# On the Use of Masking Models for Image and Audio Watermarking

Arnaud Robert and Justin Picard

***Abstract*—In most watermarking systems, masking models, inherited from data compression algorithms, are used to preserve fidelity by controlling the perceived distortion resulting from adding the watermark to the original signal. So far, little attention has been paid to the consequences of using such models on a key design parameter: the robustness of the watermark to intentional attacks. The goal of this paper is to demonstrate that by considering fidelity alone, key information on the location and strength of the watermark may become available to an attacker; the latter can exploit such knowledge to build an effective *mask attack*. First, defining a theoretical framework in which analytical expressions for masking and watermarking are laid, a relation between the decrease of the detection statistic and the introduced perceptual distortion is found for the mask attack. The latter is compared to the Wiener filter attack. Then, considering masking models widely used in watermarking, experiments on both simulated and real data (audio and images) demonstrate how knowledge on the mask enables to greatly reduce the detection statistic, even for small perceptual distortion costs. The critical tradeoff between robustness and distortion is further discussed, and conclusions on the use of masking models in watermarking drawn.**

***Index Terms*—Attacks, mask attack, masking models, robustness, watermarking, Wiener attack.**

## I. Introduction

DIGITAL watermarking techniques are used to embed an imperceptible and generally encoded/encrypted message, the watermark, into a host content—digital data (audio, image, video). Watermarking may serve different purposes such as data hiding, copyright protection, integrity check and so on.

### A. Background

In its own context, watermarking can be regarded as seeking the best tradeoff between three critical design parameters: robustness, fidelity and capacity. Robustness measures the watermark's ability to resist malicious or unintentional attacks in the scope of the considered watermarking application. Fidelity is an essential property of watermarking systems; it asserts how perceptually similar the watermarked and original content are. Fidelity can be evaluated using a measure of distance between the original and watermarked content; although the signal-to-noise ratio (SNR) is often used, it is known to be a rather poor indicator of fidelity. Transparency implicitly means that the fidelity constraint was successfully attained. Finally, capacity (or payload) reflects the number of useful information embedded into the original signal—from a single bit when determining the presence (or absence) of a watermark to many bits when conveying a more complex message such as an identification number or the ASCII transcryption of a web site address. This paper specifically addresses the relationship between the two parameters, fidelity and robustness.

The most successful step toward minimizing the perceived distortion in watermarking was the adaptation of masking models at the embedding process—those utilized in data compression to shape the quantization noise in agreement with perceptual findings on human hearing and visual systems. Masking models determine the intensity level required, as a function of time or frequency, for a signal to be perceived in presence of other stimuli—and thus determine which intensity levels are not perceived. This function is represented by a set of values referred to as the *mask*; the latter is usually computed over successive segments of data. For example, in JPEG compression, the image is divided into $8 \times 8$ pixel blocks and the mask is computed independently for each block. The mask then either modulates the watermark to ensure it is not perceived [1], [14], [17], or serves to increase the watermark's energy for a given fidelity constraint by allowing maximal imperceptible signal energy to be embedded. Masking models have been essential to ensure one property of watermarks: transparency. But by having given priority to fidelity, little attention has been paid to the potential vulnerability of mask-shaped watermarks to attacks.

Robustness, or detection performance, of watermarking techniques has become increasingly important as more copyright applications were foreseen. Attacks on suggested techniques have been popular and have become increasingly sophisticated over the past few years. In particular, the estimate-and-remove class of attacks has gained momentum: the watermark is estimated and then subtracted from the watermarked signal. Early work can be found in [9]. More recently, the Wiener filter has been successfully implemented [8], [13], [15]; it assumes that the embedded watermark has a zero-mean Gaussian distribution with an estimated standard deviation.

### B. Scope of the Paper

Several watermarking schemes use masking models to shape a white spectrum message in order to guarantee the watermark's transparency. The scope of this paper is to determine and quantify how, using knowledge on the masking model, one can de-

Manuscript received August 27, 2001; revised December 18, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Harrick M. Vin.

A. Robert was with the Audio-Visual Communications Laboratoy (LCAV), Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland. He is now with Thomson-Technicolor, Burbank, CA 91504 USA (e-mail: arnaud.robert@thomson.net).

J. Picard was with the Laboratory of Nonlinear Systems (LANOS), Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland. He is now with Thomson-MediaSec, Essen 45127, Germany (e-mail: jpicard@mediasec.com).

Digital Object Identifier 10.1109/TMM.2005.846781

TABLE I
NOTATIONS USED IN THE TEXT

| entity | meaning | type |
|---|---|---|
| $m$ | message | vector |
| $w$ | watermark | vector |
| $x$ | host signal | vector |
| $y$ | watermarked host | vector |
| $y'$ | attacked signal | vector |
| $\alpha$ | mask | vector |
| $\beta$ | weighting factor | vector |
| $\gamma$ | average perceptual distortion | scalar |
| $d(a,b)$ | distortion between $a$ and $b$ | scalar |
| $DDS$ | decrease in detection statistic | scalar |

rive a new estimate and remove attack, hereafter referred to as the *mask attack*. The starting point is finding an analytical relation between the masking model and the robustness of the watermarking system.

### C. Outline

A description of the masking models used in this study is found at Appendix. The first section of this paper provides an analytical form for a generic masking model and an additive watermarking scheme. Studying the relation between the decrease in detection statistic and the introduced perceptual distortion, in the defined framework, the mask attack is derived. A comparison with the Wiener filter attack is made and results of theoretical simulations given to formalize the theoretical behavior of the attack. Next, experimental results on real sounds and images illustrate the effectiveness of the attack. Finally, a discussion on the tradeoff between robustness and distortion is addressed, and conclusions on the use of masking models in watermarking drawn.

## II. THE MASK ATTACK—THEORY

This section introduces a theoretical framework for an additive watermarking scheme which makes use of a masking model to shape the watermark. Studying the relation between perceptual distortion and robustness, the mask attack is derived. The latter is compared to the Wiener attack; while both attacks are based on the estimation and further subtraction of the estimates watermark given a perceptual distortion constraint, they differ with respect to the knowledge that is utilized: the Wiener attack makes an assumption on the global statistics of the watermark while the mask attack assumes knowledge on the mask and therefore local characteristics of the watermarked signal. The assumptions needed to derive the mask attack will prove correct in Section III, where experimental results are given.

### A. The Model

Before defining the embedding and detection processes and the utilized measure of perceptual distortion, let us introduce in Table I the notation of, and the assumptions on, the signals and entities used in this study. Each signal is represented as a vector (a matrix can be re-written as a vector).

The *host signal* is the original content (audio sample, image). The *message* is a sequence of bits of information to be conveyed by the watermark; it could be a series of random numbers, an optimal coding sequence, etc. The *watermark* is the shaped message that is ultimately added to the host signal. Each data vector, in the considered transform domain, has a total length $N$. Each vector is decomposed in subvectors (blocks) of length $M\,(<N)$ for processing purposes; considering the masking models used in this paper (see Appendix), a block corresponds to an $8\times 8$ pixel matrix for images and samples corresponding to a duration of 20 ms for sounds. A Gaussian distribution of mean $\mu$ and variance $\sigma^2$ is denoted as $N(\mu,\sigma^2)$. The notation $\mathbf{a}\cdot\mathbf{b}$ indicates that each dimension of $\mathbf{a}$ is multiplied by the corresponding dimension of $\mathbf{b}$: $a_i\cdot b_i$.

**Embedding.** The watermark embedding process is described by the following relations:

$$y = x + w$$
$$w = \alpha \cdot m.$$

In order to be able to derive theoretical results, three hypotheses on the signals are made, as follows:

$$\begin{aligned} &\text{H1:} \quad x \sim N\left(0,\sigma_x^2\right)\\ &\text{H2:} \quad m \sim N(0,1)\\ &\text{H3:} \quad \alpha \sim N\left(\mu_\alpha,\sigma_\alpha^2\right). \end{aligned} \tag{1}$$

The mask $\alpha$ gives the maximum allowed perceptual distortion for each coefficient; it is used to modulate the message, $m$, such that $E[w^2]=\alpha^2$. Since the distribution of the mask can not be described in a closedform, the hypothesis of a Gaussian distribution was adopted. Observations show that, in general, the distribution of mask values can be better approximated by a mixture of Gaussians, as illustrated at Fig. 1.

The distribution of the mask values for two DCT coefficients (normalized to the standard deviation of the corresponding DCT coefficient, i.e., divided by $\sigma_x$), is shown in Fig. 2 for the reference image "Girl". Coefficients indexed (4,4) and (4,7), with computed mean and standard deviation of $\mu_\alpha = 0.28$, $\sigma_\alpha = 0.277$ and $\mu_\alpha = 0.506$, $\sigma_\alpha = 0.387$ are shown on the left and right side of the figure, respectively. Mask values greater than 1.5 are spread over a long interval and are not shown; they represent only a small fraction of the overall distribution, although they can indeed be very useful to the attacker. It appears from the computed data that the Gaussian hypothesis is adequate; it is further validated in Section III where empirical results are given. It is worth noting that this hypothesis is not stronger than the widely accepted Gaussian distribution approximation for the DCT coefficients of an image—which vary much locally in a given image.

**Detection.** The presence (or absence) of a known watermark in a received signal is assessed using a standard detection statistic method: the correlation between the received signal $r$ and the embedded message $m$. This computation is supported by experiments from Zeng and colleagues [17]. However, recent work suggest that detectors based on the generalized Gaussian distribution yield better detection results [5]. The generic detection function is expressed as

$$f(r,m) = r \cdot m.$$

When detection theory is applied to watermarking, one of two hypotheses—the presence or the absence of a known watermark—is verified by comparing the output of $f(\cdot)$ to a threshold. The latter is computed with respect to the system's design (cost functions, etc.). In the maximum-likelihood case, the normalized threshold is equal to one half.

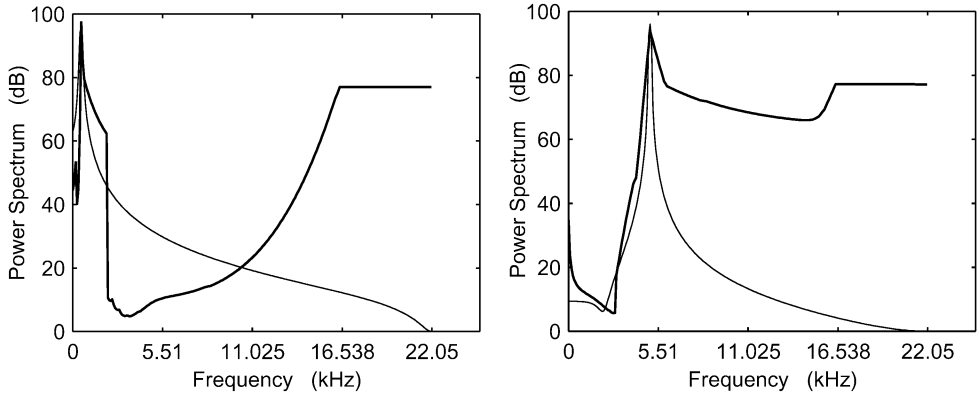



Fig. 1. Original spectrum (normal line) and masking threshold (bold line) for sine waves at frequency 500 (left) and 5000 (right) Hz.

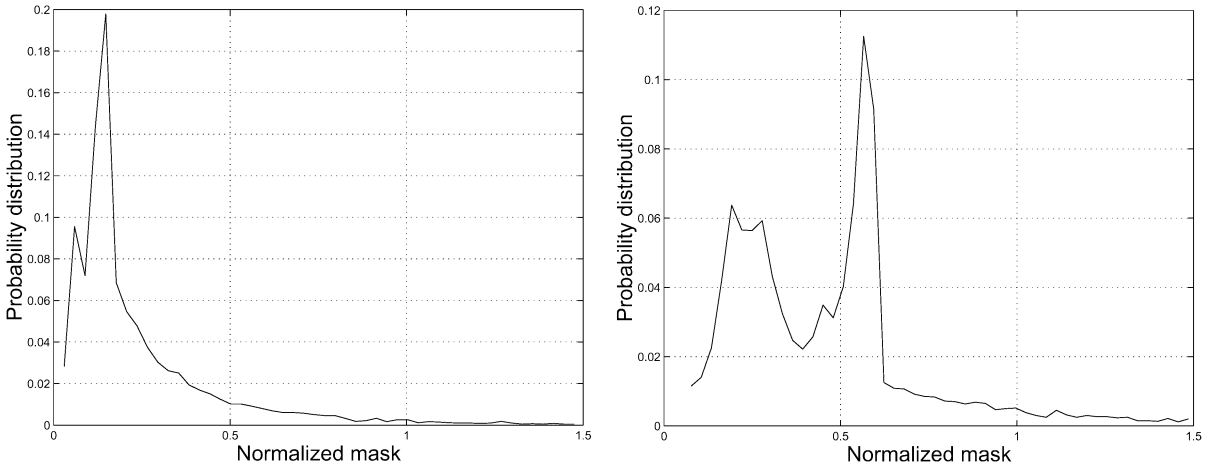



Fig. 2. Distribution of the normalized mask values for two DCT coefficient of the reference image "Girl". The two coefficients are left (4,4) and right (4,7)

**Measure of fidelity.** Distortion is a scalar measure of the difference between two signals. The distortion between the original and watermarked content is relevant to the watermarker while the distortion between the attacked and original content is relevant to the attacker. The distortion is usually computed using one of two methods: 1) the standard SNR value, computed as the sum of the squared differences $d(x_1, x_2) = (x_1 - x_2)^2$ or 2) a perceptual distortion value, computed as an SNR value weighted by a masking model so that regions of the signal where humans are less sensitive to are little considered, and vice-versa. In the present context, the perceptual distortion is defined by

$$d(r, x) = \frac{1}{N} \cdot \frac{(r - x)^2}{\alpha^2}. \tag{2}$$

If $d(r, x) = 0$ the original data was unaltered while if $d(r, x) = 1$ the maximal allowable perceptual distortion (distortion "just" not perceived) was introduced.

**Measure of the effectiveness of an attack.** An attack is considered *efficient* if it produces a significant decrease in the detection statistic $(DDS)$ for a given perceptual distortion constraint. The effectiveness can be measured by normalizing the difference in detection statistics before and after the attack, as expressed by the variable $DDS$:

$$DDS = \frac{f(y, m) - f(y', m)}{f(y, m)}. \tag{3}$$

If $DDS = 0$, the attacked image was unaltered and the watermark can be retrieved. If $DSS = 1$ the attack was successful and the watermark can not be retrieved, just as if it was removed. Clearly, it is not exactly the case: in general, the detection statistic drops severely at components significant to the detection process and is little altered at the other components. In the case of a maximum-likelihood detector the watermark would be considered *NOT* present if $DDS > 1/2$.

### *B. The Mask Attack*

In estimate-and-remove attacks, a weighting factor usually controls the strength of the attack and can somewhat be considered as a measure of the introduced distortion. Instead of (or in addition to) exploiting statistics on the signal as done with the Wiener filter, the attacker can advantageously exploit of knowledge on the mask and implement the *mask attack*. One way to derive a theoretical framework for the latter is to find the relation between the mask and the attack's efficiency.

A first, critical, assumption is that the mask $\alpha$ is known to the attacker. Since $y \simeq x$ by definition, one could argue that the mask of the watermarked content is a good approximation of that of the original content: $\alpha(y) \simeq \alpha(x)$. This intuitive assumption was previously used in the literature [5], [15]. To further validate this assumption, Fig. 2 illustrates the correlation between the mask of the original and watermarked content, for the DCT coefficient (4,3) of the reference image "Girl" computed over

6144 blocks of data; the value of the correlation factor is 0.91. Computing the correlation factor for other coefficients yielded similar values, crediting further the assumption. If $\alpha$ is known (deterministic), it can be considered as a constant when deriving the statistics of $w$. Hence, we have $w \sim N(0, \alpha^2)$. The Wiener estimate $\hat{w}$ of the watermark is [6, p. 147]

$$\hat{w} = \frac{\alpha^2}{\sigma_x^2 + \alpha^2} \cdot y. \tag{4}$$

The attack consists of subtracting the estimate of the watermark from the watermarked content. Introducing a weight factor $\beta'$, the attack is defined by the following equations (for $\beta, \beta' > 0$):

$$\begin{aligned} y' &= y - \beta' \hat{w} \\ \beta &= \beta' \cdot \frac{\alpha^2}{\sigma_x^2 + \alpha^2} \\ y' &= y - \beta y. \end{aligned} \tag{5}$$

The parameter $\beta$ represents the strength of the attack. It can be related to the perceptual distortion constraint, as well as to the expected reduction of detection statistic as shown next. Indeed

$$\begin{aligned} y' &= (1-\beta)x + (1-\beta)\alpha m \\ E[f(y,m)] &= \alpha \\ E[f(y',m)] &= (1-\beta)\alpha. \end{aligned}$$

Thus, the expected $DDS$, denoted $\overline{DDS}$, is

$$\overline{DDS} = E\left[\frac{f(y,m) - f(y',m)}{f(y,m)}\right] = \beta. \tag{6}$$

Now, let us derive the expected perceptual distortion for a given $\beta$. We note that

$$\begin{aligned} w &\sim N(0, \alpha^2) \\ y' &\sim N\left(0, (1-\beta)^2\left(\sigma_x^2 + \alpha^2\right)\right) \\ (y' - x) &\sim N\left(0, \beta^2\sigma_x^2 + (1-\beta)^2\alpha^2\right). \end{aligned}$$

The expected perceptual distortion is

$$\begin{aligned} \gamma &= E\left[\frac{(y'-x)^2}{\alpha^2}\right] \\ &= Var\frac{\left[\left(N\left(0, \beta^2\sigma_x^2 + (1-\beta)^2\alpha^2\right)\right)\right]}{\alpha^2} \\ &= \frac{\beta^2\sigma_x^2}{\alpha^2} + (1-\beta)^2. \end{aligned} \tag{7}$$

These results are also valid for binary messages: if $m$ follows a binary equiprobable distribution ($m \sim \{-1, +1\}$), then $w \sim \{-\alpha, +\alpha\}$ and it can be shown that the expected $DDS$ and the expected $\gamma$ will be equal to the found expressions.

Using $nsr = \sigma_x^2/\alpha^2$, we can compute $\beta$ from (7) as

$$\begin{aligned} \beta &= \alpha \cdot \frac{\alpha + \sqrt{\gamma\alpha^2 + (\gamma-1)\sigma_x^2}}{\alpha^2 + \sigma_x^2} \\ &= \frac{1 + \sqrt{\gamma + (\gamma-1)\cdot nsr}}{1 + nsr}. \end{aligned} \tag{8}$$

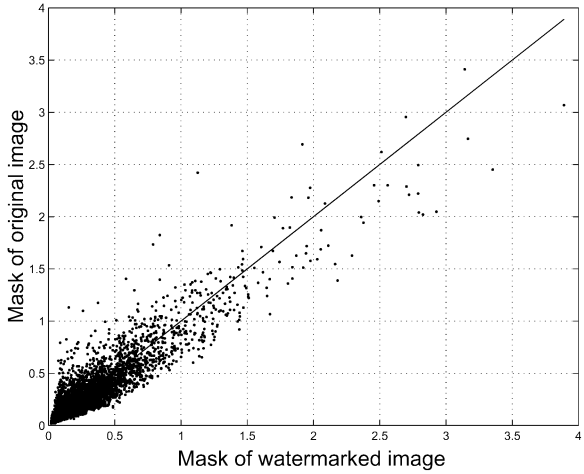



Fig. 3. Mask values for original versus watermarked images ("Girl") for the DCT coefficient (4,3).

Let us make a few comments. The parameter $\gamma$ is a global attack parameter and a constant scaling factor (over all $M$ dimensions) set by the attacker. The mask $\alpha$, a vector, determines the amount by which each of the $M$ dimensions can be modified. Consequently $\beta$ is a vector that parameterizes the local attack and indicates the amount by which the detection statistic will be decreased in each of the $M$ dimension. The value of $\beta$ increases with relative mask energy. Finally, the expected perceptual distortion $\gamma$ is linked to the expected attack effectiveness $DDS$ by the preceding equation in which $\beta$ is replaced by $\overline{DDS}$.

To illustrate the behavior of this equation (i.e., of the mask attack), two graphs are shown in Fig. 3, for $\gamma = 1, 2, 3, 4$: the $DDS$ as a function of the normalized mask (top) and the expected detection statistic $\alpha(1-\beta)$ as a function of the normalized mask (bottomt). It can be seen that: 1) the $DDS$ increases with $\gamma$ for a given value of the mask; 2) the $DDS$ increases with the value of the mask—the attack is more efficient; 3) as the value of the mask increases, the expected detection statistic first increases, reaches a maximum and then decreases; and 4) when considering the attack which introduces the smallest perceptual distortion ($\gamma = 1$), the optimal value of the mask is $\alpha = 0.5$; furthermore, considering a higher perceptual distortion value ($\gamma = 4$) for which the signal is still of reasonable quality, the watermark should *not* be embedded in regions where the mask value exceeds 0.5.

Three cases are of particular interest.

- **Zero perceptual distortion attack.**

  If $\gamma = 1$, the attacker does not introduce additional perceptual distortion. The corresponding value of $\overline{DDS}$ is

  $$\overline{DDS} = \frac{1 + \sqrt{\gamma}}{1 + nsr}. \tag{9}$$

  In theory, the detection statistic can be decreased at little or no cost by the attacker. This is confirmed by experiments reported in Section III, where in some cases $\overline{DDS}$ is significantly reduced at *no* perceptual distortion cost. Thus, it is suggested that the derivation of embedding rules that consider the robustness parameter should take into account the $DDS$ factor.

- **Approximation for small mask energy.**
  When $\sigma_x^2 \gg \alpha^2$ (large $nsr$) the $DDS$ can be approximated by

$$\overline{DDS} \simeq \sqrt{\gamma - 1}\frac{\alpha}{\sigma_x}$$

  and the mean reduction of detection statistic $\beta$ for an expected $\gamma$ becomes

$$\overline{DDS}_M \simeq \sqrt{\gamma - 1} \cdot \frac{\mu_\alpha}{\sigma_x}. \tag{10}$$

  Therefore, the reduction in detection statistic is proportional to the mean mask energy. This corresponds to the linearization of the graph shown in Fig. 3 (top).
- **Approximation for high mask energy—or small** $nsr$. $\left(\sigma_x^2 \ll \alpha^2\right)$

$$\beta = 1 + \sqrt{\gamma}. \tag{11}$$

  In some cases, the amount of energy that can be introduced at given dimensions, using the mask, is significantly higher than the energy of the host signal at those dimensions; this is especially seen for sounds at higher frequencies. Using masking models would lead us to believe that the watermark would be more robust since "more signal energy means better detection performance". Conversely, embedding the watermark in these specific dimensions seems undesirable since the presence of the watermark is no longer a secret when the mask is known. For example, according to (11), $\beta$ can be greater than 2 (no distortion attack), which means the detection statistic is strongly negative at these dimensions.

### C. Comparison With the Wiener Attack

The Wiener attack, an estimate-and-remove attack, does not take into consideration local values of the mask $\alpha$. The single (yet strong) assumption of the Wiener attack is that $w \sim N\left(0, \sigma_w^2\right)$. Let us recall the attack

$$y' = y - \beta y = (1-\beta)(x+w) = (1-\beta)(x+\alpha m) \tag{12}$$

where $\alpha$ is now a random variable (unknown). The value of $\beta$ can not be adjusted to take into account local perceptual distortion.

Let us compute the average perceptual distortion $\gamma_{wiener}$ for a given, fixed, $\beta$

$$\begin{aligned}\gamma_{wiener} &= E\left[\frac{(y'-x)^2}{\alpha^2}\right] \\ &= E\Big[\beta^2 \cdot \frac{x^2}{\alpha^2} + (1-\beta)^2 \cdot m^2 - 2\beta \\ &\qquad \cdot (1-\beta) \cdot \frac{mx}{\alpha}\Big] \\ &= \beta^2 E\left[\frac{x^2}{\alpha^2}\right] + E[m^2](1-\beta)^2 \\ &\quad - 2E\left[\frac{mx}{\alpha}\right]\beta(1-\beta).\end{aligned} \tag{13}$$

Clearly, $E[m^2] = 1$. However, one can only obtain estimates for $E[x^2/\alpha^2]$ and $E[mx/\alpha]$ in which $\alpha$ is a random variable. One can still use the first order approximation, noting that it is only valid when $Var[Y]$ is small compared to $(E[Y])^2$, as follows:

$$\begin{aligned}E\left[\frac{X}{Y}\right] &\approx \frac{E[X]}{E[Y]}\left(1 + \frac{Var[Y]}{E[Y]^2} - \frac{cov(X,Y)}{E[X]\cdot E[Y]}\right) \\ &= \frac{E[X]}{E[Y]}\left(1 + \frac{Var[Y]}{E[Y]^2}\right)\end{aligned} \tag{14}$$

where $cov(X,Y) = 0$, since $x$, $\alpha$, and $m$ are mutually independent.

Assuming that $corr(\alpha, \alpha^2) = 1$ (almost true), we find: $Var[\alpha^2] = 4\mu_\alpha^2\sigma_\alpha^2 + 2\sigma_\alpha^2$. Using this result with (14), we obtain

$$\begin{aligned}E\left[\frac{x^2}{\alpha^2}\right] &= \frac{E[X^2]}{E[\alpha^2]}\left(1 + \frac{Var[\alpha^2]}{E[\alpha^2]^2}\right) \\ &= \frac{\sigma_x^2}{\mu_\alpha^2 + \sigma_\alpha^2}\left(1 + \frac{4\mu_\alpha^2\sigma_\alpha^2 + 2\sigma_\alpha^4}{\left(\mu_\alpha^2 + \sigma_\alpha^2\right)^2}\right)\end{aligned} \tag{15}$$

$$\begin{aligned}E\left[\frac{mx}{\alpha}\right] &= \frac{E[mx]}{E[\alpha]}\left(1 + \frac{\sigma_\alpha^2}{\mu_\alpha^2}\right) \\ &= \frac{E[m]E[x]}{E[\alpha]}\left(1 + \frac{\sigma_\alpha^2}{\mu_\alpha^2}\right) = 0.\end{aligned} \tag{16}$$

Finally, by importing (15) and (16) in (13), one obtains

$$\begin{aligned}\gamma_{wiener} &\approx \beta^2\frac{\sigma_x^2}{\mu_\alpha^2 + \sigma_\alpha^2}\left(1 + \frac{4\mu_\alpha^2\sigma_\alpha^2 + 2\sigma_\alpha^4}{\left(\mu_\alpha^2 + \sigma_\alpha^2\right)^2}\right) + (1-\beta)^2 \\ &= \beta^2\frac{\sigma_x^2}{\mu_\alpha^2}\left(1 + \frac{1 + 6c^2 + 3c^4}{(1+c^2)^3}\right) + (1-\beta)^2\end{aligned} \tag{17}$$

where $c = \sigma_\alpha/\mu_\alpha$ is a measure for the spreading of mask values. Note that the previous equation is only valid for $\sigma_\alpha$ values that are small compared to $\mu_\alpha$; hence, for a small $c$. In that case, (17) can be resolved for $\beta$

$$\beta = \frac{1 + \sqrt{\gamma + \left(\frac{\sigma_x^2}{\mu_\alpha^2}\right)c'(\gamma - 1)}}{1 + \left(\frac{\sigma_x^2}{\mu_\alpha^2}\right)c'}; \quad c' = \frac{1 + 6c^2 + 3c^4}{(1+c^2)^3}. \tag{18}$$

A comparison with the Mask attack is possible for small mask values, i.e., $\sigma_x^2 \gg \mu_\alpha^2$, in which case $\beta \approx \sqrt{(\gamma - 1/c')} \cdot (\mu_\alpha/\sigma_\alpha)$. Recalling that the expected decrease in detection statistic is equal to the parameter $\beta(\overline{DDS_W} = \beta)$, we can make a comparison with the Mask attack, using (10)

$$\overline{DDS_W} = \frac{1}{\sqrt{c'}}\overline{DDS_M}. \tag{19}$$

For $c$ varying from 0 to 0.5 ($\sigma_\alpha = 0$ to $\sigma_\alpha = \mu_\alpha/2$), we note that $1/\sqrt{c'}$ decreases from 1 to 0.85. Therefore, the following conclusions can be drawn.

- It is not surprising that both the mask and the Wiener attacks have the same result for a constant mask value $(c = 0)$, since there is no point of locally adjusting the attack in that case.
- As the mask values $c$ become more disperse, the Wiener attack becomes decreasingly effective compared to the mask attack, reaching a factor of 15% for $c = 0.5$. This

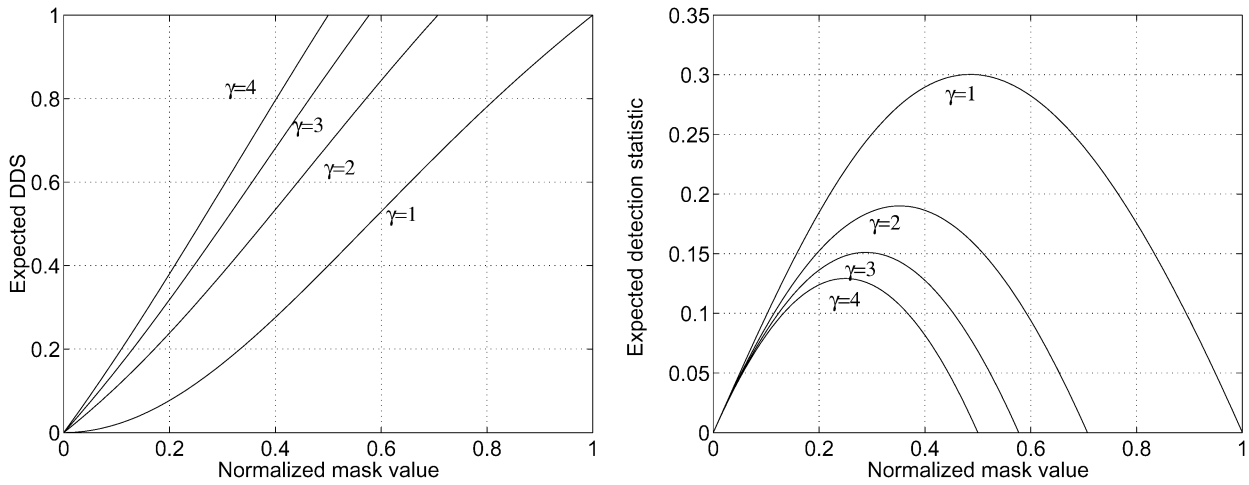



Fig. 4. $DDS$ versus the normalized mask (left), and the expected detection statistic $\alpha(1-\beta)$ versus the normalized mask (right).

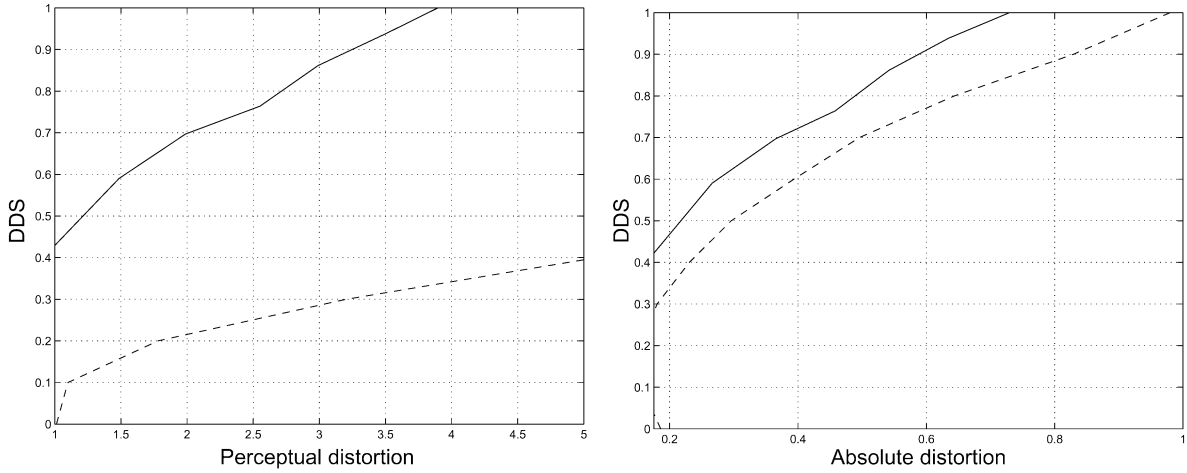



Fig. 5. Value of the $DDS$ as function of the introduced perceptual (left) and absolute (right) distortions for the Wiener and masking attacks. Here $\mu_\alpha = 0.3$, $\sigma_\alpha = 0,3$. Wiener attack: dashed line. Mask attack: solid line.

demonstrates how knowledge on the mask value allows a better targeted attack.

- For higher values of $c$, the first-order approximation which was used is no longer validated and a higher order approximation would be needed to readily compare the two attacks. However, it is expected that the difference in effectiveness between the two will increase: the Wiener attack may result in higher perceptual distortions as the mask value gets locally more unpredictable. The theoretical simulations provided next confirm this intuition.

### D. Simulations

*1) Comparison Between Wiener and Mask Attacks:* The theoretical behavior of the mask attack can be simulated and compared to the Wiener attacks, for different mask values. One important characteristic is the decrease in detection statistic $(DDS)$ as a function of the introduced perceptual distortion. Artificial signals corresponding to hypotheses H1-H3 of (1) are generated. Both perceptual and absolute (SNR) distortion measures are computed. The simulations were conducted with the values $\mu_\alpha = 0.3$, $\sigma_\alpha = 0.3$, $\sigma_x = 1$, which correspond approximately to the average and standard deviation of the normalized (with respect to $\sigma_x$) coefficient (4,4) in the block discrete cosine transform (DCT) of the three reference images. Mask values below 0.1 yield artificially high perceptual distortions for the Wiener attack and were rejected. The value of $\beta$ was taken as a constant in the case of the Wiener case, and set according to (8) in the case of the mask attack. The simulation results are presented in Fig. 4.

Taking into account knowledge on the mask $\alpha$ clearly yields higher $DDS$ values, for any given distortion, than using the Gaussian assumption on $w$. Such a result must be taken in its context: the watermarking techniques considered, the assumptions on the signals and so on.

*2) Finding the Most Robust DCT Coefficients:* Finding the more robust DCT coefficients is motivated by two factors: first, to better understand the dynamics of the mask attack, and second to help derive new watermark embedding rules. The procedure utilized here comprises several steps: 1) estimate $\mu_\alpha$ and $\sigma_\alpha$ for each DCT coefficient of the $8 \times 8$ matrix and averaged over the three reference images—see Fig. 5; 2) generate the signals $x$, $m$ and $\alpha$, for each of the 100 000 simulation runs, according to the average values given by the previous

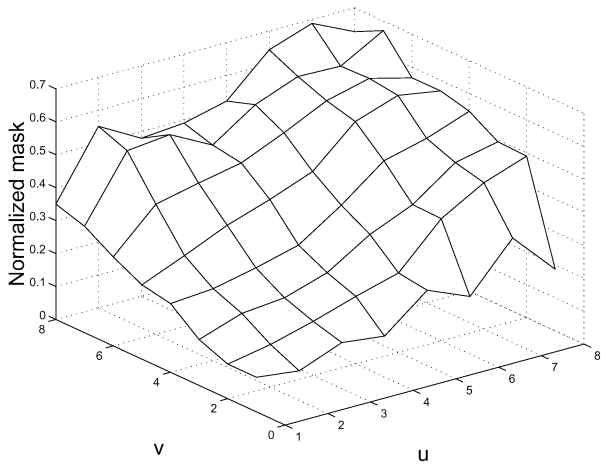



Fig. 6. Average values of $\alpha$ over the three reference images. The indices $u$ and $v$ refer to the spatial frequencies.

step; 3) compute the global detection statistic and the contribution of each of the DCT coefficient to this value; 4) increase $\beta$ until the value of $DDS$ reaches 0.5 and then 1, so that all DCT coefficients that have a negative contribution to the detection statistic (therefore finding the maximal possible detection statistic value) are removed; and 5) keep the set of the largest DCT coefficients corresponding to 80% of the maximal possible detection statistic value. The value 80% was set arbitrarily but serves as a good indication of which coefficient contribute the most to the detection statistic.

The set of the DCT coefficients identified using the above methodology is shown in Fig. 6. The two values of $DDS$ (0.5 and 1) in the experiment correspond to the threshold value of a maximum-likelihood detection, and the value at which the detection statistic is null.

A number of comments can be made: 1) high-frequency DCT coefficients are robust to weak attacks while low-frequency coefficients are robust to stronger attacks; 2) considering the case where $DDS = 0.5$ the watermark should be embedded in all coefficients, excluding the two or three coefficients at the extrema; 3) considering the case where $DDS = 1$, the watermark should be embedded in the first half of the DCT coefficients, excluding the DC coefficient for obvious perceptual distortion reasons; 4) the coefficients common to the two preceeding cases form the diagonals three to six of the DCT block matrix; and 5) the first diagonals of the DCT matrix contribute little to the detection statistic, but usually contribute much to perceptual distortions. This experiment confirms that middle frequencies form the most suited region to embed the watermark, as indirectly suggested in [5] and [17].

### E. Conclusion

The use of a mask in watermarking gives valuable information to the attacker on the location and eventually the strength of the watermark. Taking realistic hypotheses, it is possible to derive a theoretical relation between the decrease in detection statistic and the introduced perceptual distortion; the former depends mostly on the signal energy and the mask. A similar relation can be derived for the Wiener attack and, when comparing the two, the theory shows that the mask attack can be more efficient. This is confirmed through simulations using generated signals. Furthermore, for images, the use of the DCT coefficients in diagonals 3 to 6, suggested intuitively in previous literature, was validated by theoretical results. Also, it was concluded that the watermark should not be embedded in dimensions with high mask values, where knowledge on the mask is very useful to the attacker. As a rule of thumb, embedding the watermark in frequencies where the normalized mask values is above 0.5 is not recommended. Once again, these conclusions must be considered within the specified context and working assumptions.

## III. Experiments on Real Data

This section provides experimental results on the effectiveness of the mask attack on real data—sounds and images. Experiments are based on the theoretical framework laid in the previous section. The focus of these experiments is the decrease in detection statistic $(DDS)$ as a function of the introduced perceptual distortion.

### A. Data

In order to represent the variety of characteristics found in different sounds and images, a selection of three audio extracts and three image representative samples was made.

1) "Handel": extract from Handel's *Fireworks*
2) "Emma": extract from Emma Shapplin
3) "Song": *Torn* from pop singer Natalie Imbruglia

- Audio (all 16 bits at 44.1 kHz, of ms in duration):

1) "Handel": extract from Handel's *Fireworks*
2) "Emma": extract from Emma Shapplin
3) "Song": *Torn* from pop singer Natalie Imbruglia

- Images:

1) "Benz": synthetic image of an old Mercedes Benz
2) "Girl": human face
3) "Mandrill": chimpanzee with important contrast

Furthermore, for all experiments, the message is a pseudo-noise sequence of arbitrary length—yet much greater than the size of the host signal.

### B. Attacks on Sounds

The reference audio watermarking technique utilized in these experiments is that of Swanson and colleagues [14]; it makes use of both a temporal (envelope-based) and a spectral masking model. At the embedding process no energy is put into the first ten (out of 256) spectral coefficients; this would induce important perceptual distortion. Two cases are considered next: embedding the watermark in all remaining spectral coefficients (type I), or only in the first half (type II).

A trivial attack would be to low-pass filter the watermarked signal at a cutoff frequency of 10 kHz or so: in high frequencies, the global masking threshold suggests to embed significant watermark energy but most audio samples have no meaningful components there (see Fig. 12). Low-pass filtering significantly decreases the detection statistic at a very low perceptual distortion cost. Not embedding watermark at high frequencies obviously circumvents this attack but decreases considerably the watermark payload. Experimental results are not reported here.

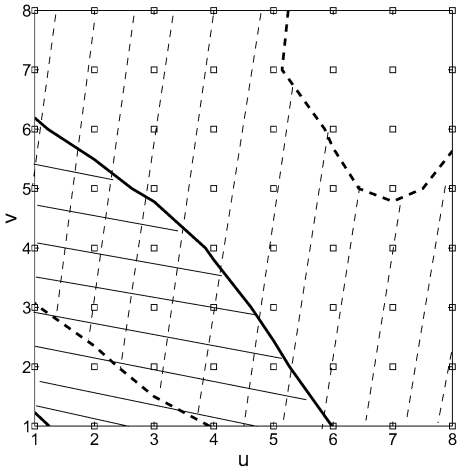



Fig. 7. Ensemble of DCT coefficients which correspond to 80% of the maximal possible detection statistic, after the mask attack. Normal lines: $DDS = 1$, dashed lines: $DDS = 0.5$

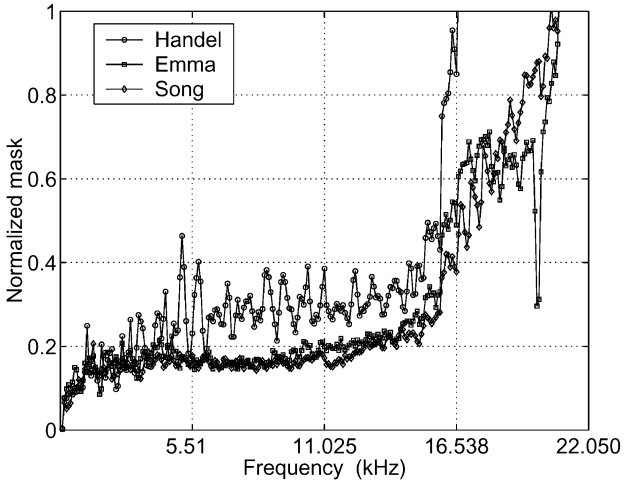



Fig. 8. Mask for the three reference audio segments. The mask is normalized, and computed as the average over all blocks.

A second attack would take advantage of the fact that tonal components identified by the mask process remain unchanged once the watermark is embedded (see Appendix). Knowledge on tone identification process allows to wisely remove the watermark in the tonal regions at little perceptual distortion cost. We do not foresee any trivial counter-measure other than computing the tonal components differently or not embedding the watermark in such regions. Experimental results are not reported here.

The third suggested attack is the mask attack introduced in Section II. The attacker's challenge is to find the tradeoff between the decrease in detection statistic and the introduced perceptual distortion.

The average normalized mask for the three reference audio samples is shown in Fig. 7; the mask values were normalized to the value of $\sigma_x$ and averaged over all blocks. As a consequence of the masking model properties, the mask values are large in the high frequencies, as expected. The mask attack is then perpetrated on the watermarked signals. The decrease in detection statistic $DDS$ as a function of the introduced perceptual distortion is shown in Fig. 8. A perceptual distortion of 4 is not heard, but moving beyond this value results in noticeable

TABLE II
EMBEDDING DISTORTIONS FOR THE THREE REFERENCE SOUNDS WHEN THE EMBEDDING WAS ON ALL COEFFICIENTS, OR WHEN ONLY THE FIRST HALF WERE SELECTED

| audio sample | $d$ with all coefs | $d$ with half the coefs |
|---|---|---|
| 'Handel' | -11.1 dB | -23.9 dB |
| 'Emma' | -9.4 dB | -22.8 dB |
| 'Song' | -18.2 dB | -29.7 dB |

perceptual distortion. No $DDS$ values greater than unity are reported since, at this value, the detection statistic is already null (the received content contains no detectable watermark). The shape of the mask in higher frequencies was already a strong indication of the expected efficiency of the mask attack. Since high frequencies contribute little to the perceptual distortion cost and can be readily attacked, one could anticipate that type-I embedding would be less robust than type-II embedding; this was confirmed by the reported results.

For the samples "Handel" and "Emma" the value of $DDS$ reaches unity for a low perceptual distortion cost while for the sample "Song" much greater perceptual distortion is necessary to obtain similar results. The mask attack was particularly effective on the sample "Handel" because mask values are out of scale. Also, much higher watermark energy was embedded in samples "Handel" and "Emma" than in "Song", as reported in Table II. The belief that "more watermark energy means better robustness", stated in a number of studies on robustness, suggests that the robustness of the watermarks embedded in the first two audio samples, since it has greater energy, would be greater than that of the third sample. Yet the reported experiments, in agreement with the theoretical framework of Section II, show the opposite.

### C. Attack on Images

The reference image watermarking technique utilized in the experiments is that of Zeng and colleagues [17]. The watermark is embedded in the DCT coefficients and the spectral contrast masking model is computed. In order to take into consideration different masking model implementations, two embedding processes are considered: embedding the watermark only in diagonals four to six of the $8 \times 8$ DCT block (type I) and embedding the watermark in all but the first three diagonals (type II).

The mask attack, taking into account a local estimation of $\sigma_x$, was successful on all reference images. The decrease in detection statistic $DDS$ is shown for different values of the introduced perceptual distortion in Fig. 9.

Let us make a few comments, validated on all three reference images: 1) for a reasonable perceptual distortion cost, the detection statistic is decreased to a point where the detector detects no watermark: the mask attack is successful; 2) the detection statistic can be made null even for reasonable introduced perceptual distortion; 3) the masking attack is most efficient for the Type-II embedding than for type I embedding; and 4) one can find perceptual distortion values below 1 as described in (8). The original, watermarked and attacked samples of "Girl" are shown in Fig. 10 when using the type-I embedding; the perceptual distortion on the attacked image was equal to 2.7.

Perpetuating the mask attack on a series of images revealed two visible artifacts, **only** when considering a $DSS$ close to

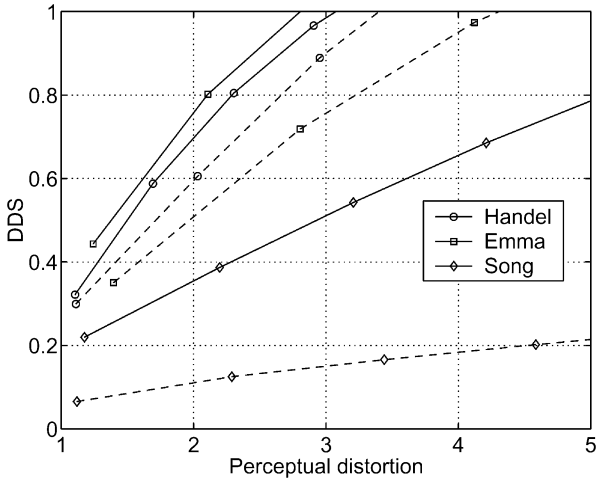



Fig. 9. Experimental results: $DDS$ versus introduced distortion for the three reference sounds. Solid and dashed lines refer to Type-I and Type-II embedding, respectively.

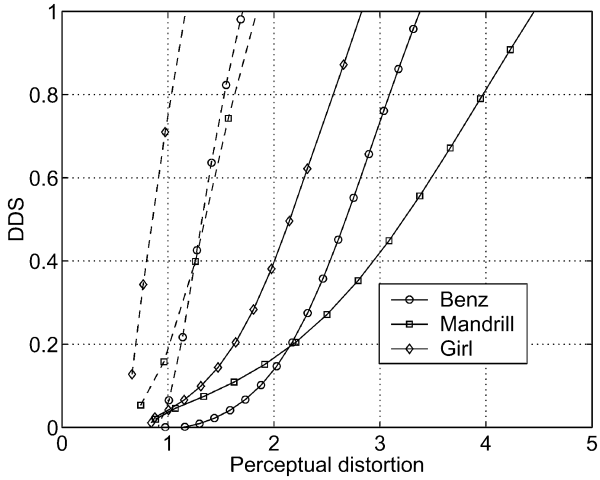



Fig. 10. Experimental results: $DDS$ versus introduced distortion for the three reference iamgess. Solid and dashed lines refer to Type-I and Type-II embedding, respectively.

unity: 1) there is a blurring of the image, which can be corrected using standard algorithms such as an asymmetric high-pass filters and 2) one can observe a slight change in luminance; no simple correction is foreseen. This is particularly true for the image "Mandrill" and can be partially explained by results shown in Fig. 9: the perceptual distortion is greatest for this image when the $DDS$ is close to unity.

Contour plots of the set of DCT coefficients that contribute to 80% of the maximal detection statistic value are shown in Fig. 11 for each reference images. One can deduce from this graph that the "best" general location for watermark embedding, with respect to the mask attack, is the middle frequencies. From our experiments we can conclude: a) little watermark energy should be embedded at low-frequencies because the introduced perceptual distortion is important; b) considerable energy can be embedded at high frequencies but the watermark can be removed (and even "reversed"); therefore c) placing the watermark in the middle frequencies seems to optimize the tradeoff between embedded energy, introduced distortion and robustness. This result, based on experiments with real data, concurs with the theoretically derived "optimal" embedding place (Section II) and with insights given in [17].

### D. Conclusions

Experiments on real data, considering additive watermarking techniques using masking models, confirmed the theoretical calculation of Section II: the mask attack successfully decreases the detection statistic at a low perceptual distortion cost. Even more so, the reeiver was sometimes unable to detect any trace of the watermark. These results reasonably question the systematic use of masking models in watermarking, and shed some light on the design tradeoff between distortion and robustness.

## IV. Discussion

### A. Forgetting the Origin of Masking Models

The publicly available masking models utilized in watermarking are inherited from data compression applications. Using the same models to shape a watermark in order to have maximal SNR while remaining imperceptible seems a reasonably good strategy; we have however shown in this paper that, in most cases, too much information on the watermark is revealed to a potential attacker.

Masking models identify perceptually significant regions of the signal—regions that should not be altered when embedding the watermark. It is reasonable to assume that the mask of the watermarked signal is very similar to the one of the original signal (this assumption was validated in our experiments). In other words, the attacker gains knowledge on the location of the watermark by computing the mask of the watermarked signal. Also, since masking models are usually derived locally (they are computed over short data segments and are based on local properties of the signal) the attacker can target precise and confined regions within the signal where high watermark energy can be expected.

As our understanding of audio and image perception improves and as the quest for higher data compression ratio continues, masking models will become increasingly accurate and will continue to improve. Two points are worth noting. First, watermarking, which uses imperceptible regions of the signal to hide information, will trail behind these increasingly accurate masking models which, in turn, aim at defining regions that are useless to the observer. For example, an ideal lossless compression scheme would ensure perfect content fidelity and leave no room to embed a watermark. Second, there is a legacy issue: watermarks embedded using today's state-of-the-art masking models may be removed (or severely damaged) when next-generation models are used to further compress existing watermarked data.

Using masking models allows the embedding of high SNR watermarks, especially in certain locations, while preserving the fidelity of the watermarked content. While this a very attractive strategy, one must not forget that it is the cover content that hides the watermark; by allowing high SNRs in certain locations, the watermark is no longer "hidden" in the content and therefore becomes exposed and vulnerable to an attacker.

Fig. 11. From left to right: original, watermarked, and attacked image.

### B. The Efficiency of the Mask Attack

A number of successful attacks on watermarking systems have been reported in the literature. This includes tailored attacks against selected techniques, general attacks such as cropping, scaling, distorting the data (i.e., Stirmark) and, more recently, estimate and removal attacks—sometimes called copy attacks [8], [15]. Although the assumptions on the signals are not always clearly stated, these attacks are successful. This paper introduced a new estimate-and-remove attack: the mask attack.

The mask attack was shown to significantly decrease the detection statistic for small perceptual distortion cost for watermarking techniques that make use of publicly available masking models. What makes the attack very effective is the use of the knowledge on the mask (the mask of the original and watermarked content are much correlated): not only does the attacker gain privilege information on the most probable location of the watermark, he can also locally evaluate its strength in each dimension (each region). This enables a better estimate of the watermark than, for example, when the Wiener filter assumes a zero-mean, Gaussian distribution watermark. In addition, the perceptual distortion introduced by attack can be better controlled.

The mask attack was successful on audio and image data alike. The results indicate that the watermark detection fails (values of $DDS$ approach unity) for reasonable perceptual distortion levels. The agnosticism of the attack to the different masking methodology is not surprising since the working assumptions are general enough to cope with both models considered in this study; in fact similar conclusions are expected for any masking model. The mask attack was more successful on images than on audio signals. This can be explained, in part, by three factors: 1) audio coefficients in the transform domain have much larger dynamics than their image counterpart; 2) in audio, the number of coefficients that contributes to the detection statistic is large, conversely to the case of images in which the contribution mostly comes from the middle frequency coefficients; and 3) there was more watermark energy embedded in audio than there was in images resulting in the mask assumption—$\alpha(y) = \alpha(x)$—being not as strongly validated.

### C. The Fidelity-Robustness Tradeoff

The tradeoff between robustness and fidelity is not properly addressed in the literature. On one hand, *ad-hoc* methods are used to embed the watermark in regions where it will be both transparent and "quite robust". On the other hand, theoretically-based techniques which are optimally robust with respect to some criteria (function of the hypotheses used to derive them) are usually based on a SNR distortion measure which does not guarantee, in any way, transparency. Therefore, there seems to be a missing theoretical link between robustness and distortion. A key factor is finding the analytical expression of perceptual distortion in relation to theoretical hypotheses on the signals.

The common belief that more watermark energy means more robust watermark, often stated in the literature, was challenged by the reported experiments on real data. For example, the belief would be that the watermark embedded in the two audio samples "Handel" and "Emma" be more robust than the one embedded in the sample "Song" because more energy was embedded. Yet, in agreement with the theoretical framework presented in Section II, the opposite conclusion was verified. This shows that a link between robustness and perceptual distortion must definitely be found.

Accordingly, a first step in that direction was suggested in this paper: an analytical closed-form equation was found for the general masking model and results derived on the effectiveness of the mask attack. Theoretical simulations as well as experiments on real data allowed to identify the set of dimensions which are most robust to the attack. The mid-frequencies are best for images (see Fig. 7) while audio signals benefited from a watermark spread over a larger number of dimensions (there is no specific region for sounds as the variety of local energy with respect to dimension is much more random than it is for images). This is in agreement with *ad-hoc* suggested techniques [17].

A number of theoretical work and results giving priority to robustness exist. For example, given a SNR distortion constraint, one can derive an optimal watermarking rule that minimizes the detection error probability [12]. The tradeoff solution indicates the watermark should be embedded in selected regions of the signal for which the watermark to signal ratio will be optimized at the detector, hence optimizing the detection statistic. There is no guarantee of transparency. A masking model could eventually be used to assess if the introduced watermark is "too" noticeable. An interesting fact in that framework is that the attacker has no choice but to distort significantly the content in order to remove the watermark. Other similar methods are presented in [4], [13]. These work could likely integrate perceptual distortion measures to better take into account the fidelity aspect of watermarking.

## V. Conclusion

Watermarking is a tradeoff between fidelity, robustness, and payload. The relation between the last two parameters has been studied in previous literature and solutions have been suggested. Fidelity is often attained by using masking models, but, when doing so, this papers shows that not considering robustness allows the attacker to benefit from key information on the location and strength of the watermark. The end result is the derivation of the *mask attack*. Let us conclude with the following comments.

- Masking models were introduced in data compression to shape the quantization noise—a meaningless signal. The consequence of using such models in watermarking to control perceptible distortion when shaping the embedded watermark, a meaningful and potentially vulnerable signal, was not studied so far. The scope of the paper was to link the usage of a mask with the system's robustness.
- The mask attack makes use of global and local knowledge on the signals, or perceptual properties deduced from the masking models, to estimate and remove the embedded watermark. The mask attack is derived from a theoretical framework whose main battle ground lies in the relation between the decrease in detection statistic and the introduced perceptual distortion. Theoretical simulations and experiments on real data successfully demonstrated the efficiency of the attack for audio and images alike; the attack was shown more efficient than the Wiener attack in most cases. To best counteract the mask attack, image watermarks should be embedded in the middle frequencies (when the DCT transform is used). For audio signals, however, the only conclusion was to avoid high frequencies and dominant tonal components.
- Embedding a robust, imperceptible watermark that carries a lot of information is indeed quite a challenge. There is a need for theoretical grounds to derive analytical tradeoff equations that globally take into accounts all three parameters robustness, capacity and fidelity. The authors hope to have achieved a first step in that direction here. However, given the variety of application scenarios and attacks, the numerous types of watermarking methods, and the difficulty to determine "objective" perceptual models and distances, many other research avenues are possible.

## Appendix
## Masking Models for Sounds and Images

This appendix describes the masking models most commonly used in audio and image watermarking, and specifically used in reference techniques in Section III. This should not be regarded as a tutorial and references are given for further reading.

### A. The Mask

The *mask* is a matrix of values computed on individual processing blocks; it modulates a typical white spectrum watermark at the embedding process. In its simplest form, the mask is the identity matrix multiplied by a small constant; the watermark is spread over the entire transform domain but has very small energy: its detection is impaired by any small modification to the signal. Using a slightly more sophisticated mask, the watermark would be selectively embeded in particular regions in the transform domain (the energy of the watermark in that region can be large), and put as "zero" elsewhere (the mask has the null values); taking the example of images, the watermark could be embedded only in spectral regions where the values of the DCT coefficients are above a given threshold.

Watermarking techniques usually make use of much more sophisticated masks, adapted from data compression applications and based on findings on our hearing or visual systems. The mask identifies regions of the signal that are not perceived when in presence of the main stimulus (the sound or the image); these regions may correspond to temporal windows or spectral intervals, depending on the masking model. Typicaly, masks are computed on individual processing blocks of 20 milliseconds for audio samples, and $8 \times 8$ pixels for images. Considering an additive watermarking scheme, a mask is computed for each successive blocks and used to modulate the watermark. This method ensures that the embedded watermark be transparent (not perceived).

### B. Audio Masking Model

After years of perceptual experiments on humans and golden ears, the MPEG group has developed the reference temporal and spectral audio masking models. They are use in the MPEG compression algorithms [2], [3].

The temporal audio masking model identifies temporal windows, before and after the occurrence of a tonal stimuli, within which no other tone (and more generally no other stimuli) of lower intensity can be heard. The pre-echo (before the tonal stimuli) and the post echo (after the tonal stimuli) can be modeled respectively as rising and falling exponentials with different time constants. One can approximate the MPEG model by computing the short term amplitude envelope of the original signal. In a first watermarking technique exploiting the temporal masking, artificial echoes are added within the original signal at the appropriate time and retrieved at the receiver side since their timing is known [1]. Because of the predictability of the scheme a tailored attack was soon suggested [10]. In a second technique based on temporal masks, the short-term envelope of the original signal modulates the watermark [14]; one advantage of this technique is that no watermark is embedded into silence segments.

The spectral audio masking model is based on the following observation: in presence of a tone at frequency $f$, humans do not hear tones at frequencies below and above $f$ if their intensity level is below a masking threshold. The shape of this threshold function was determined by perceptual experiments using tonal stimuli; additional experiments with noise were also conducted for wider spectrum stimuli. The spectral mask must be computed on segments of audio signals which can be considered as stationary, typically 20–30 ms. Spectal masking thresholds for tones at 500 (left) and 5000 (right) Hz are illustrated in Fig. 12. The most widely used spectral masking model is computed according to the following steps (detailed in [14]):

1) divide the audio signal into nonoverlapping segments of 20 ms in duration; then, for each segment;
2) compute the short-term power spectrum;

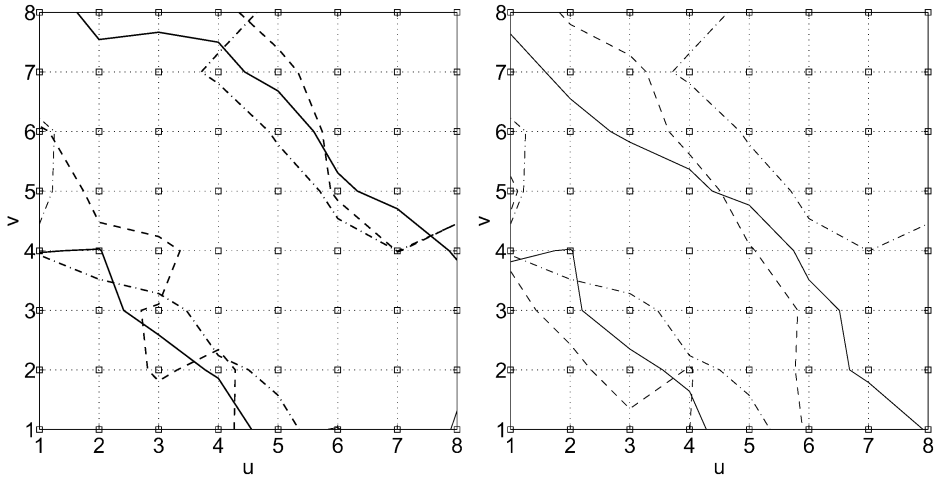



Fig. 12. Contour plots of the ensemble of DCT coefficients which contribute to 80% of the maximal detection statistic value at the receiver, for the three reference images. Indexes $u$ and $v$ correspond to the spatial frequencies. Left: $\overline{DDS} = 0.5$. Right: $\overline{DDS} = 1$.

3) identify the tonal (pure tones) and nontonal (noise) components;
4) remove the masked components, including all sounds below the absolute hearing threshold and those tonal components that are too close to one another;
5) compute the individual masking thresholds, by accounting for the frequency masking effects of the auditory system;
6) deduce the global masking threshold, which is function of the individual and of the absolute ones;
7) repeat from step 2.

As shown in Fig. 12, the masked region is usually significant. As most implementations see the watermark embedded at 3–6 dB below the global masking threshold, to guarantee "complete" inaudibility of the watermark, the WSR (watermark to signal ratio) is usually high. Yet, the spectral masking model was not designed for use in watermarking applications leading to inherent drawbacks if the model is used as such. For example the high mask values at high frequencies enables effective tailored attacks, as shown in Section III.

### C. Image Masking Model

Image masking models also result from perceptual studies and from our understanding of the human visual system—better known than the hearing system. Both spatial (pixel) and transform (DFT, wavelet) domain masking models were proposed in the literature.

Spatial image masking models rely on two independent observations. First, as humans unequally detect changes in colors, primary colors can be altered (watermarked) differently; for example, a high energy watermark can be embedded in the blue channel of an RGB image with little perceptual distortion since humans are significantly less sensitive to changes in that channel. Second, exploiting image contrast properties, one can embed a watermark in regions where there are large changes in color gradients. This method, conversely to the first, is image-dependent. Both these models have been exploited in image watermarking techniques [7], [11].

Spectral image masking models have also been developed based on a block-wise (usually $8 \times 8$ pixels) decomposition of the image. In the first approach the transform coefficients with the least significant contributions (usually those with smallest value) were retained to embed the watermark according to the following rule in the transform domain: $\hat{I} = I \cdot [1 + \alpha w]$. The coefficient $\alpha$ may be a constant, or be dependent on the coefficient's position, or value. Later on, the *frequency sensibility* was taken into account according to the modulation transfer functions reported at the human visual system level. These functions describe the sensitivity of our eyes to sine waves of different energy and spectral location. From these modulation functions, given the viewing distance, one may determine the *just noticeable difference-JND* threshold at each frequency bin. These thresholds serve both quantization and bit-allocation purposes in compression algorithms. The resulting masking model is image independent.

Two refinements using image dependent information improved spectral masking. First, luminance sensitivity—or the ability of the eye to detect noise on an uniform background—strongly depends on the average luminance of the image and that of the noise. The spectral mask can be re-adjusted by computing the ratio between local luminance values (estimated by the DC coefficient of each block) and the average luminance of the image (the average of all DC coefficients). Second, contrast sensitivity—or the detectibility of an image component in presence of another component—is strongest when the two components have similar frequency, location and orientation. The contrast masking allows to take into account our particular perception of the high frequencies and the texture regions of the image.

The widely used image spectral masking model is determined by computing the following steps:

1) divide the image into blocks of $8 \times 8$ pixels;
2) for each block, compute the spectral coefficients in the luminance domain;
3) for each block, compute the threshold values $T_f(u, v)$ given for example by the difference between the value of the spectral coefficients of the original image and the image compressed with JPEG at a quality factor of 75;
4) compute the luminance sensitivity $T_l(u, v, b) = T_f(u, v) \cdot (X_{DC,b}/\bar{X}_{DC})^a$ where $X_{DC}$ is the DC coefficient and

$a$ is a parameter which controls the degree of luminance sensitivity, usually taken as 0.649;

5) compute the contrast masking threshold (referred to as the JND) as: $T_c(u,v,b) = \max[T_l(u,v,b), T_l(u,v,b) \cdot (X_{u,v,b}/T_l(u,v,b))^w]$, where $0 < w < 1$ and is typically taken as 0.7.

This model, as well as others, are detailed in [16] and [17]. In most image watermarking implementations, the embedding is limited to spectral coefficients in the middle frequencies. Changes in the DC component would cause significant artifacts, and high frequencies are subject to drastic manipulations from standard compression algorithms.

## REFERENCES


[1] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Syst. J.* , vol. 35, 1996.

[2] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "Iso/iec mpeg-2 advanced audio coding," in *Audio Engineering Society 101st Conv.*, Los Angeles, CA, Nov. 1996.

[3] ——, "ISO/IEC MPEG-2 advanced audio coding," *J. AES*, vol. 45, pp. 789–814, 1997.

[4] B. Chen and G. Wornell, "Dither modulation: A new approach to digital watermarking and information embedding," in *SPIE—Security and Watermarking of Multimedia Content*, Los Angeles, CA, Jan. 1999.

[5] J. R. Hernandez, M. amado, and F. Perez-Gonzalez, "DCT-domain watermarking techniques for still images: Detector performance analysis and a new structure," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 55–68, Jan. 2000.

[6] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[7] M. Kutter, "Watermarking resisting to translation, rotation, and scaling," in *Proc. SPIE—Security and Watermarking of Multimedia Content*, San Jose, CA, Jan. 1998.

[8] M. Kutter, S. Voloshynovski, and A. Herrigel, "The watermark copy attack," in *Proc. SPIE—Security and Watermarking of Multimedia Content*, San Jose, CA, Jan. 2000.

[9] G. Langelaar, R. Lagendijk, and J. Biemond, "Removing spatial spread spectrum watermarks by nonlinear filtering," in *Proc. European Signal Processing Conf. (EUSIPCO 98)*, Rhodes, Greece, 1998.

[10] F. Petitcolas, R. Anderson, and M. Kuhn, "Attacks on copyright marking systems," in *Proc. 2nd Workshop on Information Hiding*, OR, May 1998.

[11] C. I. Podilchuk, "Digital image watermarking using visual models," in *Proc. Electronic Imaging*, San Jose, CA, 1996.

[12] A. Robert and R. Knopp, "Watermarking and detection theory," in *Proc. SPIE—Security and Watermarking of Multimedia Content*, San Jose, CA, Jan. 2000.

[13] J. Su and B. Girod, "Fundamental performance limits of power-spectrum condition-compliant watermarks," in *Proc. SPIE—Security and Watermarking of Multimedia Content*, San Jose, CA, Jan. 2000.

[14] M. Swanson, B. Zhu, A. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Process.*, vol. 66, pp. 337–355, 1998.

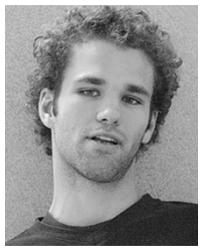
[15] S. Voloshynovski, S. Pereira, A. Herrigel, N. Baumgartner, and T. Pun, "Generalized watermarking attack based on watermark estimation and perceptual remodulation," in *Proc. SPIE—Security and Watermarking of Multimedia Content*, San Jose, CA, Jan. 2000.

[16] R. Wolfgang, C. Podilchuk, and E. Delp, "Perceptual watermarks for digital image and video," *IEEE Trans. Image Process.*, vol. 87, no. 7, pp. 1108–1126, Jul. 1999.

[17] W. Zeng and B. Liu, "A statistical watermark detection technique without using original images for resolving rightful ownerships of digital images," *IEEE Trans. Image Process.*, vol. 8, no. 11, pp. 1534–1548, Nov. 1999.


**Arnaud Robert** received the B.Sc. degree from Ecole Polytechnique de Montreal, Montreal, QC, Canada, in 1996, and both the M.Sc (signal processing, 1996) and Ph.D. (computer science, 1999) degrees from the Swiss Federal Institute of Technology Lausanne (EPFL).

He then spent a year at the Audio-Visual Communications Laboratoy (EPFL) as a First Assistant, where he focused on watermarking. He has since been involved in many areas of content protection—including watermarking, conditional access, content protection, DRM and CD/DVD protection—with positions at NagraVision-Kudleski, Microsoft, and recently Thomson-Technicolor, Los Angeles, CA, where he is the Vice President of Content Security. He has (co-)authored over 30 papers in content protection and perceptual models and is the inventor of eight patents in content security.



**Justin Picard** received the M.Sc.A. degree in electronics from the Ecole Polytechnique de Montreal, Montreal, QC, Canada, in 1997 and the Ph.D. degree in computer science from the University of Neuchâtel, Neuchâtel, Switzerland, in 2000.

He is now Head of Research at Thomson-MediaSec, Essen, Germany. He previously was with the Laboratory of Nonlinear Systems, Swiss Federal Institute of Technology, Lausanne. He has (co-)authored 20 papers in the areas of digital watermarking, information retrieval, and uncertain reasoning. He is currently active in the area of printed document authentication using signal processing and digital watermarking techniques.