ROBUST AUDIO WATERMARK DECODING BY NONLINEAR CLASSIFICATION

S. Kırbız, Y. Yaslan, B. Günsel

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

ABSTRACT

This paper introduces an audio watermark (WM) decoding scheme that performs a Support Vector Machine (SVM) based supervised learning followed by a blind decoding. The decoding process is modelled as a two-class classification procedure. Initially, wavelet decomposition is performed on the training audio signals, and the decomposed audio frames watermarked with +1 and -1 constitute the training sets for Class 1 and Class 2, respectively. The developed system enables to extract embedded WM data at lower than -40dB Watermark-to-Signal-Ratio (WSR) levels with more than 95% accuracy and it is robust to degradations including audio compression (MP3, AAC), and additive noise. It is shown that the proposed audio WM decoder eliminates the drawbacks of correlation-based methods.

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, correlation-based decision rules are used in most of the WM decoding methods, 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 unextracted WM data. On the other hand, there is a trade-off between the correlation and the audibility.

In this paper, supervised learning of embedded WM data is proposed and it is shown that performance of the developed SVM-based audio WM decoder outperforms the existing correlation-based decoders. Due to the good learning capability, SVMs 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 [6], higher-order statistical deviations that give information about the embedded data are obtained by Quadrature Mirror Filters (QMF) and then these statistics are used for the SVM training and classification of watermarked images. In [7], without extracting the WM information, the SVMs are used for classifying the watermarked and un-watermarked audio signals, based on some audio quality features.

Unlike the existing methods, this paper proposes a SVM based audio watermark decoding scheme which is capable of correctly extracting the WM bits. Test results demonstrate that performance of the introduced WM decoding technique outperforms state-of-the-art correlation-based decoding techniques [2, 4] and it is robust to attacks such as additive noise and audio compression, i.e., mp3 and AAC. Obtained results encourage its usage in on-line monitoring and authentication applications.

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 $\mathbf{s}_i$ refers 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),

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

where Refresh Period (*RP*) refers to the number of block insertions. 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 watermark generation. Our encoder applies an iterative approach that allows specifying a nonlinear $f(.)$ in a data adaptive way [2, 3]. In [4], an analytic approach to analyze a linear $f(.)$ is introduced. In Eq.(1), $w_j \mathbf{k}_{m_i}$ models the nonlinear distortions, where $\mathbf{k}_{m_i}$ is the modulated key

embedded into audio frame *i* after multiplied by w_j. The WM encoder generates $\mathbf{k}_{m_i}$ by shaping the secret key sequence **k** according to masking thresholds obtained by psychoacoustic masking of $\mathbf{s}_i$.

3. AUDIO WATERMARK DECODING BY SVMs

The developed SVM-based decoding scheme describes the WM decoding as a pattern recognition problem, and brings a new approach to the audio WM extraction.

3.2. 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 that the embedded WM data are dominant in the detail parts of the wavelet transformed signal [2]. Consequently, the *N* dimensional *i*th training vector $\mathbf{t}_i$ can be obtained by taking the inverse wavelet transform of the detail coefficients described as:

$$\mathbf{t}_i = W^{-1}\left(\mathbf{d}_{\mathbf{s}_{i_{WM}}}\right), \qquad i=1,..,l \tag{2}$$

where W^{-1} denotes the inverse Wavelet transform, and $\mathbf{d}_{\mathbf{s}_{i_{WM}}}$ refers to the detail coefficients of the watermarked audio signal.

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.

3.2. Training the SVM for Watermark Decoding

Due to the good learning capability, SVMs are used in the training stage. Originally, the SVM classifier is designed for two-class classification [8]. 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 [8] of the Lagrange functional given in Eq.(3),

$$\max_{\alpha} W(\alpha) = \max_{\alpha}\left(\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{t}_i,\mathbf{t}_j)\right) \tag{3}$$

subject to constraints

$$\sum_{i=1}^{l}\alpha_i y_i = 0, \quad 0 \le \alpha_i \le C, \quad i = 1,\acute{E}, l, \tag{4}$$

where α_i is the *i*th Lagrange multiplier corresponding to the *i*th 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 multiplier, α_i, $i = 1,\acute{E}, l$.

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.

In the proposed WM decoding method, the decoding process is modeled as a two-class classification procedure, i.e., audio frames watermarked by +1, by -1 are labeled as Class 1 and Class 2, respectively. The training set $\mathbf{T}=\{(\mathbf{t}_1,y_1),...,(\mathbf{t}_l,y_l)\}$ is formed by assigning the class label $y_i \in \{+1,-1\}$ to each training vector $\mathbf{t}_i$, obtained by the wavelet decomposition of *i*th audio frame. The SVM classifier is trained with the training vectors coming from two classes. The hyper-plane parameters **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 **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 shown that, performance of the introduced audio WM decoder does not rely on the selection of the training samples.

3.3. Classification of the Audio Frames

Let $S = \{\mathbf{t}_1, ..., \mathbf{t}_u\}$ denote our test set where $\mathbf{t}_v$, $v = 1, ..., u$, is an *N*-dimensional test vector. In order to obtain the test vector $\mathbf{t}_v$, the received signal $\mathbf{s}_{v_R}$ is first decomposed into its detail $\mathbf{d}_{\mathbf{s}_{v_R}}$ and approximation $\mathbf{e}_{\mathbf{s}_{v_R}}$ parts by wavelet transform. In order to eliminate channel noise, the detail coefficients of decomposed signal, $\mathbf{d}_{\mathbf{s}_{v_R}}$, are thresholded before taking the inverse wavelet transform as in Eq. (5);

$$\mathbf{t}_v = W^{-1}\left(\Lambda_h\left(\mathbf{d}_{\mathbf{s}_{v_R}}\right)\right), \quad v = 1,\ldots,u \tag{5}$$

where Λ_h refers to the thresholding operation thus eliminates the coefficients less than a threshold *h*.

The classification of the test vectors is performed according to Eq.(6),

$$F(\mathbf{t}_v) = \operatorname{sgn}\left(\sum_{s \in SV} \alpha_s y_s K(\mathbf{t}_s, \mathbf{t}_v) + \bar{b} \right) \quad (6)$$

where $F(.)$ describes the decision rule of the binary classifier, $\mathbf{t}_v$ 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 correspond to $\alpha_s > 0$, and $\bar{b}$ is the bias term obtained by the SVM training.

4. CORRELATION-BASED AUDIO WM DECODING

This section briefly describes the state-of-the-art audio WM decoder that uses the correlation-based decision rule [2,4]. Eq.(7) defines the correlation function between the received audio frame, $\mathbf{s}_{i_R}$, and the secret key signal **k,** for *i*th frame;

$$\begin{aligned} r_i &= \sum_{n=1}^{N} k(n) s_{i_R}(n) \\ &= \sum_{n=1}^{N} k(n) s_i(n) + \sum_{n=1}^{N} w_j k(n) k_{m_i}(n) + \sum_{n=1}^{N} k(n) \mathrm{n}(n) \end{aligned} \quad (7)$$

Since **k** is a PN signal which should be un-correlated with $\mathbf{s}_i$ and **n,** in ideal $\sum_{n=1}^{N} k(n) s_i(n) \approx 0$ and $\sum_{n=1}^{N} k(n) \mathrm{n}(n) \approx 0$. Therefore, Eq.(7) can be simplified as:

$$r_i \approx w_j \sum_{n=1}^{N} k(n) k_{m_i}(n) \quad (8)$$

Consequently, w_j, the WM bit embedded into frame *i* can be estimated according to the decision rule given in Eq.(9):

$$w_j = \begin{cases} 1, & \text{if } w_j \sum_{n=1}^{N} k(n) \hat{k}_{m_i}(n) \geq 0 \\ -1, & \text{if } w_j \sum_{n=1}^{N} k(n) \hat{k}_{m_i}(n) \leq 0 \end{cases} \quad (9)$$

where $w_j \hat{\mathbf{k}}_{m_i}$ is estimated by using Wavelet denoising. In Eq.(10), if the correlation value is greater than zero, w_j is extracted as +1, if it is lower than zero, w_j is extracted as -1. Thus the extracted WM bit highly depends on the threshold value. Furthermore, in practice, neither **k** and $\mathbf{s}_i$, nor **k** and **n** can be chosen as uncorrelated, that also reduces the WM extraction accuracy of the decoder. However, these fundamental problems of the existing correlation-based decoders are eliminated by the introduced SVM-based decoder.

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. Robustness to compression is evaluated on the same test data after MP3 and AAC compression (96kbps). Watermark embedding within a 2-22050 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 True Classification ratio (TC) and False Classification ratio (FC) versus WSR and SNR. TC, FC, WSR and SNR are described by equations (10) through (13).

$$TC = \frac{Number\_of\_Correctly\_Classified\_Bits}{Number\_of\_Total\_Bits} \quad (10)$$

$$FC = \frac{Number\_of\_False\_Positive\_Bits}{Number\_of\_Total\_Bits} \quad (11)$$

$$WSR\left(s_i, s_{i_{WM}}\right) = 10 \log_{10} \left(\frac{\sum_{n=0}^{N-1} [s_i(n) - s_{i_{WM}}(n)]^2}{\sum_{n=0}^{N-1} [s_i(n)]^2} \right) dB \quad (12)$$

$$SNR = 10 \log_{10} \left(\frac{\sum_{n=1}^{N} \left[s_{i_{WM}}(n) \right]^2}{\sum_{n=1}^{N} \left[\mathrm{n}(n) \right]^2} \right) dB \quad (13)$$

The SVM based classification has been performed by using RBF kernel with the parameters σ = 22 and C = 1.

5.2. Performance versus WSR and SNR

In order to observe WM decoding performance, performance at different WSRs has been examined for two-class classification. The SVM classifier is trained by an audio file of length about 139 sec. It is observed that the decoding performance does not depend on selection of the training data. Thus, test results reported in this section are obtained by the training data collected from a single audio clip. Distribution of TC versus WSR for un-compressed, MP3 compressed and AAC compressed files are compared and reported in Fig. 1 and Fig. 2. Note that, the FC ratios of Class 1(+1) and Class 2(-1) are obtained nearly the same, thus we reported the arithmetic mean of them. The SVM-based and correlation-based results are obtained by using a test set of length about 2.5 hours. As it is observed from Fig. 1 and Fig. 2, for un-compressed audio files, true classification performance of the correlation method and the SVM based method are similar and the decoding accuracy remains greater than 95% when WSR>-40 dB. However, the superiority of the proposed scheme can be seen in compressed domain. For the mp3 compressed audio, the proposed SVM–based decoding provides about 10% gain at WSR = -45 dB, while it is about 6% for the AAC compressed audio. In both compressed and un-compressed domain, TC ratio reaches to 100% for both decoding scheme when WSR ≥ -20 dB.

In order to evaluate the WM decoding accuracy at noisy communication channels, the same test audio files are distorted by i.i.d. Gaussian noise. The reported performance is obtained on a test set of length about three hours. The SVM training is performed on a training set of length about 278 sec. As it is seen from Fig. 3, distribution of TC versus SNR is almost the same for correlation

based and SVM based decoding schemes. Decoding accuracy exceeds 90% at SNR = 15 dB, and reaches 99% at SNR = 20 dB.

6. CONCLUSION

This paper proposes a blind audio watermark decoding scheme based on supervised learning of the watermarked audio signals. Performance of the proposed decoder is superior to the classical correlation based method in both uncompressed and compressed domains. The developed watermark decoding method is also robust to channel noise.

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*, Cambridge, UK, 2004, pp. 879-882.

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

[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.Lyu and H. Fardi, "Detecting Hidden Messages Using Higher-Order Statistics and Support Vector Machines," *Lecture Notes in Computer Science*, vol.2578, pp.340-354, 2002.

[7] H. Ozer, I. Avcibas, B. Sankur, and N. Memon, "Steganalysis of audio based on audio quality metrics," *Proceedings of the IS&T/SPIE's 15th Annual Symposium on Electronic Imaging*, vol. 5020, Santa Clara, CA, US, January 2003, pp.55-66.

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

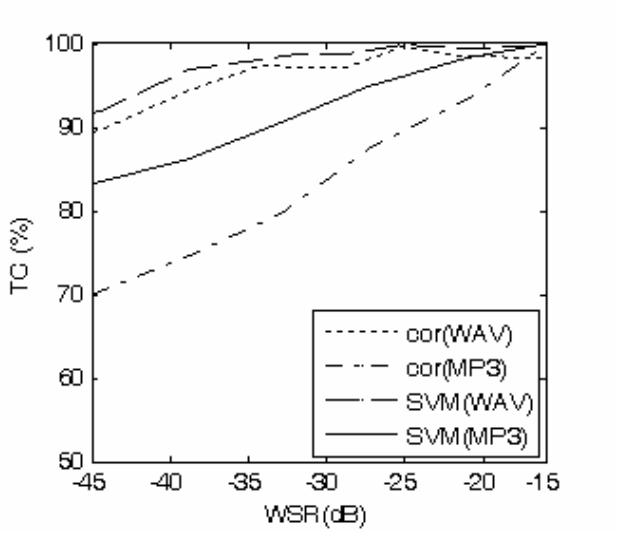

Fig. 1. TC versus WSR for wav and mp3 files

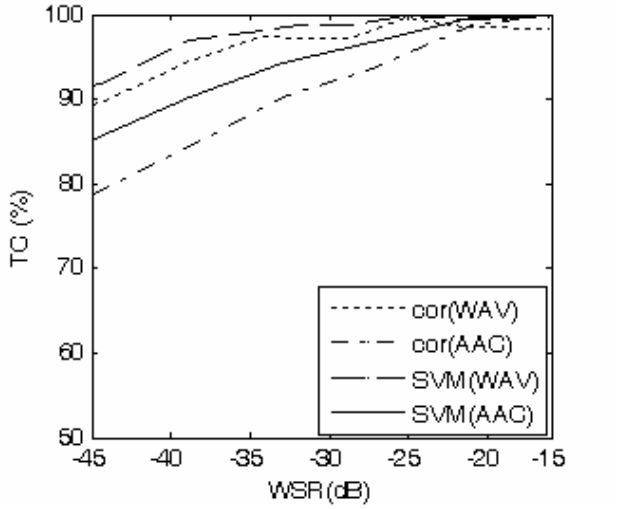

Fig. 2. TC versus WSR for wav and AAC files.

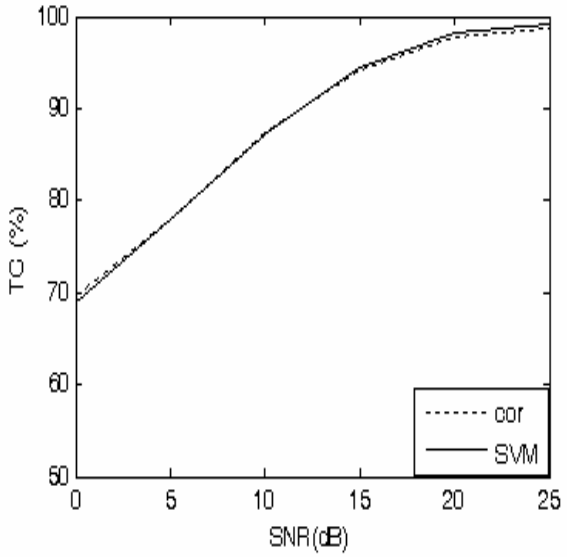

Fig. 3. TC versus SNR.