INFORMED ALGORITHMS FOR WATERMARK AND SYNCHRONIZATION SIGNAL EMBEDDING IN AUDIO SIGNAL

Przemysław Dymarski and Robert Markiewicz

Institute of Telecommunications, Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665, Warsaw, Poland
email: dymarski@tele.pw.edu.pl, R.Markiewicz@tele.pw.edu.pl

ABSTRACT

This work presents audio watermarking algorithms robust to D/A and A/D conversion, scaling and quantization noise. The watermark is embedded in log-spectrum domain using a differential scheme and a quaternary bi-orthogonal code. An informed embedding algorithm is proposed, minimizing bit error rate while keeping distortion level below the masking threshold. An informed coding scheme, based on Costa's approach, is also used. Synchronization is based on two pilot signals, and the informed synchronization scheme is proposed, in order to speed up a synchronization process.

Index Terms— *Audio watermarking, informed embedding*

1. INTRODUCTION

The audio watermark signal must survive transmission distortions which depend on the specific application. In the copyright protection applications there are deliberate attacks of impostors, contrary to the annotation watermarking (e.g. coding of lyrics of a song or its translation to the other language). On the other side, the annotation watermarking requires higher bit rate (50-100 bps) than the copyright protection watermarking (less than 10 bps), it must be robust to quantization noise (introduced by audio compression) and distortions caused by D/A and A/D conversion (amplitude scaling, time and frequency offset due to the difference of sampling frequencies in D/A and A/D devices [1]). Linear distortions and background noise appear in the analog transmission and in the acoustic channel (sonic watermarking [2]).

Watermarking is a problem of channel coding in presence of side information (the audio signal is known at the transmitter but not known at the receiver [3]). Informed watermark embedding algorithms tend to adapt the watermark to the audio signal by simulating the watermark reception at the transmitter's side [4]. Informed watermark coding consists in selecting a proper vector from a set of vectors representing the same message (Costa's scheme [3]).

An important issue is a choice of a watermark embedding domain (i.e. in which domain the measurements of the received signal are taken and compared with stored patterns). In time domain embedding [5], [6] special measures must be taken to keep the watermark below the masking threshold (which is defined in frequency domain). Moreover, the time domain watermarking systems require very accurate symbol synchronization algorithms, contrary to the frequency domain embedding. In [2] watermark is embedded in a difference of log spectra of audio signal, measured in neighbouring frames. This approach yields immunity to strong peaks in the spectrum of audio signal and conformity with the psychoacoustic analysis (in log-spectrum domain, watermark is spread more uniformly within the frequency band chosen for its transmission). The watermarking system presented in this paper is partially based on [2], [7] but it is optimized using informed embedding and coding. Generally, frequency (and DWT) domain watermarking systems rather use the blind embedding [1], [2], [8], [9]. The aim of this paper is to show that informed embedding yields some improvement of the frequency domain watermarking system.

Symbol synchronization system must be robust to sampling frequency shift (in a D/A and A/D conversion [1]). In low bit rate watermarking systems self-synchronization is usually used [8], [9]. In annotation watermarking pilot signals may be used [10], that have proven useful in OFDM technique [11]. The same pilot signals may be used to estimate the sampling frequency shift or the amount of time scale modification [10], [11], [12]. In this paper the "informed" pilot signals are embedded, in order to speed up the synchronization process and to increase its accuracy.

This paper is organized as follows: In Sect.2 the blind watermark embedding is described. In Sect.3 the informed embedding and coding algorithms are presented. In Sect.4 the blind and informed synchronization algorithms are proposed and in Sect.5 final conclusions are drawn.

2. BLIND WATERMARK EMBEDDING IN LOG-SPECTRUM DOMAIN

The watermark should not be perceived, so its power spectral density (psd) should not exceed the masking threshold. At a first approximation, the masking threshold is proportional to the psd of the audio signal, estimated in short intervals (about 20 ms). This suggests a multiplicative watermark model: psd(watermark) = α psd(audio), which becomes additive in log-spectrum domain.

Thus some log-frequency watermark patterns (symbols) may be defined and a correlation receiver may be used in log-spectrum domain:

$$i_{received} = \arg\max_i < \log|\bar{Y}|, \ker_i > \qquad (1)$$

where $\bar{Y}$ - spectrum of the audio signal with the embedded watermark, $\ker_i$ - the i^{th} symbol, called a kernel [7], $< >$ - scalar product (correlation).

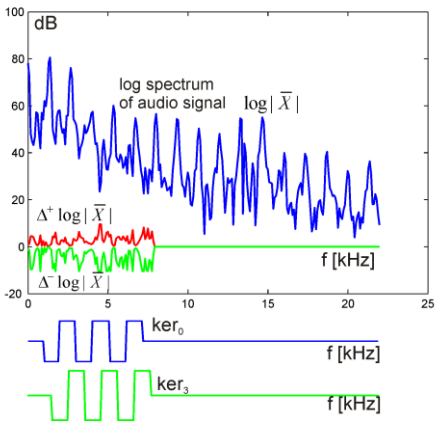

Fig.1. Modification of the signal spectrum – audio spectrum, maximal changes allowed by masking threshold and two orthogonal kernels ($\ker_1 = -\ker_0$ and $\ker_2 = -\ker_3$ are not shown)

The watermark is embedded by increasing or decreasing the log spectrum in subbands indicated by the sign of the kernel. E.g. if the logical 1 is transmitted the log spectrum has to be increased in subbands fulfilling the condition $\ker_1 > 0$, and decreased if $\ker_1 < 0$. The amplification and attenuation of subband signals should not exceed the masking threshold. To fulfill this requirement, the masking threshold is calculated in log scale (like in the MPEG1-audio coding), then the maximal changes allowed by masking threshold are obtained ($\Delta^+ \log|\bar{X}|$ and $\Delta^- \log|\bar{X}|$, where $\bar{X} = \{X_0,\ldots,X_{N-1}\}$ - DFT of audio signal frame). The modifications of spectrum are limited to ± 10 dB. The maximum amplitude spectrum of the added (or subtracted) signal ($|W_i|$) is then expressed in linear scale. This process of spectrum modification is explained in Fig.1.

The main problem in the watermarking systems of this kind is the presence of strong spectral peaks in the audio signal. E.g. if the spectral peak appears in a subband with a positive value of $\ker_1$, the reception of logical 1 is forced, despite of the log spectrum modification (limited by the masking threshold). This problem was solved in [2], with application of the mirrored watermarks. The time slot used for transmission of one bit is split into two subintervals (subframes). In each subframe a kernel of opposite sign is used, e.g. -1,+1 for the logical 1 and +1, -1 for the logical 0. In telecommunications such scheme is called the Manchester signaling [7] (a suite of negative and positive pulse represents logical 1 and vice versa). Decision is based on the comparison of correlations obtained in both subframes:

$$\Delta C_i = < \log|\bar{Y}^2|, \ker_i > - < \log|\bar{Y}^1|, \ker_i > = \\ = < (\log|\bar{Y}^2| - \log|\bar{Y}^1|), \ker_i > = < \Delta \log|\bar{Y}|, \ker_i > \qquad (2)$$

Thus a watermark is embedded in log-spectrum domain using a differential scheme.

The watermark amplitude spectrum $|W_i|$ is calculated using the masking threshold and the phase spectrum ϕ_i is that of the audio signal. More precisely, $\phi_i = \arg(X_i)$ in subbands, where the audio spectrum is to be increased, and $\phi_i = \arg(X_i) + \pi$ in subbands, where the audio spectrum is to be decreased. In order to increase the robustness to symbol shifts we have applied the Hanning window at the receiver [10]. However, a window applied in time domain yields a convolution in frequency domain, which modifies the phase of audio and watermark signals in different ways. This would be partially compensated if the embedded watermark had the phase of the windowed audio signal: $\phi_i = \arg(\bar{X} * \bar{H})_i$ or $\phi_i = \arg(\bar{X} * \bar{H})_i + \pi$, where $\bar{H}$ – DFT of the Hanning window and * denotes convolution. Indeed, using the Hanning window and this phase adjustment considerably improved robustness of the watermarking system [10].

The embedding strategy using two antipodal kernels ($\ker_1 = -\ker_0$) corresponds to a binary modulation. In order to double the bit rate, the quaternary modulation is proposed, based on bi-orthogonal kernels: $\ker_1 = -\ker_0$, $\ker_2 = -\ker_3$, where $\ker_0$ is orthogonal to $\ker_3$ – Fig.1.

Orthogonality is obtained by a frequency shift of 50% of the subbands. The bit rate is doubled at a cost of increasing the BER. The number of bi-orthogonal symbols (kernels) may be increased, as it is used in watermarking systems operating in time domain [5], [6]. In log-spectrum domain, however, the number of different symbols is limited, for the reasons discussed in section 3.2. We use in our system either binary or quaternary modulation. Moreover, we insert our watermarks in the band 1-8 kHz. The whole audio signal band cannot be used for watermark embedding, because the high frequencies may be eliminated e.g. by the MP3 coder.

3. INFORMED WATERMARK EMBEDDING AND CODING

3.1. Informed watermark embedding in log-spectrum domain

Blind embedding yields a watermark collinear with a kernel representing the information to be transmitted (without considering the audio signal). If a watermark is embedded in time domain, a proper symbol (spread spectrum signal perceptually filtered with a filter modeling the masking curve [6]) is just added to the signal. In log-spectrum domain blind watermark is obtained by filling up a kernel with values $\Delta^+ \log|\bar{X}|$ or

$\Delta^{-}\log|\overline{X}|$, representing maximal changes below the masking threshold (Fig.1). Such watermark has the maximal strength, but it is no longer collinear with its kernel.

Informed watermark is usually calculated as a linear combination of blind watermarks representing all the symbols (kernels) available in the watermarking system [4], [6]. Indeed, only these kernels are used in the receiver to find the maximally correlated one, so there is no sense to construct a watermark stemming out of the subspace spanned with the kernels. The optimal watermark should minimize the probability of erroneous reception P_e.

The watermarked audio frame may be described, in log-spectrum domain, with a vector: $\Delta\log|\overline{Y}| \approx \approx \Delta\log|\overline{X}| + \Delta\log|\overline{W}|$. Its correlation with $\ker_i$ equals:

$$\Delta C_i \approx < \Delta\log|\overline{X}|, \ker_i > + < \Delta\log|\overline{W}|, \ker_i > \quad (3)$$

If the watermark is embedded in time domain, the terms $\Delta\log|\overline{X}|$ and $\Delta\log|\overline{W}|$ are replaced with the audio and watermark signal waveforms, ($\overline{x}$ and $\overline{w}$ correspondingly). In order to minimize P_e, suboptimal algorithms are used: in [5] the pairwise error probability is minimized, in [6] two iterative approaches are proposed: the first one considers, at each iteration, the proper and the most "menacing" kernel, and the second one calculates gradient of P_e and appends to the watermark a small vector of direction opposite to the gradient.

For the quaternary bi-orthogonal modulation (Fig.2) finding the optimal watermark causes no problems, if embedding is performed in time domain. For the audio signal vector $\overline{x}$ the most "menacing" kernel is found (it is a kernel representing an improper pair of bits and exhibiting the greatest correlation with $\overline{x}$). Then the difference of correlations between the watermarked audio and the proper $\ker_i$ and menacing $\ker_j$ kernels is maximized:

$$\begin{aligned} &\max_{\overline{w}} (< \overline{x}, \ker_i > + < \overline{w}, \ker_i > - < \overline{x}, \ker_j > - < \overline{w}, \ker_j >) \\ &= \max_{\overline{w}} (< \overline{x}, (\ker_i - \ker_j) > + < \overline{w}, (\ker_i - \ker_j) >) \end{aligned} \quad (4)$$

Considering that the watermark energy (i.e. the squared Euclidean norm $||\overline{w}|_2^2$) is bounded (this is the inaudibility condition in time domain), the optimal watermark should be collinear with $\ker_i - \ker_j$. If the "menacing" correlation is zeroed and the maximum watermark norm not achieved yet, a vector collinear with a proper kernel should be added to the watermark. Thus a watermark is a linear combination of a "good" vector $\ker_i$ and a "good minus wrong" vector $\ker_i - \ker_j$ (Fig.2, "informed(L_2)" watermark).

In log-spectrum domain, similarly, we aim at maximizing the difference of correlations $< \Delta\log|\overline{W}|, (\ker_i - \ker_j) >$. There are, however, some obstacles. Firstly, as it has been mentioned before, the blind watermarks are no longer collinear with the kernels used for watermark detection. Secondly, the Euclidean norm is no longer appropriate for description of constraints. Note that any component of the blind watermark vector in log spectral domain (i.e. $\Delta\log|\overline{W}|$) attains the extreme value allowed by the masking threshold. Linear combinations of blind watermark vectors may be constructed, but the masking threshold must be respected. This yields L_1 norm rather than Euclidean one. E.g. a combination of two blind watermark vectors $\alpha(\Delta\log|\overline{W_1}|) + \beta(\Delta\log|\overline{W_2}|)$ does not violate the inaudibility constraint if $\alpha + \beta \le 1$.

The linear combinations of blind watermark vectors have thus less strength than the blind watermarks and the watermark optimization may be questioned. It is shown in Fig.2: $\ker_0$ represents the proper symbol (i.e. 2 bits to be transmitted), $\ker_2$ is the most "menacing" symbol, $\ker_1 = -\ker_0$, and $\ker_3 = -\ker_2$ are not explicitly shown. The blind watermark is added to the audio signal in direction $\ker_0$. The informed watermark respecting constraints described with L_2 norm (embedding in time domain) is shown as well as its two components: $\ker_0 - \ker_2$ and $\ker_0$. It is evident that the difference of correlations between the watermarked audio and the proper ($\ker_0$) and menacing ($\ker_2$) kernels is higher than for the blind watermark. On the other hand, the watermark respecting constraints described with L_1 norm (embedding in log-spectrum domain) is not better than the blind one (the same difference of correlations).

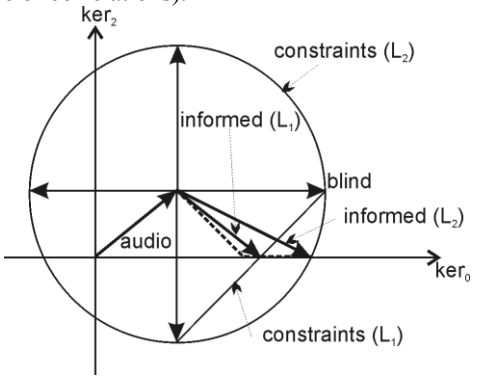

Fig.2. L_1 and L_2 norms in watermark embedding

Does it mean that the blind watermark is optimal under L_1 constraints in any case? Not necessarily, because the blind watermarks, serving as components of the linear combination, are not orthogonal (as it is assumed, for simplicity, in Fig.2). If the line describing L_1 constraints were rotated a little bit counterclockwise, then the informed watermark would be better than the blind one. Such a case is shown, using vectors obtained from the real audio signals (violin) in Fig.3. Here four nonorthogonal blind watermarks $\Delta\log|\overline{W_i}|$ may be seen. The informed watermark is a linear combination of two vectors: the blind watermark (here $\Delta\log|\overline{W_0}|$) and a vector approximating a difference between the blind watermark and the most menacing watermark - here $\frac{1}{2}(\Delta\log|\overline{W_0}| - \Delta\log|\overline{W_2}|)$. A linear combination may be also built using $\Delta\log|\overline{W_0}|$ and

$\Delta \log |\overline{W}_3|$ vectors. The watermarked audio ("audio + informed" in Fig.3) is almost collinear with ker_0 and yields greater difference of correlations with the proper and menacing kernels than the blind watermark.

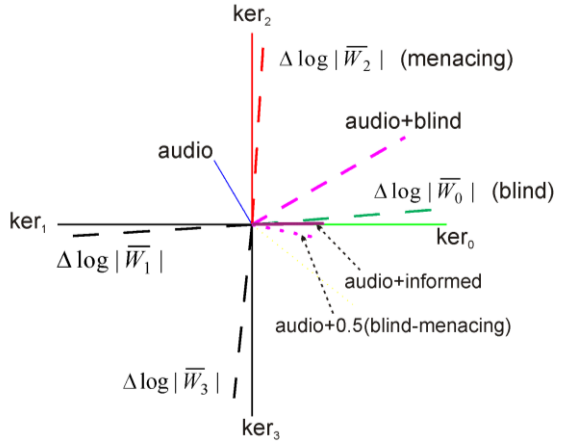

Fig.3. Blind and informed watermark embedding in log-spectrum domain: real audio signals

3.2. Informed watermark coding

The informed watermark embedding consists in using Costa scheme [3], in which the same information is represented by several symbols. In bi-orthogonal quaternary modulation the same binary value is transmitted in opposite kernels: e.g. ker_0 and ker_1 represent logical "1" whereas ker_2 and ker_3 represent logical "0". Of course the bit rate is thus reduced to one bit per signal frame.

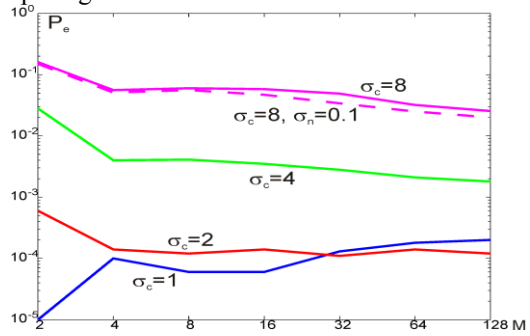

Fig.4. Costa scheme based on M-ary bi-orthogonal modulation in N=64-dimensional space (for M=2 a simple bipolar code is obtained: $\text{ker}_1 = -\text{ker}_0$)

We were interested (as in [5] for orthogonal codewords) in checking the influence of the number of bi-orthogonal codewords (kernels) on bit error rate (BER). Simulations were made using the artificial signals (uncorrelated, gaussian, N=64-dim. vectors) and watermark embedding in time domain. The root mean square (rms) value of watermark signal was set to σ_w=1, rms value of channel noise (representing e.g. a quantization noise) σ_n=1.5, and rms value of cover signal σ_c (representing the audio signal) was a parameter. The suboptimal informed watermark embedding algorithm, described in [6], was applied. Results (Fig.4) for a quaternary modulation were satisfactory, so we use 4 kernels, ker_0 and $\text{ker}_1 = -\text{ker}_0$ representing logical "1" and ker_2 and $\text{ker}_3 = -\text{ker}_2$ representing logical "0". Given a logical value, the optimal kernel is chosen (for minimum angle with the audio signal vector), then the informed embedding algorithm, described in previous section, is applied.

3.3. Testing

Simulations were performed in the following conditions: Two bits or one bit (if Costa scheme is used) are transmitted in a frame of N=1024 samples (sampling frequency 44100 Hz), yielding the bit rates of about 86 bits/s or 43 bits/s, correspondingly. Each frame consists of 2 subframes of N'=512 samples. Each of 4 kernels consists of 6 subbands in a frequency range 1033-7235 Hz (Fig.1). Seven audio files of duration 6-7 seconds were used. (i.e. 250-500 bits per file).

Tab.1. Simulations results: BER - bit error rate, SWR - signal to watermark ratio, T – transmitter with a simulated receiver, R – remote receiver (processing the watermarked audio after MPEG1 coding/decoding at 128 kbits/s)

	quaternary modulation, 86 bits/s		Costa scheme, 43 bits/s	
	blind embedding	informed	blind embedding	informed
BER (T)	1.53%	1.20%	0.60%	0.40%
margin (T)	0.062	0.073	0.081	0.094
Σcor (T)	12.6	2.5	12.4	2.6
BER (R)	2.3%	1.7%	1.3%	0.7%
margin (R)	0.037	0.041	0.051	0.054
SWR	33.8 dB	34.4 dB	36.8 dB	37.4 dB

At SWR=34 dB watwermark is inaudible, according to informal testing. In order to test the Costa scheme, SWR was increased to 37 dB, so as to increase the number of errors and accuracy of BER estimator. If the number of observed errors is zero, margin is calculated as a function of a distance between the "good" correlation and the "bad" one. The average of 10% of the worst results is a margin. Σcor is a sum of correlations of watermarked audio frames with kernels orthogonal to the kernel carrying the transmitted information - in an ideal case Σcor should be equal to zero. The results (Tab.1) show improvement due to informed embedding: BER and Σcor is decreased while the margin and SWR increased for both systems.

4. INFORMED ALGORITHM FOR EMBEDDING OF SYNCHRONIZATION SIGNALS

For symbol synchronization two pilot signals are used, of frequencies f_k=689 and f_{k+1}=732 Hz (the 16th and 17th base

function of the DFT, N=1024). The "blind" synchronization signal, $s_n = -a_n \sin(2\pi n f_k / f_s) + a_n \sin(2\pi n f_{k+1} / f_s)$, ($f_s$ = 44100 Hz is the sampling frequency), has the envelope which attains zero at the edges of the frame, which enables a smooth adjustment of its amplitude a_n (it should not exceed the masking threshold). The angle between the 16th and 17th DFT coefficients is a function of the symbol shift m: $\psi = \pi + 2\pi m / N$. Both DFT coefficients are cumulated for at least 3 s, yielding the symbol alignment accuracy of 10-20 samples [10].

The same pilot signals are used for sampling frequency offset measurement and correction (in the case of time scale modification or A/D – D/A sampling frequency shift). The algorithm, based on techniques used in OFDM [11], has been used for watermarking purposes in [12] and [10].

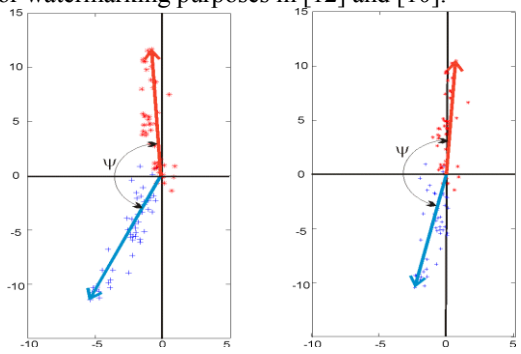

Fig.5. Cumulated DFT coefficients for the blind (left) and informed (right) synchronization

Knowing the frequency of a sinusoid hidden in gaussian noise, the optimal (in Cramer-Rao sense) estimator of its phase is a correlation with $\exp(j2\pi n f_k / f_s)$. In audio watermarking the role of noise plays the audio signal, which is not stationary and it is known at the transmitter. Therefore the other "informed" phase estimators may be considered. Our proposal is to modulate the phase of the synchronizing pilot signals in order to compensate for a phase drift caused by the audio signal:

$$s_n = -a_n \sin(2\pi n \frac{f_k}{f_s} + \Delta\psi_k(n)) + a_n \sin(2\pi n \frac{f_k}{f_s} + \Delta\psi_{k+1}(n)) .$$

The values of $\Delta\psi_k(n)$ and $\Delta\psi_{k+1}(n)$ attain zero at the edges of the frame and do not exceed 0.5 rd in the middle. In Fig.5 the cumulated DFT coefficients are shown for the frequencies f_k (lower side) and f_{k+1} (upper side). The symbol shift was equal to zero, so the true phase shift ψ should be equal to π. The measured phase shifts for the blind and the informed synchronization signals are shown – the informed embedding yields more accurate result.

The informed synchronization algorithm shortens the synchronization process. E.g. the informed synchronizer was able to find the proper symbol position (error of 2 samples) in 1.2 s of a song, whereas the error of the blind synchronizer was, in the same conditions, about 60 samples. However, in several seconds both algorithms yield symbol synchronization accuracy, sufficient for audio watermarking in frequency domain.

5. CONCLUSIONS

Informed algorithms of audio watermark embedding in log-spectrum domain are proposed. Due to L_1 norm used for description of constraints, these algorithms are less efficient than their time domain counterparts, however they still offer some improvement as compared to the blind algorithms. Informed (but of low complexity) algorithm of symbol synchronization is also presented, that speeds up the synchronization process and increases its accuracy. These algorithms were tested in an audio annotation watermarking system robust to D/A and A/D conversion, scaling and quantization noise. Direct comparison with the other systems is difficult, because of different bit rate (e.g. 20 bps in [1], 2 bps in [2], 170 bps in [8], 4 bps in [9]) and different testing conditions. Anyway, our aim was not to compete with ready to use watermarking systems, we just wanted to show that informed embedding may improve the frequency domain watermarking systems.

6. REFERENCES

[1] S. Xiang, "Audio watermarking robust against D/A and A/D conversions", *EURASIP J. on Advances in Signal Proc.,* 2011:3

[2] R. Tachibana, S. Shimizu, T. Nakamura, and S. Kobayashi, "An audio watermarking method robust against time- and frequency-fluctuation," in *SPIE Conf. on Security and Watermarking of Multimedia Contents III*, San Jose, USA, January 2001, vol. 4314, pp. 104–115.

[3] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 3, pp. 439–441, May 1983.

[4] M. L. Miller, G. J. Doerr and I. J. Cox, "Applying informed coding and embedding to design a robust high capacity watermark" *IEEE Trans. Image Proc.*, vol.13, no.6, pp.792–807, Jun. 2004.

[5] A. Abrardo and M. Barni "Informed Watermarking by Means of Orthogonal and Quasi-Orthogonal Dirty Paper Coding", *IEEE Trans. on Signal Proc.*, vol.53, no 2, Feb. 2005, pp. 824-833

[6] C. Baras, N. Moreau, and P. Dymarski, "Controlling the inaudibility and maximizing the robustness in an audio data hiding", *IEEE Trans. on Audio, Speech and Language Processing*, Sep. 2006, vol.14, No 5, pp 1772-1782

[7] P. Dymarski "Watermarking of audio signals using adaptive subband filtering and Manchester signaling" *Proc. of IWSSIP*, Maribor, June 2007

[8] S.Wu, J. Huang, D. Huang and Y.Q. Shi "Efficiently self-synchronized audio watermarking for assured audio data transmission", *IEEE Trans. on Broadcasting*, vol.51, no.1, pp. 69-76, March 2005

[9] Y. Wang, S. Wu, and J. Huang "Audio watermarking scheme robust against desynchronization based on the DyadicWavelet Transform", *EURASIP Journal on Advances in Signal Processing* Vol. 2010, Article ID 232616

[10] P. Dymarski and R. Markiewicz, "Time and sampling frequency offset correction in audio watermarking", *Proc. of IWSSIP*, Sarajevo, June 2011

[11] M. Sliskovic, "Sampling frequency offset estimation and correction in OFDM systems," *Proc. ICECS*, pp. 437-440, 2001.

[12] Z. Piotrowski, "Drift Correction Modulation Scheme for Digital Signal Processing" *Mathematical and Computer Modelling*, 2011, doi: 10.1016/j.mcm.2011.09.016.