ROBUST AUDIO WATERMARK DECODING BY SUPERVISED LEARNING†

Serap Kırbız *Bilge Günsel*

Multimedia Signal Processing and Pattern Recognition Lab.
Dept. of Electronics and Communications Eng. Istanbul Technical University
34469 Istanbul, Turkey
web: http://www.ehb.itu.edu.tr/~bgunsel/mspr

ABSTRACT

Most of the watermark (WM) decoding schemes use correlation-based methods because of their simplicity. In these methods, the WM signal embedded through a secret key is assumed as uncorrelated with the host signal. This is a hard restriction that can never be achieved and correlation between the received signal and the secret key becomes greater than zero even though the received signal is un-watermarked. Mostly a decision threshold specified semi-automatically is used at the decoding site. Since the audio watermarking is a nonlinear process that guarantees the inaudibility, there is no analytic way of determining an optimal threshold value that makes the WM decoding problem harder. This paper introduces a learning scheme followed by a nonlinear classification thus eliminates the threshold specification problem. The decoding process is modelled as a three-class classification problem and Support Vector Machines (SVMs) are used in the learning of the embedded data. The decoding and detection performances of the developed system are greater than 98% and 95%, respectively. When the Watermark-to-Signal-Ratio (WSR) is higher than -30dB, system false alarm ratios remain less than 2%. It is shown that the introduced WM decoding method is robust to additive noise and most of add/remove and filter attacks of *Stirmark*.

1. INTRODUCTION

Recently, distribution of audio data in digital form became easier and more extensive, that makes the copyright protection much more difficult. Audio watermarking techniques are proposed to ensure the IP rights by embedding ownership information into the host data, while preserving originality. Accurate decoding of the embedded watermark (WM) information is a challenging problem in audio watermarking and many techniques have been proposed for this.

In the literature, watermark decoding [1,2,3,4,5,6] and watermark detection are often considered as separate problems. In most of the decoding methods correlation-based decision rules are used because of their simplicity [1,2,3,4]. The lack of these systems is that, the WM decoding performance relies on the accuracy of the calculated correlation between watermarked and embedded key signals. Higher the correlation, lower the un-extracted WM data. On the other hand, there is a trade-off between the correlation and the audibility. Another difficulty with the correlation-based methods is that, they do not allow accurate identification of the watermarked and un-watermarked audio clips that is required in many applications, i.e., on-line broadcast monitoring and royalty tracking.

In this paper, the WM decoding and detection problems are integrated into a unique classification problem and supervised learning of the embedded WM data is introduced. Due to the good learning capability, SVMs[7] are used in the training stage. In the literature, there are some preliminary works that use SVMs for image watermark decoding, i.e. it is used for logo detection where the intensity level differences of the pixels' blue components are used for the training of SVMs [5]. In our previous work [6], a binary SVM classifier is proposed for audio watermark decoding but not detection.

Unlike the existing methods, this paper proposes a learning-based audio watermark classification scheme which is capable of correctly extracting the WM bits while simultaneously detecting the un-watermarked audio frames. Test results demonstrate that performance of the introduced integrated technique outperforms state-of-the-art correlation-based techniques [2, 3, 4] and it is robust to noise attacks as well as several Stirmark [8] attacks.

2. ADAPTIVE WATERMARK EMBEDDING

An adaptive spread spectrum audio watermarking scheme [2, 3] that is compatible to MPEG Layer 3 Model 2 (mp3) audio compression standard is used for embedding the WM information.

Let $\mathbf{s}_i$ refer to the i*th* frame of the input audio signal. At each instant, the encoder takes an original audio frame, $\mathbf{s}_i$, as its input and transmits the corresponding watermarked frame, $\mathbf{s}_{i_{WM}}$, over the communication channel. The watermarked audio frame is formulated as in Eq.(1),

$$\mathbf{s}_{i_{WM}} = \mathbf{s}_i + w_j \lambda f(\mathbf{s}_i, \mathbf{k}) = \mathbf{s}_i + w_j \mathbf{k}_i ,$$
$$i = 1, ..., (LxRP) , \; j = 1, ..., L . \qquad (1)$$

where Refresh Period (*RP*) refers to the number of block insertions and λ is a parameter that controls the power of the embedded watermark. In Eq.(1), WM bit w_j can be either +1 or -1 where j=1,…L and L is the length of the watermark block. $\mathbf{k}$ refers to the secret key sequence with zero mean generated by a Pseudo Noise generator (PN). $f(\mathbf{s}_i, \mathbf{k})$ is a nonlinear function of the input audio signal, $\mathbf{s}_i$, and the secret key $\mathbf{k}$ that models the embedded data. Our encoder applies an iterative approach that allows specifying a nonlinear $f(.)$ in a data adaptive way [2, 3]. In Eq.(1), $w_j \mathbf{k}_i$ models the nonlinear distortions, where $\mathbf{k}_i$ is the shaped key signal that is embedded into the audio frame *i,* after multiplied by w_j. The WM encoder generates $\mathbf{k}_i$ by shaping the secret key sequence $\mathbf{k}$ according to masking thresholds obtained by psychoacoustic masking of $\mathbf{s}_i$. In [4], an analytic approach to analyze a linear $f(.)$ is introduced.

3. DIFFICULTIES WITH TRADITIONAL METHODS

The traditional audio watermark decoding schemes mostly use correlation-based methods. In these methods, the watermark detection and decoding are often considered separately [1,2,3,4,5,6] and mostly decoding performance is declared. This is because of the difficulties in the specification of detection threshold. In the con-

† This work is supported by TUBITAK BAYG and TUBITAK EEAG.

text of this paper, we make the distinction between WM detection and decoding. The term WM detection is used to denote the ability of the decoding algorithm to declare the presence or absence of a watermark on an audio. Whenever the algorithm declares the audio is watermarked, the embedded WM is decoded. This is important for the considered applications, i.e., broadcast monitoring, royalty tracking, etc.

WM detection can be considered as a hypothesis testing problem, and the two hypotheses are being:

H_0: the audio under test does not host the watermark under investigation

H_1: the audio under test hosts the watermark under investigation

In spread spectrum watermarking, hypothesis H_1 can be further divided into two sub-hypotheses:

H_{1a} : the audio under test is watermarked by +1

H_{1b} : the audio under test is watermarked by -1

Eq.(2) defines the correlation function between the received audio frame, $\mathbf{s}_{i_R} = \mathbf{s}_i + w_j \mathbf{k}_i + \mathbf{r}$, and the secret key signal **k,** for *i*th frame;

$$c_i = \sum_{n=1}^{N} k(n) s_{i_R}(n) = \sum_{n=1}^{N} k(n) s_i(n) + \sum_{n=1}^{N} w_j k(n) k_i(n) + \sum_{n=1}^{N} k(n) r(n) \quad (2)$$

Since **k** is a PN signal which should be un-correlated with $\mathbf{s}_i$ and the additive channel noise **r**, in ideal case, $\sum_{n=1}^{N} k(n) s_i(n) \approx 0$ and $\sum_{n=1}^{N} k(n) r(n) \approx 0$.

In order to eliminate the noise we applied wavelet-denoising on the received signal, therefore the correlation is calculated as;

$$c_i = \sum_{n=1}^{N} k(n) t_i(n) \quad (3)$$

where

$$\mathbf{t}_i = W^{-1}\left(\Lambda_h\left(\mathbf{W}(\mathbf{s}_{i_R})\right)\right) \quad (4)$$

Detailed explanation of Eq.(4) is given in Section 4.

Consequently, w_j , the WM bit embedded into frame *i* can be estimated according to the decision rule given in Eq.(5):

$$F(\mathbf{s}_{i_R}) = \begin{cases} 0, & \text{if } |c_i| < thr \\ w_j = \text{sgn}(c_i), & \text{if } |c_i| \geq thr \end{cases} \quad (5)$$

In Eq.(5), *thr* refers to the decision threshold that means; if the correlation value is less than *thr,* H_0 is accepted; if it is greater than *thr,* H_1 is accepted. If the watermark is detected, the sign of c_i specifies the embedded WM bit. Thus the decision highly depends on the threshold value. There is no analytical way to specify the optimal value of *thr*, therefore it is either taken as equal to zero or chosen heuristically. If *thr is* specified as equal to zero, the correlation-based decoder will not be able to detect the un-watermarked audio frames. On the other hand, since there is a trade off between the correct decision probability and false alarm ratio, it is difficult to specify the threshold values heuristically. In order to minimize the false alarms, the value of *thr* is set to a small value that makes the detection of un-watermarked clips unfeasible. Furthermore, in practice, neither **k** and $\mathbf{s}_i$, nor **k** and **r** can be chosen as uncorrelated, that also reduces the WM extraction accuracy of the decoder. In order to eliminate these fundamental problems of the existing correlation-based audio WM decoders, in the next section, we introduce a SVM-based learning and classification method for audio WM decoding.

4. AUDIO WATERMARK DECODING BY SVMS

In this paper, an integrated audio WM detection and decoding scheme that performs a SVM-based supervised learning followed by a blind decoding is introduced. The decoding process is modelled as a three-class classification procedure. Initially, wavelet decomposition is performed on the audio signals, and the decomposed audio frames watermarked with +1 and -1 constitute Class 1 and Class 2, respectively. The proposed method is not only capable of correctly decoding the embedded WM bits but is also capable of detecting un-watermarked audio frames defined as Class 3.

4.1. Extraction of Training Vectors

The proposed decoding algorithm first performs wavelet decomposition on the audio signals collected in the training data set. The idea behind using the wavelet decomposition is twofold: first, elimination of noise by wavelet denoising; second, reducing the computational complexity, because the embedded WM data are dominant in the high frequency components, thus in detail coefficients of the wavelet transformed signal [2].

Let $\mathbf{s}_{i_{WM}}$, the watermarked audio frame is first decomposed into approximation and detail parts by using the Daubechies-4 wavelets:

$$\boldsymbol{W}\left(\mathbf{s}_{i_{WM}}\right) = \mathbf{e}_{i_{WM}} + \mathbf{d}_{i_{WM}}, \qquad i = 1, ..., l \quad (6)$$

where *W* denotes the wavelet transform, *l* is the number of the training vectors, $\mathbf{e}_{i_{WM}}$ and $\mathbf{d}_{i_{WM}}$ refer to the approximation and detail coefficients of the watermarked signal, respectively. Note that, for un-watermarked case, there is no WM information, so $\mathbf{d}_i$ is the detail coefficients of the decomposed original signal $\mathbf{s}_i$.

Let the feature vectors, $\mathbf{t}_i$, *i*=1,..,*l*, constitute the *N* dimensional training vectors for the SVM classifier, where *l* refers to the number of training vectors. Then, *i*th training vector can be obtained by taking the inverse wavelet transform of the detail coefficients as described by Eq.(7):

$$\mathbf{t}_i = \begin{cases} W^{-1}\left(\mathbf{d}_{i_{WM}}\right), & \textit{for Class1 and Class2} \\ W^{-1}\left(\mathbf{d}_i\right), & \textit{for Class3} \end{cases}, \quad i=1,..,l \quad (7)$$

where W^{-1} denotes the inverse Wavelet transform.

4.2. Training the SVM for Three-Class Classification

Due to the good learning capability, SVMs are used in the training stage. Originally, the SVM classifier is designed for binary classification [7]. Given a training set $\mathbf{T}=\{(\mathbf{t}_1, y_1), ..., (\mathbf{t}_l, y_l)\}$, where $\mathbf{t}_i \in R^N$ is an *N*-dimensional feature vector and $y_i \in \{-1, +1\}$ is a class label, the aim of the SVM training is to find an optimal hyper-plane, $\mathbf{a} \cdot \mathbf{t} + b = 0$, where **a** is normal to the decision hyper-plane, $2 / |\mathbf{a}|$ is the margin, and $|b| / ||\mathbf{a}||$ is the perpendicular distance from the decision hyper-plane to the origin. The optimal SVM classifier that maximizes the margin is designed by maximizing the Wolfe dual [7] of the Lagrange functional given in Eq.(8),

$$\max_{\alpha} W(\alpha) = \max_{\alpha}\left(\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \mathrm{K}(\mathbf{t}_i \cdot \mathbf{t}_j)\right) \quad (8)$$

subject to constraints

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C, \quad i = 1, ..., l, \tag{9}$$

where α_i is the ith Lagrange multiplier corresponding to the ith training vector. If the training set is not separable, deviations of the misclassified samples from the decision boundary is controlled by the misclassification cost parameter C where C defines an upper bound for the Lagrange multipliers, α_i, $i = 1, ..., l$.

In this work, because of the nonlinear nature of the audio watermark decoding problem, a nonlinear SVM classifier is designed by using a Gaussian Radial Basis Function (RBF) kernel. The Gaussian RBF kernel is defined as $K(\mathbf{t}_i, \mathbf{t}_j) = e^{-\|\mathbf{t}_i - \mathbf{t}_j\|^2 / 2\sigma^2}$, where σ is the width of the RBF kernel [7].

Since the considered problem *is* a M=3-class classification problem, the classification is achieved by transforming it into $(M-1)M/2$ binary classification based on the one-against-one method [7]. Therefore, the training set $\mathbf{T}=\{(\mathbf{t}_1, y_1), ..., (\mathbf{t}_l, y_l)\}$ is formed by assigning the class label $y_i \in \{+1, -1, 0\}$ to each training vector $\mathbf{t}_i$, obtained by the wavelet decomposition of ith audio frame. For each class pair, the SVM classifier is trained with the training vectors coming from two corresponding classes. The hyper-plane parameters $\mathbf{a}$ and b, that determine the decision surface, and the support vectors $\mathbf{t}_s \in SV$, that correspond to $\alpha_s > 0$ where $SV \subseteq \mathbf{T}$ are obtained.

In order to evaluate the classification performance tendency to selection of the training vectors, the training set $\mathbf{T}$ is formed in two different ways. In the first case, all of the training vectors are collected from a single audio clip, and training of the SVM classifier is achieved where l is determined with the best adaptation between the classification accuracy and the computational complexity. In the second case, l / 10 training vectors are collected from 10 different audio files. It is observed that, performance of the introduced audio WM decoder does not rely on the selection of the training samples.

4.3. Classification of the Audio Frames

Let $S = \{\mathbf{t}_1, ..., \mathbf{t}_u\}$ denote our test set where $\mathbf{t}_i$, $i = 1, ..., u$, is an N-dimensional test vector. In order to obtain the test vector $\mathbf{t}_i$, the received signal $\mathbf{s}_{i_R}$ is first decomposed into its detail $\mathbf{d}_{i_R}$ and approximation $\mathbf{e}_{i_R}$ parts by applying wavelet transform. In order to eliminate channel noise, the detail coefficients of decomposed signal, $\mathbf{d}_{i_R}$, are thresholded before taking the inverse wavelet transform as in Eq. (10);

$$\mathbf{t}_i = W^{-1}\left(\Lambda_h\left(\mathbf{d}_{i_R}\right)\right), \quad i = 1, ..., u \tag{10}$$

where Λ_h refers to the thresholding operation thus eliminates the coefficients less than a threshold h. In this work, Λ_h simply eliminates half of the coefficients.

The pair-wise classification of the test vectors is performed according to Eq.(11),

$$F(\mathbf{t}_i) = \text{sgn}\left(\sum_{s \in SV} \alpha_s y_s K(\mathbf{t}_s, \mathbf{t}_i) + \bar{b}\right) \tag{11}$$

where $F(.)$ describes the decision rule of the binary classifier, $\mathbf{t}_i$ is the considered test vector, SV is the support vector set determined at the training stage, $\mathbf{t}_s \in SV$, is the support vector that corresponds to $\alpha_s > 0$, and $\bar{b}$ is the bias term obtained by the SVM training. After the pair-wise classification, the final decision has been made by voting using the "Max-Wins" strategy which assigns the test vector to the class which has the largest number of votes [7].

5. TEST RESULTS

5.1. Test Data and Performance Measures

A test data set is prepared by sampling various speech and music files at 44.1 kHz (16 bits/sample, N=1024). The test set consists of watermarked and un-watermarked audio files in total length of 15 hours. Watermark embedding within a 0-22050 Hz frequency band is achieved by using the adaptive WM encoder with a WM sequence of length L = 15 bits.

Watermark decoding performance is reported in terms of the ratio of False Positives (*FP*) and False Negatives (*FN*) versus WSR and SNR. *FP* and *FN* can be defined as:

$$FP(i) = P(H_i \mid H_j) + P(H_i \mid H_k) \tag{12}$$

$$FN(i) = P(H_j \mid H_i) + P(H_k \mid H_i) \quad . \tag{13}$$

In Eq. (12) and Eq.(13), $FP(i)$ ($FN(i)$) denotes the ratio of *FP* (*FN*) for the hypothesis H_i, where i, j, k, refer to the hypothesis $0, 1a,$ or $1b$, defined in Sec. 3, and $i \neq j \neq k$.

WSR can be defined as the ratio of the watermarked signal power to the original signal power. In a similar way, SNR can be defined as the ratio of the original signal power to the noise power. The decision threshold *thr* in Eq. (5) is set to 0.01 for the correlation based classifier.

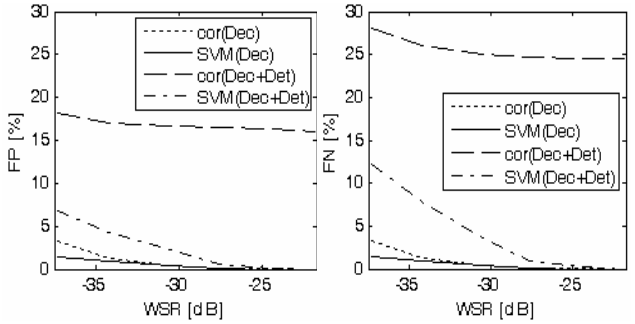
The SVM based classification has been performed by using RBF kernel with the parameters $\sigma = 22$ and C = 1. The SVM classifier is trained by an audio file of length about 417 sec which consists of l=6000 audio frames per class, where each frame is of length N=1024 samples. It is observed that the decoding performance does not rely on the selection of training set. Thus, test results reported in this section are obtained by the training data collected from a single audio clip. The training is performed only once and it takes about 21 minutes on a computer which utilizes a 2.8 GHz Pentium IV machine. As a result of training, 1880, 1875, and 3792 support vectors are obtained for Class 1, Class 2 and Class 3, respectively. The support vectors are used for classifying the test vectors. Thus, training is an offline process. The classification of a test vector takes about 0.1 sec where the length of the test vector is N = 1024 (0.023sec).

5.2. Performance versus WSR and SNR

The performance of the proposed method at different WSRs has been examined for watermark decoding and detection. The distributions of *FP* and *FN* versus WSR, obtained for only Decoding (Dec) and both Decoding and Detection (Dec + Det), are presented in Fig 1.(a) and (b), respectively. In these figures, index "SVM" refers to the introduced decoder and index "cor" refers to the traditional correlation-based audio WM decoder. The length of the test set is about six hours. In Fig.1, it is seen that the decoding performances of both decoders are nearly the same. In terms of Dec + Det performance, the *FP* and *FN* of the introduced SVM-based decoder remain less than 5% when WSR ≥ -30 dB. Although the correlation-based decoder can decode the WM bits with high accuracy, its Dec + Det performance drops significantly, and the *FP* and *FN* values increase to around 17% and 25 % at all WSRs. This indicates that the correlation-based decoder is not capable of detecting the un-watermarked

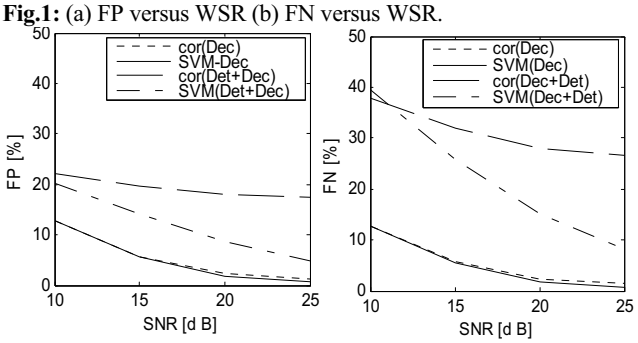
audio clips while decoding. Note that, we tried to improve the performance by using a correlation threshold which is different from zero. However, it is not possible to specify a good threshold adequate for all of the audio clips.

Robustness to channel noise is also evaluated for decoding and detection and results are reported in Fig.2. The length of the test set is about six hours. As it is seen from Fig.2(a) and (b), the decoding results obtained by both decoders are nearly the same. After SNR = 15 dB, the SVM based classifier achieves Dec+Det with low *FP*s and *FN*s. *FP*s and *FN*s obtained by the correlation-based classifier for Dec+Det at SNR = 15 dB are about 20% and 30%, respectively. Similar to the results obtained above, the high *FP* and *FN* values obtained by the correlation-based method denote that the correlation method is not capable of detecting un-watermarked audio clips. This reduces its overall decoding performance. On the other hand, learning process provides a powerful alternative to the adaptive specification of the decision boundaries rather than specification of a decision threshold.



(a) (b)

Fig.1: (a) FP versus WSR (b) FN versus WSR.



(a) (b)

Fig.2: (a) FP versus SNR (b) FN versus SNR.

5.3. Robustness to Stirmark Attacks

In order to quantify the robustness of the proposed decoding method to the standardized audio watermarking attacks, the *Stirmark* benchmark program is used [8]. For this experiment, an 6 min audio clip including un-watermarked and watermarked (WSR = -32 dB) portions is used. Decoding and detection is performed on the original, the marked copy, and all 46 clips created by the Stirmark Audio program. As the *FP* and *FN* values reported in Table 1 demonstrate, the SVM-based decoder is robust to several add/remove (add_brumm_10100, add_noise, addsinus, dynnoise, lsbzero), filter (compressor, rc_higpass, rc_lowpass) and modification (zerocross) attacks. Although the performance decreases in "rc_lowpass" and "zerocross" attacks, the *FP* and *FN* values still remain in an acceptable range. However, traditional correlation-based decoder is not robust to these Stirmark attacks.

Table 1: *FP*s and *FN*s obtained by the correlation and SVM based classifiers for Stirmark attacks.

Classifier	**SVM**		**Correlation**	
Attack	**FP(%)**	**FN(%)**	**FP(%)**	**FN(%)**
original	0.105	0.192	16.594	25.082
addbrumm_10100	0.105	0.192	16.69	25.183
addnoise_900	0.476	0.831	16.709	25.208
addsinus	0.112	0.206	17.119	25.799
compressor	0.326	0.593	16.596	25.028
dynnoise	0.535	0.928	16.511	24.893
fft_real_reverse	0.105	0.192	16.615	25.058
lsbzero	0.105	0.192	16.695	25.185
normalize	0.273	0.428	16.69	25.174
rc_highpass	0.103	0.186	16.585	25.027
rc_lowpass	3.772	7.564	16.493	24.912
zerocross	2.653	5.177	16.943	26.069

6. CONCLUSION

This paper proposes a blind audio watermark decoding scheme based on supervised learning of the watermarked audio signals which combines the watermark decoding and detection problems into a single classification problem. Performance of the proposed decoder is superior to the classical correlation based method in both distorted and non-distorted environment.

In order to reduce the computational complexity, we are working on the selection of optimal feature space dimension. Since the watermarking is a nonlinear process, principal component analysis is not an adequate method to reduce the dimensionality. Currently, we are investigating the learning performance on the wavelet detail coefficients. Preliminary results are promising.

REFERENCES

[1] F. Hartung and M. Kutter, "Multimedia Watermarking Techniques," in Proc. of the IEEE, vol 87, no 7, pp. 1079-1107, 1999.

[2] Y. Yaslan and B. Gunsel, "An Integrated Decoding Framework for Audio Watermark Extraction," *Proc. of the ICPR 2004*, pp. 879-882, Cambridge, UK, August 2004.

[3] S. Sener and B. Gunsel, "Blind Audio Watermark Decoding Using Independent Component Analysis," *Proc. of the ICPR 2004*, pp. 875-878, Cambridge, UK, August 2004.

[4] H. S. Malvar and D. F. Florencio, "Improved Spread Spectrum: A New Modulation Technique for Robust Watermarking," *IEEE Trans. On Signal Processing*, vol. 51, no. 4, pp. 898-905, 2003.

[5] Y. Fu, R. Shen and H. Lu, "Optimal Watermark Detection Based on Support Vector Machines," *Lecture Notes in Computer Science*, vol. 3137, pp. 552-557, 2004.

[6] S. Kirbiz, Y.Yaslan and B.Gunsel, "Robust Audio Watermark Decoding by Nonlinear Classification," *Proc. of the 13th European Signal Processing Conference,* Turkey, Sept, 2005.

[7] V. N. Vapnik, *Statistical Learning Theory*: John Wiley, New York, 1998.

[8] Steinebach, M., Lang, A., Dittmann, J. and Petitcolas, F. A. P., "Stirmark benchmark: Audio watermarking attacks based on lossy compression", *Proc. SPIE Security Watermarking Multimedia*, vol. 4675, pp. 79–90, San Jose, CA, Jan. 2002.