EFFICIENT COHERENT PHASE QUANTIZATION FOR AUDIO WATERMARKING

Xiao-Ming Chen, Gwenaël Doërr, Michael Arnold, and Peter G. Baum

Technicolor – Security & Content Protection Labs

ABSTRACT
The weighted overlap add (WOLA) technique is routinely used in audio processing to avoid introducing audible clicks. This process introduces inter-dependencies between WOLA coefficients and it is thus necessary to revisit watermarking algorithms to accommodate for such dependencies and prevent introducing self-inflicted interferences. In this paper, we focus on an existing phase quantization based audio watermarking system. Instead of the original computationally expensive ad-hoc process for dealing with the inter-dependencies between WOLA coefficients, we introduce an alternative strategy. Experimental results show a notable reduction of the complexity while maintaining robustness performance.

Index Terms— Signal coherent watermarking, phase quantization, WOLA transform.

1. INTRODUCTION

Most watermarking systems can be reduced to a three steps process: (i) apply some transform (e.g. DCT, DWT, etc) to an original content and pseudo randomly select a set of transform coefficients, (ii) apply a watermark embedding technique to the selected transform coefficients, and (iii) incorporate the modified coefficients and apply the inverse transform to obtain the final watermarked content. It is commonly accepted that state-of-the-art watermarking techniques, such as spread-spectrum watermarking [1] or binning schemes [2], are readily applicable when the transform produces independent transform coefficients. Though manageable, dealing with correlated coefficients is much more complex.

In practice, however, it may is needed to use transforms that introduce inter-dependencies. In audio processing for instance, the weighted overlap add (WOLA) mechanism inherently produces correlated transform coefficients. It is routinely used to prevent the appearance of audible clicks after processing an audio signal. Using such transforms requires revisiting baseline watermarking techniques so as to guarantee that the modifications introduced by the watermarking process are coherent with the natural inter-dependencies induced by the transform. This is a vital element to avoid introducing undesired self-inflicted interferences between watermark samples.

In this paper, we focus on a previously published audio watermarking algorithm [3]. Section 2 details the characteristics of this algorithm based on quantizing the phase of WOLA coefficients. It also emphasizes an ad-hoc reference angles generation strategy necessary to reduce the impact of the WOLA transform on the embedding effectiveness of the watermarking algorithm. Section 3 then analytically surveys the impact of the WOLA transform and derives a simplified strategy to generate reference angles. Experimental results reported in Section 4 clearly demonstrates that the proposed strategy reduces the computational complexity notably while maintaining robustness performances. Finally, Section 5 reviews the findings of the article and suggests further routes for investigations.

2. PREVIOUS WORK

Over the last few years, a number of audio watermarking techniques exploiting the analysis-synthesis framework have been proposed [4, 5]. In the remainder of this paper, we will focus on a previously published audio watermarking system relying on phase quantization in the WOLA domain [3].

2.1. The WOLA Domain

In audio processing, it is common practice to combine (i) filtering along a temporally sliding window and (ii) a reconstruction mechanism in an attempt to keep control over the psychoacoustic impact of audio processing primitives. This is commonly referred to as the analysis-synthesis framework, which the WOLA mechanism [6] is simply an example of.

During the analysis phase of the WOLA process, the input audio signal $x$ is segmented in $B$-samples long blocks $x^{(m)}$ using a sliding window with a hop-size of $R$ samples. The discrete Fourier transform (DFT) is then applied to each block. To avoid potential cross-talk in the estimated spectrum, each block $x^{(m)}$ is multiplied by an analysis window $w_{A}$ to obtain a windowed block $\tilde{x}^{(m)}$ before applying the DFT transform. The resulting collection of DFT-transformed windowed blocks $\tilde{X}^{(m)}$ can subsequently be input to an audio processing algorithm, e.g. watermarking.

During the synthesis phase of the WOLA process, the modified blocks $\tilde{X}^{(m)}$ output by the audio algorithm are used to reconstruct the audio signal. First, the inverse discrete Fourier transform (IDFT) is applied to the blocks in order to get back in the time domain. In general, audio applications may involve non-linear spectral processing and it is thus recommended to apply a synthesis window $w_{S}$ to the blocks $\tilde{Y}^{(m)}$ after the IDFT, so as to suppress audible artifacts by fading out spectral modifications at block boundaries. Finally, the resulting blocks $\tilde{Y}^{(m)}$ are aligned according to the hop-size $R$ before being added together to obtain the output audio signal $y$.

In the remainder of this paper, we will abusively refer to embedding the watermark in between the analysis and synthesis phases as watermarking in the WOLA domain. Moreover, we will use a half-window hop size ($R = B/2$) and the same sine windows $w$ for analysis and synthesis:

$$w[n] = w_{A}[n] = w_{S}[n] = \sin \left( \frac{\pi n}{B} \right), \ 0 \leq n < B.$$  \hspace{1cm} (1)

This setup guarantees perfect reconstruction if the signal is not modified between the analysis and synthesis phases.

2.2. Watermarking in the WOLA Domain

The baseline audio watermarking system detailed in [3] basically reduces to embedding a watermark by quantizing the phase of WOLA coefficients while maintaining psychoacoustic fidelity. Moreover, a
correlation-based detector in the time domain is employed at the receiver side to retrieve the embedded watermark.

2.2.1. Phase Quantization-based Embedding

The embedding process consists in (i) extracting the phase\(^3\) of WOLA coefficients from incoming transformed blocks \(X^{(n)}\) and arranging them sequentially in a 1-D signal \(\varphi\). (ii) applying a quantization based embedding algorithm to obtain the watermarked phases \(\psi\), and (iii) segmenting the resulting signal in \(B\)-samples long blocks to reconstruct the watermarked transformed blocks \(\tilde{Y}^{(n)}\) which is inverse transformed back to the time domain.

Assuming that the system embeds symbols taken from an \(A\)-ary alphabet \(\mathcal{A}\), the embedding process can be written:

\[ \psi[i] = \theta_{a,K}[i], \quad a \in \mathcal{A}, \quad i \in S.B.N + [0 : S.B], \quad (2) \]

where \(\theta_{a,K}\) is a sequence of reference angles in \([-\pi, \pi]\) associated to the symbol \(a\), pseudo-randomly generated from a secret key \(K\). The parameter \(S\) indicates that each symbol may be spread across several WOLA blocks to guarantee robustness. This is a pure replacement watermarking strategy: the output angles \(\psi\) are independent of the input angles \(\varphi\) derived from the host signal \(x\).

Psycho-acoustic adaptation. Unfortunately, this straightforward embedding strategy produces very audible artifacts and it is necessary to slightly adjust the embedding protocol in practice, so as to accommodate for the sensitivity of human auditory system. First, the embedding process is limited to a specified frequency band \(\mathcal{F} = [f_s, f_s]\). Angles below frequency tap \(f_s\) are discarded due to their high audibility whereas angles above frequency tap \(f_s\) are ignored because of their high variability. In addition, within \(\mathcal{F}\), the embedding process is modified so that the angle difference \(\delta[i] = |\psi[i] - \varphi[i]|\) remains below psycho-acoustic slacks \(\psi[i] \in [0, \pi]\) obtained after spectra analysis [3, 7]. This can be formally written:

\[ \psi[i] = \varphi[i] + \Delta[i], \quad |\Delta[i]| \leq \min \{\delta[i], \psi[i]\}, \quad i \in B.N + [f_i : f_s] \]

with \(\Delta[i] = \theta_{a,K}[i] - \varphi[i]\). (3)

Controlling distortion this way guarantees that the introduced changes are strictly inaudible.

Reference angles generation strategies. To generate the reference angles \(\theta_{a,K}\), one could be tempted to generate for each symbol \(a\) a sequence of angles in \([-\pi, \pi]\) using a pseudo-random number generator seeded with the secret key \(K\). Although intuitive and straightforward, this strategy results in very poor performances: the angles set after embedding are severely distorted by the WOLA reconstruction process i.e. without any subsequent attack.

To circumvent this issue, the original paper [3] used the following ad-hoc generation strategy for each symbol \(a\):

1. Generate \(\frac{L}{2}\) in \(B\)-samples long white Gaussian noise patterns directly in the frequency domain by combining a unit magnitude over the full spectrum and pseudo-randomly generated angles \(\theta_{a,K}^{(n)}, 0 \leq n < \frac{L}{2}\), uniformly distributed over \([-\pi, \pi]\) respecting the symmetry property of the DFT of real signals;

2. Apply the IDFT to all the patterns and concatenate the resulting time-domain signals \(r_{a,K}^{(n)}\) to obtain a \(L = \frac{L}{2}\) in \(B\)-samples long reference signal \(r_{a,K}\) in the time domain.

\(^3\)In this paper, all angles or angle differences lie in the interval \([-\pi, \pi]\) after appropriate modulo-\(2\pi\) operations, if not otherwise stated.

3. Apply the forward WOLA transform and retrieve the phase of the resulting WOLA coefficients to obtain the \(S.B\) reference angles \(\theta_{a,K}\) to be used for watermarking.

In other words, the reference angles \(\theta_{a,K}\) have already experienced the WOLA transform and are thus less impacted by the subsequent WOLA reconstruction process.

2.2.2. Correlation-based Detection in the Time Domain

At the receiver side, an audio signal \(z\) is presented to the detector that may, or may not contain a watermark. Since the embedder relies on quantization, a natural detection strategy would be to use some nearest codeword identification technique. Still, this approach is extremely sensitive to desynchronization and is computationally expensive if it needs to performed over some sliding window. In their original paper [3], the authors suggested using a practical time domain detector using cross-correlation.

For each potential symbol \(a\), the following array of correlation scores is computed:

\[ \rho_{a,B}[l] = \frac{1}{L} \sum_{i=0}^{L-1} z[i]\hat{z}_{a,K}[l+i], \quad -L + 1 \leq l \leq L - 1 \quad (4) \]

where \(\hat{z}\) is the whitened version of the tested audio signal \(z\), \(\hat{z}_{a,K}\) is the whitened version of time-domain reference signal \(r_{a,K}\), and \(l\) is the correlation lag. Since the watermark is embedded in the WOLA phase domain, spectral amplitudes are irrelevant for watermark detection. This is the reason why the tested/reference signals are whitened prior to detection in order to eliminate potential interferences from the host signal. The whitening process reduces to (i) applying the forward WOLA transform to the tested/reference signal, (ii) normalizing the magnitude of all WOLA coefficients to 1, and (iii) applying the inverse WOLA transform to obtain the whitened signal.

Eventually, the detector isolates the symbol \(\hat{a}\) exhibiting the highest correlation in absolute value:

\[ \hat{a} = \arg \max_{a \in \mathcal{A}} \rho_{a,B}, \quad \text{with} \quad \rho_{a,B} = \max_{l} |\rho_{a,B}[l]|. \quad (5) \]

If the correlation score \(\rho_{a,B}\) exceeds a specified threshold \(\tau\), then the tested signal \(z\) is considered to be watermarked and to carry the symbol \(a\). Otherwise, the tested signal is considered not to convey any watermarking information.

3. EFFICIENT REFERENCE ANGLE GENERATION

The reference angles generation strategy described in Subsection 2.2.1 involves a number of (1)DFT transforms, which induce a significant computational load. More precisely, computing the \(S.B\)-samples long sequence \(\theta_{a,K}\) requires \(\frac{3L}{2}\) times length-\(B\) IDFTs, \(S.B\) multiplications for windowing and \(S\) times length-\(B\) DFTs. Assuming that \(B = 2^p\) is a power of 2, performing a length-\(B\) (1)DFT involves \(2.B.p\) real multiplications and \(2.B.p\) additions with the Cooley-Tukey algorithm [8]. In other words, each WOLA coefficients roughly needs \(6.p\) real additions and \(6.p + 2\) real multiplications to be computed.

This section carefully analyzes the impact of the WOLA reconstruction process to highlight an alternative means to generate the reference angles \(\theta_{a,K}\) efficiently so as to facilitate the incorporation of this watermarking algorithm onto embedded systems. For clarity,
Fig. 1. Illustration of the reference angles generation in the original audio watermarking system [3].

Figure 1 illustrates the previously described phase generation process. It is straightforward to realize that the reference angles are derived either from a single reference signal segment $r_{u,K}^{(l)}$ (WIN case) or from two segments (WOL case).

3.1. WIN Case

In this setup, the WOLA transform reduce to the DFT of the multiplication of some pseudo-random reference signal $r_{u,K}^{(l)}$ and the sine window $w$. Using an elementary property of the DFT, this is equivalent to computing the circular convolution between the DFT of the reference signal and the DFT of the sine window:

$$R_{u,K}^{(m)}[u] = \frac{1}{B} \sum_{v=0}^{B-1} c^{(k)}_{u,K}[v] W[u - v], \quad (6)$$

where the notation $< n > = n - \lfloor \frac{n}{B} \rfloor B$ indicates the modulo operation which maps any integer to the interval $[0, B)$. The DFT of the sine window defined in Equation (1) is real valued and equal to:

$$W[u] = \frac{-\sin(\frac{\pi}{B})}{\cos(\frac{\pi}{B}) - \cos(\frac{2\pi}{B} u)}, \quad 0 \leq u \leq B. \quad (7)$$

This spectrum is characterized by the fact that most its energy is concentrated around $u = 0$, leading to the following approximation:

$$R_{u,K}^{(m)}[u] \approx \frac{1}{B} \sum_{v=0}^{B-1} c^{(k)}_{u,K}[v] W[u], \quad (8)$$

when only the $2L_u + 1$ dominant terms of $W$ are kept. Moreover, $W$ being symmetric, the previous equation can be further reduced to:

$$R_{u,K}^{(m)}[u] \approx \frac{W[0]}{B} c^{(k)}_{u,K}[0] + \sum_{v=1}^{L_u} \frac{W[v]}{B} \left( c^{(k)}_{u,K}[v] + c^{(k)}_{u,K}[v+1] \right). \quad (9)$$

The $L_u + 1$ coefficients $W[u]$ do not depend of the secret key $K$, and can thus be precomputed and stored in a lookup table for later use. In the WIN setup, each WOLA coefficients can then be computed with $2L_u + 1$ real multiplications and $4L_u$ real additions.

3.2. WOL Case

In this setup, the block $r_{u,K}^{(m-1)}$ input to the WOLA process is constructed by concatenating segments of the sequences derived from the angles $\phi_{u,K}^{(n-1)}$ and $\phi_{u,K}^{(n)}$. The spectrum of this block prior to WOLA transform can be written:

$$R_{u,K}^{(m-1)}[u] = \frac{(-1)^v}{2} \left( c^{(k)}_{u,K}[v] + c^{(k)}_{u,K}[v+1] \right) + \frac{1}{B} \sum_{v=0}^{B-1} F[v](1-\cos^2(\phi_{u,K}^{(n-1)})) + \frac{1}{B} \sum_{v=0}^{B-1} G[v](1-\cos^2(\phi_{u,K}^{(n)})), \quad (10)$$

where the spectrum $F$ is defined as follows:

$$F[v] = \left\{ \begin{array}{cl} \frac{1-\cos^2(\phi_{u,K}^{(n-1)})}{1-\cos^2(\phi_{u,K}^{(n)})}, & \text{if } v \neq 0, \\ 0, & \text{if } v = 0. \end{array} \right. \quad (11)$$

As in the WIN case, the WOLA process then reduces to a circular convolution with $W$ and the WOLA coefficients can be rewritten:

$$R_{u,K}^{(m)}[u] = \frac{1}{B} \sum_{v=0}^{B-1} W[v](1-\cos^2(\phi_{u,K}^{(n-1)})) + \frac{1}{B} \sum_{v=0}^{B-1} G[v](1-\cos^2(\phi_{u,K}^{(n)})), \quad (12)$$

where $G$ is the circular convolution of $W$ and $F$. For the sine window defined in Equation (1), the expression of $G$ can be analytically derived and is given by:

$$G[v] = \frac{(-1)^v}{2} \left( 1 + \frac{\sin(\frac{2\pi}{B})}{\sin(\frac{\pi}{B})} \right) \cos(\frac{\pi v}{B}) \quad (13)$$

Again, most of the energy of $G$ is concentrated around $v = 0$ and derivations can be simplified by keeping only the $2L_u + 1$ most significant terms. Moreover, the spectrum $G$ in this narrow band is heavily dominated by the imaginary component, which allows to ignore the real component in the previous equation. Putting everything altogether, WOLA coefficients can be expressed as follows:

$$R_{u,K}^{(m)}[u] \approx \frac{1}{B} \sum_{v=0}^{L_u} W[u](1-\cos^2(\phi_{u,K}^{(n-1)})) + \frac{1}{B} \sum_{v=0}^{L_u} G[u](1-\cos^2(\phi_{u,K}^{(n)})), \quad (14)$$

By storing appropriate pre-computed coefficients in lookup tables and exploiting the (a)symmetric properties of the spectra $W$ and $G$, it is then possible to compute each WOLA coefficient in the WIN setup with $2L_u + 2L_d + 2$ real multiplications and $8L_u + 8L_d + 2$ real additions.

4. EXPERIMENTAL RESULTS

For experiments, the block length $B$ has been set to 1,024 samples (i.e. $p = 10$) and each embedded symbol $a$ is spread over $S = 31$ blocks. In other words, each watermarked symbol requires 16 k samples for embedding. Additionally, the frequency range for embedding has been limited to 300–10,000 Hz. For PCM audio sampled at 48 kHz, it implies that 418 WOLA coefficients are modified by the embedding process instead of 1,024. The next subsections clearly demonstrate that the proposed phase generation technique succeeds in decreasing the computational load while maintaining similar robustness performances.
Table 1. Benchmarking results. Detection rate (%) for the original scheme are given in brackets. No false positive was observed during benchmarking.

<table>
<thead>
<tr>
<th>Lossy compression</th>
<th>mp3</th>
<th>mp3Pro</th>
<th>AAC</th>
<th>ACC+</th>
</tr>
</thead>
<tbody>
<tr>
<td>64 kb/s</td>
<td>79 (80)</td>
<td>94 (94)</td>
<td>85 (85)</td>
<td>95 (95)</td>
</tr>
<tr>
<td>48 kb/s</td>
<td>86 (86)</td>
<td>61 (65)</td>
<td>49 (50)</td>
<td>58 (58)</td>
</tr>
<tr>
<td>32 kb/s</td>
<td>27 (28)</td>
<td>27 (27)</td>
<td>22 (23)</td>
<td>9 (8)</td>
</tr>
</tbody>
</table>

Additive white Gaussian noise (AWGN)

<table>
<thead>
<tr>
<th>SNR</th>
<th>20 dB</th>
<th>10 dB</th>
<th>3 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>65 (65)</td>
<td>33 (32)</td>
<td>12 (12)</td>
</tr>
</tbody>
</table>

4.1. Detection SNR Preservation

To evaluate the impact of the WOLA reconstruction process on the generated reference sequences, the detection signal-to-noise ratio ($SNR_{det}$) is introduced:

$$SNR_{det} = \frac{\text{var}[r_1[n]]}{\text{var}[r_2[n] - z[n]]}. \quad (15)$$

For calibration, the psycho-acoustic adaptation module has been disabled and a zero-mean Gaussian signals have been used as host signals. While using purely random reference angles leads to a detection SNR of 3.3 dB, the original phase generation process actually raises this performances metric up to 12.2 dB. It clearly highlights that the WOLA process naturally introduces inter-dependencies between coefficients and that if, the watermarking process do not mimic them in some way, the WOLA reconstruction process will self-damage the embedded watermark signal even in the absence of any attack. Using the newly proposed reference angles generation mechanism with parameters $L_w = 2$ and $L_g = 3$ has been found to yield a similar detection SNR i.e. around 12 dB.

4.2. Computational Complexity Reduction

Using the original scheme, computing a WOLA coefficient requires 60 real additions and 62 real multiplications. In contrast, the proposed strategy only requires in average around 24 real additions and 8 real multiplications per coefficient with the parameters setup obtained after calibration ($L_w = 2$ and $L_g = 3$). This is a computational gain of roughly a factor 4. However, in practice in our implementations on embedded systems, we observed a reduction from 34 kCycles to 13 kCycles corresponding to a computational gain of a factor of approximately 2.6. The difference is due to the fact that our implementation is not optimized assembler but C code. Furthermore we observed that the not optimal random memory access in the embedded device has a significant influence on the computational load. Nevertheless the achieved computational gain enables the real-time generation of the references, which would otherwise not be possible in the set-up used.

4.3. Benchmarking

As a sanity check, both variants of the audio watermarking scheme were submitted to an intensive benchmarking campaign including a wide scope of signal processing primitives (AWGN, filtering, lossy compression, denoising, desynchronization, etc) and involving 100 different audio files with a total play length of more than 7 hours.

Due to the lack of space, we only report in Table 1 the detection rates under lossy compression using a variety of audio codecs with bit rates up to 64 kb/s and under additive white Gaussian noise. The findings are similar in all cases: the proposed algorithm for reference angles generation does not incur any notable performances loss.

5. CONCLUSION

To facilitate integration onto embedded systems, this paper revisited a previously published audio watermarking system relying on some reference angles in the WOLA domain for embedding [3]. The WOLA reconstruction process introduces some inter-dependencies between transform coefficients which need to be properly taken into account during watermarking to avoid introducing self-inflicted interferences. This phenomenon was originally addressed with a computationally expensive procedure and the paper proposed an alternate strategy by reducing the original procedure to a couple of frequency-domain short-length convolutions. Experimental results clearly demonstrate that this new strategy reduces computational complexity while maintaining robustness.

This paper focuses on making the embedded watermark coherent with the inter-dependencies imposed by a transform and nicely complements other works aiming at making the embedded watermark coherent with the natural redundancy of the host signal itself, either for higher security issues [9] or for guaranteed embedding effectiveness [10]. Future work will be dedicated to analyzing in depth the effects of the WOLA process in an attempt to further decrease the amount of self-inflicted interferences. For instance, the magnitude of the host signal spectrum is likely to play a role at embedding time which is not yet fully understood.

6. REFERENCES