Self-synchronization for Spread Spectrum Audio Watermarks after Time Scale Modification

Andrew Nadeau and Gaurav Sharma

Dept. of Electrical and Computer Engineering, Rochester NY, USA

December 20, 2013

ABSTRACT

De-synchronizing operations such as insertion, deletion, and warping pose significant challenges for watermarking. Because these operations are not typical for classical communications, watermarking techniques such as spread spectrum can perform poorly. Conversely, specialized synchronization solutions can be challenging to analyze/optimize. This paper addresses desynchronization for *blind* spread spectrum watermarks, detected without reference to any unmodified signal, using the robustness properties of short blocks. Synchronization relies on dynamic time warping to search over block alignments to find a sequence with maximum correlation to the watermark. This differs from synchronization schemes that must first locate invariant features of the original signal, or estimate and reverse desynchronization before detection. Without these extra synchronization steps, analysis for the proposed scheme builds on classical SS concepts and allows characterizes the relationship between the size of search space (number of detection alignment tests) and intrinsic robustness (continuous search space region covered by each individual detection test). The critical metrics that determine the search space, robustness, and performance are: time-frequency resolution of the watermarking transform, and blocklength resolution of the alignment. Simultaneous robustness to (a) MP3 compression, (b) insertion/deletion, and (c) time-scale modification is also demonstrated for a practical audio watermarking scheme developed in the proposed framework.

1. INTRODUCTION

Many audio watermarking applications require robust detection performance despite attacks. Among the attacks on watermarking systems, desynchronization attacks pose a particularly difficult challenge. Desynchronizing operations, such as time-scale modification (TSM) for audio, can be imperceptible to humans, but severely impact watermark detection performance. Desynchronization does not directly remove or jam a signal, but shifts and warps the watermarked content so that it is no longer recognizable to the detector. Spread spectrum (SS) watermarking schemes are specifically vulnerable to desynchronization because detecting low-power embedded watermark signals typically requires correlation with long SS sequences to overcome interference from the original host signal and other non-desynchronizing operations, such as MP3 compression.

Because vulnerability to desynchronization increases with watermark length, there has been only limited use of SS watermarks for applications where desynchronization is a concern. Typical solutions to desynchronization vulnerabilities include:[1] watermark repetition or structure, which can increase vulnerability to watermark estimation and removal;[2] exhaustive search,[3,4] which can be computationally challenging for a meaningful range of attacks; or semantic feature based synchronization techniques,[5,6] which can depend on specific characteristics of the original host signal. Recent work introduced an alternative synchronization technique for SS watermarks by segmenting the watermark into short blocks and using an efficient DTW search for sequences of detections.[7] The short blocks are robust to insertions and deletions (ID) which would break apart and desynchronize longer SS watermark sequences. Contrary to many exhaustive search and feature based synchronization schemes which aim to find and compensate any desynchronizing modifications, DTW is incorporated directly into the correlation estimation procedure used for SS watermark detection. The absence of a preprocessing step to extract special features, or estimate and reverse modifications makes more general analysis of the scheme possible and increases the ability to extend results to different host audio signals with varying characteristics and features. Although the DTW based detection in Ref. 7 successfully addresses ID desynchronization, the work does not address TSM.

On the theoretical side, it is difficult to characterize attacks that modify the time scale of the signal, or analyze corresponding watermarks' robustness. Most theoretical work focuses on performance bounds for detection tests

Media Watermarking, Security, and Forensics 2014, edited by Adnan M. Alattar, Nasir D. Memon, Chad D. Heitzenrater, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 9028, 90280E
 · doi: 10.1117/12.2039486

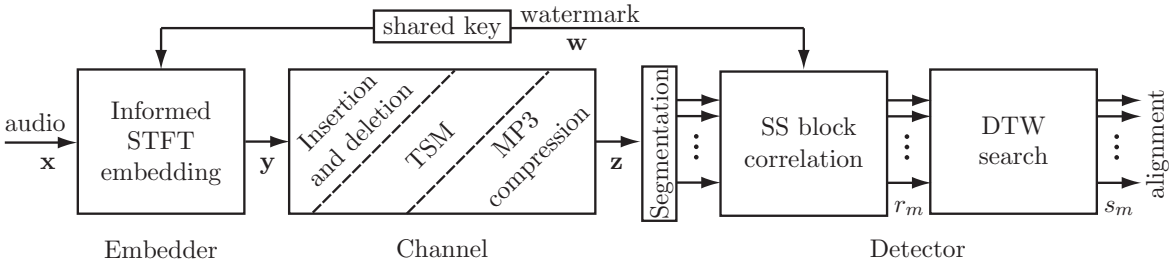


Figure 1. Watermarked audio y, is produced by embedding watermarks w, into original audio x. Modified audio z, is segmented into blocks before detection aligns each block to the portion of watermark signal embedded in it. As a blind detection scheme, the only information shared between the embedder and detector is the shared key that generates pseudo-random watermark chips.

using a single detection statistic. These bounds rely on mappings from the signal space to the scalar detection statistic and memoryless constraints on the distribution of attack distortion and host signals.[2] Such memoryless constraints encounter difficulty accounting for desynchronization because shifting the signal introduces introduce sequential dependencies. More relevant work for desynchronization robustness considers multiple detection tests over a finite search space of delays.[8,9] The proposed analysis builds on the concept of search space and its influence on detection performance in a more practical setting where desynchronization, especially TSM, is not limited to discrete intervals or a finite search space.

This paper presents: (a) an audio watermarking scheme that builds upon Ref. 7 to address TSM in addition to ID desynchronization; and (b) performance analysis for the interaction between intrinsic robustness and search space. The analysis shows that intrinsic robustness of individual detection tests determines how many detection tests are needed to cover an adequate search space for reliable detection after desynchronization. Intrinsic robustness is determined in part by the time-frequency resolution of the transform used for watermarking. Prior time-frequency transforms used for audio watermarking include the modulated complex lapped transform[3] (MCLT), wavelet transform,[4] and full signal Fourier transform with log coordinate mapping[10] (LCM). While these transforms are well motivated individually, comparisons and benchmarking between them is difficult and often unsatisfying. We introduce the use of frequency resolution as the pertinent metric for any given transform. For example, early work in Ref. 3 uses an intermediate frequency resolution and detection requires a grid search over both time shifts and frequency scalings. Alternatively, a single Fourier transform of the entire signal[10] gives maximal frequency resolution, and limits the search space to frequency scalings only. Analysis for the proposed short time Fourier transform (STFT) varies the frequency resolution and characterizes the impact on the detection search space in a more constrained framework. Analysis also demonstrates the importance of embedding in the transform's magnitude spectrum, such that detection disregards phase and avoids shifts and offsets.

2. WATERMARK IMPLEMENTATION

In addition to robustness analysis, this work implements an operational audio watermarking scheme as shown in Fig. 1. The basic infrastructure for desynchronization robustness is the segmentation of STFT, $\mathbf{z}$ of the audio signal at the detector into short blocks. Synchronization is achieved by locating which portion of the watermark signal is present in each of these short blocks. Synchronization may alternatively detect no watermark is present in a block, corresponding to insertion of unwatermarked audio. Alignments s_m, between each block, indexed by m, and locations in the watermark signal, as shown in Fig. 2, reconstruct how insertion, deletion, and TSM have modified the audio signal.

In order to enable alignment of the modified audio at the detector, embedding inserts pseudo-random ± 1 SS chip sequences throughout the original audio signal by modifying its STFT magnitude spectrum. Positive and negative chips are embedded by respectively increasing or decreasing the signal magnitude in each STFT bin. Perceptual quality of the watermarked signal is insured by using a perceptual model[11] which limits embedding distortion to an imperceptible level below masking thresholds. For the audio signal shown in Fig. 2, the watermark signal power is limited to $-22.5dB$ below the level of the audio content by perceptual thresholds. For reference,

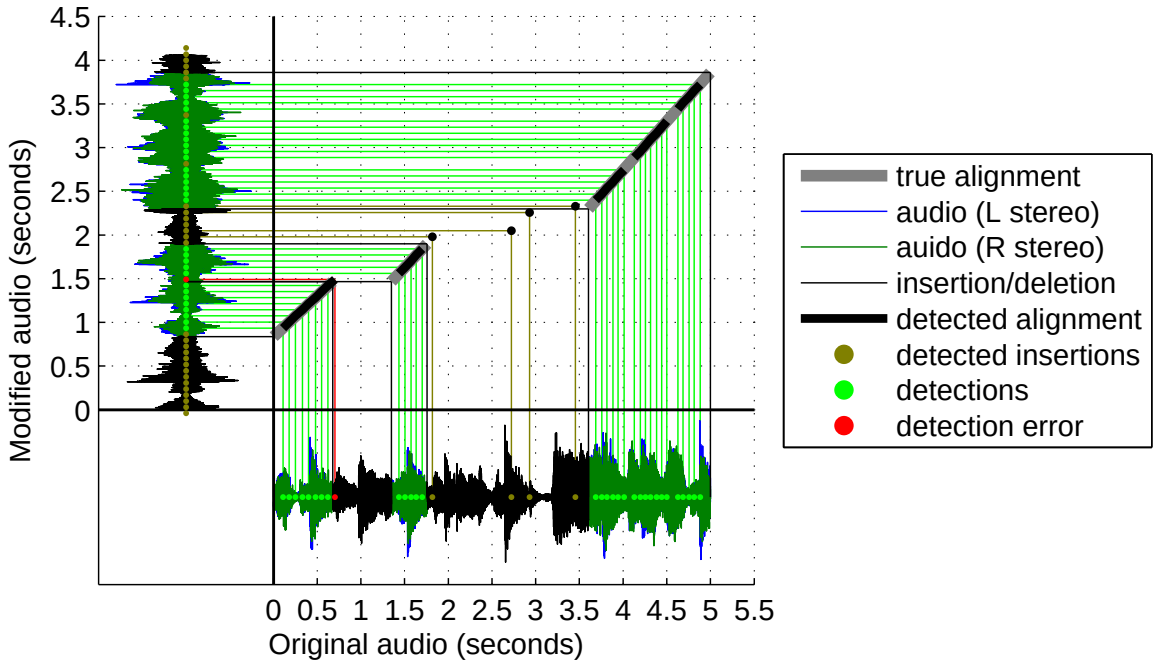


Figure 2. Alignment detects the correspondence between the blocks of $\mathbf{z}$ (vertical-axis), and the watermark embedded in $\mathbf{y}$ (horizontal-axis). Before MP3 compression to the modified audio, desynchronization operations, given numerically in Table 1, include: (a) *deletions*, black portions of the original waveform; (b) *insertions*, black portion of the modified waveform; and (c) *TSM*, portions of the alignment sloped at -7%, $+6\%$, and $+12\%$ from the 45° diagonal.

Segment number	TSM percent	Original signal interval (sec)	Original signal length (sec)	Modified signal interval (sec)	Modified signal length (sec)	Block detection errors	Block detection accuracy
insertion	$\sim$	$\sim$	0	0.00 to 0.84	0.84	$0/_{12}$	100%
#1	-7%	0.00 to 0.68	0.68	0.84 to 1.47	0.63	$1/_{9}$	88%
deletion	$\sim$	0.68 to 1.35	0.67	$\sim$	0	$\sim$	
#2	+6%	1.35 to 1.76	0.41	1.47 to 1.90	0.43	$1/_{6}$	83%
deletion	$\sim$	1.76 to 3.60	1.85	$\sim$	0	$\sim$	
insertion	$\sim$	$\sim$	0	1.90 to 2.30	0.40	$0/_{6}$	100%
#3	+12%	3.60 to 5.00	1.40	2.30 to 3.86	1.56	$4/_{22}$	81%
insertion	$\sim$	$\sim$	0	3.86 to 4.06	0.20	$0/_{5}$	100%

Table 1. Deletions, insertions, and TSM for the ground truth alignment in Fig. 2. Detection accuracies give false positives within *insertion* intervals, and total misses/misalignments within preserved intervals. TSM alters the intervals' lengths.

MP3 compression uses a more sophisticated perceptual model and is able to introduce distortion at a level of $-15dB$. The time-frequency transform resolution controls robustness by determining SS chip duration and bandwidth. Robustness to TSM requires broad frequency bandwidth for each chip, while robustness to time shifts require long time durations for each chip. Bandwidths for each chip, determined by the number of frequency bins it is repeated over, increase in a geometric progression with frequency. Increasing bandwidths provide robustness to pitch scaling, and are equivalent to using a log-scale for frequency.[10] Modifying only the magnitude spectrum is necessary for cases when desynchronization offsets the STFT analysis windows between embedding and detection. Otherwise, if the transform windows are offset, there will be significant phase shifts for high frequency STFT bins and loss of detection peaks in the correlation scores. Because extensive searches such as DTW can detect watermarks after such a wide range of possible modifications, false positive detection errors are an important consideration. To address false positive errors, multiple watermarks are embedded simultaneously using POCS (projections on convex sets) such that separate watermarks detected in overlapping

blocks can be cross-validated.[7] With validation, the occurrence of unwatermarked audio that coincidentally matches one version of a modified watermark signal would also need match all other watermarks modified in an identical manner. As in Fig. 2, individual blocks from insertion portions of the modified audio, and aligned out of regular sequence can be corrected by validation.

In order to align each block of audio to the portion of watermark signal present in it, detection first calculates cross-correlations, r_m between blocks of STFT coefficients and the watermark chips, and then uses DTW to search over the range of alignments possible after desynchronization. Analysis shows that the spreading gain for the correlation peak at the watermark portion actually present in the block, and random interference will be small due to the short block-lengths used, but robust to desynchronization. This low correlation SNR can prevent simple thresholds from detecting which portion of a SS watermark signal is present in individual blocks. However, DTW allows the scheme to find sequences of blocks that contain consecutive portions of the watermark signals. These consecutive detections sum together constructively using DTW, while interference is uncorrelated between blocks in much the same way as longer SS block-lengths increase SNR. Unlike using larger audio blocks to reduce interference, the DTW search technique can survive desynchronization such as insertion/deletion or TSM, but at the cost of increased false positive errors when unwatermarked audio coincidentally matches one of the many alignments to the watermark signal that DTW could have detected. Blocklength is set to $B = 18$ in the general region where best performance was found. This region of B agrees with predicted performance seen in the modeled marginal distributions of the detection statistic.

3. DESYNCHRONIZATION ROBUSTNESS ANALYSIS

In the absence of desynchronization, performance of a single, classic detection test is characterized by a receiver operating characteristic (ROC) curve. For a given noise level, the trade-off between false positive (false-alarm), and false negative (missed-detection) error is governed by the threshold set for the detection statistic. With desynchronization, exhaustive search techniques introduce a new parameter: number of tests, or size of the search space. Extending the search space allows robust detection after desynchronization, but also can increase false positive detection errors. Characterizing this performance loss requires extending analysis from single tests to multiple tests that comprehend the set of possible desynchronization operations. This approach has been adopted in prior work[9] though for the rather limited class of desynchronizations characterized by a cyclic shift. The development here considers a more comprehensive set of desynchronizations for continuous time audio signals, including insertions, deletions, and TSM. Insertions, deletions, and TSM disrupt long embedded watermarks, eliminating the ability to rely on asymptotic performance for long sequences, which was the basis of the analysis in Ref. 9. Because the wider set of desynchronization operations considered in this paper are not bounded to discrete sample intervals, it is not possible to cover the entire space with a finite number of detection tests as in Ref. 9. Consequently, watermark modulation must provide a degree of robustness to small time shifts and TSM. This is referred to as *intrinsic robustness*, and allows discrete detection tests to cover continuous regions in the search space. This section characterizes intrinsic robustness and search space for the proposed detection scheme, and shows how they interact in practice.

3.1 Problem Formulation

Follow the system shown in Fig. 1, let $\mathbf{x} = \{x[0], \ldots, x[N-1]\}$ denote the feature vector extracted from the unwatermarked host signal. Assuming adaptive whitening and normalization during feature extraction, elements of $\mathbf{x}$ are modeled as i.i.d. and normally distributed with zero mean and unit variance. Informed embedding adds a pseudo-random watermark vector, $\mathbf{w} = \{w[0], \ldots, w[N-1]\}$ to $\mathbf{x}$ while obeying perceptual thresholds that are roughly proportional to the power of the host content. Consequently, embedding produces,

$$\mathbf{y} = \mathbf{x} + \gamma \mathbf{w}, \tag{1}$$

where $\mathbf{y}$ is the watermarked feature vector and $-20 \log_{10} \gamma$ is a constant document to watermark ratio (DWR). Detection first extracts the feature vector, $\mathbf{z} = \{z[0], \ldots, z[M-1]\}$ from an unknown audio signal, and then segments $\mathbf{z}$ into $\frac{M}{B}$ blocks of B features each. DTW searches over the space of possible alignments, $\mathcal{S}$ for the true alignment of the blocks after insertion and deletion. Each alignment is denoted by a sequence $\mathbf{s} = \{s_0, \ldots, s_{\frac{M}{B}-1}\}$,

where s_m corresponds to the m^{th} block of $\mathbf{z}$, and either gives an alignment index into $\mathbf{w}$ or indicates an insertion in $\mathbf{z}$ that does not correlate to any portion of the watermark vector. The detection statistic,

$$\rho(\mathbf{s}) = \frac{B}{M} \sum_{m=0}^{\frac{M}{B}-1} r_m(s_m), \tag{2}$$

$$r_m(s_m) = \begin{cases} \frac{1}{B} \sum_{n=0}^{B-1} z[mB+n] \cdot w[s_m+n], & s_m \text{ does not indicate } \textit{insertion} \\ r_{insert}, & s_m \text{ indicates } \textit{insertion} \end{cases}. \tag{3}$$

is the sum of correlations, r_m for each block aligned to $\mathbf{w}$ at s_m. The term r_{insert} is set to $2\frac{\sigma_z}{\sqrt{B}}$ to favor $\mathbf{s}$ containing insertions over false alignment of unwatermarked blocks to $\mathbf{w}$.

3.2 Intrinsic Robustness for SS Blocks and Missed Detection Probability

The ROC for detection performance is found by characterizing conditional probability densities for the detection statistic, $\rho(\mathbf{s})$. For a watermarked feature vector, $\mathbf{z}$ of length M, the portion $\frac{N'}{M}$ of the original N watermarked features remain after desynchronization.* Given the true alignment, $\mathbf{s}_{\text{true}}$, $\rho(\mathbf{s}_{\text{true}})$ is normally distributed, $\mathcal{N}(\mu, \sigma^2)$:

$$\rho(\mathbf{s}_{\text{true}}) \sim \mathcal{N}\left(\frac{N'}{M}\gamma' + \left(1 - \frac{N'}{M}\right) r_{insert}, \frac{N'}{M}\frac{1}{M}\right). \tag{4}$$

Only the portion, $\frac{N'}{M}$ of blocks from $\mathbf{z}$ are watermarked; the remaining portion, $1 - \frac{N'}{M}$ are classified as insertions by $\mathbf{s}_{\text{true}}$. Each watermarked and inserted block respectively contribute γ' or r_{insert} to μ in (4), but only the former contributes to σ^2, because r_{insert} is a fixed value for inserted blocks. The function, $\gamma' = f_\lambda(\gamma)$ gives the expected correlation for each watermarked block, and represents intrinsic robustness with warping, λ.

Intrinsic robustness and the function, f_λ are determined by the time domain watermark signal, and the individual correlations tests, r for each block. For the proposed frequency domain watermark, each feature, $x[n]$ is the spectrum of the audio within the n^{th} STFT analysis window. $\mathbf{w}$ is a grid of ± 1 chips, and $w_k[n]$ is the watermark chip embedded into the k^{th} frequency line of the n^{th} analysis window. The STFT with resolution, N_{fft} uses analysis windows, $h_0 = \cos \pi \frac{t}{N_{\text{fft}}}$ with 50% overlap between windows. In (6), the STFT breaks the host into $\frac{N_{\text{fft}}}{2}$ channels, each with impulse response, $h_k(t)$. Channels for $k > \frac{Nfft}{2}$ are folded back to represent the quadrature components. The embedded time domain watermark signal, $w(t)$ can be produced by modulating the watermark chips with $h_k(t)$:

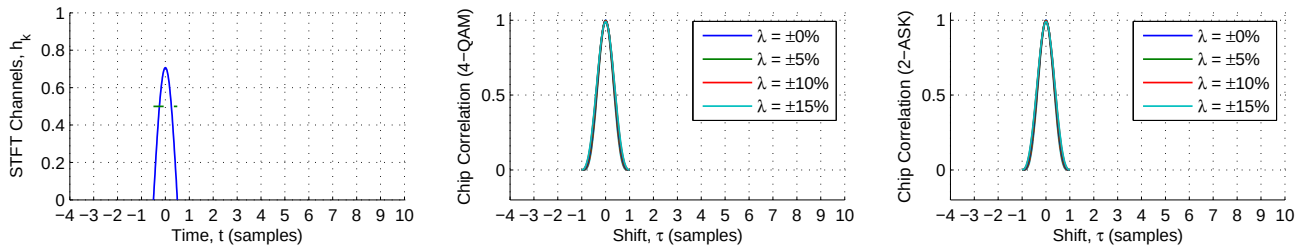
$$w(t) = \sum_{n=0}^{N-1} \sum_{k=0}^{N_{\text{fft}}-1} \gamma w_k[n] \cdot h_k(t - n\tfrac{N_{\text{fft}}}{2}), \tag{5}$$

$$h_k(t) = \frac{1}{\sqrt{N_{\text{fft}}}} \cos \pi \frac{1}{N_{\text{fft}}} t \cdot \begin{cases} \cos 2\pi \frac{k}{N_{\text{fft}}} t, & k \le \frac{N_{\text{fft}}}{2}, \\ \sin 2\pi \frac{N_{\text{fft}}-k}{N_{\text{fft}}} t, & k > \frac{N_{\text{fft}}}{2}, \end{cases} \tag{6}$$

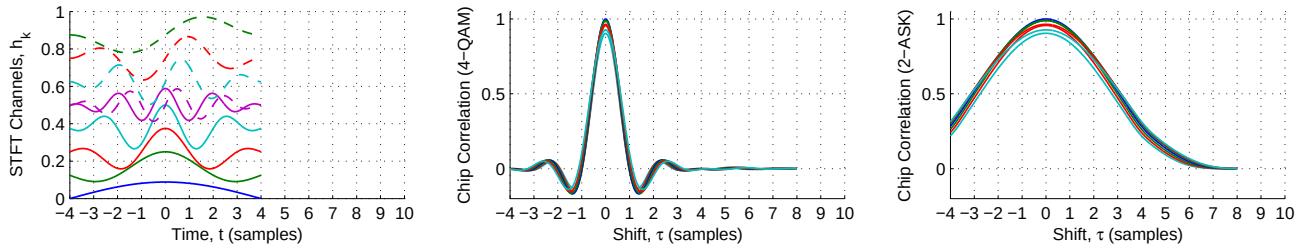
The optimal SS detector is a matched filter (inverse STFT) that performs a correlation between the received signal and the reference embedded signal. The signal power from the correlation detector with time shift, τ, and warping, λ, is given by the height of the autocorrelations summed across the STFT frequency lines,

$$r(\tau, \lambda) = \sum_{k=0}^{N_{\text{fft}}-1} \int_{\frac{-N_{\text{fft}}}{2}}^{\frac{N_{\text{fft}}}{2}} \gamma h_k(t) h_k(\lambda t - \tau)\, dt. \tag{7}$$

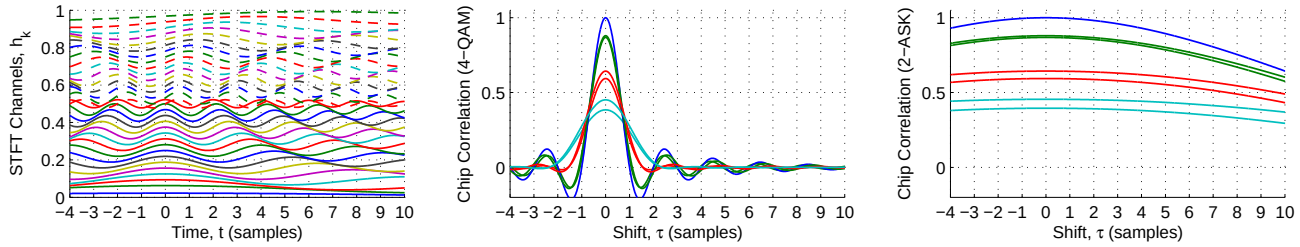
*As implemented, $\mathbf{x}$ is produced by a STFT filter bank with 512X downsampling, and N is ≈430 for 5 second durations of 44.1kHz uncompressed audio. Desynchronization advances through the watermarked signal inserting, deleting, and preserving segments with exponentially distributed durations. Before each preserved segment there is a 25% chance of insertion, 25% chance of deleting a segment of the original, and 50% chance of both. An insertion is always added after the last preserved segment. Expected durations are set to 1.13 seconds for insertions and deletions, and 2.26 seconds for preserved segments. The expected numbers of inserted, deleted, and preserved segments are 2.6, 1.6, and 1.6, and the resulting expectations for M and N' are ≈528 and ≈313.



(a) $N_{\text{fft}} = 1$ (time domain watermark)



(b) $N_{\text{fft}} = 8$



(c) $N_{\text{fft}} = 32$

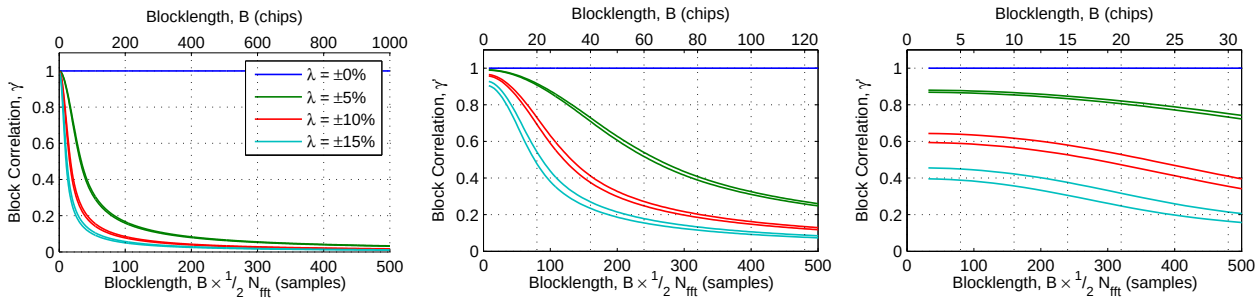
Figure 3. STFT frequency channel impulse responses $h_k(t)$ are shown in the left plots for varying number, N_{fft} of frequency channels. Quadrature responses for $k > \frac{N_{\text{fft}}}{2}$ are dashed. Autocorrelations, $r(\tau, \lambda)$ plotted in the second second column for 4-QAM use both phase and magnitude. Autocorrelations plotted in the third column for 2-ASK use magnitude only.

Because the STFT frequency channels, including the in-phase and quadrature responses, $h_k(t)$ are just an alternative basis set to the time domain, directly summing all the h_k autocorrelations as in (7), and plotted in the middle column of Fig. 3, shows no robustness benefit† as compared to the time domain watermark in the top row of Fig. 3. The autocorrelation peaks are even lowered as the frequency resolution increases because watermark power can spill over the narrow frequency bands in the presence of warping.

The benefit of frequency domain embedding is seen in the broadening of the autocorrelations in the third column of plots in Fig. 3, depicting embedding in the STFT magnitude spectrum without altering phase. Embedding in the magnitude spectrum combines the in-phase and quadrature responses such that $w(t)$ matches the phase of the host signal, $\mathbf{x}$. However, because the quadrature channels are used, the number of chips and interference tolerance are both decreased by a factor of two. Detection in the power spectrum is a Costas receiver that can demodulate the watermarked frequency channels even after phase ambiguity due to τ [12, pp. 365]:

$$r(\tau, \lambda) = \sum_{k=0}^{\frac{N_{\text{fft}}}{2}} \sqrt{\left(\int_{\frac{-N_{\text{fft}}}{2}}^{\frac{N_{\text{fft}}}{2}} \gamma h_k(t) \cdot h_k(\lambda t - \tau)\, dt \right)^2 + \left(\int_{\frac{-N_{\text{fft}}}{2}}^{\frac{N_{\text{fft}}}{2}} \gamma h_{(N_{\text{fft}}-k)}(t) \cdot h_k(\lambda t - \tau)\, dt \right)^2} \tag{8}$$

†Autocorrelations for the low frequency h_k show robust broad autocorrelation,[4] but the capacity and spreading gain are limited by the low bandwidth and number of chips that can be embedded in only these channels.



(a) $N_{\text{fft}} = 1$ (time domain) (b) $N_{\text{fft}} = 8$ (c) $N_{\text{fft}} = 32$

Figure 4. The function $\gamma' = f_\lambda$ shows that for short blocklengths, the mean value of the detection statistic, γ' is close to the value, $\gamma = 1$ possible in the absence of desynchronization. λ gives the severity of warping. Watermark structure is disregarded in these figures, resulting in significant underestimate of γ' for high N_{fft} and severe warping.

The most important implication of broadening the autocorrelation is for the size of the detection search space. If the autocorrelation peak is broad over a region of time shifts, τ, then exhaustive correlations over the audio length can be down-sampled by a factor proportional to the breadth of the peak without losing significant detection SNR. While higher frequency resolution, N_{fft} increases the intrinsic robustness of SS chips to time shifts, this benefit comes at the cost of intrinsic robustness to frequency scaling.

In addition to using multiple frequency channels to improve robustness, watermark modulation can also introduce structure or repetition. For SS modulation, repeating each plus or minus chip over multiple time steps or frequency channels broadens the autocorrelation peak by a factor proportional to the number of repetitions. The proposed scheme employs repetition in the frequency domain only, in order to mitigate the loss of warping robustness due to increasing the frequency resolution. This repetition improves robustness because modulation allows repeated SS chips to interfere constructively when warping causes embedded signal in one frequency band to spill to adjacent bands. The two trade-offs involved in introducing structure are the loss spreading gain, and vulnerability to the estimation and remove attack addressed in Ref. 3.

Autocorrelations plotted in Fig. 3 show that correlation detection of a single feature is robust to significant time shift, allowing the search space to covered more sparsely with fewer correlation tests. However, to further decrease the search space and false positive detections, detection segments $\mathbf{z}$ into blocks rather than aligning each feature individually. For large blocks, warping can result in large time misalignment, even if one portion of the block is aligned well. This effect of TSM can be modeled by assuming the center of a block is aligned with $\tau = 0$, and calculating the resulting offset, τ_m for each of the other features within the block,

$$\gamma' = f_\lambda(\gamma) = \frac{1}{B}\sum_{m=0}^{B-1} r(\tau_m, \lambda), \tag{9}$$

$$\tau_m = \left(m - \tfrac{B-1}{2}\right) \cdot (\lambda - 1) \cdot \tfrac{N_{\text{fft}}}{2}, \tag{10}$$

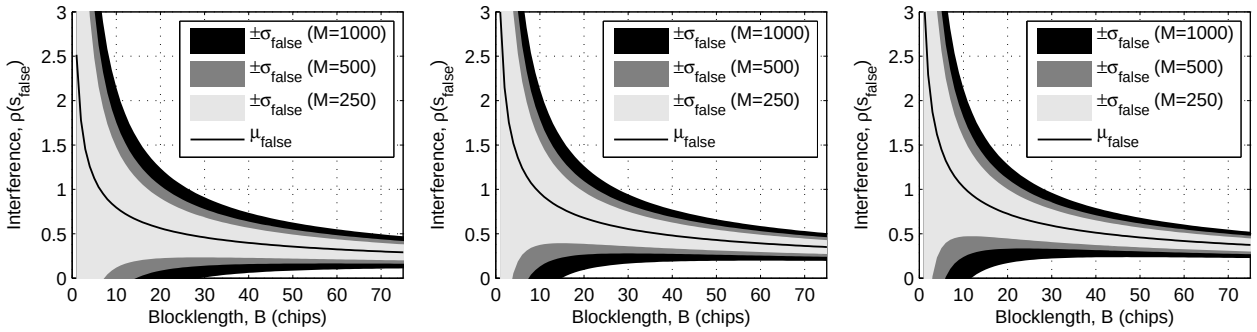
where $m \in \{0, \ldots, B-1\}$ is the index for each chip in the block, and $(\lambda - 1)$ is portion by which the offset grows for each chip father from the center, $\frac{B-1}{2}$. $\frac{N_{\text{fft}}}{2}$ is the width of each chip in audio samples, and B is the number of chips per block such that the blocklength is $B\frac{N_{\text{fft}}}{2}$. The function f_λ is plotted in Fig. 4, showing that as warping becomes more severe, it is important to use as small blocklength as possible to maximize γ' and $\rho(\mathbf{s}_{\text{true}})$ in (4).

3.3 DTW Search Space and False Alarm Probability

The probability of false alarm is determined by the distribution of the detection statistic,

$$\rho(\mathbf{s}_{\text{false}}) = \max_{\mathbf{s}\in\mathcal{S}} \rho(\mathbf{s}), \tag{11}$$

for the maximal alignment, $\mathbf{s}_{\text{false}}$ of $\mathbf{w}$ to an unwatermarked feature vector of length M. While DTW uses dynamic programing to build a maximum alignment in polynomial time, the search space, $\mathcal{S}$ grows exponentially for full



(a) $N = 100$ (b) $N = 500$ (c) $N = 1000$

Figure 5. Interference is characterized by the distribution for false positive detections, $\rho(\mathbf{s}_{\text{false}}) \sim \mathcal{N}(\mu_{\text{false}}, \sigma^2_{\text{false}})$. Increasing the blocklength, B decreases the search space by decreasing the total number of blocks that the M element feature vector is segmented into. Increasing the watermark length, N decreases σ_{false}, but also increases μ_{false}.

alignments between the $\frac{M}{B}$ blocks and the N watermark elements. Partial alignments errors are disregarded because these can be considered as complete misalignment of subsections of $\mathbf{z}$ and $\mathbf{w}$. The DTW constraint that each block be aligned in sequence is also relaxed, breaking (2) into a sum of independent block correlations,

$$\rho(\mathbf{s}_{\text{false}}) \leq \frac{B}{M} \sum_{i=0}^{\frac{M}{B}-1} \max_{s_i \in \left\{ \substack{0,\ldots,N-1,\\ insert} \right\}} r_i(s_i) \tag{12}$$

Because individual $r_i(0, \ldots, N-1)$ for the i^{block} block are each normally distributed with zero mean and variance, $\sigma^2_{r_i} = \frac{1}{B}\sigma^2_x$, the distribution of the maximum r_i within the summation in (12) that contributes to $\rho(\mathbf{s}_{\text{false}})$ is,

$$P_{r_{\max}}(r) = \begin{cases} \Phi(\frac{\sqrt{B}}{\sigma_x} r_{insert})^N \cdot \delta(r - r_{insert}), & \text{for } r \leq r_{insert}, \\ N \cdot \Phi(\frac{\sqrt{B}}{\sigma_x} r)^{N-1} \cdot \dfrac{\sqrt{B}}{\sigma_x \sqrt{2\pi}} e^{-\frac{Br^2}{2\sigma_x^2}}, & \text{for } r > r_{insert}. \end{cases} \tag{13}$$

$\Phi(\frac{1}{\sigma} r)$ is the the cumulative distribution of a Normal random variable, and $\delta(r)$ is a Dirac δ-function. $P_{r_{\max}}(r)$ converges in shape to an Extreme Value type I (Gumbel) distribution for $r_i > r_{insert}$ as the watermark length, N becomes large [13, pp. 23]. For $r_i \leq r_{insert}$ the distribution is limited to the minimum r_{insert}. The distribution for the interference, approximated in (12) converges to a Normal distribution as M becomes large. The mean, μ_{false} and variance, σ^2_{false} are calculated from $P_{r_{\max}}(r)$ using the central limit theorem for averaging $\max r_i$ over large numbers of blocks:

$$\mu_{\text{false}} = \int r \cdot P_{r_{\max}}(r)\, dr \tag{14}$$

$$\sigma^2_{\text{false}} = \frac{M}{B} \cdot \left(\int r^2 \cdot P_{r_{\max}}(r)\, dr - \mu^2_{\text{false}} \right) \tag{15}$$

$$\tag{16}$$

As shown in Fig. 5, decreasing the size of the search space determined by B, M, and N, decreases the interference distribution for $\rho(\mathbf{s}_{\text{false}})$. Lower values for $\rho(\mathbf{s}_{\text{false}})$ decrease the probability of false positive errors.

Figures 4 and 5 show a trade off for the size of the search space can be governed by blocklength. Smaller blocklengths provide finer resolution for the DTW alignment and allow greater robustness to TSM. However, finer DTW alignment resolution also increases the number of possible alignments within the search space, and increases the probability of false positive errors. Greater robustness is achieved by sacrificing baseline performance possible in the absence of desynchronization – a trade-off that is demonstrated in the results. The importance of time-frequency domain watermarking and of watermark structure, aka repetition, in providing TSM robustness

is also apparent from the results. These techniques provide intrinsic robustness, such that individual detection tests remain sensitive to an embedded watermark over wider ranges of desynchronization. By allowing individual detection tests to cover continuous ranges of time offsets and frequency scalings, intrinsic robustness decreases the number of detection tests and the probability of false positive errors.

4. CONCLUSIONS

Building on prior work,[7] this paper introduces a framework for SS audio watermarking in the presence of desynchronizing attacks including insertions, deletions, and time scale modification (TSM). The presented analysis introduces the useful notion of intrinsic robustness and characterizes the trade-off between this parameter and the search space. The practical audio watermark implemented in the framework shows promising simultaneous robustness to insertion/deletion, TSM, and MP3 compression.

REFERENCES

[1] Moulin, P. and Koetter, R., "Data-hiding codes," *Proceedings of the IEEE* **93**, 2083 –2126 (Dec. 2005).

[2] Cox, I., Miller, M., Bloom, J., Fridrich, J., and Kalker, T., [*Digital Watermarking and Steganography*], Morgan Kaufmann Publishers, San Francisco, CA, USA, second ed. (2007).

[3] Kirovski, D. and Malvar, H. S., "Spread-spectrum watermarking of audio signals," *Signal Processing* **51**, 1020–1033 (Apr. 2003).

[4] Wu, S., Huang, J., Huang, D., and Shi, Y., "Efficiently self-synchronized audio watermarking for assured audio data transmission," *IEEE Transactions on Broadcasting* **51**, 69 – 76 (Mar. 2005).

[5] Kirovski, D. and Attias, H., "Audio watermark robustness to desynchronization via beat detection," in [*Information Hiding*], Petitcolas, F., ed., *Lecture Notes in Computer Science* **2578**, 160–176, Springer Berlin / Heidelberg (2003).

[6] Coumou, D. J. and Sharma, G., "Insertion, deletion codes with feature-based embedding: A new paradigm for watermark synchronization with applications to speech watermarking," *IEEE Transactions on Information Forensics and Security* **3**, 153–165 (June 2008).

[7] Nadeau, A. and Sharma, G., "Insertion deletion robust audio watermarking: a set theoretic, dynamic programming approach," in [*Proc. SPIE: Media Watermarking, Security and Forensics*], Alattar, A. M., Memon, N. D., and Heitzenrater, C. D., eds., **8665**, 866503 (Feb. 2013).

[8] Tchamkerten, A., Chandar, V., and Wornell, G. W., "Communication under strong asynchronism," *IEEE Transactions Information Theory* **55**, 4508–4528 (Oct. 2009).

[9] Barni, M., "Effectiveness of exhaustive search and template matching against watermark desynchronization," *IEEE Signal Processing Letters* **12**, 158–161 (Feb. 2005).

[10] Kang, X., Yang, R., and Huang, J., "Geometric invariant audio watermarking based on an LCM feature," *IEEE Transactions on Multimedia* **13**(2), 181–190 (2011).

[11] ISO/IEC, "ISO/IEC 11172-3:1993 - Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio." Padrão (1993).

[12] Proakis, J. G., [*Digital Communications*], McGraw-Hill, New York, fourth ed. (2001).

[13] Reiss, R. D., [*Approximate distributions of order statistics: with applications to nonparametric statistics*], Springer-Verlag, New York (1989).