About zero bit watermarking error exponents

Teddy Furon

ABOUT ZERO BIT WATERMARKING ERROR EXPONENTS

Teddy Furon

Inria
Linkmedia
Campus de Beaulieu, Rennes, France

ABSTRACT

This paper aims to motivate more research works on the design of zero-bit watermarking schemes by showing an upper bound of the performances that known solutions failed to reach. To this end, an upper bound of error exponent characteristic is derived by translating Costa's rationale to zero-bit watermarking with side information. Three schemes are then considered: the dual-cone detection region originally proposed by Cox *et al.* and improved in Merhav *et al.* papers, ISS (Improved Spread Spectrum), and ZATT (Zero Attraction). It turns out that in certain conditions the latter performs better than the first one, which questions the optimality claimed Merhav *et al.* Nevertheless, the main conclusion is that these schemes are in general far away from the upper bound in the region of practical interest.

***Index Terms*—** Watermarking, Hypotheses test, Error exponent

1. INTRODUCTION

In a *zero-bit* watermarking scheme, one is solely interested in distinguishing watermarked from non watermarked content. The embedding does not hide any message, and there is no decoding. Zero-bit watermarking just embeds and detects a mark. There has been some confusion about the terminology. Some misused the term 'one-bit' [1, 2]: A 'one-bit' watermarking scheme is when one detects and then decodes a message of a single bit.

Zero-bit watermarking is a hypotheses test problem with two specifities:

- Under hypothesis $\mathcal{H}_0$, the received signal is *given by Nature*: it is a signal extracted from an original content possibly corrupted by some noise.
- Under hypothesis $\mathcal{H}_1$, the received signal has been modified by the embedding: it is a signal extracted from a content with the addition of a watermark signal *dependent on the host signal* and a secret key, and possibly corrupted by some noise.

There are two types of error: Detection of a mark whereas the received signal has not been watermarked. This is a false positive whose probability is denoted by $\mathbb{P}_{\text{fp}}$. A false negative misses the presence of the mark whereas the received signal has been watermarked. Its probability is denoted by $\mathbb{P}_{\text{fn}}$.

As in [1], we are interested in the error exponents, which are the exponential decay rates of these probabilities as the size n of the received signals goes to infinity:

$$E_{\text{fp}} := \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_{\text{fp}}, \quad E_{\text{fn}} := \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_{\text{fn}}. \quad (1)$$

This paper investigates the mapping $E_{\text{fn}} = F(E_{\text{fp}})$. These exponents are non negative, and for a given setup and E_{fp}, the bigger E_{fn} the better a watermarking scheme performs.

We focus on two points of the characteristic: the left and the right endpoints. The graph of the function $E_{\text{fn}} = F(E_{\text{fp}})$ starts on the left by the point $(E_{\text{fp}}, E_{\text{fn}}) = (0, E_{\text{fn}}^L)$ where $E_{\text{fn}}^L := \lim_{E_{\text{fp}}\to 0^+} F(E_{\text{fp}})$. On the other hand, the graph ends on the right at $(E_{\text{fp}}, E_{\text{fn}}) = (E_{\text{fp}}^R, 0)$. Larger false positive rates are achievable but then $E_{\text{fn}} = 0$: Probability $\mathbb{P}_{\text{fn}}$ may still vanish to zero but not exponentially as $n \to \infty$.

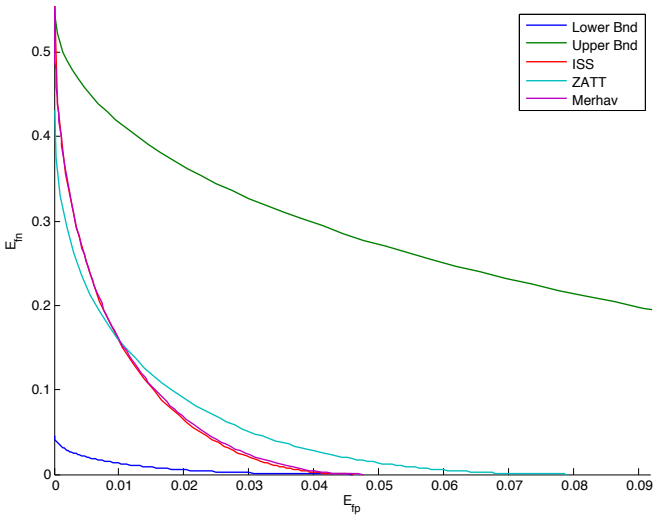

Fig. 1. Comparison of the error exponent characteristics. Setup: $\sigma_X = 1$, $P = 0.1$, $\sigma_Z = 0.3$. Lower and upper bounds (8), ISS (9), ZATT (10), and Merhav *et al.* (17)

2. UPPER AND LOWER BOUNDS

This section proposes lower and upper bounds of the characteristic $E_{\text{fn}} = F(E_{\text{fp}})$ in the Gaussian setup. It is a pastiche of Costa's paper [3] and generalizes the study presented in [4].

A feature vector in $\mathbb{R}^n$ is extracted from the content. Vectors $\mathbf{x}$ and $\mathbf{r}$ denote respectively the extracted feature from an original content, so-called the host, and from the content received by the detector. The embedder transforms $\mathbf{x}$ into $\mathbf{y}$ by adding a watermark $\mathbf{w}$: $\mathbf{y} = \mathbf{x} + \mathbf{w}(\mathbf{x})$. This vector depends on the host (for a side-informed watermarking scheme) and on a secret key (not indicated to keep notations simple). We consider a power constraint watermark problem where the energy of the watermark per sample is limited: $\|\mathbf{w}(\mathbf{x})\|^2/n \leq P$. An attack is modeled by the addition of a noise vector $\mathbf{z}$: $\mathbf{r} = \mathbf{y} + \mathbf{z}$.

The theoretical setup models the host and noise vector by random vectors $\mathbf{X}$ and $\mathbf{Z}$ distributed as Gaussian white signal: $\mathbf{X} \sim \mathcal{N}(\mathbf{0}_n, \sigma_X^2 \mathbf{I}_n)$ and $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_n, \sigma_Z^2 \mathbf{I}_n)$. Vector $\mathbf{0}_n$ denotes the vector of n zero components, and $\mathbf{I}_n$ the identity matrix of size n. This stems into the following statistical model of the received vector $\mathbf{R}$:

$$\mathcal{H}_0 : \mathbf{R} = \mathbf{X} + \mathbf{Z}, \quad \mathcal{H}_1 : \mathbf{R} = \mathbf{X} + \mathbf{w}(\mathbf{X}) + \mathbf{Z} \tag{2}$$

2.1. Derivations of the error exponents

We resort to the classical derivation of the error exponents (see for instance [5, Sec. 2.7]). It consists in pretending that the detector knows the distributions of the received signal $\mathbf{R}$ under both hypothesis. The optimal detector is then the Neyman-Pearson test which compares the log likelihood ratio $s(\mathbf{R})$ to a threshold τ. The Chernoff bound provides inequalities for both error probabilities via the cumulant-generating function $\mu_n(t) := \log \mathbb{E}(e^{ts(\mathbf{R})}|\mathcal{H}_0)$. For instance, $\mathbb{P}_{\text{fp}} \leq e^{\mu_n(t) - t\tau}$, $\forall t > 0$. The tightest bound is given for t s.t. $\mu_n'(t) = \tau$. A concentration inequality shows that this Chernoff bound is asymptotically tight as $n \to \infty$ in the Gaussian setup so that:

$$E_{\text{fp}} = \lim_{n\to\infty} \frac{-\mu_n(t) + t\mu_n'(t)}{n}, \tag{3}$$

$$E_{\text{fn}} = \lim_{n\to\infty} \frac{-\mu_n(t) - (1-t)\mu_n'(t)}{n}. \tag{4}$$

2.2. Lower bound

The following presents a lower bound in the sense that a skilled watermarker cannot do worse than designing an embedder not taking into account the side information. In other words, $\mathbf{w}(\mathbf{X}) = \mathbf{w}$, a fixed secret reference signal s.t. $\|\mathbf{w}\|^2 = nP$. For this additive spread spectrum scheme, $\mathbf{R} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$ under $\mathcal{H}_0$ and $\mathbf{R} = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_n)$ under $\mathcal{H}_1$ with $\sigma^2 := \sigma_X^2 + \sigma_Z^2$. Easy calculation finds $\mu_n(t) = nPt(t-1)/2\sigma^2$ leading to:

$$(E_{\text{fp}}, E_{\text{fn}}) = \frac{P}{2\sigma^2} \left(t^2, (1-t)^2\right), \quad \forall t \in (0,1) \tag{5}$$

which can be rewritten as $E_{\text{fn}} = \mathcal{F}(E_{\text{fp}}, \sigma_X^2 + \sigma_Z^2)$ with

$$\mathcal{F}(E_{\text{fp}}, \sigma^2) := \left(\left| \sqrt{P/2\sigma^2} - \sqrt{E_{\text{fp}}} \right|_+ \right)^2 \tag{6}$$

and $|a|_+ := a$ if $a \geq 0$ and 0 otherwise.

2.3. Upper bound

The upper bound is derived by giving a clear advantage to the detector: it knows both $\mathbf{X}$ and $\mathbf{w}(\mathbf{X})$. Then, $\mathbf{R} - \mathbf{X} \sim \mathcal{N}(\mathbf{0}_n, \sigma_Z^2 \mathbf{I}_n)$ under $\mathcal{H}_0$ and $\mathbf{R} - \mathbf{X} = \mathcal{N}(\mathbf{w}(\mathbf{X}), \sigma_Z^2 \mathbf{I}_n)$ under $\mathcal{H}_1$. Similar calculations end up with

$$(E_{\text{fp}}, E_{\text{fn}}) = \frac{P}{2\sigma_Z^2} \left(t^2, (1-t)^2\right), \quad \forall t \in (0,1) \tag{7}$$

which can be rewritten as $E_{\text{fn}} = \mathcal{F}(E_{\text{fp}}, \sigma_Z^2)$.

2.4. Conclusion

When side-information (*i.e.* the knowledge of the host) is available at the embedder only, the scheme should perform better or equal than a scheme without side-information. At the same time, it should perform worse or equal than a scheme with side-information at the embedding and the detection sides. In other words, its characteristic function should lie in between the two previous ones:

$$\mathcal{F}(E_{\text{fp}}, \sigma_X^2 + \sigma_Z^2) \leq F(E_{\text{fp}}) \leq \mathcal{F}(E_{\text{fp}}, \sigma_Z^2). \tag{8}$$

Surprisingly, this result is new as [4] derived similar bounds only for one specific point of the characteristc: E_{fn}^L.

3. APPLICATION TO TWO SIMPLE SCHEMES

The previous study raises the question of achievability: is it possible to design a scheme whose characteristic is close to the upper bound? We first analyze two very simple schemes.

3.1. ISS (Improved spread spectrum)

The embedding reduces the interference with the host by creating $\mathbf{w}(\mathbf{x}) = (\alpha - \lambda(\mathbf{x}^\top \mathbf{u}))\mathbf{u}$, where $\mathbf{u}$ is a secret direction in $\mathbb{R}^n$ with $\|\mathbf{u}\| = 1$, and λ is the host rejection coefficient. The power constraint imposes that $\alpha^2 + \lambda^2 \sigma_X^2 \leq nP$. Asymptotically as $n \to \infty$, we can set $\lambda = 1$ so that $\alpha^2 = nP - \sigma_X^2 > 0$.

As the received signal is still Gaussian distributed under both hypotheses, similar calculations give the following parametric definition of the characteristic: $\forall t \in (0,1)$

$$(E_{\text{fp}}, E_{\text{fn}}) = \frac{P}{2\sigma_X^2} \frac{\rho}{(1+t\rho)^2} \left(t^2(1+\rho), (1-t)^2\right), \tag{9}$$

with the SNR $\rho := {}^{\sigma_X^2}/_{\sigma_Z^2}$. This characteristic function tends to the upper bound as $t \to 0^+$ achieving the left endpoint $E_{\text{fn}}^L = {}^{P}/_{2\sigma_Z^2}$ as claimed in [4]. But, it tends to the lower bound as $t \to 1^-$ reaching the right endpoint $E_{\text{fp}}^R = {}^{P}/_{2(\sigma_Z^2+\sigma_X^2)}$. This result is depicted in Fig. 1.

3.2. ZATT - Zero Attraction

We now consider the ZATT scheme [6, Sect. 3.3], for which the derivation of the error exponents is simple and insightful.

The embedder cancels the k first components[1] of $\mathbf{x}$ by adding $\mathbf{w}(\mathbf{x}) = -(x(1), \cdots, x(k), 0, \cdots, 0)^\top$. This is possible if $k\sigma_X^2 \leq nP$. We assume here that $P < \sigma_X^2$, which is relevant in most practical watermarking scenarios. Asymptotically, as $n \to +\infty$, k can go to infinity as well, scaling as $\lfloor nP/\sigma_X^2 \rfloor$. Under both hypotheses, the $n-k$ last components of $\mathbf{R}$ follow the same distribution. Therefore, they do not contribute to error exponents. Under the Gaussian setup, we obtain:

$$\begin{aligned} E_{\text{fp}} &= \frac{P}{2\sigma_X^2}\left(\log(1+t\rho) - \frac{t\rho}{1+t\rho}\right), \\ E_{\text{fn}} &= \frac{P}{2\sigma_X^2}\left(\log\frac{1+t\rho}{1+\rho} + \frac{(1-t)\rho}{1+t\rho}\right). \end{aligned} \tag{10}$$

For any $t \in (0,1)$, the point of the characteristic $(E_{\text{fp}}, E_{\text{fn}})$ tends to $(+\infty, \log t - 1 + {}^1/_t)$ when $\rho \to +\infty$, *i.e.* $\sigma_Z^2 \to 0$. Then, E_{fn} can be set as big as possible by driving t close to 0. This is not surprising: detecting the watermark iff $r(1)^2 = 0$ gives a perfect test (*i.e.* $\mathbb{P}_{\text{fp}} = \mathbb{P}_{\text{fn}} = 0$) in the noiseless setup where $\sigma_Z^2 = 0$. This even holds when n is finite provided that $n \geq \sigma_X^2/P$. This scheme has little interest in practice, but it stresses the fact that, under the noiseless setup, it is very easy to achieve perfect performances.

In the noisy setup, the left and right endpoints are achieved for $t = 0$ and $t = 1$ respectively:

$$E_{\text{fn}}^L = \frac{P}{2\sigma_Z^2}\left(\rho - \log(1+\rho)\right), \tag{11}$$

$$E_{\text{fp}}^R = \frac{P}{2\sigma_Z^2}\left(\log(1+\rho) - \frac{\rho}{1+\rho}\right). \tag{12}$$

This shows that ZATT fails reaching the upper bounds ${}^{P}/_{2\sigma_Z^2}$ on both endpoints as soon as $\sigma_Z^2 > 0$. We will see that, despite these shortcomings, ZATT is not devoid of interest.

4. MERHAV *et al.* APPROACH

Sabbag and Merhav [7] show that, under limited resources, the optimal detector thresholds the absolute value of the cosine, *i.e.* the normalized correlation, between $\mathbf{r}$ and $\mathbf{u}$. This defines the acceptance region $\mathcal{C}$ as a dual hypercone of axis $\mathbf{u}$ and angle β: The watermark is detected if

$$\frac{|\mathbf{r}^\top \mathbf{u}|}{\sqrt{n}\|\mathbf{r}\|} \geq \cos\beta. \tag{13}$$

This theoretically justifies a long tradition in the history of watermarking since the seminal paper of Cox *et al.* [8].

4.1. False positive error exponents

Under $\mathcal{H}_0$, $\mathbf{R}$ has an isotropic distribution and the probability of false alarm is given by $\mathbb{P}_{\text{fp}} = 1 - I_{\cos^2\beta}(1/2; (n-1)/2)$, where $I_x(a;b)$ is the incomplete beta function [9]. For a fixed angle β, this yields the error exponent [7]:

$$E_{\text{fp}} = -\log(\sin\beta). \tag{14}$$

This exponent is bigger as the aperture of the dual hypercone is smaller. Indeed, angle β will play the role of the auxiliary variable defining the parametric characteristic function.

4.2. False negative error exponent

Later on, Comesaña *et al.* show that the optimum embedding creates $\mathbf{y} = a(\mathbf{x})\mathbf{x} + b(\mathbf{x})\mathbf{u}$ (see [10] for the expressions of functions a and b), which provides in the 'high SNR regime' the following false negative exponent [10, Th.2]:

$$E_{\text{fn}} = S(\max\{1, {}^{P}/_{\sigma_X^2 \cos^2\beta}\}), \tag{15}$$

with $S(x) := (x - 1 - \log x)/2 \geq 0, \forall x > 0$.

5. CRITICAL ANALYSIS

We now humbly criticize the work of [10]. We acknowledge the remarkable quality of this work: the derivation of the false negative error exponent is indeed very technical and much more complicated than the easy schemes presented in Sec. 3. However, there are two pitfalls in this article.

5.1. 'High SNR regime'

The authors of [10] misuse the wording 'high SNR regime'. Parameter σ_Z is missing in (15). Indeed, this equation must be understood as the limit of E_{fn} when $\sigma_Z \to 0$. Therefore, this is a zero-order approximation of E_{fn} in the 'high SNR regime', or more rigorously it is the expression of E_{fn} in the noiseless scenario.

5.2. Optimality

The optimality of this scheme claimed in [10] is also arguable. In the noiseless setup and if $P < \sigma_X^2$, $0 < E_{\text{fn}} < \infty$ when $E_{\text{fp}} \in [0, -{}^1/_2 \log(1 - {}^{P}/_{\sigma_X^2}))$. This means that, in this range, both $\mathbb{P}_{\text{fp}}$ and $\mathbb{P}_{\text{fn}}$ vanish exponentially as $n \to \infty$. However, Sec. 3.2 shows that, under the same conditions, the ZATT scheme provides a perfect test ($\mathbb{P}_{\text{fn}} = \mathbb{P}_{\text{fp}} = 0$). This holds when n is finite provided that $n \geq \sigma_X^2/P$.

[1] To provide security, this is indeed done on $k < n$ secret orthogonal projections.

5.3. Extensions

The journal version of this paper proves that the characteristic of this scheme achieves the following right endpoint:

$$E_{\text{fp}}^R = \frac{1}{2}\log\frac{2\sigma_X^2}{\sqrt{(\sigma_X^2-P-\sigma_Z^2)^2+4\sigma_X^2\sigma_Z^2}+(\sigma_X^2-P-\sigma_Z^2)}. \tag{16}$$

This complies with the result of [10]: if $P<\sigma_X^2$ and $\sigma_Z=0$, then $E_{\text{fp}}^R = -1/2\log(1-P/\sigma_X^2)$. Moreover, before this endpoint, E_{fn} is positive and can be bounded by:

$$\bar{E_{\text{fn}}} = \frac{\left(\tan^{-1}\beta\sqrt{\sigma_Z^2+\sigma_X^2\sin^4\beta}-\sqrt{P-\sigma_X^2\cos^4\beta}\right)^2}{2\sigma_Z^2}. \tag{17}$$

It appears that, as $\beta\to\pi/2$ (*i.e.* $E_{\text{fp}}\to 0$), $E_{\text{fn}}\to\bar{E_{\text{fn}}}$, which in turn converges to $P/2\sigma_Z^2$: the left endpoint of this scheme achieves the upper bound (8). On the other hand, $\bar{E_{\text{fn}}}$ cancels when $P=\sigma_X^2\cos^2\beta+\sigma_Z^2\tan^{-2}\beta$, which gives back (16).

6. COMPARISON

We make a comparison encompassing the lower and upper bounds (8), ISS (Sect. 3.1), ZATT (Sect. 3.2), and the double hypercone scheme (Sect. 4).

Fig. 1 gives an overview of the situation. Note first that this figure is given for a 'practical' setup of robust watermarking where $P<\sigma_Z^2<\sigma_X^2$. We clearly see that these schemes reach high E_{fn}, close to the upper bound (8), but only on the left endpoint (*i.e.* for very small values of E_{fp}). On the right endpoint (*i.e.* high E_{fp} for small values of E_{fn}), they completely fail getting closer to the upper bound.

We believe that this remains *the* major issue of robust zero-bit watermarking. As far as we know, no watermarking scheme achieves the upper bound on the right hand side of the characteristic. This is all the more important because this side is the most relevant in practice. In many applications, the risk of a false positive is far bigger than the risk of a false negative. In DRM, for instance, a false positive amounts to accuse an innocent user whereas a false negative results in failing to capture a dishonest user. The requirements usually set the probability of false positive to extremely weak level. Consequently we seek high values of E_{fp}.

Secondly, the scheme proposed by Merhav and co-authors outperforms ISS by a tiny margin in this setup. Indeed, ZATT performs better at high E_{fp}. This somewhat calls in question the optimality claimed in [7, 10]. Yet, the complexity of detectors is in $O(n)$ except for ZATT in $O(kn)$.

We now give a closer look to the endpoints E_{fn}^L and E_{fp}^R plotted as functions of σ_Z in Fig. 2. For the left endpoint E_{fn}^L, Fig. 2 (top) shows that the upper bound, ISS and Merhav's scheme are superposed. ZATT's endpoint is lower especially at high σ_Z. Indeed, its endpoint can be lower than $P/2(\sigma_X^2+\sigma_Z^2)$, when $\sigma_Z^2\geq 2/3\sigma_X^2$.

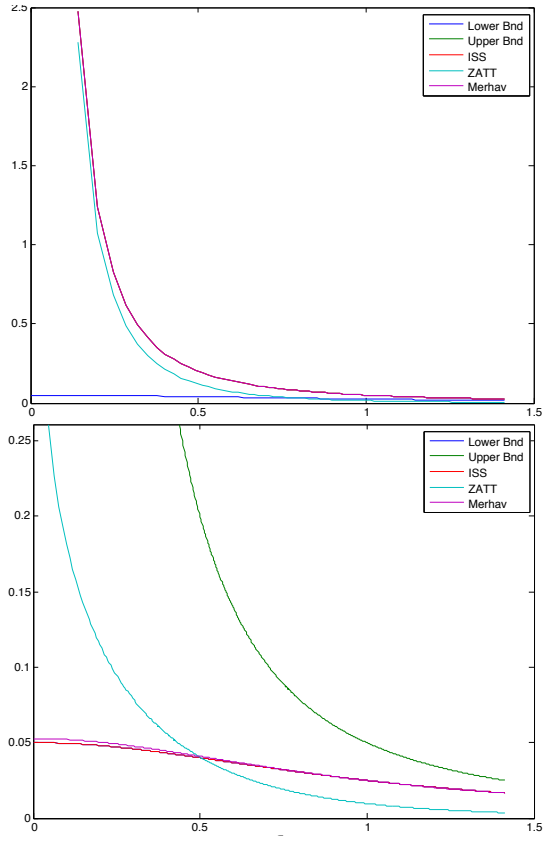

Fig. 2. Comparison of the endpoints. Setup: $\sigma_X=1$, $P=0.1$. E_{fn}^L (top) and E_{fp}^R (bottom) as functions of σ_Z.

As for the right endpoint E_{fp}^R, the other schemes perform like the lower bound (or slightly better), while the superiority of ZATT is remarkable at low σ_Z. Again, at $\sigma_Z=0$, it provides a perfect test for $k=1$, *i.e.* under limited resources. Yet, the endpoint of ZATT goes below than the lower bound when $\sigma_Z^2\geq\sigma_X^2/4$ (approximatively), which is still a useful range in practice.

7. CONCLUSION

Fig. 2 summarizes the paper:

- The ZATT scheme challenges the optimality claimed in [7, 10] even in the 'high SNR regime' and unders limited resources. It achieves a lower left endpoint but a much higher right endpoint.
- While these schemes achieve the upper bound at the left endpoint, the mismatch at the right endpoint is large. This is critical because this region matters in practice.

Contrary to channel capacity with side information [3], the optimal characteristic of zero-bit watermarking remains unknown.

8. REFERENCES

[1] Pedro Comesana, Neri Merhav, and Mauro Barni, "Asymptotically optimum embedding strategy for one-bit watermarking under gaussian attacks," in *Proc. SPIE*, 2008, vol. 6819, pp. 681909–681909–12.

[2] T. Liu and P. Moulin, "Error exponents for one-bit watermarking," in *Proc. of ICASSP*, Hong-Kong, apr 2003.

[3] M.H.M. Costa, "Writing on dirty paper (corresp.)," *Information Theory, IEEE Transactions on*, vol. 29, no. 3, pp. 439–441, May 1983.

[4] Teddy Furon, Julie Josse, and Sandrine Le Squin, "Some theoretical aspects of watermarking detection," in *Security, Steganography, and Watermarking of Multimedia Contents VIII*, E. Delp and P. W. Wong, Eds., San Jose, CA, USA, United States, Jan. 2006, SPIE, vol. 6072 of *Proc. SPIE-IS&T Electronic Imaging*.

[5] H. Van Trees, *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*, Wiley, 1968.

[6] Jonathan Delhumeau, Teddy Furon, Guénolé Silvestre, and Neil Hurley, "Improved Polynomial Detectors for Side-Informed Watermarking," in *Security and Watermarking of Multimedia Contents IV*, P. W. Wong and E. Delp, Eds., San José, CA, USA, United States, Jan. 2002, SPIE, pp. 311–321.

[7] E. Sabbag and N. Merhav, "Optimal watermark embedding and detection strategies under limited detection resources," in *Information Theory, 2006 IEEE International Symposium on*, July 2006, pp. 173–177.

[8] Ingemar J. Cox, Joe Kilian, F.T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *Image Processing, IEEE Transactions on*, vol. 6, no. 12, pp. 1673–1687, Dec 1997.

[9] Teddy Furon, Cyrille Jégourel, Arnaud Guyader, and Frédéric Cérou, "Estimating the probability of false alarm for a zero-bit watermarking technique," in *IEEE International Conference on Digital Signal Processing*, Santorini, Greece, July 2009.

[10] P. Comesana, N. Merhav, and M. Barni, "Asymptotically optimum universal watermark embedding and detection in the high-snr regime," *Information Theory, IEEE Transactions on*, vol. 56, no. 6, pp. 2804–2815, June 2010.