introduces a noise with a power of \( 7D_f^2/48 \), which is equivalent to 8.36 dB attenuation in comparison to \( D_f^2 \). On the other hand, the window \( M(k_i) \) in Eq. (2) contributes to the SNR by \(-0.46 \) dB. Combining with the 10 dB given in Eq. (4), the overall SNR resulting from the watermarking is around 17.9 dB.

For a synchronization code spanning over \( K \) frames, the resulting error signal is of the following form:

\[
e(i) = \sum_{k=0}^{K-1} \sum_{i=0}^{N_s-1} (\delta_{k_f L_f + k_s L_s + k_i}) - x(k_f L_f + k_s L_s + k_i)
\]

where \( u(\cdot) \) denotes a unit step function. As the rectangular window shown inside the last parentheses of the above equation predominates the variation of \( e(n) \), taking the discrete-time Fourier transform of \( e(i) \) reveals a windowing effect approximating to \( \sin(L_s \pi F)/\sin(\pi F) \) with \( F \) denoting the normalized frequency. Figure 1 depicts a typical power spectrum of \( e(i) \), which exhibits a series of decreasing sidelobes along the frequency axis, each with a width of slightly larger than \( 350 \) Hz (\( \approx f_s/126 \)). It appears that \( e(i) \) is a lowpass shaping signal with its energy primarily concentrated below \( 350 \) Hz. According to the theory of auditory entropy [5], [6], the masking threshold for the spectral components below \( 350 \) Hz is generally higher than \(-17 \) dB regardless of the tonality of the audio signal. Consequently, the deliberately reserved \( 17.6 \) dB SNR during the embedding process can ensure the imperceptibility of the synchronization codes.

4. Performance Evaluation

The test subjects comprised a total of 450 seconds of audio signals excerpted from ten music recordings, including vocal arrangements and ensembles of musical instruments. All audio signals were sampled at 44.1 kHz with 16-bit resolution.

The method used in [2] serves as a baseline for comparison. Here the synchronization code \( w(i) \) is a random bipolar sequence with a frame length identical to that used in the proposed scheme. The code embedding for the baseline method is carried out using

\[
y(i) = x(i) + a(W_{\text{Ham}}(i) \ast |x(i)|w(i)) \tag{8}
\]

where \( W_{\text{Ham}}(i) \) denotes a Hamming window of size 25, and
Table 1 Statistics of the measured objective difference grades. The data are interpreted as “mean [standard deviation]”.

| Method        | ODG          |
|---------------|--------------|
| Baseline      | -1.0644 [1.3279] |
| Proposed      | -0.0192 [0.1876] |

“∗” stands for the convolution operator. α is the embedding strength, which is adjusted to attain a SNR = 20 dB. As for comparison, the average SNR for data embedding using the proposed scheme is 20.73 dB with a standard deviation of 0.84 dB.

The quality of the watermarked audio signal is first evaluated using the perceptual evaluation of audio quality (PEAQ) [7], which renders an objective difference grade (ODG) between −4 and 0 to signify a perceptual impression from “very annoying” and “imperceptible”. Table 1 presents the average ODG’s for the watermarked audio signals obtained by the baseline method and proposed scheme. As shown in Table 1, the average ODG resulting from the proposed scheme is near zero, indicating that all the watermarked signals are hardly distinguishable from the original ones. On the contrary, the baseline method results in a noticeable degradation in quality.

The attacks adopted to test the robustness and effectiveness of the proposed scheme include:

A. Resampling: conducting down-sampling to 11.025 kHz and then up-sampling back to 44.1 kHz.
B. Requantization: quantizing the watermarked signal to 8 bits/sample and then back to 16 bits/sample.
C. Noise corruption (I): adding zero-mean white Gaussian noise to the watermarked audio signal with SNR = 30 dB.
D. Noise corruption (II): adding zero-mean white Gaussian noise to the watermarked audio signal with SNR = 20 dB.
E. Echo addition: adding an echo signal with a delay of 50 ms and a decay to 5% of the original amount.
F. Lowpass filtering: applying a lowpass filter with a cutoff frequency of 8 kHz.
G. Jittering: randomly deleting or adding one sample for every 100 samples within each frame.
H. MPEG compression (I): compressing and decompressing the watermarked audio signal with a MPEG layer III coder at a bit rate of 128 kbps.
I. MPEG compression (II): compressing and decompressing the watermarked audio signal with a MPEG layer III coder at a bit rate of 64 kbps.
J. Time shifting: shifting the watermarked audio signal by 50% of the frame length.

For the baseline method, the position of the embedded code can be searched by examining the correlation function between y(i) and w(i). To establish a comparative basis, this study uses a normalized correlation function r_2(i):

\[ r_2(i) = \frac{\sum_i y(i)w(i)}{\sum_i \alpha(W_{Ham}(i)|x(i)|w(i))w(i)}. \]  

(9)

Theoretically, either \( r_1(i) \) or \( r_2(i) \) will render values near 1 in the presence of the synchronization code. The efficiency is demonstrated via the means and standard deviations of \( r_1(i) \) and \( r_2(i) \) obtained from two scenarios, namely the one with the synchronization code presented and the other without. Table 2 gives the statistical results. For both the proposed and baseline schemes, whether or not the difference of the means attributable to the presence of synchronization codes is justified using one-way analysis of variance (ANOVA). The resulting p-values are found to be extremely near zero in all cases, suggesting that the two means obtained from different scenarios are significantly distinct from each other. In case I, the mean value of \( r_1(i) \)’s looks somewhat smaller than that in other cases. The corresponding standard deviation is also rather larger. This is because the 64 bps MPEG coder receives greater distortion in the low frequency region, where is used by the proposed scheme to insert synchronization codes.

As revealed by the low standard deviations of the obtained \( r_1(i) \)’s, the proposed scheme renders a much easier detection of the synchronization code since the sample populations for the presence and absence of the synchronization codes are well separated. In case the synchronization code is not present, the derived output generally exhibits a random bipolar sequence. By assuming the occurrences of +1 and -1 are of equal probability, the standard deviation of the normalized cross-correlation function is approximately \( 1/\sqrt{L_{sync}} \). Such an extent does not cause any trouble in the synchronization code identification. In contrast, using \( r_2(i) \) to detect the synchronization codes appears impracticable because large standard deviations always occur.

To illustrate how the variance affects the detection of
synchronization codes, let $\lambda_1$ and $\lambda_0$ be the samples drawn from the normalized correlation functions with and without the presence of synchronization codes respectively. Suppose $\lambda_i$ ($i = 0, 1$) holds a Gaussian distribution with a mean $\mu_i$ and a standard deviation $\sigma_i$. The probability for the occurrence $\lambda_1 \leq \lambda_0$ can be calculated as

$$P(\lambda_1 \leq \lambda_0) = \int_{-\infty}^{\lambda_0} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\lambda_1 - \mu_1)^2}{2\sigma_1^2}} d\lambda_0 \int_{\lambda_1}^{\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\lambda_0 - \mu_0)^2}{2\sigma_0^2}} d\lambda_0 d\lambda_1.$$  

(10)

As shown in Table 2, the derived $P(\lambda_1 \leq \lambda_0)$ for the proposed scheme is essentially zero, while the amount for the baseline method is generally above 0.03 under various attacks. For a probability like this, it is unlikely to discriminate a single $\lambda_1$ from a bunch of $\lambda_0$'s for samples in one frame.

5. Concluding Remarks

This study improves the watermarking approach presented in [1] by coupling the QIM with an adaptive quantization step. The proposed scheme exploits auditory masking properties to make the embedded synchronization codes inaudible. As revealed by the PEAQ evaluation, the audio signal embedded by the proposed scheme is virtually indistinguishable from the original one. The effect of the adaptive QIM can be thought of as an automatic gain adjustment with lowpass shaping in frequency. The experimental results confirm that the embedded data are not only able to cope with time-shifting situations but also robust against common signal processing attacks.

Acknowledgments

This work was supported by the National Science Council, Taiwan, ROC, under Grant NSC 101-2221-E-197-033.

References

[1] S. Wu, J. Huang, D. Huang, and Y.Q. Shi, “Efficiently self-synchronized audio watermarking for assured audio data transmission,” IEEE Trans. Broadcast., vol.51, no.1, pp.69–76, 2005.
[2] P. Bassia, I. Pitas, and N. Nikolaidis, “Robust audio watermarking in the time domain,” IEEE Trans. Multimedia, vol.3, no.2, pp.232–241, 2001.
[3] X.Y. Wang and H. Zhao, “Novel synchronization invariant audio watermarking scheme based on DWT and DCT,” IEEE Trans. Signal Process., vol.54, no.12, pp.4835–4840, 2006.
[4] B. Chen and G.W. Wornell, “Quantization index modulation: A class of provably good methods for digital watermarking and information embedding,” IEEE Trans. Inf. Theory, vol.47, no.4, pp.1423–1443, 2001.
[5] B. Carnero and A. Drygajlo, “Perceptual speech coding and enhancement using frame-synchronised fast wavelet packet transform algorithms,” IEEE Trans. Signal Process., vol.47, no.6, pp.1622–1635, 1999.
[6] T. Painter and A. Spanias, “Perceptual coding of digital audio,” Proc. IEEE, vol.88, no.4, pp.451–515, 2000.
[7] P. Kabal, “An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality,” TSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University, 2002.  
http://www.TSPECE.McGill.CA/MMSP/Documents