This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2661724, IEEE Transactions on Information Forensics and Security

An Audio Watermark Designed for Efficient and Robust Resynchronization after Analog Playback

Andrew Nadeau and Gaurav Sharma, *Fellow, IEEE,*

***Abstract*—We propose a spread spectrum (SS) audio watermark designed to withstand analog playback, including desynchronization caused by small differences between playback and recording rates. Desynchronization robustness relies on detecting short blocks of a magnitude-only watermark embedded in the frequency domain where the resolution of the SS chips can be reduced. Lost spreading gain due to the lower number of SS chips is compensated using blind dynamic time warping (DTW) detection (does not access original signal). DTW aligns sequences of blocks to improve robustness to interference while mitigating the vulnerabilities of long SS sequences to desynchronization. Results demonstrate that the proposed watermark survives analog playback and warping up to $\pm 2\%$. Additionally, compared with a recent baseline scheme that uses brute force resampling to search for resynchronization, the proposed watermark is 300 times more computationally efficient, and does not compromise robustness to either desynchronization (e.g. jitter, resampling, time warping, frequency scaling) or non-desynchronizing modifications (e. g. AAC compression, additive noise).**

***Index Terms*—Synchronization, warping, acoustic path, informed embedding, dynamic time warping (DTW)**

I. Introduction

Although audio watermarking has been extensively researched (see [1] for a recent comprehensive survey), there is renewed interest in applications enabled by the proliferation of handheld smartphone and tablet computers. For example, the so called "second screen" scenario [2] occurs when a consumer viewing primary content on a large size display is also using a handheld smartphone or tablet device at the same time. In this situation, a watermark embedded in the audio track of primary content and detected by the user's handheld device from ambient audio allows the personal device to synchronize with the broadcast content and connect to auxiliary information/media on the Internet [3]–[6]. For such applications, the watermark must survive an analog playback channel comprising of a consumer grade loudspeaker and microphone and the acoustic path between these devices [7].

The proposed audio watermark targets the key challenges of analog playback applications: resynchronization and efficiency. Specifically, the analog playback channel not only introduces additive noise and reverberation [8], but is also a source of desynchronization due to small deviations from the nominal sampling rate of each device (normally 44.1 kHz or 48 kHz). A net relative deviation between the loudspeaker and microphone speeds up or slows down the watermarked signal and causes warping similar to proposed active attacks for removing watermarks [9].

Desynchronization due to warping or other operations modifies the time scale of the audio signal and disrupts common spread spectrum (SS) watermarking techniques that rely on the cross-correlation statistics of long pseudorandom chip sequences to withstand noise and host interference [10]. Robust resynchronization for SS watermarks is alternatively accomplished in prior work by parameterizing detection with variables for time shift and frequency scaling, and then searching over the parameter space [3], [11]–[15]. However, for meaningful robustness the parameter space can be quite large, and searching can be resource intensive. Alternative resynchronization strategies for blind detection have also been proposed for audio [16], [17] and speech [18] by using signal features for detection and embedding, and for images by using either optical flow to register to an embedded pseudo-random synchronization pattern [19] or via elastic graph matching [20].

The proposed audio watermarking scheme relies on two techniques to increase the desynchronization robustness of a SS watermark without requiring a computationally expensive exhaustive search: magnitude-only frequency domain embedding, and a block based, dynamic time warping (DTW) blind detection technique. Frequency domain embedding is already commonly used in watermarking to apply perceptual shaping during embedding [1], [21], [22]. The proposed watermark demonstrates two additional benefits of magnitude only embedding: it provides robustness to small time/frequency scale modifications and reduces the resolution of the cross-correlation statistic during detection. Lower resolution means fewer samples, and is critical for computationally efficient detection and making the proposed DTW technique feasible. The proposed DTW technique detects the watermark in each block of audio by finding an alignment between the watermark and a sequence of blocks of audio. The sequence includes all past blocks (forward reinforcement) and a small fixed number of future blocks (backward reinforcement). Prior audio [23] or video [24] watermarks have used DTW for *non-blind* detection by aligning to the available original media to undo temporal modifications prior to watermark detection. The proposed watermark, on the other hand, demonstrates desynchronization robustness for efficient *blind* DTW audio watermark resynchronization using block-wise cross-correlation detection statistics for the novel magnitude-only embedding. Robustness and computational efficiency of the proposed scheme are experimentally demonstrated by comparing it against a baseline watermark based on [3]. The baseline watermark uses higher resolution (more SS chips per block), phase-only

A. Nadeau is with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, USA (e-mail: andrew.nadeau@rochester.edu).

G. Sharma is with the Department of Electrical and Computer Engineering, Department of Computer Science, and Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14627, USA (e-mail: gaurav.sharma@rochester.edu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2661724, IEEE Transactions on Information Forensics and Security

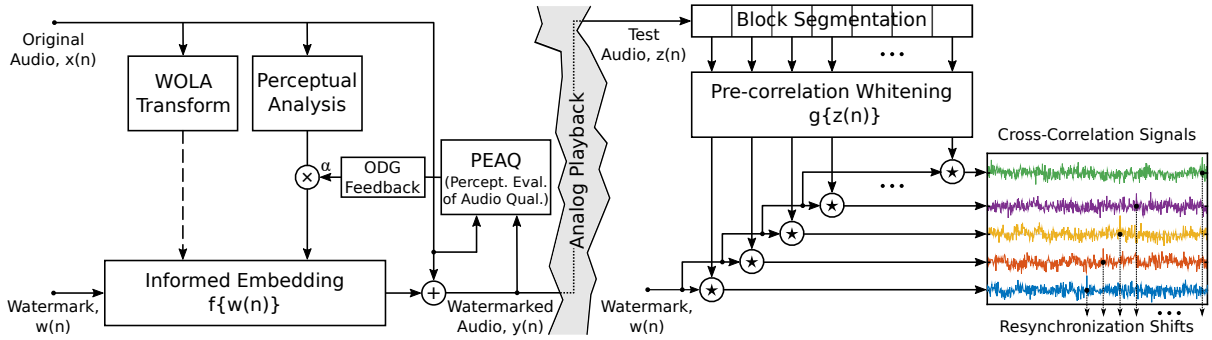


Fig. 1. The spread spectrum (SS) framework, used for both the baseline and proposed watermarks. The function f shapes the watermark for informed embedding (dashed connection shows the known host interference), and the function g whitens and attenuates the host interference prior to the calculation of the cross-correlation statistic for blind detection/resynchronization (no access to original audio signal). Note that all signals in the figure are depicted in the time domain, even though several of the operations are performed in the frequency domain, as is elaborated and clarified in the subsequent description.

embedding in the frequency domain. Higher resolution allows more aggressive pre-correlation whitening to improve robustness to noise (from cover interference and other sources), but has to rely on an exhaustive search over resampling rates to maintain robustness against desynchronization. Because of the higher resolution and exhaustive search, the baseline scheme requires a runtime over 300 times that of the proposed method (using MATLAB$^{\text{TM}}$ implementations for both). The higher resolution also makes the proposed DTW alignment rather computationally demanding for the baseline scheme and therefore unrealistic for practical implementation.

Preliminary research leading to the work proposed in this paper was presented in [25]. As compared to [25], the watermark proposed here is tuned specifically for robustness to analog playback and increased additive noise while remaining robust to time shifts. The watermark in [25] was primarily designed for desynchronization robustness, including the ability to survive insertions/deletions in the audio signal, pitch-invariant time scaling up to $\pm 15\%$, and MP3 compression.

The paper is organized as follows: Section II introduces the SS watermarking framework used by both the proposed watermark and the baseline scheme, and the commonalities between the two schemes; Section III gives the details of novel components of the proposed watermark; Section IV gives an overview of the baseline watermark used for comparison [3]; Section V describes how the watermarks are evaluated and compares the performance of the proposed watermark to the baseline; and in conclusion, Section VI summarizes the main contributions.

II. Common Spread Spectrum (SS) Framework

To demonstrate the benefits of the proposed watermark it is benchmarked against an alternative, *baseline* SS watermark based on [3], also designed for analog playback. This section introduces the common framework for both watermarks, which is shown in Fig. 1. The notational conventions $w(n)$, $x(n)$, $y(n)$, and $z(n)$ are used to denote the watermark, original audio, watermarked audio, and audio being tested at the detector, respectively; where n is the index for the time-domain samples. It is assumed that $w(n)$ and the audio signals are each N samples long. Longer audio signals are watermarked by breaking the audio into *segments* of length N, and repeatedly embedding the same $w(n)$ in each segment.

The detector processes *blocks* of the audio signal $z(n)$ (block segmentation in Fig. 1). For each block of B samples, the goal is to detect which portion of $w(n)$ is present, or, alternatively, determine that the block is unwatermarked. Detection relies on the cross-correlation signal calculated between the watermark and each block at all possible shifts. Ideally, the shift that maximizes the cross-correlation corresponds to the true portion of the watermark present in the block. The blocklength B is important because longer durations of an embedded SS watermark are less robust to desynchronization [25]. Long, high resolution sequences of SS chips can be significantly offset by warping such that the cross-correlation is no longer suitable for detection. Longer blocks also add latency to detection.

Watermarking literature provides numerous techniques to increase the power of an embedded SS watermark and decrease the interference in the cross-correlation signal: informed, transform domain embedding; perceptual shaping; and pre-correlation whitening [10], [22]. These common concepts are included in both the proposed and baseline watermarks and described in the next three subsections.

A. WOLA Transform

Both the baseline and the proposed watermarks use the same weighted overlap-add (WOLA) implementation of the short-time Fourier transform (STFT) to represent a signal's local frequency content [3], [26]. This shared time-frequency transform allows both techniques to share the same perceptual model for controlling the impact of each watermark on the audio quality.

The WOLA transform first divides an audio signal into overlapping *frames* (not to be confused with the blocks of audio used for detection, or segments used for embedding). Each frame is N_{DFT} samples long and overlaps $N_{\text{DFT}}/2$ samples with the prior frame. For each frame, the discrete Fourier transform (DFT) is used to calculate the WOLA coefficients as,

$$X(m,k) = \text{WOLA}\{x(n)\} \stackrel{\text{def}}{=} \sum_{n=0}^{N_{\text{DFT}}-1} w_{\text{sine}}(n)\, x\left(n + m\frac{N_{\text{DFT}}}{2}\right) e^{-\jmath 2\pi \frac{k}{N_{\text{DFT}}} n}, \quad (1)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2661724, IEEE Transactions on Information Forensics and Security

where: $n = 0, 1, \ldots, (N_{\text{DFT}} - 1)$ indexes the samples in each frame; $k = 0, 1, \ldots, (N_{\text{DFT}} - 1)$ indexes the N_{DFT} frequency bins; $m = 0, 1, \ldots, M$[1] indexes the $M = N \div (N_{\text{DFT}}/2)$ audio frames in the audio segment $x(n)$ and the one additional frame arising from the half frame zero-padding on either side; and $w_{\text{sine}}(n) = \sin(\pi n / N_{\text{DFT}})$ is the sine window weighting required by the weighted overlap (WO in WOLA). Due to the weighted overlap between adjacent frames, the reverse WOLA transform can reconstruct the signal as

$$\begin{aligned} x(n) &= \text{WOLA}^{-1}\{X(m,k)\} \\ &\stackrel{\text{def}}{=} \sum_{m=0}^{M} w_{\text{sine}}\left(n - m\frac{N_{\text{DFT}}}{2}\right) \\ &\quad \frac{1}{N_{\text{DFT}}}\left(\sum_{k=0}^{N_{\text{DFT}}-1} X(m,k)\, e^{\jmath 2\pi \frac{k}{N_{\text{DFT}}}(n - m\frac{N_{\text{DFT}}}{2})}\right). \end{aligned} \tag{2}$$

The weighted overlap also attenuates the audible artifacts that can occur due to discontinuities between reconstructed frames introduced by modifications to the WOLA coefficients. The WOLA-domain coefficients for the other time domain signals $y(n)$, $z(n)$, and $w(n)$ are also denoted using uppercase letters: $Y(m,k)$, $Z(m,k)$, and $W(m,k)$, respectively.

B. Perceptibility Model

Both the baseline and proposed watermark employ the same two stage process to control the perceptual impact of embedding. First, a perceptual model is applied to each segment $x(n)$ to determine masking thresholds $T(m,k)$ as detailed in Appendix A. The second stage uses a feedback process to meet a perceptual quality target specified in terms of the objective difference grade (ODG) between the watermarked signal $y(n)$ and the original signal $x(n)$. The ODG is computed using the Perceptual Evaluation of Audio Quality (PEAQ) algorithm [27]. Specifically, the magnitude frequency domain spectrum of the embedding distortion is constrained as $|Y(m,k) - X(m,k)| \leq \alpha T(m,k)$ where α is a nonnegative scaling factor, determined iteratively for each M frame segment of audio. While the embedding distortion is outside of the target ODG range, each iteration either: increases α to scale up the embedding strength if the ODG is above the target embedding range; or decreases α if the ODG is below the target embedding range; and then recalculates the watermarked audio $y(n)$ and new ODG.

C. Informed Embedding and Whitening

Informed embedding ($f\{w\}$ in Fig. 1) and whitening ($g\{z\}$ in Fig. 1) provide robustness to host interference and additive noise (as opposed to desynchronization). This subsection explains this common motivation for f and g, while the details specific to the proposed and baseline watermarks are given in Sections III and IV, respectively.

Because the host $X(m,k)$ is known at the embedder, $f\{w\}$ can shape the embedding distortion to both: maximize the possible embedding strength within the constraint $\alpha T(m,k)$ established in Section II-B; and optimize the signal space direction of the embedding distortion for robust detection [26], [28]. As shown in Fig. 2, the optimal direction for the baseline, phase-based watermark is very nearly orthogonal to the host phase $\angle X(m,k)$ and the optimal direction for proposed, magnitude-based watermark is in the same direction as $X(m,k)$. Host dependent, informed embedding has been shown to improve watermark robustness as compared to blind embedding [26], and benchmarking the proposed technique against a simple, blind, direct sequence SS watermark would be unfair.

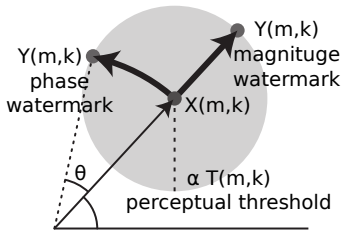


Fig. 2. Each bold arrow shows the embedding distortion added by $f\{w\}$ to one host WOLA coefficient, $X(m,k)$, by either the baseline (phase) or proposed (magnitude) watermarks. Unlike blind embedding, the magnitude and direction of the distortion depend on the perceptual threshold $\alpha T(m,k)$ and phase of $X(m,k)$, respectively (informed embedding).

The spectral characteristics of the host interference allow g to approximate Wiener filtering at the detector. Wiener filtering amplifies the portions of a signal's spectrum with high signal to noise ratio (SNR) and attenuates portions with low SNR. This equalization is especially beneficial for audio watermarks due to the perceptual thresholds during embedding. The thresholds $\alpha T(k)$ (dropping the frame index m for notational simplicity) typically allow a much greater SNR between the watermark (the signal) and host (the noise) at high frequencies than low frequencies where the audio power is higher. Assuming the signal and interference are uncorrelated (pseudorandom ± 1 watermark chips), the Weiner filter's frequency response $G(k)$ [29] is

$$G(k) = \frac{S_w(k)}{S_w(k) + S_x(k)}\,. \tag{3}$$

Although the power spectral densities (PSD) $S_w(k)$ and $S_x(k)$ for the embedded watermark and host are not available at the detector, the high dynamic range of typical audio signals makes even an approximate filtering solution highly effective. The detector assumes that the watermark has a PSD much lower than $S_x(k)$ and that embedding is approximately constant across frequencies, $S_w(k) \approx \epsilon \ll S_x(k)$. Additionally, the detector assumes that the majority of received signal is host interference, $S_x(k) \approx |Z(k)|^2$. Using these assumptions, $G(k)$ pre-whitens $z(n)$ before watermark detection,

$$G(k) \approx \frac{\epsilon}{\left|Z(k)\right|^2}\,, \tag{4}$$

where ϵ is an arbitrary constant whose value does not affect performance. Both the proposed and baseline use pre-whitening functions $g\{z\}$ motivated by (4). The specific implementation details for the proposed and baseline $g\{z\}$ are given in Sections III and IV, respectively.

[1]The $m = 0^{\text{th}}$ and M^{th} frames of $x(n)$ in (1) are padded with $N_{\text{DFT}}/2$ zeros for perfect reconstruction. For the modulo M detection operations, the 0^{th} and M^{th} frames of $W(m,k)$ are identical.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2661724, IEEE Transactions on Information Forensics and Security

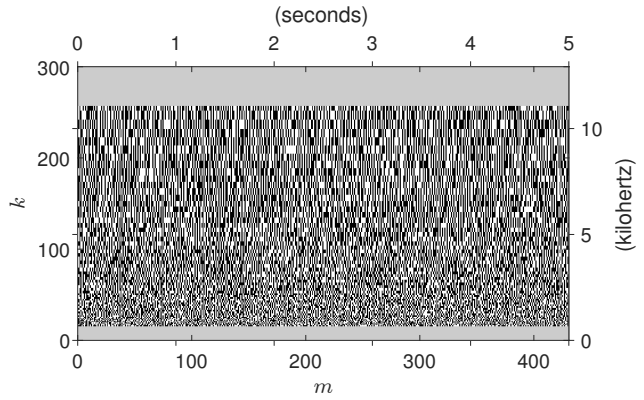


Fig. 3. One instance of the proposed watermark, $W(m,k)$ (generated for a 5 s long signal sampled at 44.1 kHz with the Nyquist frequency 22.05 kHz corresponding to the $k{=}N_{\text{DFT}}/2{=}512^{\text{th}}$ WOLA coefficient). The frequency regions of WOLA coefficients below 690 Hz ($k{=}15$) and above 10.8 kHz ($k{=}250$) and shown in gray are excluded from embedding, i.e. $W(m,k){=}0$. The remaining frequency region is divided to $N_{\text{band}}{=}59$ bands containing +1 (white) and -1 (black) chips.

III. Proposed Watermark

The proposed watermark improves synchronization efficiency by embedding and detecting the watermark in only the magnitude of the WOLA coefficients of the audio. Embedding and detection follow the framework shown in Fig. 1 with the exception that $w(n)$ and the functions f and g are defined in the WOLA-domain, and a dynamic time warping (DTW) procedure is applied to the cross-correlation signals at the detector.

The watermark signal $W(m,k)$ is a grid of pseudorandom ±1 chips in the WOLA-domain. Following the framework from Section II, $W(m,k)$ uses the same indices m and k and has the same length M as the coefficients $X(m,k)$ from each audio segment. The ±1 chips span a midrange of frequencies $k_{\text{low}} \leq k \leq k_{\text{high}}$, avoiding both low frequencies that are susceptible to perceptual distortion and high frequencies that are commonly lost to audio compression. The midrange of frequencies is grouped into N_{band} *bands*, such that each ±1 chip spans multiple frequency coefficients, but only one WOLA coefficient in the time dimension. Outside the range $k_{\text{low}} \leq k \leq k_{\text{high}}$, $W(m,k)$ is zero. The width ω_l of each chip band increases exponentially with frequency, similar to log scale embedding [12]. Figure 3 depicts the chip bands $W(m,k)$. The conjugate symmetric half of the WOLA spectrum ($k{>}N_{\text{DFT}}/2$) is not shown in Fig. 3 and is directly determined from the DFT symmetry relation [30] $W(m, N_{\text{DFT}}{-}k) = W(m,k)$.

Additionally, the proposed watermark uses chess watermarking [22]. Chess watermarking decreases the attenuation of the watermark signal during whitening by modifying the pseudorandom process used to generate $W(m,k)$ to favor an alternating "chessboard" of +1 and -1 chips. This chessboard differentiates the spectrum of $W(m,k)$ from the smoother spectrum of the host interference. The chips in each band are determined sequentially such that each chip is given the opposite sign as the previous chip in the band with 75% probability and the same sign with 25% probability. Chess watermarking reduces the low spectral and cepstral content of the watermark reference signal that whitening would attenuate.

A. Embedding

As mentioned in Section II, the watermark is embedded in the original audio signal segment by segment, embedding the same $W(m,k)$ in each segment. The signals $x(n)$ and $X(m,k)$ denote the current audio segment and its corresponding WOLA coefficients, respectively. Embedding modifies each coefficient's magnitude $|X(m,k)|$ depending on both the sign of the ±1 chip in $W(m,k)$ and the perceptual distortion limit $T(m,k)$ defined in Section II-B. Specifically, the magnitude $|Y(m,k)|$ of the watermarked coefficient is obtained as,[2]

$$|Y| = \begin{cases} |X| + \alpha T\, W & \text{if } |X| + \alpha T\, W > 0 \\ 0 & \text{otherwise} \end{cases}, \qquad (5)$$

where α is the scaling factor defined in Section II-B that controls the perceptual impact of the embedding distortion. The condition in (5) limits embedding to subtract no more than the entire magnitude of the original coefficient. The watermarked audio segment $y(n)$ is then reconstructed using the reverse WOLA transform, maintaining the original coefficients' phase $\angle X(m,k)$,

$$y(n) = \text{WOLA}^{-1}\left\{ |Y(m,k)|\, e^{\jmath \angle X(m,k)} \right\}. \qquad (6)$$

As shown in Fig. 2, the direction of the embedding distortion vector, $f\{w\} = Y(m,k) - X(m,k)$, between $y(n)$ and $x(n)$ in the WOLA transform domain depends on the phase of $X(m,k)$. Because of this phase dependence, the direction of the embedding distortion counteracts the host interference during detection, and is an instance of informed embedding [26].

B. Detection

As introduced in Section II, the detector first segments the audio $z(n)$ into N_{block} adjacent blocks $z_1(n), z_2(n), \ldots, z_{N_{\text{block}}}(n)$ of B samples each. The detector then decides whether the embedded watermark $W(m,k)$ is present in each block and, if yes, estimates the block's resynchronization shift. Specifically, an estimated shift $\hat{m}_i$ for the i^{th} block indicates that the portion of $W(m,k)$ corresponding to the indices $m = \hat{m}_i{+}1, \hat{m}_i{+}2, \ldots, \hat{m}_i{+}B_m$ is present in $z_i(n)$; where $B_m = B \div N_{\text{DFT}}/2$ is the number of WOLA analysis frames from each block of B time-domain samples. The length of each block $z_i(n)$ and the WOLA domain watermark $W(m,k)$ are depicted in Fig. 4.

Conventionally, the estimated shift between two signals is the shift that maximizes their cross-correlation. Figure 1 shows how the conventional SS watermarking framework would use the maximum cross-correlation peak from each block to estimate the resynchronization shifts. The proposed dynamic time warping (DTW) based detection works in lieu of maximizing the cross-correlation of each block individually

[2]For brevity in the equations, we drop the indices (m,k) for $W(m,k), X(m,k), Y(m,k)$ and $T(m,k)$ and denote these signals by W, X, Y and T.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2661724, IEEE Transactions on Information Forensics and Security

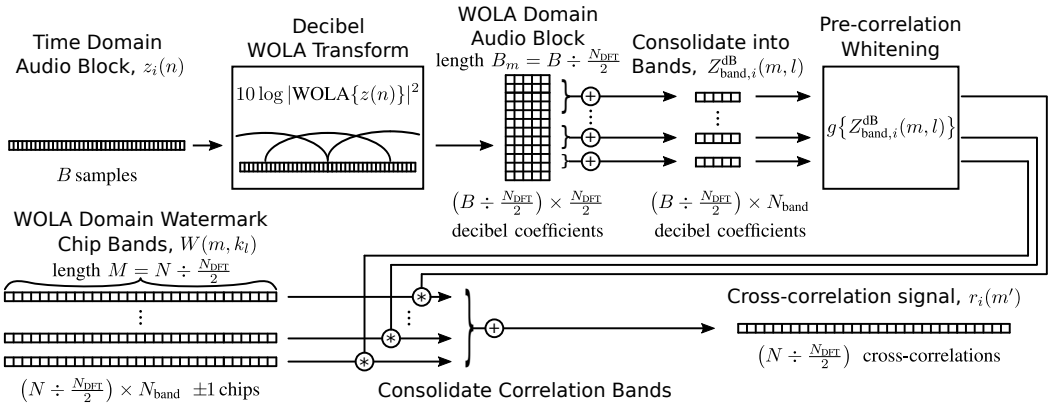


Fig. 4. Whitening and cross-correlation for the i^{th} block of audio are calculated directly in the WOLA-domain. The small grid units (not to scale) depict how the signals' length and computational complexity are reduced as compared to standard time-domain cross-correlation for the baseline watermark in Fig 6.

and provides a more robust estimate of each block's shift after desynchronization.

Before introducing the proposed DTW based detection, the next two subsections explain how the cross-correlation and pre-correlation whitening are calculated for each block.

1) Pre-correlation Whitening: The whitening function g is applied to each block $z_i(n)$ as motivated in Section II-C. Due to the proposed magnitude-only embedding technique and the arrangement of $W(m,k)$ in frequency bands, both whitening and cross-correlation are calculated in the frequency domain. As shown in Fig. 4, the detector first calculates the WOLA coefficients $Z_i(m,k)$ from $z_i(n)$ using (1), and then takes the average decibel power $Z^{\text{dB}}_{\text{band},i}(m,l)$ in each frequency band (same bands as the ± 1 chip bands in $W(m,k)$),

$$Z^{\text{dB}}_{\text{band},i}(m,l) = \frac{1}{\omega_l} \sum_{k=k_l}^{k_l+\omega_l-1} 10\log\left|Z_i(m,k)\right|^2, \qquad (7)$$

where m and k index the WOLA coefficients in time and frequency; l indexes the N_{band} frequency bands; k_l is the first WOLA frequency coefficient in the l^{th} band; and ω_l is the number of WOLA coefficients in the l^{th} band. Pre-correlation whitening $g\{z\}$ is applied to the block $Z^{\text{dB}}_{\text{band},i}(m,l)$ as

$$g\left\{Z^{\text{dB}}_{\text{band},i}(m,l)\right\} = Z^{\text{dB}}_{\text{band},i}(m,l) - \hat{S}^{\text{dB}}_{x,i}(m,l), \qquad (8)$$

where $\hat{S}^{\text{dB}}_{x,i}(m,l)$ is the estimated spectral power of the host interference. Due to the decibel scale, the subtraction of $\hat{S}^{\text{dB}}_{x,i}(m,l)$ in (8) is equivalent to applying a whitening filter to $Z_i(m,k)$ multiplicatively in the frequency domain analogous to $G(k)$ in (4). The estimate $\hat{S}^{\text{dB}}_{x,i}(m,l)$ is calculated individually for each chip band, and is the average decibel power in the bandwidths above and below each chip bandwidth,

$$\hat{S}^{\text{dB}}_{x,i}(m,l) = \frac{1}{2\omega_l} \sum_{k=k_l}^{k_l+\omega_l-1} \Big(10\log\left|Z_i(m,k+\omega_l)\right|^2 + 10\log\left|Z_i(m,k-\omega_l)\right|^2\Big), \qquad (9)$$

where i, m, k, l, k_l, ω_l are defined as in (7).

2) Cross-Correlation: Next, the detector calculates the cross-correlations between the watermark $W(m,k)$ and the block of whitened coefficients $g\{Z^{\text{dB}}_{\text{band},i}(m,l)\}$, at each possible shift. The cross-correlation signal $r_i(m')$ for the i^{th} block and shift m' is calculated following the steps depicted in Fig. 4 as,

$$r_i(m') = \sum_{l=1}^{N_{\text{band}}} \sum_{m=1}^{B_m} W\big((m+m') \bmod M, k_l\big)\, g\left\{Z^{\text{dB}}_{\text{band},i}(m,l)\right\}, \qquad (10)$$

where m, l, k_l and N_{band} are defined as in (7); $W(m,k_l)$ are the ± 1 watermark chips; $g\{Z^{\text{dB}}_{\text{band},i}(m,l)\}$ is the i^{th} block of whitened coefficients from (8); M is the temporal length of $W(m,k)$; B_m is the temporal length of $g\{Z^{\text{dB}}_{\text{band},i}(m,l)\}$; and the modulo M operation makes the cross-correlation cyclic. We use cyclic cross-correlation because the embedder repeats the same $W(m,k)$ in each segment of audio and the block $z_i(n)$ may span a transition between segments where $W(m,k)$ begins repeating again from $m=0$. Only the frequency bands $l = 1,\ldots,N_{\text{band}}$ are used calculate $r_i(m')$. Summing over the full range of watermarked frequencies k_{low} to k_{high} is not needed because each coefficient $Z^{\text{dB}}_{\text{band},i}(m,l)$ is already averaged over all the coefficients in the l^{th} chip band.

3) Dynamic Time Warping (DTW) Based Detection: As opposed to conventional SS detection (using the maximum of the cross-correlation signal $r_i(m')$ to estimate the shift $\hat{m}_i$), the proposed DTW based technique uses the cross-correlations from a sequence of blocks to estimate each $\hat{m}_i$. Sequential blocks of watermarked audio will contain sequential portions of $W(m,k)$, and produce a corresponding sequence of detection peaks in the cross-correlation signals. For example, the portion of $W(m,k)$ embedded in the i^{th} block in Fig. 5 is obscured by interference but the peaks in $r_{i-1}(m')$ and $r_{i+1}(m')$ due to preceding and following portions of $W(m,k)$ are still present.

The proposed DTW based technique estimates the i^{th} shift $\hat{m}_i$ by finding the best valid alignment path for the sequence of blocks $(z_1(n), z_2(n), \ldots, z_{i+N_{\text{buff}}}(n))$, which includes all previous blocks and a buffer of N_{buff} future blocks beyond the i^{th} block. As shown in Fig. 5, an alignment path consists of the shifts $(m_1, m_2, \ldots, m_{i+N_{\text{buff}}})$ for each block in the sequence. If the weighted sum of the cross-correlation scores for the best path exceeds the detection threshold τ, the i^{th} block is determined to be watermarked. The estimated shift $\hat{m}_i$ is then estimated using the alignment of the i^{th} block in the best path. Otherwise, if no valid path exceeds τ, the i^{th} block is declared

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2661724, IEEE Transactions on Information Forensics and Security

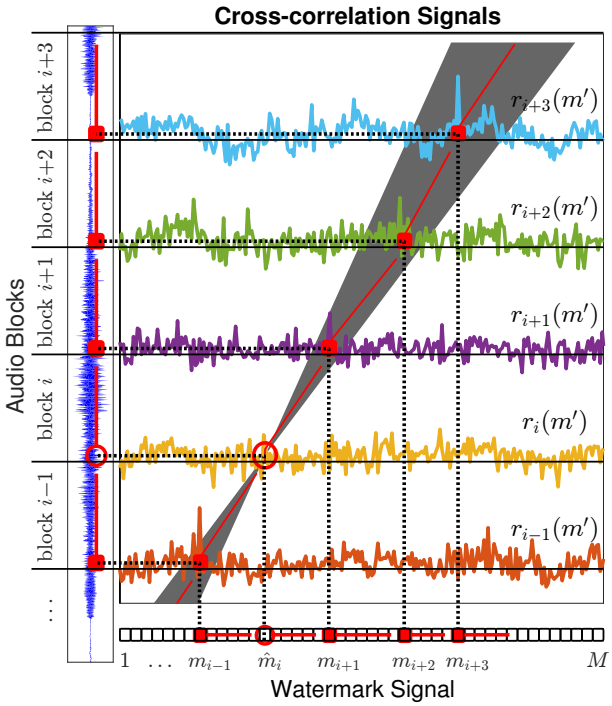


Fig. 5. Dynamic time warping (DTW) estimates the shift $\hat{m}_i$ (red circle) for the portion of $W(m,k)$ (bottom) present in the i^{th} block of $z(n)$ (left) by finding the valid alignment path $m_1, m_2, \ldots, m_{i+N_{\text{buff}}}$ (red sequence) that maximizes the sum of cross-correlations. The valid shifts for sequential blocks allow an exponential number of possible alignment paths (gray region).

unwatermarked. After estimating $\hat{m}_i$, the detector buffers an additional block of audio, recalculates a new best alignment path including the additional block, and uses the new path to estimate $\hat{m}_{i+1}$.

DTW uses three key components to find the best valid alignment path [31, Chap. 4]: 1) a warping constraint on the sequential shifts within a valid path; 2) a DTW matching score that can be summed over the shifts within a path; and 3) a recursive (dynamic programming) formulation for maximizing the total matching score for subsequences of blocks:

1) The warping constraints define a set $\mathcal{P}_i$ of all valid alignment paths for the sequence of blocks $(z_1(n), z_2(n), \ldots, z_{i+N_{\text{buff}}}(n))$ used to estimate the shift $\hat{m}_i$ for the i^{th} block. In the absence of warping, all alignment shifts within a valid path would be regularly spaced at intervals of the blocklength B_m. Due to the possibility of desynchronization, this exact spacing constraint is relaxed to the interval $[B_m^-, B_m^+]$ of allowable spacings between sequential shifts of adjacent blocks in a valid path in $\mathcal{P}_i$:

$$\mathcal{P}_i = \big\{(m_1, m_2, \ldots, m_{i+N_{\text{buff}}}) \mid B_m^- \le (m_j - m_{j-1}) \bmod M \le B_m^+, \quad \forall j \in \{2, 3, \ldots, i+N_{\text{buff}}\}\big\}, \tag{11}$$

where $(m_1, m_2, \ldots, m_{i+N_{\text{buff}}})$ is an alignment path; B_m^- and B_m^+ are the minimum and maximum valid spacings between consecutive shifts in the path; and the modulo M calculation again reflects the cyclic embedding of $W(m,k)$ as in (10).

2) The DTW matching score is defined as the sum of the cross-correlations at each shift in a path, weighted to emphasize the cross-correlations of the blocks closest to the current (i^{th}) block. Specifically, for each postulated shift m' for the i^{th} block, the DTW matching score $r_i^{\text{dtw}}(m')$ is the maximum weighted sum of cross-correlations over the subset of paths that include the shift $m_i = m'$, computed as

$$r_i^{\text{dtw}}(m') = \max_{\mathcal{P}_i | m_i = m'} \left(\sum_{j=1}^{i} \gamma_{\text{forw}}^{i-j} \, r_j(m_j) + \sum_{j=i+1}^{i+N_{\text{buff}}} \gamma_{\text{back}}^{j-i} \, r_j(m_j) \right), \tag{12}$$

where $\gamma_{\text{forw}}^{i-j}$ and $\gamma_{\text{back}}^{j-i}$ are exponentially decreasing weights that emphasize the cross-correlation scores of nearby past and future blocks, respectively.

3) A recursive DTW computation simplifies the evaluation of the score in (12). The objective function (12) is broken into two components that can be calculated recursively for

Algorithm 1: DTW based detection.

Input : Watermark $W(m,k)$; Sequential blocks $z_1(n), z_2(n), \ldots$ each containing B sample long, adjacent, non-overlapping portions of the test audio $z(n)$.

Output: Detected alignment shifts $\hat{m}_1, \hat{m}_2, \ldots$ for the portions of $W(m,k)$ embedded in each block. The i^{th} shift $\hat{m}_i$ is estimated once the $i^{\text{th}} + N_{\text{buff}}$ block is available (fixed latency).

```
//Initialize N_buff block lookahead
for j = 1, 2, ..., N_buff do
    //Cross-correlations for j^th block
    Calculate r_j(m') using (10);
end
//Initialize forward DTW scores
i ← 1; r_{i-1}^forw(m') ← 0, ∀m';
while Next i^th + N_buff block of audio exists do
    //Cross-correlations for i^th + N_buff block
    Calculate r_{i+N_buff}(m') using (10);
    //Update forward DTW scores
    Calculate r_i^forw(m') using (14);
    //Initialize backward DTW scores
    r_{i+N_buff+1}^back(m') ← 0, ∀m';
    for j = i + N_buff, i + N_buff − 1, ..., i + 1 do
        //Update backward DTW scores
        Calculate r_j^back(m') using (15);
    end
    //Detection scores for i^th block
    Calculate r_i^dtw(m') using (13);
    //Threshold i^th detection scores
    if max_m' r_i^dtw(m') > τ then
        m̂_i ← arg max_m' r_i^dtw(m');
    else
        Declare the i^th block unwatermarked;
    end
    i ← i + 1;
end
```

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2661724, IEEE Transactions on Information Forensics and Security

each possible shift m' for the i^{th} block,

$$r_i^{\text{dtw}}(m') = r_i^{\text{forw}}(m') + \max_{B_m^- \le b \le B_m^+} \gamma_{\text{back}}\, r_{i+1}^{\text{back}}(m'+b)\,, \quad (13)$$

where $r_i^{\text{forw}}(m')$ (forward score) is the summation in (12) for blocks $j = 1, 2, \ldots, i$; and $r_{i+1}^{\text{back}}(m'+b)$ (backward score) is the summation for blocks $j = i+1, i+2, \ldots, i+N_{\text{buff}}$. The forward score is calculated recursively using $r_{i-1}^{\text{forw}}(m')$ from the prior block,

$$r_i^{\text{forw}}(m') = r_i(m') + \max_{B_m^- \le b \le B_m^+} \gamma_{\text{forw}}\, r_{i-1}^{\text{forw}}(m'-b)\,, \quad (14)$$

The backward score $r_{i+1}^{\text{back}}(m')$ for the i^{th} block must be recalculated recursively in its entirety from the buffered blocks $j = i+N_{\text{buff}}, i+N_{\text{buff}}-1, \ldots, i+1$ but is limited to N_{buff} recursions,

$$r_j^{\text{back}}(m') = r_j(m') + \max_{B_m^- \le b \le B_m^+} \gamma_{\text{back}}\, r_{j+1}^{\text{back}}(m'+b)\,. \quad (15)$$

The backward recursions are initialized by setting $r_{i+N_{\text{buff}}+1}^{\text{back}}(m')$ to zero at all values of m'.

After calculating $r_i^{\text{dtw}}(m')$ using (13), the threshold τ is applied to access if the i^{th} block is watermarked. If the maximum of $r_i^{\text{dtw}}(m')$ exceeds τ, the maximizing shift is returned as the estimated resynchronization shift $\hat{m}_i$ for the i^{th} block. Algorithm 1 summarizes the overall detection procedure, including DTW.

IV. Baseline Watermark

The baseline scheme is designed to provide a benchmark for fair comparison, and draws upon concepts from [3].

A. Embedding

As shown in Fig. 2, the baseline watermark is embedded in the phase of the WOLA coefficients. The watermark signal consists of a target phase $\angle W(m,k)$ for each WOLA coefficient. Due to the overlap between WOLA frames and the possibility of destructive interference from embedding distortion in nearby WOLA coefficients, $\angle W(m,k)$ is generated by taking the WOLA transform of the random time domain signal $w(n)$ to ensure phase coherence [3]. The same interval of frequency coefficients $k = [k_{\text{low}}, k_{\text{high}}]$ is used for embedding as the proposed watermark. Outside of this range the coefficients' phase is not changed. The watermark is embedded by adjusting the phase $\angle Y(m,k)$ of the watermarked audio signal,

$$\angle Y = \begin{cases} \angle X + \theta_T & \text{if } \angle W - \angle X > \theta_T \\ \angle X - \theta_T & \text{if } \angle W - \angle X < -\theta_T \\ \angle W & \text{otherwise} \end{cases}, \quad (16)$$

where: $\angle X(m,k)$ and $\angle Y(m,k)$ are the phase of $X(m,k)$ and $Y(m,k)$; and $\theta_T(m,k)$ is the maximum embedding limit determined by the perceptual model. All phase calculations are done in the interval $[-\pi, \pi]$ radians. The limit $\theta_T(m,k)$ is related to threshold of allowable distortion $T(m,k)$ and scaling factor α from Section II-B by the trigonometric relation:

$$\theta_T(m,k) = 2 \arcsin \frac{2\,\alpha T(m,k)}{|X(m,k)|}\,. \quad (17)$$

The watermarked phase $\angle Y(m,k)$ is then used to reconstruct the time-domain watermarked audio $y(n)$, preserving the magnitudes $|X(m,k)|$ from the original audio,

$$y(n) = \text{WOLA}^{-1}\left\{ |X(m,k)|\; e^{\jmath \angle Y(m,k)} \right\}. \quad (18)$$

Similar to the proposed method, the baseline uses informed as opposed to blind embedding. As depicted in Fig. 2, the direction of the embedding distortion $f\{w\} = \text{WOLA}^{-1}\{Y(m,k) - X(m,k)\}$ is nearly orthogonal to $X(m,k)$. This differs from blind techniques such as direct sequence SS (DSSS) embedding that disregard the direction of the host [10]. DSSS embedding would add $w(n)$ to $x(n)$ such that the direction of the embedding distortion depends solely on $w(n)$.

B. Detection

An overview of the detector for the baseline watermark is shown in Fig. 6. Detection works on a block to block basis using the same intervals of B samples as the proposed detection technique. For each block $z_1(n), z_2(n), \ldots, z_{N_{\text{block}}}(n)$, detection estimates a shift $\hat{n}_i$ indicating the portion, if any, of the watermark signal present in the i^{th} block. As opposed to the proposed reduced resolution cross-correlation in Fig. 4, the baseline technique calculates the cross-correlation signal using the time-domain watermark signal $w(n)$ as shown in Fig 6. Detection relies on the maximum peak in the cross-correlation signal to estimate $\hat{n}_i$ for the i^{th} block of $z(n)$. However, to ensure the detection peak in the cross-correlation signal can withstand warping in time and frequency, cross-correlation relies on a brute force search. The search calculates the cross-correlation repeatedly, testing alternate warped versions of $w(n)$ to find resampling ratio that produces the maximum cross-correlation peak.

Before calculating the cross-correlation signal between i^{th} block $z_i(n)$ and $w(n)$ at each resampling ratio, the baseline applies a pre-correlation whitening function g to $z_i(n)$. As motivated in Section II-C, whitening smooths the audio spectrum. Specifically, whitening scales the magnitude of each WOLA-coefficient $Z_i(m,k)$ from $z_i(n)$ to one while leaving phase $\angle Z_i(m,k)$ unchanged. The whitened time-domain signal $g\{z_i(n)\}$ is reconstructed using the reverse WOLA transform,

$$g\left\{z_i(n)\right\} = \text{WOLA}^{-1}\left\{ e^{\jmath \angle Z_i(m,k)} \right\}, \quad (19)$$

where $\angle Z_i(m,k)$ is the phase of the WOLA coefficients from the i^{th} block $z_i(n)$. Before computing cross-correlations, the detector pre-computes the resampled versions of $w(n)$,

$$w_\lambda(n) = w(\lambda\, n), \quad (20)$$

where $w_\lambda(n)$ is the watermark, resampled at the ratio λ of the original sampling frequency. Interpolation is used when $\lambda\, n$ is not an integer. For each potential shift n' for the i^{th} block, the cross-correlation $r_i(n')$ is the maximum over the search space

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2661724, IEEE Transactions on Information Forensics and Security

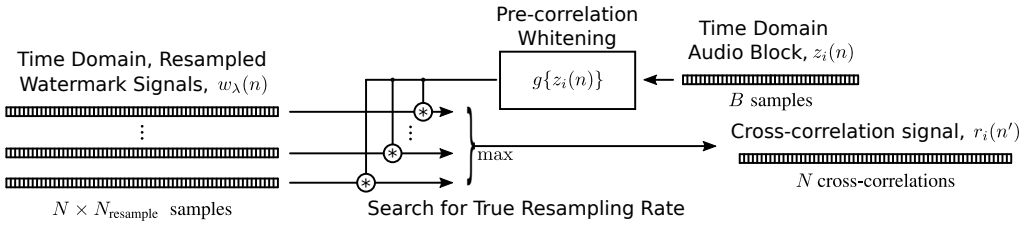


Fig. 6. The baseline detector calculates the cross-correlation between each block of audio and the time-domain watermark $w(n)$. Warping robustness requires calculating the cross-correlation using multiple versions $w_\lambda(n)$ of $w(n)$, resampled at various rates λ. Comparing this typical, time-domain cross-correlation calculation to Fig. 4 reveals the computational efficiency of the proposed WOLA frequency-domain cross-correlation.

Λ of resampling rates,

$$r_i(n') = \max_{\lambda \in \Lambda} \frac{1}{B} \sum_{n=1}^{B} w_\lambda\big((n + [n'/\lambda]) \bmod \lambda N\big)\; g\left\{z_i(n)\right\}, \tag{21}$$

where n, λ and the resampled watermark $w_\lambda(n)$ are defined as in (20); n' is the cross-correlation shift; $g\{z_i(n)\}$ is the i^{th} block of whitened audio from (19); λN is the length of $w_\lambda(n)$; B is the length of $g\{z_i(n)\}$; the square braces round n'/λ to the nearest integer; and the modulo λM operation makes cross-correlation cyclic. The watermark is detected wherever the value of $\max\{r_i(n')\}$ exceeds the detection threshold, τ. Detection is successful when $\hat{n}_i = \arg\max\{r_i(n')\}$ returns the shift for the true portion of $w(n)$ present in the i^{th} block.

V. Performance Evaluation

The following results demonstrate the desynchronization robustness and computational efficiency benefits of the proposed watermark over the baseline. Additionally, it is shown that the benefits of the proposed watermark did not degrade perceptual quality or decrease robustness to additive interference as compared to the baseline. Both schemes were also shown to be robust to analog playback.

A. Implementation Parameters

The audio test set was composed of 58 complete tracks extracted in uncompressed, 16-bit, 44.1 kHz, wav format from four compact disks (CD). The four CDs cover a wide range of genres: James Horner (movie soundtrack), Jewel (country/pop), Julianne Baird (classical), Moby (electronic/pop). Each track was converted from stereo to mono.

For each track, a different pseudorandom key was used to generate the watermark signals $w(n)$ and $W(m,k)$ for the baseline and proposed techniques, respectively. Each watermark was N=221184 samples long (corresponding to M=432 WOLA frames, an approximately 5 second duration), which also determined the length of the segments $x(n)$ and $y(n)$ used for embedding, as mentioned in Section II. The proposed and baseline watermarks both used the same N_{DFT}=1024 sample WOLA transform and perceptual model. Both techniques used up to four iterations to set the scaling factor α for the embedding distortion, targeting an ODG between -0.90 and -0.85 for each 5 second audio segment. If the ODG did not reach the target range by the fourth iteration, the audio segment from the last iteration was used. Considering the degradation inherent in analog playback this range of perceptual quality was reasonable and also found to be acceptable in limited blind subjective listening tests. After embedding, the ODG for each watermarked audio track was determined using the automated PEAQ [27] (using the entire track instead of the the individual 5 second segments used to determine α). The distribution of ODG scores for the baseline and proposed watermarks is shown in Fig. 7.

TABLE I
Chip bands for the proposed watermark. From the 512 WOLA frequency bins, bins 15 to 250 (690 Hz to 10.8 kHz) are consolidated into N_{BAND}=59 bands for embedding and detection.

Band, l	Starting Index, k_l	Bandwidth, ω_l
1	15 (690 Hz)	1 (43 Hz)
↓	...	1 (43 Hz)
10	24 (1.08 kHz)	1 (43 Hz)
11	25 (1.12 kHz)	2 (86 Hz)
↓	...	2 (86 Hz)
23	49 (2.15 kHz)	2 (86 Hz)
24	51 (7.80 kHz)	3 (129 Hz)
↓	...	3 (129 Hz)
31	72 (3.14 kHz)	3 (129 Hz)
32	75 (3.27 kHz)	4 (172 Hz)
↓	...	4 (172 Hz)
38	99 (4.31 kHz)	4 (172 Hz)
39	103 (4.48 kHz)	5 (215 Hz)
↓	...	5 (215 Hz)
43	123 (5.34 kHz)	5 (215 Hz)
44	128 (5.56 kHz)	6 (258 Hz)
↓	...	6 (258 Hz)
47	146 (6.33 kHz)	6 (258 Hz)
48	152 (6.59 kHz)	7 (301 Hz)
↓	...	7 (301 Hz)
51	173 (7.49 kHz)	7 (301 Hz)
52	180 (7.80 kHz)	8 (345 Hz)
53	188 (8.14 kHz)	8 (345 Hz)
54	196 (8.48 kHz)	8 (345 Hz)
55	204 (8.83 kHz)	9 (388 Hz)
56	213 (9.22 kHz)	9 (388 Hz)
57	222 (9.60 kHz)	9 (388 Hz)
58	231 (9.99 kHz)	10 (431 Hz)
59	241 (10.4 kHz)	10 (431 Hz)

The resolution of the frequency bands of the proposed watermark $W(m,k)$ is shown in Fig. 3, and the corresponding numeric intervals in terms of WOLA frequency index k are listed in Table I. The baseline watermark utilized the same total bandwidth (frequency coefficients k_{low}=15 to k_{high}=250), but modified each WOLA coefficient individually for a higher frequency resolution.

Both the proposed and baseline watermarks used 370 millisecond detection blocks (B=16384 samples, B_m=32 WOLA frames). The proposed forward-backward reinforcement was set to buffer the correlation calculations for N_{buff}=3 blocks beyond the current block. Including the current audio block

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2661724, IEEE Transactions on Information Forensics and Security

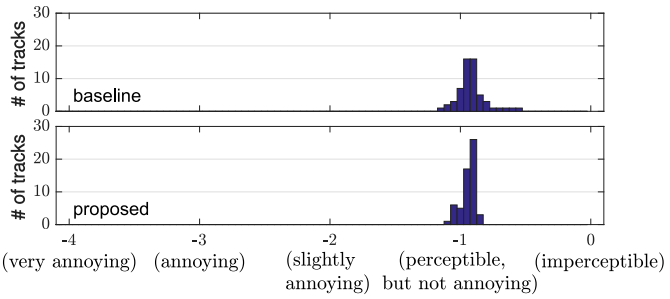


Fig. 7. Histogram of recorded Objective Difference Grades (ODG) for the watermarked audio signals using the baseline and proposed watermarks.

and buffering, the latency for watermark detection was 1.5 seconds. Comparatively, the baseline scheme processes only one block at a time, for a latency of 0.4 seconds. The additional latency of the proposed watermark is a significant drawback, but compares favorably to the computational complexity of the baseline. Considering the number of resampling ratios needed, the baseline was reported to take up to 10 seconds to regain synchronization running at 10% CPU load on an iPad2 (dual core ARM Cortex-A9@1 GHz) [3]. The attenuation factors γ_{forw} and γ_{back} used for forward and backward reinforcement were empirically set to 0.8 and 0.9, respectively. The warping limits for DTW were set to B_m^-=31 and B_m^+=33 WOLA frames. These limits corresponded to a maximum allowable relative warping of $\pm 1/32^{\text{nd}}$ to the temporal rate of the original audio. The baseline detection scheme used N_{resample}=45 resampling ratios, evenly distributed $\pm 1\%$ from the original sampling rate.

Both the proposed and baseline detection techniques were first run using the true pseudorandom key for each track to record the rate of missed detections (blocks that were either misaligned or the scores $r_i^{\text{dtw}}(m')$ or $r_i(n')$ did not exceed the detection threshold). Equal misalignment tolerance was given to the proposed and baseline watermarks accounting for the difference in resolution between m' and n'. Detection was run a second time for each track using an incorrect pseudorandom key to record the portion of blocks that triggered false alarms. Detection error tradeoff (DET) curves [32] are used to show the tradeoff between decreasing the threshold τ to avoid missing detections when the watermark was present and increasing τ to limit false alarms when the watermark was absent (or detection used a wrong key).

B. Playback Robustness

Each watermarked track was played using *Harman/Kardon HK206* speakers and recorded using an *Insignia NS-PAUM50* USB microphone. Synchronization beeps were appended at the start and end of each track and used as temporal reference to compute the ground-truth resynchronization shifts for the recorded waveform. The duration between beeps in the recorded signal also indicated the extent of the warping introduced by analog playback. Warping between $+0.55\%$ and $+0.6\%$ was observed over all tracks. This includes the recorded audio signal in Fig. 8 where a progressive offset due to warping can be seen between the signals $x(n)$ and $y(n)$.

TABLE II
COMPUTATIONAL EFFICIENCY FOR THE BASELINE AND PROPOSED DETECTION TECHNIQUES. RUNTIME RATIO (.1:1 IS 10X FASTER THAN REAL TIME) AND OVERHEAD ARE THE SLOPE AND INTERCEPT OF A LINEAR LEAST SQUARES FIT FOR THE DETECTION RUNTIME, RESPECTIVELY. THE PROPOSED RUNTIME INCLUDES DTW.

	Correlation Complexity	Runtime Ratio	Overhead
Baseline	$O(N_{\text{resample}} BN)$	3.1:1	990 ms
Proposed	$O\left(N_{\text{band}} \frac{B}{N_{\text{DFT}}/2} \frac{N}{N_{\text{DFT}}/2}\right)$	0.012:1	23 ms

The DET curves in Fig. 9 (a) depict the performance of the baseline and proposed watermarks after analog playback. The combination of warping and high additive interference was especially catastrophic for watermark detection in quiet sections of audio. In many portions of the recorded signals even the original audio signal is inaudible above the noise introduced by analog playback. Consequently, these quiet portions caused bursts of missed detection errors for both the proposed and baseline schemes. The proposed watermark correctly estimated the resynchronization shift for 70% of blocks across all 58 tracks, while the baseline watermark was correctly resynchronized in only 61% of blocks. These rates are reasonable considering the difficulty of the recorded signals and that neither technique includes any error correction coding or heuristic techniques to ensure reliability (e.g. waiting for a high power segment of audio to lock onto an alignment).

C. Desynchronization Robustness

Watermark performance was tested using four types of desynchronization: straightforward warping via resampling (modifies the pitch and duration), jitter (sinusoidal frequency modulation at an amplitude of 1 sample), time warping (pitch-invariant), and frequency scaling (time-invariant). Resampling relied on a third order anti-aliasing filter and jitter was simulated using linear interpolation to shift the temporal location of each sample. Time warping and frequency scaling used a 512 channel spectrogram to modify either the pitch or duration individually [33]. The DET curves in Fig. 9 (b), (c), (d), and (e) give the blockwise resynchronization error rates for the four operations across all 58 tracks. The limited search space of the baseline watermark allowed it to survive resampling and frequency scaling up to $\pm 1\%$, while the proposed watermark survived $\pm 2\%$ or more. Both watermarks were more robust to jitter and time warping which did not modify the audio's frequency scale as much as resampling or frequency scaling.

D. Detection Efficiency

Computational efficiency was evaluated by timing the duration of the detection process for each watermarked audio track. Both the proposed and baseline detectors were implemented in *Matlab* and optimized for speed using vectorized functions. The cross-correlation calculation had the biggest impact on runtime. The proposed detector ran the fastest while using `conv2(`$W(m,k_l)$`, rot90(`$g\left\{Z^{\text{dB}}_{\text{band},i}(m,l)\right\}$`,2), 'valid')` to calculate (10), while the baseline was fastest using `fftfilt(flipud(`$g\{z_i(n)\}$`)),` $w_\lambda(n)$`)` to calculate the cross-correlations $r_i(n')$ in equation (21). Detection runtime was recorded on an *Acer Aspire E1-572* laptop with a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2661724, IEEE Transactions on Information Forensics and Security

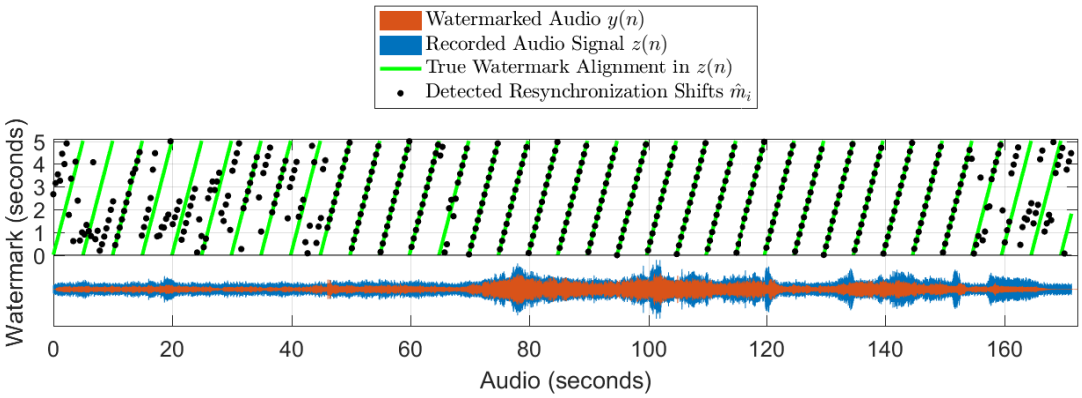
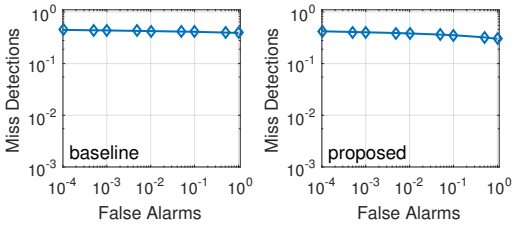
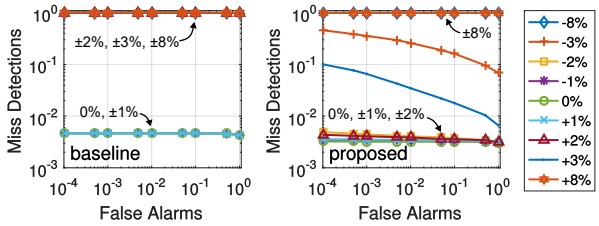


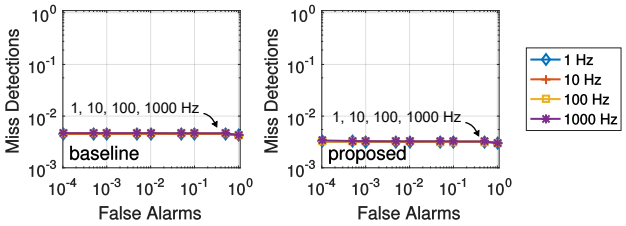
Fig. 8. The proposed watermark is used to detect the resynchronization shift of each B sample block of $z(n)$ after analog playback. The true alignment shows the repetition of the same 5 s long embedded watermark. Notice the progressively increasing offset between $y(n)$ and $z(n)$ due to the loudspeaker/microphone sampling rates. The detection threshold is set to $-\infty$, showing the resynchronization errors in the quiet portions.



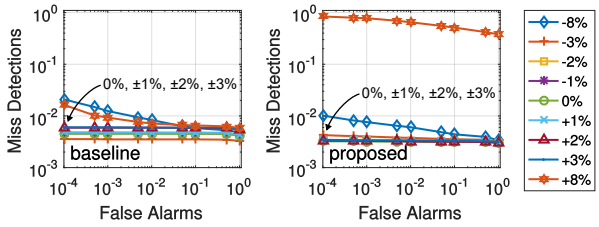
(a) Analog Playback



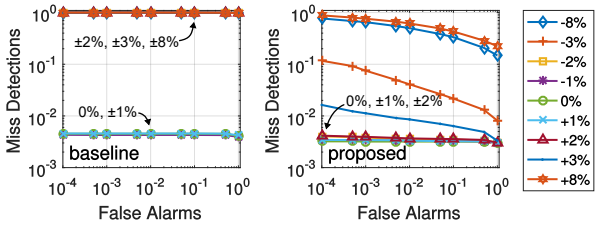
(b) Resampling (Pitch-variant and Time-variant)



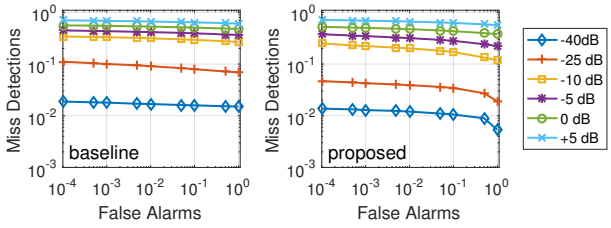
(c) Jitter



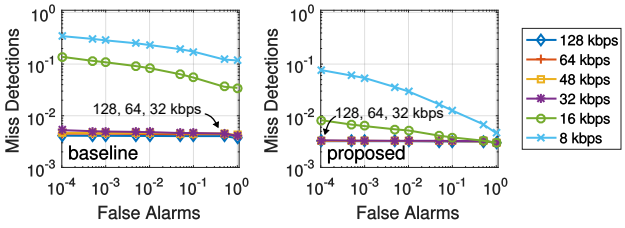
(d) Time Warping (Pitch-invariant)



(e) Frequency Scaling (Time-invariant)



(f) Additive White Gaussian Noise (AWGN)



(g) Advanced Audio Coding (ACC)

Fig. 9. DET (detection error tradeoff) curves. Starting from the right, the dots on each curve depict increasing detection thresholds. Even when the false alarm rate is 1 (rightmost dot), the miss rate is not 0; this miss rate represents blocks that are misaligned regardless of the threshold because the maximum alignment score does not occur at the true alignment. Figures are best viewed in color in the electronic version of this paper under high zoom.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2661724, IEEE Transactions on Information Forensics and Security

Intel Core i5-4200U CPU set to run at its maximum, 2.3 GHz frequency. Results are given in Table II. Using linear least squares, a line was fit to map the playback duration of each audio track to the detection runtime. The slope of the line gives the *runtime ratio* relative to 1 for running detection for each track in real time at a fixed lag, and the y-intercept gives the *overhead*. Including the DTW operation, detecting the proposed watermark was about 300X faster than the baseline, and was able to run comfortably in real time.

E. Non-desynchronization Robustness

The watermarks were also evaluated to show that robustness to desynchronization did not come at the cost of robustness to additive noise, interference or other modifications that do not introduce time warping or pitch shifting. Specifically, detection robustness was tested after additive white Gaussian noise (AWGN) and Advanced Audio Coding (AAC) [34]. AAC was performed using *FFmpeg* [35] at bitrates from 8 to 128 kbps. The DET curves in Fig. 9 (f) and (g) give the detection results for both watermarking techniques. AWGN was particularly difficult for the watermark to withstand, because it introduced noise at high frequencies where the host interference was low and the thresholds $T(m,k)$ of allowable watermark power were high. Additionally, the difficulty of distinguishing random noise from the watermark signal, prevented pre-correlation whitening from attenuating AWGN as effectively as the host interference. The proposed watermark outperformed the baseline after both AWGN and AAC.

F. Analysis of Individual Proposed Components

The robustness of the proposed watermark relies on two innovations: magnitude-only embedding and DTW based detection. To highlight the role of each component, Fig. 10 shows the cross-correlation detection statistics for the proposed and baseline schemes and for modified versions of these obtained by dropping the resampling search for the baseline and the DTW for the proposed scheme. The key difference between the proposed magnitude-only embedding and the baseline is the resolution of the SS chips in the watermark signal as shown in Figs. 4 and 6. Several comparisons among the subfigures in corresponding rows and columns of Fig. 10 are instructive. Subfigures (a) and (b) demonstrate that the lower resolution of the proposed watermark sacrifices robustness to additive noise as compared to the baseline, but improves robustness to resampling (resampling distorts both the pitch and time scales). Subfigures (a) and (c) show that the resampling search for the baseline enables it to survive survive desynchronization. Subfigures (b) and (d) shows that the DTW in the proposed watermark complements the magnitude-only embedding and restores robustness to additive noise. Because DTW sums cross-correlations over multiple blocks it is analogous to using a longer blocklength to increase the SS spreading gain, but without incurring the vulnerabilities of longer SS sequences to desynchronization.

VI. Conclusion and Discussion

This paper proposes a novel audio watermarking technique specifically designed to provide computationally efficient resynchronization after analog playback. The proposed watermark also demonstrates robustness to other non-desynchronizing and desynchronizing signal modifications. However, the primary contribution of this work over prior robust watermarks is the computational efficiency of watermark detection, and the combination of robustness to both analog playback and other signal modifications, e.g. AAC compression and non-pitch-invariant warping.

The efficiency benefits of the proposed watermark are a consequence of embedding the watermark in only the magnitude of the host audio spectral content as opposed to a classic additive SS watermark which would be embedded in both the phase and magnitude information. This embedding strategy allows the detector to use a reduced resolution cross-correlation calculation and eliminates the need for exhaustive search to survive desynchronization. Robustness also benefits from a novel forward-backward DTW detection strategy introduced in the proposed scheme.

The proposed watermark does not included data embedding. However, due to computationally efficient detection, a small payload could be transmitted by allowing one of a few different pseudo-random sequences to be embedded and running multiple detectors simultaneously to determine which sequence was embedded. Knowledge of the intended application would allow performance to be further enhanced in practical deployments of the proposed techniques. For example, reliability can be significantly increased by waiting for a high power segment of audio to lock-on to an alignment or by applying error correcting coding in conjunction with the proposed alignment techniques.

Acknowledgment

The authors thank the Center for Integrated Research Computing, University of Rochester, for providing the computational resources used to obtain the results in this paper.

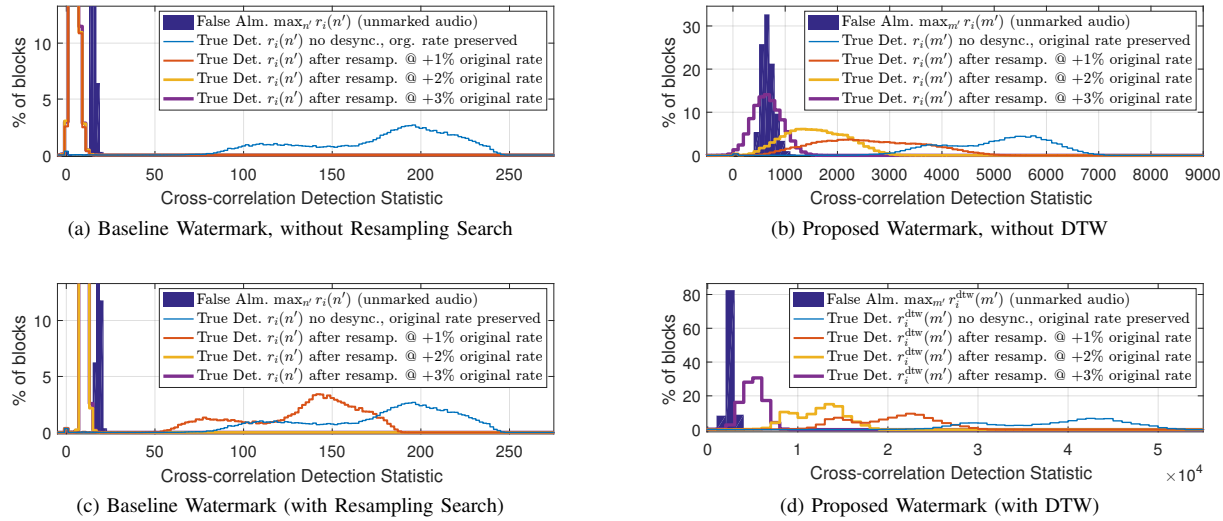
Appendix A
Masking Tresholds

The masking threshold $T(m,k)$ is calculated by finding the sets $\mathcal{K}_{\text{t}}(m)$ and $\mathcal{K}_{\text{n}}(m)$ of frequencies in the m^{th} frame that are the significant tonal and noise like maskers [36],

$$T(m,k) = T_{\text{quiet}}(k) + \sum_{\kappa\in\mathcal{K}_{\text{t}}(m)} h_{\text{tonal}}\big(k-\kappa, P_{\text{tonal}}(m,\kappa)\big) + \sum_{\kappa\in\mathcal{K}_{\text{n}}(m)} h_{\text{noise}}\big(k-\kappa, P_{\text{noise}}(m,\kappa)\big), \tag{22}$$

where $T_{\text{quiet}}(k)$ is the threshold of hearing in silence and the summations over $\mathcal{K}_{\text{t}}(m)$ and $\mathcal{K}_{\text{n}}(m)$ represent the masking contributions of the tonal and noise like maskers; $P_{\text{tonal}}(m,\kappa)$ and $P_{\text{noise}}(m,\kappa)$ are the power of a masker located at the frequency κ in the m^{th} frame; and the profiles h_{tonal} and h_{noise} represent the masking contribution to $T(m,k)$ from each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2661724, IEEE Transactions on Information Forensics and Security



(a) Baseline Watermark, without Resampling Search

(b) Proposed Watermark, without DTW

(c) Baseline Watermark (with Resampling Search)

(d) Proposed Watermark (with DTW)

Fig. 10. Histograms of the cross-correlation detection statistics for: (a) a modified version of the baseline watermark with the resampling search disabled, (b) a modified version of the proposed watermark without DTW, (c) the baseline watermark (with resampling search), and (d) the proposed watermark (with DTW). The phased-based cross-correlation for the baseline watermark uses a greater number of higher resolution SS chips per block, which improves robustness to additive noise (greater separation between the *True Detections* (Det.) and *False Alarms* (Alm.) distributions) but sacrifices robustness to warping when the resampling search is omitted (Subfigures (a) and (c)). The magnitude-only SS chips in the proposed watermark are lower resolution and more robust to warping even before applying DTW (Subfigure (b)). However, the lower number of chips deceases the spreading gain and relies on DTW to restore robustness to additive noise (Subfigure (d)). Note that the *False Alarms* do not have a zero mean due to the maximization over the shifts n' and m' in (21) and (12). Histograms in Subfigures (c) and (d) correspond to the DET curves in Fig. 9 (b).

individual tonal or noise like masker. The profile for each masker varies over the distance $k - \kappa$ and depends on both the power and frequency of the masker. Detailed definitions of P_{tonal}, P_{noise}, $\mathcal{K}_{\text{t}}(m)$, $\mathcal{K}_{\text{n}}(m)$, h_{tonal} and h_{noise} are available in [36], [37].

Calculating $T(m,k)$ also uses a simple acoustic transient detection technique to avoid audible *pre-echo*. Possible pre-echo due to transients is detected when the energy in a frame exceeds the energy in the previous frame by 5 dB or more. In these frames, $T(m,k)$ is set to the same values as the quieter, preceding frame instead of using (22).

REFERENCES

[1] G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. Thing, "Twenty years of digital audio watermarking-a comprehensive review," *Signal Process.*, vol. 128, pp. 222 – 242, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168416300263

[2] P. Cesar, D. C. A. Bulterman, and J. Jansen, " Leveraging user impact: an architecture for secondary screens usage in interactive television," *Multimedia Syst.*, vol. 15, no. 3, pp. 127–142, Jun. 2009. [Online]. Available: http://dx.doi.org/10.1007/s00530-009-0159-z

[3] M. Arnold, X.-M. Chen, P. Baum, U. Gries, and G. Doërr, "A phase-based audio watermarking system robust to acoustic path propagation," *IEEE Trans. Info. Forensics and Security*, vol. 9, no. 3, pp. 411–425, Mar. 2014.

[4] M. Arnold, P. G. Baum, and W. Voeßing, " A phase modulation audio watermarking technique," in *Info. Hiding*, ser. Lecture Notes in Computer Science, S. Katzenbeisser and A.-R. Sadeghi, Eds. Springer Berlin Heidelberg, 2009, vol. 5806, pp. 102–116. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04431-1_8

[5] G. Calixto, A. Angeluci, C. Kurashima, R. de Deus Lopes, and M. Zuffo, "Effectiveness analysis of audio watermark tags for IPTV second screen applications and synchronization," in *Intl. Telecomm. Symposium (ITS)*, Aug. 2014, pp. 1–5.

[6] C. Howson, E. Gautier, P. Gilberton, A. Laurent, and Y. Legallais, "Second screen TV synchronization," in *Proc. IEEE Int. Conf. on Consumer Electron.*, Sept. 2011, pp. 361–365.

[7] M. Arnold, P. G. Baum, M. Alonso, U. Gries, and G. Doërr, "Simulating large scale acoustic path benchmarking," in *Proc. SPIE: Media Watermarking, Security and Forensics*, N. D. Memon, A. M. Alattar, and E. J. D. III, Eds., vol. 8303, Jan. 2012, pp. 83 030T–1–83 030T–12.

[8] E. Wolff, C. Bams, and C. Siclet, "Toward robustness of audio watermarking systems to acoustic channels," in *18th Euro. Sig. Proc. Conf.*, Aug. 2010, pp. 1257–1261.

[9] M. Steinebach, F. A. Petitcolas, F. Rayna, J. Dittmann, C. Fontaine, C. Seibel, N. Fates, and L. C. Ferri, "Stirmark benchmark: audio watermarking attacks," in *Proc. Intl. Conf. on Info. Tech.: Coding and Computing*. IEEE, 2001, pp. 49–54.

[10] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital watermarking and steganography*. Morgan Kaufmann, 2007.

[11] Y. Xiang, I. Natgunanathan, S. Guo, W. Zhou, and S. Nahavandi, "Patchwork-based audio watermarking method robust to de-synchronization attacks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 9, pp. 1413–1423, 2014.

[12] X. Kang, R. Yang, and J. Huang, "Geometric invariant audio watermarking based on an LCM feature," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 181–190, 2011.

[13] X.-Y. Wang and H. Zhao, "A novel synchronization invariant audio watermarking scheme based on DWT and DCT," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4835–4840, 2006.

[14] S. Wu, J. Huang, D. Huang, and Y. Q. Shi, "Efficiently self-synchronized audio watermarking for assured audio data transmission," *IEEE Trans. Broadcasting*, vol. 51, no. 1, pp. 69–76, 2005.

[15] M. Barni, "Effectiveness of exhaustive search and template matching against watermark desynchronization," *IEEE Signal Process. Lett.*, vol. 12, no. 2, pp. 158–161, Feb. 2005.

[16] C.-M. Pun and X.-C. Yuan, "Robust segments detector for de-synchronization resilient audio watermarking," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 11, pp. 2412–2424, 2013.

[17] H.-Y. Yang, D.-W. Bao, X.-Y. Wang, and P.-P. Niu, " A robust content based audio watermarking using UDWT and invariant histogram," *Multimedia Tools and Apps.*, vol. 57, no. 3, pp. 453–476, 2012. [Online]. Available: http://dx.doi.org/10.1007/s11042-010-0644-6

[18] D. J. Coumou and G. Sharma, "Insertion, deletion codes with feature-based embedding: A new paradigm for watermark synchronization with applications to speech watermarking," *IEEE Trans. Info. Forensics and Security*, vol. 3, no. 2, pp. 153–165, Jun. 2008.

[19] R. Caldelli, A. De Rosa, R. Becarelli, and M. Barni, "Coping with local geometric attacks by means of optic-flow-based resynchronization

for robust watermarking," pp. 164–174, 2005. [Online]. Available: http://dx.doi.org/10.1117/12.589106

[20] G. Doërr, C. Rey, and J.-L. Dugelay, "Watermark resynchronization based on elastic graph matching," in *Proc. of the Intl. Conf. on Sciences of Electron., Tech. of Info. and Telecomm.* IEEE, Mar. 2005.

[21] Y. Xiang, I. Natgunanathan, Y. Rong, and S. Guo, "Spread spectrum-based high embedding capacity watermarking method for audio signals," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 12, pp. 2228–2237, Dec. 2015.

[22] D. Kirovski and H. S. Malvar, "Spread-spectrum watermarking of audio signals," *IEEE Trans. Signal Process.*, vol. 51, no. 4, pp. 1020–1033, Apr. 2003.

[23] C. Xu, Y. Lim, and D. D. Feng, "Recovering modified watermarked audio based on dynamic time-warping technique," in *Digital Image Computing Techniques and Applications*, 2002, pp. 241–245.

[24] B. Chupeau, L. Oisel, and P. Jouet, "Temporal video registration for watermark detection," in *Proc. IEEE Intl. Conf. Acoustics Speech and Sig. Proc.*, vol. 2, Mar. 2005, pp. 157–160.

[25] A. Nadeau and G. Sharma, "Self-synchronization for spread spectrum audio watermarks after time scale modification," in *Proc. SPIE: Media Watermarking, Security and Forensics 2014*, A. M. Alattar, N. D. Memon, and C. D. Heitzenrater, Eds., vol. 9028, Feb. 2014, pp. 902 813–1–902 813–9.

[26] H. S. Malvar and D. A. Florêncio, "Improved spread spectrum: a new modulation technique for robust watermarking," *IEEE Trans. Signal Process.*, vol. 51, no. 4, pp. 898–905, Apr. 2003.

[27] P. Kabal, "An examination and interpretation of ITU-R BS. 1387: Perceptual evaluation of audio quality," *Dept. Electr. & Compu. Eng., McGill Univ., Montreal, QC, Canada*, Dec. 2003.

[28] T. Furon and P. Bas, "Broken arrows," *EURASIP J. Inf. Secur.*, vol. 2008, pp. 3:1–3:13, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1155/2008/597040

[29] J. A. Gubner, *Probability and random processes for electrical and computer engineers*. Cambridge University Press, 2006.

[30] A. V. Oppenheim and R. W. Schafer, "Discrete-time signal processing," 2009.

[31] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, ser. Prentice Hall Signal Processing Series. Englewood Cliffs, New Jersey: Prentice Hall PTR, 1993.

[32] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," DTIC Document, Tech. Rep., 1997.

[33] L. R. Rabiner, *Multirate digital signal processing*. Prentice Hall PTR, 1996.

[34] ISO/IEC, "Information technology – Coding of audio-visual objects – Part 3: Audio," no. ISO/IEC 14496-3:2009, 2009.

[35] "FFmpeg," https://www.ffmpeg.org, 2000–2016.

[36] ISO/IEC, "Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio," no. ISO/IEC 11172-3:1993, 1993.

[37] F. A. P. Petitcolas, "MPEG psychoacoustic model 1 for MATLAB," Aug. 2003. [Online]. Available: www.cl.cam.ac.uk/ fapp2/software/mpeg/

Andrew Nadeau is pursuing his PhD in Electrical and Computer Engineering at the University of Rochester, where his research interests include audio signal processing, watermarking, and signal processing for energy aware systems. He graduated from Clarkson University in 2010 with B.S. degrees in Electrical Engineering and Physics where his honors thesis was on biometric identification using iris images. Previously Andrew has worked in the fields of steganography with the AFRL (Air Force Research Laboratory), radar systems engineering with SRC Inc., and controls and software validation with GE Transportation.

Gaurav Sharma (S'88–M'96–SM'00–F'13) is a professor at the University of Rochester in the Department of Electrical and Computer Engineering, in the Department of Computer Science and in the Department of Biostatistics and Computational Biology. From 2008–2010, he served as the Director for the Center for Emerging and Innovative Sciences (CEIS), a New York state funded center for promoting joint university-industry research and technology development, which is housed at the University of Rochester. He received the BE degree in Electronics and Communication Engineering from Indian Institute of Technology Roorkee (formerly Univ. of Roorkee), India in 1990; the ME degree in Electrical Communication Engineering from the Indian Institute of Science, Bangalore, India in 1992; and the MS degree in Applied Mathematics and PhD degree in Electrical and Computer Engineering from North Carolina State University, Raleigh in 1995 and 1996, respectively. From Aug. 1996 through Aug. 2003, he was with Xerox Research and Technology, in Webster, NY, initially as a Member of Research Staff and subsequently at the position of Principal Scientist.

Dr. Sharma's research interests include media security, bioinformatics, image processing and computer vision, color science and imaging, and distributed signal processing. He is the editor of the "Color Imaging Handbook", published by CRC press in 2003. He is a fellow of the IEEE, of SPIE, and of the Society of Imaging Science and Technology (IS&T) and a member of Sigma Xi. He has served as a Technical Program Chair for the 2016 and 2012 IEEE International Conferences on Image Processing (ICIP), as the Symposium Chair for the 2013 SPIE/IS&T Electronic Imaging symposium, as the 2010-2011 Chair IEEE Signal Processing Society's Image Video and Multi-dimensional Signal Processing (IVMSP) technical committee, the 2007 chair for the Rochester section of the IEEE and the 2003 chair for the Rochester chapter of the IEEE Signal Processing Society. From 2011 through 2015, he served as the Editor-in-Chief for the Journal of Electronic Imaging and in the past has served as an associate editor for the Journal of Electronic Imaging, IEEE Transactions on Image Processing, and IEEE Transactions on Information Forensics and Security.