TOWARD ROBUSTNESS OF AUDIO WATERMARKING SYSTEMS TO ACOUSTIC CHANNELS

Emmanuel Wolff, Cléo Baras, and Cyrille Siclet

GIPSA-Lab, DIS
Domaine Universitaire, 961 rue de la Houille Blanche, BP46 F-38402 St-Martin d'Hères CEDEX, France
phone: +33 (0)4 76 82 64 22, fax : +33 (0)4 76 57 47 90
email: {emmanuel.wolff, cleo.baras, cyrille.siclet}@gipsa-lab.grenoble-inp.fr

ABSTRACT

This article deals with blind audio watermarking systems dedicated to data transmission applications, where high embedded information bitrates are prospected. We present an original way to improve the robustness of such systems to perturbations yielded by acoustic convolutive channels. The proposed method, using the analogy between watermarking and digital communication, relies on 1) a channel estimation stage based on an original adaptation of the trained RICE algorithm to watermark inaudibility constraint and 2) a dedicated equalizer, built to invert the convolutive channel effects before data extraction. The efficiency of the proposed method is evaluated through simulations conducted for various real acoustic channels and audio signals. It is shown that the system bit error rate can be decreased from 0.2 to 9.10^{-4} thanks to our contribution when the bitrate transmission is 100 bps and the channel is the acoustic one.

1. INTRODUCTION

Audio watermarking permits to embed inaudible information into audio digital content. Practical implementations fall into two categories:

1. those oriented toward the copyright and intellectual property protection, pursuing watermark robustness and security to pirate attacks, that is, preventing the watermark to be erased or even estimated by pirates [1];
2. those related to data transmission, that aim at embedding high-capacity information with purpose to adding value to digital contents. Their design is subject to standard constraints, namely embedding transparency and system robustness to classical audio manipulations (like low-pass filtering or lossy compressions [2]), but no more to pirate attacks.

The proposed contribution stands on this second range of applications and tackles the issue of watermarking system robustness when the watermarked audio signal is emitted with a loudspeaker and recorded by a microphone. Several distortions that deeply impact the decoder performance have to be considered, including [3], [4]:

- **desynchronization**, both due to the signal propagation delay between the emitter and the receiver and to (time or frequency) stretching effects;
- **acoustic channel** effects, that significantly modify the signal frequency response and is usually modeled by a convolutive filter with finite impulse response (FIR).

Regarding desynchronization, various efficient solutions have already been proposed in the literature: the ones [4] achieve system insensibility to delays or stretching by embedding symbols no more on single embedding locations but on larger time-frequency blocks repeating them over several MCLT[1] coefficients; the decoding can then be processed anywhere in the neighborhood of the central symbol locations. Others [5] embed equally-spaced synchronization patterns and profit from the periodical structure exhibited by the autocorrelation spectrum of the watermarked signal to estimate the desynchronization parameters and invert its effects before decoding.

On the contrary, no watermarking study proposes specific solutions to the acoustic channel problem, whereas it represents a major challenge regarding watermarking system robustness: systems proposed by [6, 7, 8] address the camcorder piracy, but their designs assume the watermark is sufficiently repeated to be robust to acoustic channel effects; the achieved useful bitrates (around 5 bit per second (bps)) are therefore too low for high-capacity watermarking applications.

Therefore, the proposed contribution aims at improving audio watermarking system robustness to acoustic convolutive channels. Thus, synchronization will be considered as perfectly carried out in order to focus on the acoustic convolution. What is more, acoustic convolutive channels will be assumed non time-selective between two consecutive channel estimations, that is to say over around 5 seconds. An original strategy is proposed to compensate acoustic channel effects: it first involves an acoustic channel estimation step based on embedded training data and then a dedicated equalizer, built to inverse acoustic channel effects before watermark decoding. The proposed strategy is integrated into a State-Of-The-Art audio watermarking system with purpose to evaluate its performance through simulations with true audio signals.

This article is organized as follows. Section 2 presents the design principles of the considered audio watermarking system and details the acoustic channel effects. Section 3 describes the acoustic channel compensation strategy, detailing the channel estimation module and the equalization method. Simulation results in terms of channel estimation efficiency and system ro-

[1] Modulated Complex Lapped Transform

ISSN 2076-1465

bustness to acoustic channels are given in section 4. Finally, section 5 draws conclusions and tackles ways of improvement for future works.

2. AUDIO WATERMARKING SYSTEM FACING ACOUSTIC CHANNELS

2.1 Watermarking system principles

The considered audio watermarking scheme, presently an additive Spread-Spectrum (SS) system for digital signals sampled at frequency F_s as proposed in [2], is presented in figure 1.

At the embedder, the modulation interface maps the emitted binary sequence $\{b_l\}$ (with length L_b) into the modulated signal $v(n)$ thanks to a SS waveform $d(n)$ with duration N_b and unit power. $v(n)$ can then be expressed during the l-th bit interval as:

$$\forall n \in [(l-1)N_b; lN_b - 1], v(n) = (2b_l - 1)d(n). \quad (1)$$

To satisfy the inaudibility constraint, the watermark signal $t(n)$ is constructed by filtering $v(n)$ with an adaptive perceptual shaping filter $H(f)$. $H(f)$ is designed according to a PsychoAcoustical Model (PAM)[2] to make the watermark Power Spectral Density (PSD) equal to the masking threshold of the audio signal $x(n)$. The watermark power is then maximized under the inaudibility constraint. The watermarked audio signal $y(n)$ is finally obtained by adding the watermark $t(n)$ and the audio signal $x(n)$.

The receiver, a.k.a the extractor, first filters the received watermarked signal $\hat{y}(n)$ by a zero-forcing filter $1/\hat{H}(f)$, aiming at compensating the perceptual shaping filter $H(f)$. Since the original audio signal $x(n)$ is not available at the receiver, $H(f)$ is estimated according to the masking properties of the received watermarked signal $\hat{y}(n)$. A second filtering stage, involving a non-causal Wiener filter $W(f)$ that minimizes the mean square error MSE $= E[v^2(n)]$, is then performed, yielding the estimated modulated signal $\hat{v}(n)$. Finally, the decoder exploits a correlation demodulator, comparing $\hat{v}(n)$ and the SS waveform $d(n)$ on each l-th bit interval; the sign of the obtained correlation decides the received bit $\hat{b}_l$.

System performance is therefore related to the Bit Error Rate (BER) with respect to the embedding rate $R = F_s/N_b$ and is mainly dependent on the watermark to signal ratio at the receiver.

2.2 Problem formulation including acoustic channel effects

Supposing that resynchronization has already been performed, the considered acoustic channel can be modeled by a convolutive filter $C(f)$ with impulse response $c(n)$, assumed to be time-invariant. The watermarked signal $y(n)$ is then distorted in such a way that:

$$\hat{y}(n) = c(n) \star y(n) = c(n) \star (h(n) \star v(n) + x(n)) , \quad (2)$$

where $\star$ denotes the convolution product.

In such a context, achieving the system robustness to acoustic channels deals with maintaining the BER obtained by the system when acoustic channel perturbs the decoding to the value obtained when the system is free from perturbation.

The receiver (zero-forcing and Wiener filters) must now include an additional stage, aiming at inverting the effects of the convolutive acoustic channel $C(f)$. Since this channel is unknown from the receiver, the proposed solution depicted in figure 2 is based on the following two-steps strategy:

- first, a channel estimation stage;
- second, an additional equalization, preliminary to the two reception filters and the correlation demodulator.

3. COMPENSATING ACOUSTIC CHANNEL EFFECTS WITH CHANNEL ESTIMATION AND EQUALIZATION PROCEDURES

The proposed strategy for compensating acoustic channel consists in a training stage based on an adaptation of the training RICE[3] algorithm aiming at estimating the acoustic channel under the watermark inaudibility constraint, then a dedicated equalization, with purpose to improve the decoding performance. These two steps are detailed bellow.

3.1 Acoustic channel estimation

3.1.1 The RICE algorithm

Standard channel estimation methods are mainly split into blind estimation techniques and trained ones. Blind estimation methods could be very attractive for high-capacity watermarking applications since the channel is directly estimated using the received signal without bitrate increase. Nevertheless the estimation efficiency is directly linked to the number of recorded observations obtained with several microphones. Since the considered application supposes an unique recorded version of the watermarked audio signal, trained methods are then more suitable.

Among State-Of-The-Art trained techniques, we focus on the RICE algorithm [9], since this technique is specifically designed to estimate the frequency response of the acoustic channel $C(f)$ for audio dereverberation. A periodic SS training pattern $p(n)$ (with duration N_p) is added to the audio signal before the loudspeaker emission. At the receiver, the received signal $x(n) + p(n)$ is averaged over the set of the L_p periods to get the convolved version of the pattern $c(n) \star p(n)$ while decreasing audio interference. The pattern transparency is controlled by maintaining the average power of the training data relatively low in comparison to the audio signal power. Unfortunately, it does not prevent from introducing local audible distortions. Thus, we propose to adapt the RICE algorithm to the watermarking scenario, paying much attention to the inaudibility constraint by introducing the perceptual filtering stage.

3.1.2 RICE adaptation to the watermarking system

At the training embedder, the SS training pattern $p(n)$ is still intended to be periodically added into L_p train-

[2] derived [5] from the classical model used in the MPEG 1 Layer 1 codec

[3] Reduced Interference Channel Estimation

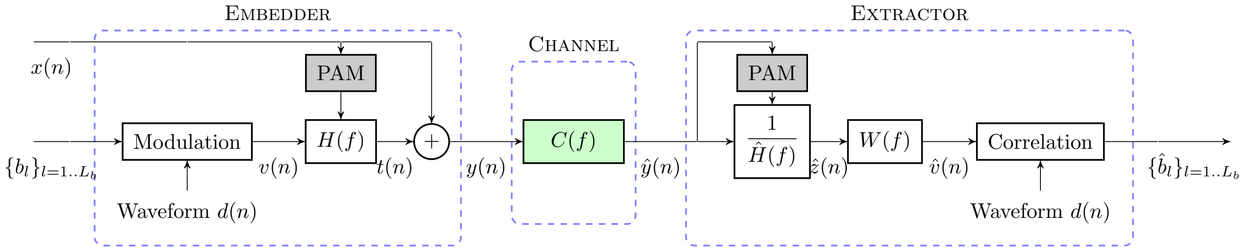

Figure 1: Principles of the considered audio watermarking system.

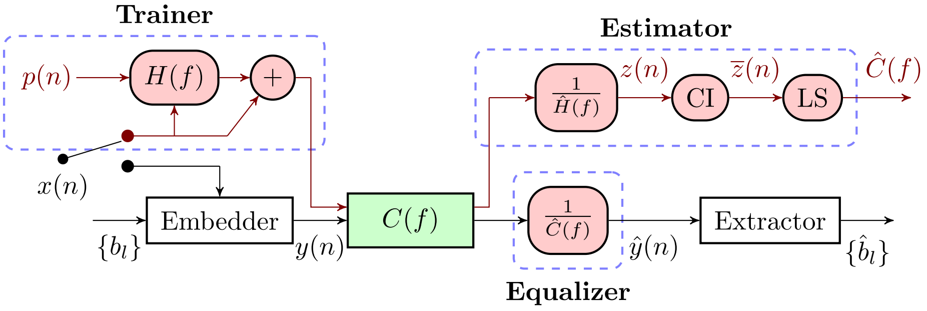

Figure 2: The proposed strategy for compensating acoustic channels: CI=Coherent Integration, LS=Least Square.

ing interval with duration N_p. For each l-th training-interval, we propose to shape $p(n)$ according to the perceptual shaping filter $H_l(f)$ derived from the audio signal PAM (as any watermark information in section 2.1). Thus, the training pattern PSD matches the audio masking threshold, having the maximized authorized power under local inaudibility constraint. The watermarked audio signal during the training stage is finally :

$$\forall n \in [(l-1)N_p; lN_p-1], y_l(n) = x_l(n)+h_l(n)\star p(n), \quad (3)$$

where $h_l(n)$ is the impulse response of $H_l(f)$.

At the receiver, the channel estimation module receives the convolved watermarked audio signal:

$$\hat{y}_l(n) = c(n) \star x_l(n) + c(n) \star h_l(n) \star p(n) \quad (4)$$

As in section 2.1, the perceptual shaping is first inverted using an estimated version $\hat{h}_l^{-1}(n)$ of $h_l(n)$ computed by applying the PAM to the received watermarked audio signal $\hat{y}_l(n)$. The psychoacoustical properties of $\hat{y}(n)$ can be assumed to be equal to those of $x(n)$ and to be independent of the acoustic channel $c(n)$, so that:

$$\hat{z}_l(n) = h_l^{-1}(n) \star \hat{y}_l(n) \simeq c(n) \star p(n) + a_l(n), \quad (5)$$

swapping $c(n)$ and $\hat{h}_l^{-1}(n)$, considering $\hat{h}_l(n)$ equals $h_l(n)$ and introducing $a_l(n) = h_l^{-1}(n) \star c(n) \star x_l(n)$ the residual audio contribution. Since the frequency response $\hat{H}(f)$ (matching the audio masking threshold) is close to the audio PSD envelope, $a_l(n)$ is a partially whitened version of the audio signal.

The original RICE estimation procedure [9] is finally carried on: the Coherent Integration over training-intervals is performed, yielding in time:

$$\underline{z}(n) = c(n) \star p(n) + \underline{a}(n), \text{ with } \underline{a}(n) = \sum_{l=1}^{N_p} \frac{a_l(n)}{N_p} \quad (6)$$

then in frequency (without any edge effect due to the periodicity of the pilot emission):

$$\underline{Z}(f) = C(f)P(f) + \underline{A}(f), \quad (7)$$

where $\underline{Z}(f)$ (resp. $P(f)$, $\underline{A}(f)$) is the Discrete Fourier Transform (DFT) of $\underline{z}(n)$ (resp. $p(n)$, $\underline{a}(n)$) and f varies from 0 to $N_p/2-1$. Denoting by $*$ the conjugate operator, the estimated acoustic channel impulse response with length N_c is finally given by:

$$\hat{c}(n) = \Re\left(DFT^{-1}\{\hat{C}(f)\}\right) \text{ with } \hat{C}(f) = \frac{\underline{Z}(f)P^*(f)}{|P(f)|^2+\alpha}, \quad (8)$$

following a Least Square method in the frequency domain and introducing the regularization factor α, that prevents noise enhancement in weak frequency components.

3.2 Acoustic Channel Equalization

Considering the acoustic channel has been estimated as $\hat{c}(n)$, we now aim at designing a dedicated equalizer, integrated in the watermarking chain before the hidden information extraction to invert the channel effects and make the system performance invariant to acoustic channels.

Since acoustic channels are often non-minimum phase, they are difficult to equalize with stable filters. Thus, the proposed solution is a zero-forcing linear equalizer with Finite Impulse Response (FIR) $\hat{c}^{-1}(n)$.

$\hat{c}^{-1}(n)$ is designed to be the non-causal least-square optimal estimation of the inverse filter $1/\hat{C}(f)$ that suppresses the Inter-Symbol Interference (ISI) due to the acoustic channel effects. Suppressing the ISI requires to have:

$$\hat{c}(n) \star \hat{c}^{-1}(n) = \delta(n) \tag{9}$$

with $\delta(n)$ the unit impulse. Let N_c' be the length of $\hat{c}^{-1}(n)$. The former equation can then be rewritten with the following matrix form:

$$\hat{\mathbf{C}} \begin{bmatrix} \hat{c}^{-1}(0) \\ \vdots \\ \hat{c}^{-1}(N_c'-1) \end{bmatrix} = \mathbf{d} \tag{10}$$

with $\hat{\mathbf{C}}$ the $(N_c' + N_c) \times N_c'$ Toeplitz matrix built from the estimated acoustic channel response $\hat{c}(n)$ and $\mathbf{d} = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}^t$ is the vectorial representation of the unit impulse delayed by $N_c'/2 + 1$.

The least-square solution of this problem is finally given by:

$$\begin{bmatrix} \hat{c}^{-1}(0) \\ \vdots \\ \hat{c}^{-1}(N_c'-1) \end{bmatrix} = \left(\hat{\mathbf{C}}^t\hat{\mathbf{C}}\right)^{-1} \hat{\mathbf{C}}^t \mathbf{d} \tag{11}$$

4. SIMULATION RESULTS

4.1 Test plan and parameters choice

The proposed acoustic channel compensation method has been tested on five different acoustic channels. Their impulse responses were first recorded with $N_c = 300$ samples in a room environment for five different loudspeaker-microphone dispositions detailed in table 1; they have then been applied to the watermarking system to simulate the acoustic channel attack on watermarked signal.

Channel	Speaker/Microphone	
	distance	angle
1	1 m	0°
2	1 m	45°
3	15 cm	0°
4	20 cm	45°
5	50 cm	0°

Table 1: Parameters of the tested acoustic channels.

The compensation module parameters were chosen as follows: the training sequence length is $N_b = 1024$ samples (taking into account that PAM is applied on frames shorter than 20 ms), the lengths of the impulse responses are $N_c = 300$ and $N_c' = 200$ samples.

The proposed method performance is evaluated through the BER measurement over a set of 10 audio signals, sampled at $F_s = 44.1$ kHz, with various styles (jazz, man voice, classical music). 2000 bits are watermarked into each music, so that the obtained BER reliability is around 5.10^{-4}.

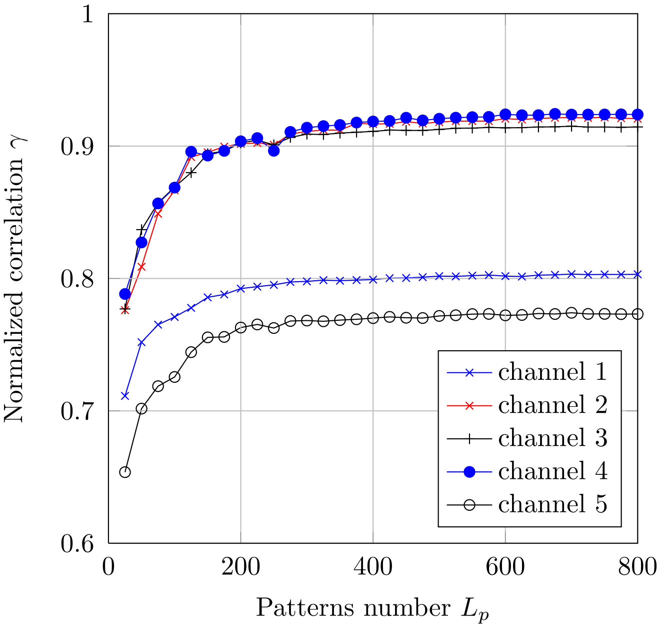

Figure 3: Normalized correlation between real and estimated acoustic channels with respect to the patterns number.

4.2 Acoustic channel estimation performance

The performance of the proposed acoustic channel estimation procedure is evaluated through a normalized correlation criterion. It is computed as the normalized correlation between the impulse response $c(n)$ of the prerecorded acoustic channel and the estimated one $\hat{c}(n)$, that is: $\gamma = \frac{\langle \mathbf{c}, \hat{\mathbf{c}} \rangle}{\|\mathbf{c}\| \, \|\hat{\mathbf{c}}\|}$, with $\langle \mathbf{c}, \hat{\mathbf{c}} \rangle = \sum_{n=0}^{N_c-1} c(n)\hat{c}(n)$, with $\hat{c}(n)$ is padded with zeros so that $c(n)$ and $\hat{c}(n)$ have the same length, $\|\mathbf{c}\| = \sqrt{\langle \mathbf{c}, \mathbf{c} \rangle}$ and $\|\hat{\mathbf{c}}\| = \sqrt{\langle \hat{\mathbf{c}}, \hat{\mathbf{c}} \rangle}$. The higher γ is, the more similar $c(n)$ and $\hat{c}(n)$ are.

The obtained normalized correlations for the five considered acoustic channels are presented in figure 3 with respect to the number L_p of embedded patterns involved in the training procedure.

The obtained results show that the estimation performance strongly depends on the acoustic channel, since for instance channel 2 is well estimated (with $\gamma = 0.9$) when the number of training patterns is high, whereas the estimation of channel 5 is acceptable with $\gamma = 0.77$. Note that no relation between the distance or the angle between the loudspeaker and the microphone and the estimation performance is displayed. The estimation procedure exhibits a systematic error, since the normalized correlation metric stagnates with high training pilot number. This bias comes mainly from the fact that the convolutive channel introduces slight differences between the perceptual shaping filter at the embedder $H(f)$ and at the receiver $\hat{H}(f)$ so that the approximation $\hat{H}(f) \approx H(f)$ no more holds; thus, the frequency shaping inversion makes the estimation of the channel-convoluted pilot imperfect.

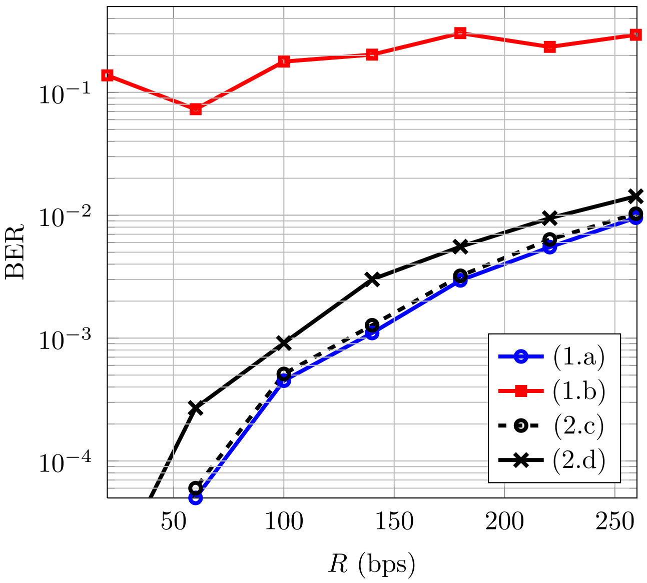

Figure 4: Averaged BER with respect to the useful bitrate R with 4 configurations: (1) the compensation module is turned off and (a) the channel is free from perturbation then (b) acoustic channels are applied; (2) acoustic channels are applied and the compensation module is turned on using, for equalization, (c) the real acoustic channel then (d) the estimated acoustic channel.

4.3 Acoustic channel equalization performance

The acoustic channel equalization efficiency is evaluated through the whole system performance in terms of BER: the mean BERs, obtained with the five considered channels, are presented in figure 4 with respect to the useful transmission bitrate R (in bps). The acoustic channel compensation module is turned on with $L_p = 100$ patterns, yielding a decrease of the transmission bitrate from 10%.

The obtained results are compared to performance of the reference system (without the acoustic channel equalization stage) when: 1) the channel is free from perturbations and 2) the channel is one of the five considered acoustic channels, but also when the equalization procedure uses the known acoustic channel $C(f)$ instead of the estimated one $\hat{C}(f)$.

These results prove the equalization procedure efficiency. System performance strongly decreases when no dedicated equalization is performed, but are quite equal to the BERs obtained with a free-from-perturbation channel when the real acoustic channel $C(f)$ is used in the dedicated equalization stage. BERs obtained with the estimated channel $\hat{C}(f)$ are slightly higher than those with the real channel, since the channel estimation is imperfect; but the system transparency to acoustic channels is almost achieved, since for instance at $R = 100$ bps the transmission achieves a BER equal to 9.10^{-4} with the proposed acoustic channel compensation module (compared to 5.10^{-4} when the channel is free from perturbation).

5. CONCLUSION

In this article, we have introduced a new method to face performance degradations of audio watermarking system in presence of acoustic channel perturbations. Our method is based on a two-stage procedure, including an estimation module and an equalization block added in amount of the system extractor. Simulations have shown the contribution efficiency with a decrease of the BER from 0.2 to 9.10^{-4} when the transmission bitrate is 100 bps bitrate whereas the acoustic channel estimation is biased and imperfect.

Future work will focus on reducing the estimation bias and on the acoustic channel time variability: the estimation stage could be replaced with a joint estimation/equalization one and be made adaptive so that the channel estimation is regularly updated with regards to the acoustic environment variations.

REFERENCES

[1] F. Cayre, C. Fontaine, and T. Furon, “Watermarking security: Theory and practice,” *IEEE Trans. Signal Processing*, vol. 53, no. 10, pp. 3976–3987, 2005.

[2] S. Larbi, M. Jaïdane, and N. Moreau, “A new Wiener filtering based detection scheme for time domain perceptual audio watermarking,” in *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 5, may 2004, pp. 949–952.

[3] M. Steinebach, A. Lang, J. Dittmann, and C. Neubauer, “Audio watermarking quality evaluation: robustness to DA/AD processes,” in *Proc. of Int. Conf. on Information Technology: Coding and Computing*, April 2002, p. 0100.

[4] D. Kirovski and H. Malvar, “Spread-spectrum watermarking of audio signals,” *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1020–1033, april 2003.

[5] C. Baras, N. Moreau, and P. Dymarski, “Controlling the inaudibility and maximizing the robustness in an audio annotation watermarking,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1772–1782, September 2006.

[6] Y. Nakashima, R. Tachibana, and N. Babaguchi, “Watermarked movie soundtrack finds the position of the camcorder in a theater,” *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 443–454, 2009.

[7] N. Lazic and P. Aarabi, “Communication over an acoustic channel using data hiding techniques,” *IEEE Transactions on Multimedia*, vol. 8, no. 5, pp. 918–924, 2006.

[8] R. Tachibana, “Audio watermarking for live performance,” in *Proc. of Security and watermarking of multimedia contents V)*, vol. 5020, January 2003, pp. 32–43.

[9] J. Chen, R. Hudson, and K. Yao, “Fast frequency-domain acoustic channel estimation with interference cancellation,” in *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2, May 2002, pp. 1709–1712.