DOPING AND WITNESS WATERMARKING FOR AUDIO PROCESSING

S. Djaziri Larbi[(1)], *G. Mahé*[(2)], *I. Marrakchi*[(1,2)], *M. Turki*[(1)], *M. Jaïdane*[(1)]

[(1)]Signals and Systems Laboratory
Ecole Nationale d'Ingénieurs de Tunis
Université Tunis El Manar
BP37 Le Belvédère-1002 Tunis(ia)

[(2)]LIPADE
Université Paris Descartes
45, rue des Saints-Pères 75270 Paris cedex 06

ABSTRACT

In the last years the research in the field of digital watermarking evolved toward non conventional watermarking applications, which we refer to as watermark aided audio processing. In this paper, we present some of our contributions to this new concept of watermarking for audio processing, namely *doping watermark* and *witness watermark*. The doping watermarking modifies the statistical characteristics of signals in order to satisfy particular conditions of algorithms or systems: it may stationarize, gaussianize a signal or make its characteristic function band-limited. We further show that a watermark can be used as a channel witness, to reproduce channel distortions and hence facilitates channel identification.

1. INTRODUCTION

Digital watermarking was proposed as a powerful way to protect digital media. Watermarking is a technique that embeds binary data, *the watermark*, in the media to be protected, *the host*, without introducing degradation, *i.e.* in an imperceptible manner.
Recently, and as watermarking is a subcategory of data hiding, many researchers presented watermarking systems thought as data transmission channels, where the channel is the host signal. In these applications, the channel, or the watermarked media, conveys hidden digital side information (which may be totally independent from the host), intended for a dedicated receiver. Under this communication model, some watermarking systems have been proposed, where the watermark conveys audio annotation/ indexation information [1] or information related to bandwidth extension for narrow band speech transmission [2].

In this paper we present a novel and non conventional aspect of audio digital watermarking, which lies within the framework of audio signal enhancement and which we refer to as watermark aided audio processing. Indeed, audio signal processing systems have to manage the difficult audio signals characteristics, *e.g.* their non stationarity and high correlation. Watermark aided audio processing can be split into two categories: *doping watermarking* and *witness watermarking*. The former is intended to change the audio signal characteristics in order to enhance the efficiency of a particular audio processing system. For example, acoustic echo cancellation (AEC), nonlinear loudspeaker identification and application of the quantization theorem require signals with respectively stationarity, Gaussianity and band-limited characteristic function. The signal may gain these properties through doping watermarking [3, 4, 5, 6]. Witness watermarking embeds a watermark in the signal to exploit its ability to identify the channel conveying the signal better than the signal itself, thanks to "good properties", mainly whiteness and stationarity [7].

An interesting particularity of this new watermarking concept is that, except for inaudibility, no further constraints are necessary, *i.e.* neither robustness nor high embedding rates are required.

The paper summarizes our work in the field of watermark aided audio processing and it is organized as follows. In sections 2 to 4 we present doping watermark concepts, respectively the stationarization, the gaussianization and the low-pass filtering of the probability density function. Section 5 details an application of the witness watermark concept in Acoustic Echo Cancelation (AEC).

2. DOPING WATERMARKING FOR AUDIO STATIONARIZATION

The considered time domain spread spectrum watermarking scheme is shown on Fig.1 (dashed part). The embedded watermark w_n is obtained by spectral shaping of the stationary and white sequence v_n. This is done through an all-pole filter $H(z)$, whose gain matches the attenuated LPC[1] spectral envelope of signal x_n[2]. Note that $H(z)$ is updated each N samples frame of x_n. The doping watermark w_n is then added to the host x_n to obtain the watermarked audio x_n^w.

Since w_n is stationary over each frame, the non stationarity of x_n^w is noticeably reduced. We address here non stationarities (NS) that consist in abrupt changes over short durations in the spectral characteristics of signals. Speech signals are made of rapid succession of noise periods (unvoiced consonants), periods of relative stability (vowels) and periods of silence. The transitions between unvoiced and voiced zones, and *vice versa*, are considered as transients. The NS is measured by stationarity indices (SIs) based on time frequency representations (TFRs). The SI was proposed in [8] as an efficient NS measure. It

This work is part of the project WaRRIS granted by the French National Research Agency (project n° ANR-06-JCJC-0009)

[1]LPC: Linear Predictive Coding.
[2]An auditory model is used in case of music signals

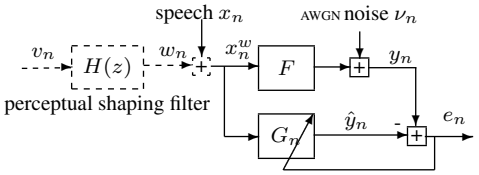


Fig. 1. Time domain watermarking system (dashed) as a preprocessing step of a generic AEC (full line).

is computed as a distance between different TFRs of the signal: at each instant n, two TFR subimages $I_1(n;\tau,f)$ and $I_2(n;\tau,f)$ with equal duration p are extracted on both sides of the instant n, then normalized and compared through a distance, here the Küllback divergence [4]:

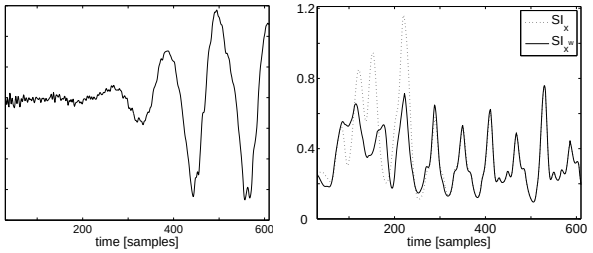
$$SI(n) = \int_{\tau=0}^{p} \int_{-\infty}^{+\infty} (I_1(n;\tau,f) - I_2(n;\tau,f)) \log \frac{I_1(n;\tau,f)}{I_2(n;\tau,f)} df d\tau.$$

If the signal characteristics have no changes at instant n, $SI(n)$ is near zero, while it peaks otherwise.
We depict on Fig.2b the SIs of an original and a watermarked speech sequence containing an unvoiced/voiced transition (Fig.2a). As expected, the SI value at the transition area (about $n = 220$) has been significantly reduced. The same *stationarizing* effect is obtained with watermarked music [4]. Thus, the watermark acts as a doping signal to enhance audio stationarity.

The proposed preprocessing by adding a doping watermark was tested with a monophonic AEC as depicted on Fig.1. In audio-conferencing, the communication quality is altered by the acoustic coupling between loudspeakers and microphones, which results in an echo transmitted through the microphones. The role of an AEC is to identify the echo path impulse response (IR) to reduce the echo.
On Fig.1, the echo path is modeled by a FIR filter F. The AEC taps G_n are updated with the NLMS algorithm according to the residual echo $e_n = y_n - \hat{y}_n$, where y_n is the echo and $\hat{y}_n$ is the estimated echo. The AEC input is the received audio signal x_w.
Adaptive algorithms are very sensitive to the non stationarity and correlation of the audio input: they interpret peaky variations of the residual echo as channel variations, which results in an altered quality of the transmitted signal after echo removal.
The AEC of Fig.1 reached an ERLE[3] enhancement of *ca.* 5 dB in the steady state [4] compared to the conventional AEC. The preprocessing improved the AEC robustness to short segment transients and led to the stabilization of the ERLE, insuring a nearly constant audio quality.

Note that w_n does not convey any particular information, it has just to be white and stationary. In the following, the watermark design is entirely focused on shaping the signal properties, released from the conventional concept of binary information.

[3]ERLE: Error Return Loss Enhancement



(a) Transient speech segment (b) Stationarity Indices SI

Fig. 2. (a) Unvoiced/voiced transition (syllable *"sa"*). (b) SI_x (dotted) and SI_{x^w}: stationarity improvement to ca 50%.

3. DOPING WATERMARKING FOR AUDIO GAUSSIANIZATION

Why audio gaussianization?

For nonlinear system identification (*e.g.* loudspeaker), several algorithms need input Gaussianity. For example, the robustness of the polynomial identification is closely related to the conditioning of the observation matrix $R_x = E[X_n X_n^t]$ (where $X_n = [1, x_n, ..., x_n^p]^t$, x_n denotes the input signal and p is the nonlinearity order) which depends on the Probability Density Function (PDF) of the input signal [5]. For all nonlinearity orders, R_x is better conditioned for a Gaussian PDF than for a Laplacian one. For nonlinear systems with memory, an identification method was proposed in [9]. Based on orthogonalization using the Hermite polynomials basis, this method relies on the hypothesis of input Gaussianity.

Usually, simple inputs such as one or two tones, white Gaussian noises and sine sweep signals [10] are used for nonlinear identification or characterization. But to capture different aspects of the nonlinear distortions in audio systems, which are physically input dependent [10], the identification or characterization should be done with audio signals. However, the latter have Generalized Gaussian distributions [11]. Therefore, we proposed an algorithm that makes audio signals more Gaussian [5].

Audio Gaussianization method

The proposed gaussianization method, detailed in [5], is based on slight changes of audio sample values (preserving inaudibility) so that the obtained PDF matches a normal distribution. We proposed a transformation of speech signals from the empirical distribution to a Gaussian distribution, performed over non overlapping frames. Each sample x_n of the audio signal is transformed into $x_n^g = x_n + w_n$, where w_n is the additive gaussianizing watermark, so that

$$F^{target}(x_n^g) = F^{emp}(x_n), \quad (1)$$

where F^{target} and F^{emp} are respectively the target normal cumulative distribution function and the empirical cumulative distribution function.
The signal w_n added through the gaussianization process is still audible because the PDF of audio is much higher than the Gaussian one around zero. The noise is particularly noticeable for silent and unvoiced areas. In order to reduce the gaussianization signal, we proposed to exclude

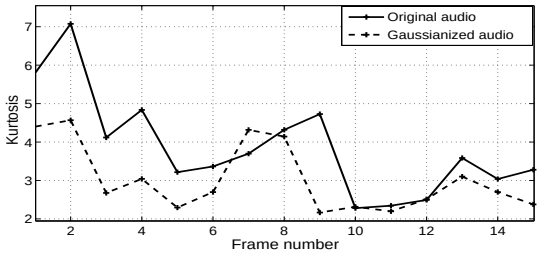


Fig. 3. Kurtosis evolution of original and gaussianized (dashed) signals (frame size=10^4 samples).

the silent and unvoiced segments from the gaussianization process and we achieved a perceptual frequency masking by an iterative limitation of the maximum amplitude of w_n. This embedded signal represents the doping watermark for audio gaussianization.
The Gaussianity is measured by estimating the Kurtosis (which equals 3 for a Gaussian input) for original and gaussianized audio signals. The Kurtosis was evaluated for a pop-music signal by disjoint frames of 10.000 samples. As shown in Fig.3, the Kurtosis of x_n^g is closer to 3 than that of x_n for most of the frames. Consequently, the condition number of R_x is between those of a Gaussian signal and the original speech. This leads to a more robust identification and a significant reduction of the estimation error (25 dB in the studied example), while identifying a memoryless nonlinear system [5].

4. DOPING WATERMARKING AND QUANTIZATION THEOREM

According to the quantization theorem [12], the PDF of a sampled signal can be recovered from the signal quantized with step q if $1/q$ is lower than two times the maximum frequency of the characteristic function[4]. If the original signal is digital, *i.e.* with a discrete PDF, the quantization theorem turns into the sub-quantization theorem [6]: if the characteristic function of a quantized signal x equals zero for frequencies $|\nu| > \frac{1}{2K}$ in $\left[-\frac{1}{2};\frac{1}{2}\right]$, then the PDF of x can be recovered from that of x_Q resulting from the sub-quantization of x with factor K.

Characteristic functions of audio signals generally spread all over the frequency space. Thus, the PDF should be low-pass filtered, with cut-off frequency $\frac{1}{2K}$, in order to apply the quantization or the sub-quantization theorem. In [6], we proposed a doping watermarking scheme for digital signals aiming at limiting the characteristic function to $[-\frac{1}{2K};\frac{1}{2K}]$. The principle is to add an inaudible noise w to x, so that the PDF of $x^w = x + w$ equals the target low-pass filtered PDF.
In practice, instead of PDF we consider histograms of long frames (*e.g.*10.000 samples) of audio signal. The low-pass filtered histogram is rounded to integer values so that its total number of samples equals the length N of the frame. We transform iteratively x into a signal x^w so that the histogram of x^w, h_{x^w}, equals the target histogram h_{target}. Initially, $x^w = x$. Then, for $i = \min x \rightarrow \max x$:
– If $h_{x^w}(i) > h_{target}(i)$, then we select randomly the excess samples x_n^w of value i, give them the value $i+1$ and actualize $h_{x^w}(i+1)$.
– If $h_{x^w}(i) < h_{target}(i)$, then we select randomly the missing samples x_n^w in the class $i+1$, give them the value i and actualize $h_{x^w}(i+1)$. If there are not enough samples in the class $i+1$ we select the remainder in the class $i+2$ and so on.
At the end of the algorithm, $h_{x^w} = h_{target}$.
The doping watermark w is all the more audible as K is high. For speech, w is almost inaudible for $K = 4$ (PESQ [13] Mean Opinion Score > 4), whereas for violin K can be set up to 8 (PEAQ [14] Objective Difference Grade $\simeq -0.5$) [6].

[4]Fourier transform of the PDF

signal	K	$d_{KS} \times 10^4$	
		x	z
speech	4	69	4.3
	8	197	5.9
violin	4	7.3	1.4
	8	13	1.3

Table 1. PDF recovery error.

The algorithm used for gaussianization (section 3) could have been used too. For a given value of K, the perceptual result is identical. The main difference is that the previous algorithm, based on the minimization of the difference between the actual and the target cumulative functions, does not lead exactly to $h_{x^w} = h_{target}$, which is guaranteed by construction here. This performance is reached at the expense of a higher complexity, so that the choice between the two algorithms depends on the required accuracy of the PDF transformation.

The PDF recovery error after sub-quantization with rate K is measured, for the original signal x and the watermarked signal x^w, through the Kolmogorov-Smirnov distance d_{KS} between the original and the recovered PDF using the sub-quantization theorem. As illustrated in table 1 [6], the doping watermark reduces significantly the recovery error.

5. WITNESS WATERMARKING FOR AEC

In section 2, the sensitivity of adaptive algorithms to the non-stationarity of the input was faced by reducing the input non-stationarity through doping watermarking. Here, we propose to identify the echo path directly from the watermark itself, in order to take advantage of its good properties. In other words, the watermark is used as a *witness* of the acoustic channel.

The AEC of Fig.1 is then so modified to obtain the proposed Watermarked AEC (WAEC) of Fig.4, which is based on the coupling of two adaptive filters. The input to the first stage is the watermarked speech x_n^w. w_n is obtained by filtering the white and stationary sequence v_n through $H_n(z)$, an adaptive perceptual filter derived from the NLMS adapted whitening filter of x_n.
The first stage is a conventional AEC using the prewhitening filter $\alpha_n H_n^{-1}(z)$, where α_n is the numerator of $H_n(z)$.

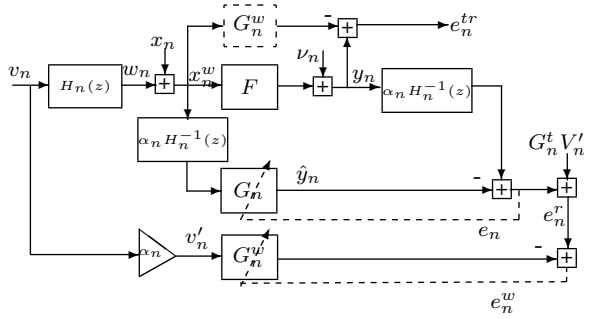


Fig. 4. Proposed Watermarked AEC

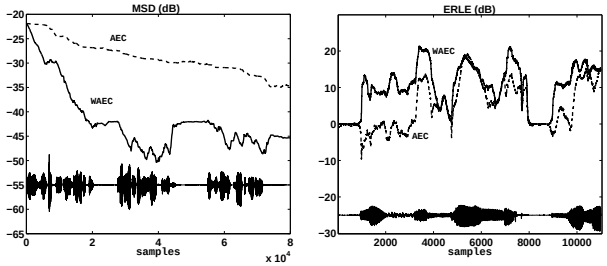


Fig. 5. MSD and ERLE variation in both cases: WAEC and conventional AEC (dotted) with watermarked input (Fig.1).

It extracts a rough estimation G_n of the echo path F. Substracting $G_n^t V_n'$ from the residual echo of this first stage provides a new reference echo signal for the second adaptive filter G_n^w

$$e_n^r = F^t V_n' + \nu_n^f + (F - G_n)^t X_n^f, \tag{2}$$

which is the echo of the watermark v_n plus a noise that should be low if G_n is close enough to F. In this equation, vectors are denoted by capital letters, *i.e.* $V_n' = [v_n', v_{n-1}', \cdots, v_{n-P}']^t$ and superscripts ^{f} and ^{t} denote respectively signals filtered by $\alpha_n H_n^{-1}(z)$ and vector transposition. Filters G_n and G_n^w have P taps and are supposed to be the same length as the echo path F.

Driven by the known watermark $v_n' = \alpha_n v_n$, the second adaptive filter G_n^w estimates then the actual echo path, taking advantage of the optimal properties of v_n. The residual echo of the second stage is then

$$e_n^w = (F - G_n^w)^t V_n' + (F - G_n)^t X_n^f + \nu_n^f \tag{3}$$

Since the adaptive filter G_n^w is driven by the nearly white and stationary sequence v_n', it provides a better identification of the echo path F.
Finally, the actual transmitted residual echo is

$$e_n^{tr} = \underbrace{(F - G_n^w)^t}_{\text{deviation } D} X_n^w + \nu_n. \tag{4}$$

The performance of the WAEC and the AEC of Fig.1 are compared on Fig.5. The WAEC performs better than the AEC, as the MSD[5] is much lower (ca -15 dB in the mean) and the ERLE is significantly higher. These results indicate the significant AEC enhancement reached by exploiting the watermark as a channel witness.

6. CONCLUSION

In this paper, some of our contributions to the new concept of watermark aided signal processing were presented. The doping watermark modifies imperceptibly audio characteristics, namely Gaussianity, stationarity or characteristic function bandwidth to fulfill assumptions required by audio processing algorithms. The witness watermark undergoes the same distortion as the audio source and it is then used more efficiently to reduce the distortion. The application field of this new concept is large, it concerns as well linear and nonlinear identification -such as AEC and loudspeaker distortions analysis- as audio quantization. Doping and witness watermarking, for watermark aided audio processing, seem to be new concepts which deserve further theoretical and practical investigations.

[5]Mean squared deviation MSD=$E[D^t D]$.

7. REFERENCES

[1] C. Baras, N. Moreau, and P. Dymarski, "Controlling the inaudibility and maximizing the robustness in an audio annotation watermarking system," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 14, no. 5, 2006.

[2] A. Sagi and D. Malah, "Bandwith extension of telephone speech aided by data embedding," *EURASIP Journal on Advances in Signal Processing*, 2007.

[3] A. Gilloire and V. Turbin, "Using auditory properties to improve the behaviour of stereophonic acoustic echo cancellers," in *ICASSP*, 1998.

[4] S. Djaziri Larbi and M. Jaïdane, "Audio watermarking: a way to modify audio statistics," *IEEE Trans. on Signal Processing*, vol. 53, no. 2, 2005.

[5] I. Marrakchi, G. Mahé, M. Jaïdane-Saïdane, S. Djaziri-Larbi, and M. Turki-Hadj Alouane, "Gaussianization method for identification of memoryless nonlinear audio systems," in *EUSIPCO*, Poland, 2007.

[6] H. Halalchi, G. Mahé, and M. Jaïdane, "Revisiting quantization theorem through audiowatermarking," in *ICASSP*, 2009.

[7] I. Marrakchi, M. Turki-Hadj Alouane, S. Djaziri-Larbi, M. Jaïdane-Saïdane, and G. Mahé, "Speech processing in the watermarked domain: Application in adaptive echo cancellation," in *EUSIPCO*, Italy, 2006.

[8] H. Laurent and C. Doncarli, "Abrupt changes detection in the time-frequency plane," in *Proc. IEEE Int. Symp. on Time-Frequency and Time-Scale Analysis*, 1996.

[9] V. J. Mathews, "Orthogonalization of correlated gaussian signals for volterra system identification," *IEEE Signal Proc. Letters*, vol. 2, no. 10, 1995.

[10] W. Klippel, "Assessing large signal performance of transducers," in *11th Regional Convention of the Audio Eng. Society*, Japan, 2003.

[11] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Proc. Letters*, vol. 10, no. 7, 2003.

[12] B. Widrow, "A Study of rough amplitude quantization by means of Nyquist sampling theory," *IRE Trans. on Circuit Theory*, vol. 3, no. 4, 1956.

[13] ITU, "Perceptual evaluation of speech quality," ITU-T Rec. P.862, 2001.

[14] ——, "Method for objective measurement of perceived audio quality," ITU-R Rec. BS.1387, 1998.