Watermark-Driven Acoustic Echo Cancellation

Sonia Djaziri-Larbi, Gael Mahé, Imen Mezghani, Monia Turki, Mériem Jaidane

Watermark driven Acoustic Echo Cancellation

Sonia Djaziri-Larbi, Gaël Mahé, Imen Mezghani, Monia Turki, and Mériem Jaïdane

***Abstract*—The performance of adaptive acoustic echo cancelers (AEC) is sensitive to the non-stationarity and correlation of speech signals. In this article, we explore a new approach based on an adaptive AEC driven by data hidden in speech, to enhance the AEC robustness. We propose a two-stage AEC, where the first stage is a classical NLMS-based AEC driven by the far-end speech. In the signal, we embed -in an extended conception of data hiding- an imperceptible white and stationary signal, *i.e.* a watermark. The goal of the second stage AEC is to identify the misalignment of the first stage. It is driven by the watermark solely, and takes advantage of its appropriate properties (stationary and white) to improve the robustness of the two-stage AEC to the non-stationarity and correlation of speech, and thus reduce the overall system misadjustment. We test two kinds of implementations: in the first implementation, referred to as A-WdAEC (Adaptive Watermark driven AEC), the watermark is a white stationary Gaussian noise. Driven by this signal, the second stage converges faster than the classical AEC and provides better performance in steady state. In the second implementation, referred to as MLS-WdAEC, the watermark is built from maximum length sequences (MLS). Thus, the second stage performs a block identification of the first stage misalignment, given by the circular correlation watermark/pre-processed version of the first stage residual echo. The advantage of this implementation lies in its robustness against noise and under-modeling. Simulation results show the relevance of the "watermark-driven AEC" approach, compared to the classical "error driven AEC".**

***Index Terms*—Adaptive Acoustic Echo Cancellation, data hiding, speech watermarking, MLS sequences, perceptual masking.**

I. Introduction

In audio- and video-conferencing, the communication quality is altered by the acoustic coupling between loudspeakers and microphones, which results in an echo transmitted through the microphones. The echo is a sound caused by the reflection of sound waves from a surface back to the listener or speaker. It is generally modeled as the convolution of the original sound (here the loudspeaker output) with the impulse response (IR) of the conference room. The role of an adaptive acoustic echo canceler (AEC) is to identify the IR in order to reduce the echo in a robust manner, even in presence of non-stationary input and ambient noise.

The limits of conventional AECs have been widely covered by researchers in the field (see for example [1]), and many enhancements were proposed to address AEC sensitivity to the correlation and non-stationarity of the input.

Power normalized adaptive algorithms (*e.g.* the NLMS[1]) reduce the impact of amplitude variation only in the mean, and not locally in the transient zone [1], [2]. Similarly, the affine projection algorithm (APA) [3] reduces the effect of variations of the frequency content only in the mean.

Several approaches were proposed in the literature to strengthen the robustness of adaptive algorithms against correlation and non-stationarity of the speech input. They are essentially based on two principles: either pre-processing the speech to soften the inappropriate variabilities of the adaptive AEC (e.g. pre-whitening techniques [4]–[6]); or using a "smart" step-size that scans the dynamics of the global adaptive AEC (e.g. gradient adaptive step size [7] and a variety of variable step size algorithms [8]–[11]).

While all of these methods use the far-end speech as the driving signal, the aim of the proposed system is to implicitly drive the AEC with a white and stationary auxiliary signal in order to cope with the sensitivity of adaptive AEC algorithms to the correlation and non-stationarity of the input. The main idea of this study is borrowed from recent applications of data hiding, where information is embedded to enhance or assist a particular processing system. These applications address a variety of signal processing issues as source separation [12]–[15], speech bandwidth extension [16], [17], packet loss concealment for wireless communications [18], [19], voicing of animated GIF [20], pre-echo reduction in audio coding [21], audio statistics modification [22]–[25], synchronization and channel equalization [26], and watermark-aided processing for linear and nonlinear audio system identification [27], [28].

In this article, we address the latter application, namely watermark-aided system identification, and particularly acoustic echo cancellation aided by data embedded in the driving audio input. This *watermark-driven AEC* (WdAEC) has its origins in [24], [29], where inserting a stationary watermark in the input audio signal enhances its stationarity and thus the performance of the AEC. Our goal is to take full advantage of the appropriate characteristics of a watermark: we propose a two-stage AEC, where the first stage is a classical NLMS-based AEC driven by the watermarked far-end speech, and the second stage is driven by the watermark solely and adaptively identifies the first stage misalignment.

The proposed WdAEC is an enhanced version of the two-stage AEC of [28], especially in terms of perceptual quality of the watermarked speech. Indeed, in this study, a frame-adaptive perceptual spectral shaping is used, and low-energy speech frames are not watermarked. Also, the input whitening filter used in the AEC structure of [28] has been removed: the

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

S. Djaziri Larbi, M. Turki and M. Jaïdane are with the Systems and Signals Lab (U2S), National Engineering School of Tunis, University of Tunis El Manar, Tunisia, e-mail:{sonia.djaziri-larbi, monia.turki, meriem.jaidane}@enit.utm.tn

G. Mahé is with LIPADE, Paris-Descartes University, France. e-mail: gael.mahe@mi.parisdescartes.fr

I. Mezghani is with the National Engineering School of Sousse, Tunisia, e-mail: mezghani_imen@yahoo.fr

[1]NLMS: Normalized Least Mean Squares.

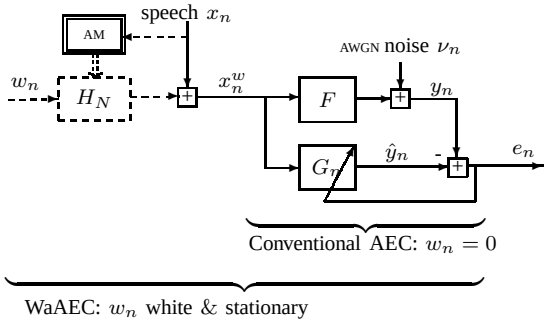

Fig. 1. WaAEC (Watermark aided AEC [29]): Watermark embedding system (dashed) as a pre-processing step of a conventional AEC (full line: $w_n = 0$). AM: Auditory Model, H_N perceptual filter.

performance enhancement achieved by the WdAEC is due to the watermarking solely.

The article is structured as follows. The principles and the structure of the proposed WdAEC are described in Section II. Two versions of the WdAEC are proposed and analyzed in Sections III and IV. In the first version, referred to as Adaptive WdAEC (A-WdAEC), the second stage is adaptive and driven by a stationary white Gaussian noise. In the second one, referred to as MLS-WdAEC, the second stage is based on a block identification using maximum length sequences, known for their performance in linear identification.

Finally, we compare and discuss the performance of A-WdAEC and MLS-WdAEC in Section V,where we also present a brief comparison with a state-of-the-art algorithm [8].

II. Methodology: Principles of Watermark driven AEC (WdAEC)

We propose a new AEC concept, driven by a white and stationary watermark that is embedded in the AEC speech input. This concept, hereafter referred to as Watermark driven AEC (WdAEC), takes advantage of the appropriate characteristics of the watermark signal w_n to enhance the echo cancellation performance. It is important to note that the auxiliary signal w_n does not convey any particular information (even if it could be the case), it just has to be white and stationary. It is referred to as watermark only because of the embedding technique borrowed from data hiding/steganography principles.

In this section, we put forward the methodology behind the design of the proposed AEC system, from which two different design versions are presented in Sections III and IV. We first briefly remind the conventional NLMS-adapted AEC and its limits. We then describe a previously developed AEC system [24], [29] driven by a watermarked speech input, here referred to as the Watermark aided AEC (WaAEC), and we explain the advantages of its concept. These steps finally lead to the design of a watermark-driven -and not only aided- AEC, the WdAEC, which is driven by the watermark solely and thus fully exploits both its properties, stationarity and whiteness.

In this work, the NLMS algorithm has been chosen to illustrate the methodology and the simulation results of the proposed system. Nevertheless, the proposed system may be applied to any conventional AEC algorithm.

A. Conventional AEC

The principle of a conventional time domain monophonic AEC is depicted in Fig. 1 (full line scheme: $w_n = 0$). The IR of the echo path to be identified is assumed to be time invariant and denoted by the taps vector $F = [f_0, f_1, \cdots, f_{p-1}]^t$, where p is the length of the IR F and $(\cdot)^t$ is the vector transpose operator. The AEC input is the received far-end speech x_n. The AEC taps $G_n = [g_0(n), \cdots, g_{p-1}(n)]^t$ are updated with the NLMS algorithm according to the residual echo $e_n = y_n - \hat{y}_n$ as follows:

$$\begin{cases} G_{n+1} &= G_n + \mu_n e_n X_n, \\ e_n &= (F - G_n) * x_n + \nu_n, \end{cases} \quad (1)$$

where n is the discrete time index, $X_n = [x_n, x_{n-1}, \cdots, x_{n-p+1}]^t$ the input signal vector, and $\mu_n = \mu / \|X_n\|^2$ the normalized adaptation step size with μ a fixed step size. ν_n is an additive white Gaussian noise (AWGN), y_n is the echo with noise and $\hat{y}_n$ is the estimated echo ($'*'$ denotes convolution).

Conventional AEC systems are driven by the speech signal, which is non-stationary and highly correlated. The performance of AEC systems suffers from these unsuitable speech characteristics as adaptive algorithms converge faster if the input signal is white. In the steady state, adaptive algorithms are very sensitive to the non-stationarity of the input. Indeed, peaky variations of the residual echo are interpreted as channel variations, and the algorithm gets into tracking mode to pursue those variations [1]. This situation results in the degradation of the AEC performance, which is equivalent to an altered quality of the transmitted signal after echo removal. Several improvements were proposed to address this problem, mainly focusing on fitting algorithms with the input signal properties [4]–[11]. In this study, we consider a conventional identification algorithm, the NLMS, and we focus on adapting the input signal to the algorithm.

B. Watermark aided AEC: the WaAEC

As a first step -and to address the non-stationarity of the input x_n- the authors proposed in [24] to add a stationary (and white) watermark w_n to the speech input, as depicted in Fig. 1 by the dashed lines. H_N is a perceptual filter (its gain approximates the frequency masking threshold of the analyzed speech frame) and is updated every N-samples frame (cf. Appendix A). Since the spectrally shaped watermark $w_n * h_N$ -where h_N is the IR of H_N- is stationary over each N-samples frame, the non-stationarity of the watermarked signal

$$x_n^w = x_n + w_n * h_N \quad (2)$$

is noticeably reduced. This was reported in [24], [25] where time-frequency stationarity indices [25], [30] were used to assess abrupt changes in signal characteristics.

The AEC G_n of Fig. 1 is adapted by the NLMS algorithm according to (1), as in the conventional case, except that X_n is replaced by the watermarked speech vector X_n^w:

$$\begin{cases} G_{n+1} &= G_n + \mu_n e_n X_n^w, \\ e_n &= (F - G_n) * x_n^w + \nu_n, \end{cases} \quad (3)$$

where $X_n^w = [x_n^w, x_{n-1}^w, \cdots, x_{n-p+1}^w]^t$.

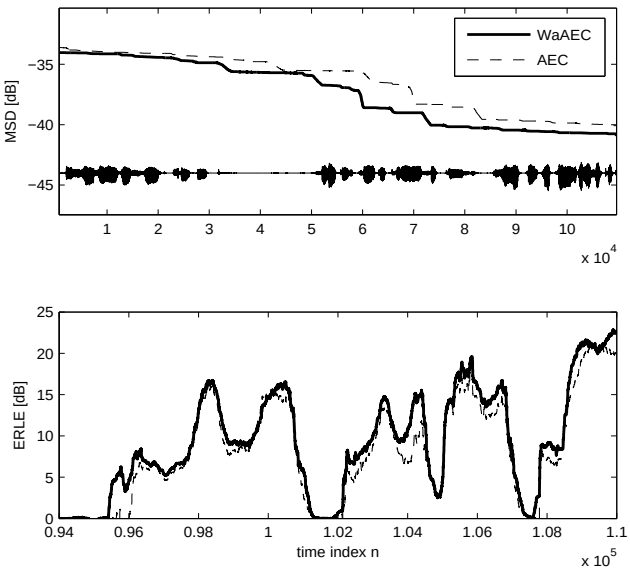

Fig. 2. Performance of the WaAEC and of the conventional AEC (SNR=30dB, $\mu = 0.02$, echo path and adaptive filter with $p = 200$ taps). Top panel: mean squared deviation (MSD). Bottom: zoom of the echo return loss enhancement (ERLE) in steady state.

The WaAEC performance described by (3) is given in Fig. 2, where it is compared to the performance of a conventional AEC: the WaAEC reaches an ERLE[2] improvement of *ca.* 2 to 5dB in the steady state as compared to the conventional AEC.

C. Principles of the Watermark driven AEC (WdAEC)

The performance enhancement reached by the WaAEC suggests the design of an AEC that fully exploits both stationarity and whiteness of the watermark: an AEC driven by the watermark itself. If we filter the output e_n of the diagram of Fig. 1 by the inverse of H_N, denoted by H'_N, we get

$$\begin{aligned} e'_n &= e_n * h'_N \\ &= \underbrace{(F - G_n)}_{D_n} * w_n + \xi_n, \end{aligned} \tag{4}$$

where h'_N is the IR of H'_N and

$$\xi_n = [(F - G_n) * x_n + \nu_n] * h'_N. \tag{5}$$

Hence, the identification task comes back to identifying the 1st stage misalignment $D_n = F - G_n$, where the input is the stationary and white signal w_n and the background noise is ξ_n. This system, named WdAEC, is depicted in Fig. 3: the 1st stage is the WaAEC of Fig. 1 driven by the watermarked speech x_n^w, and the 2nd stage AEC is driven by w_n and uses the filtered 1st stage residual echo e'_n as reference signal. In this configuration, the new actually transmitted echo is:

$$\begin{aligned} e_n^{tr} &= e_n - \hat{D}_n * x_n^w \\ &= [D_n - \hat{D}_n] * x_n^w + \nu_n, \end{aligned} \tag{6}$$

where $\hat{D}_n$ is the 2nd stage estimate of $D_n = F - G_n$. Hence, the goal of the 2nd stage is to estimate the 1st stage misalignment in order to further reduce the overall system's residual echo.

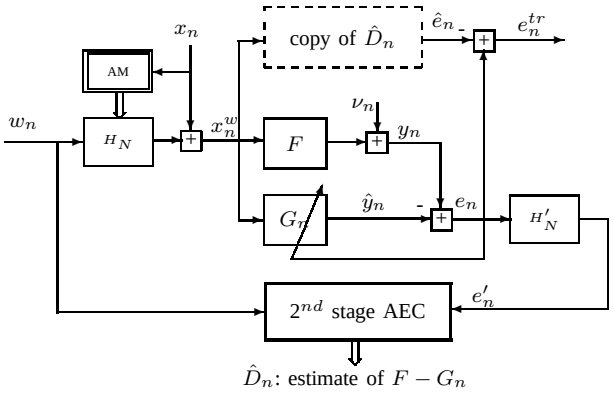

Fig. 3. WdAEC: two-stage Watermark driven AEC with spectrally shaped watermark.

Fig. 4 is an equivalent block diagram of the WdAEC system, where the 2nd stage identification may be performed adaptively (*e.g.* by the NLMS algorithm) or on a frame-by-frame basis, as proposed in Sections III and IV respectively. In both cases, it is expected that the 2nd AEC stage takes full advantage of the whiteness and stationarity of its input w_n to perform an efficient identification of the misadjustment $F - G_n$.

It is worth noting that both stages do not identify the same path. In addition, the second stage identifies a time varying misalignment $F - G_n$. We also emphasize that the first stage AEC, i.e. the WaAEC, has proven to perform better than the classical AEC (cf. Section II-B). For this reason, the performance of the WaAEC is used as a reference for all following performance evaluations and comparisons.

Before introducing both implementations of the WdAEC in Sections III and IV, we explain in the following the watermark embedding procedure and how it is adapted to meet the requirements of the proposed AEC system.

D. The watermark embedding process of the WdAEC

The embedding process used in this study -and detailed in Appendix A- is carried out in real time on a frame-by-frame basis (N samples per frame, i.e. 20msec). However, for the purpose of the proposed WdAEC, only speech frames where the signal energy is sufficiently high are watermarked. This is done mainly for two reasons:

- the perceptual filter H_N of (17) is an approximation of the frequency masking threshold of the analyzed speech frame. Hence, embedding watermarks in frames where

[2]ERLE: Echo Return Loss Enhancement, cf. Appendix C

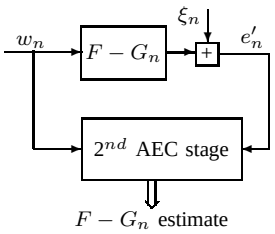

Fig. 4. WdAEC: Equivalent block diagram of Fig. 3.

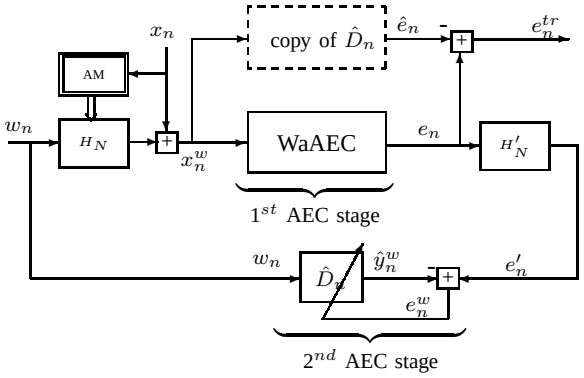

Fig. 5. A-WdAEC (Adaptive WdAEC): the 1st stage AEC is the WaAEC and the 2nd stage is a NLMS adapted filter $\hat{D}_n$.

voice activity is low introduces audible distortions and the listening quality of the far-end speech is impaired,

- low-energy speech frames yield very small λ_N values (cf. Appendix A-A). This leads to a high-energy noise ξ_n when filtering the residual echo e_n by the inverse filter H'_N. This noise amplification considerably penalizes the performance of the second stage AEC.

To address these issues, only frames with λ_N greater than a given threshold, denoted here by $\mathbf{\Lambda}$, are watermarked. This acts as a pseudo voice activity detector: during non watermarked signal segments, where the energy coefficient λ_N is smaller than $\mathbf{\Lambda}$, the 2nd stage AEC freezes the identification and the most recent estimation of $F - G_n$ is used.

This trade-off between speech quality and AEC performance implicitly controls the overall embedding rate ($\%w$), which is computed as the ratio between the number of watermarked frames and the total number of frames in the speech sample: the smaller $\mathbf{\Lambda}$ is, the higher the embedding rate is. On the other hand, when the embedding rate $\%w$ is too low, the watermark signal is not sufficiently present in the reference signal e'_n for the 2nd stage identification which may decrease its performance. However, it is worth noting that high embedding rates do not necessarily yield high AEC performance. Indeed as mentioned above, when the threshold $\mathbf{\Lambda}$ is low, even though the embedding rate is relatively high, the inverse filtered noise ξ_n is so amplified that the second stage AEC performs poorly.

The relationship between embedding rate and listening quality of the watermarked speech is given in Table I of Appendix A for different languages and voice activity thresholds $\mathbf{\Lambda}$. For the simulations presented in this paper, we used a French speech sample, with $\mathbf{\Lambda} = 0.003$ as a trade-off between listening quality and AEC performance. This setting yields an average PESQ MOS[3] of 3.5 and *ca.* 44% embedding rate.

III. Adaptive Watermark driven AEC: the A-WdAEC

In this Section, the second AEC stage is implemented as a NLMS adapted filter $\hat{D}_n$, as shown in Fig. 5. The watermark w_n is white Gaussian with unit variance. The AEC taps of $\hat{D}_n$ are adapted according to:

$$\hat{D}_{n+1} = \hat{D}_n + \mu_n^w e_n^w W_n, \tag{7}$$

where $\mu_n^w = \mu^w / \|W_n\|^2$ is the normalized step size and $W_n = [w_n, w_{n-1}, \cdots, w_{n-p+1}]^t$ is the input vector. The signal e_n^w controlling the second stage is obtained using (4) and is expressed by:

$$e_n^w = \underbrace{\left[(F - G_n) - \hat{D}_n\right]}_{D_n^w} * w_n + \xi_n, \tag{8}$$

where D_n^w is the misalignment vector related to $\hat{D}_n$, and ξ_n is the filtered equivalent noise of (5).

A. A-WdAEC: performance analysis

The transient and steady states of both the WaAEC and the A-WdAEC are described by analyzing the deviation vectors $D_n = F - G_n$ and $D_n^w = (F - G_n) - \hat{D}_n$, derived from (8), and (3) and (7) respectively (see for example [1]):

$$D_{n+1} = \left(\mathbf{I} - \mu \frac{X_n^w (X_n^w)^t}{\|X_n^w\|^2}\right) D_n - \underbrace{\mu \nu_n \frac{X_n^w}{\|X_n^w\|^2}}_{a_n} \tag{9}$$

$$D_{n+1}^w = \left(\mathbf{I} - \mu^w \frac{W_n (W_n)^t}{\|W_n\|^2}\right) D_n^w - \underbrace{\mu^w \xi_n \frac{W_n}{\|W_n\|^2}}_{b_n} \tag{10}$$

where $\mathbf{I}$ is the identity matrix. Classically, the transient and steady state behaviors of adaptive systems are described by the instantaneous mean deviation and the mean squared deviation (MSD) [1]. The mean deviation vectors of WaAEC and A-WdAEC are given by

$$\mathrm{E}[D_{n+1}] = (\mathbf{I} - \mu \mathbf{R}_{\mathbf{xw}}(n))\, \mathrm{E}[D_n], \tag{11}$$

$$\mathrm{E}[D_{n+1}^w] = (\mathbf{I} - \mu^w \mathbf{R}_{\mathbf{w}}(n))\, \mathrm{E}[D_n^w], \tag{12}$$

where $\mathrm{E}[\cdot]$ denotes the mathematical expectation. $\mathbf{R}_{\mathbf{w}}(n) = \mathrm{E}\left[\frac{W_n (W_n)^t}{\|W_n\|^2}\right]$ and $\mathbf{R}_{\mathbf{xw}}(n) = \mathrm{E}\left[\frac{X_n^w (X_n^w)^t}{\|X_n^w\|^2}\right]$ are the autocorrelation matrices of w_n and x_n^w.

Equation (11) is obtained using the statistical independence of ν_n and x_n^w and according to the classical hypothesis of statistical independence of X_n^w and D_n [1]. Similarly, in (12) we assumed that ξ_n has a zero mean and is statistically independent of w_n and that W_n is independent of D_n^w.

It is well known that the lower the condition number of the autocorrelation matrix of the AEC input, the faster the AEC convergence [1]. As w_n is obviously less correlated than x_n^w, the convergence speed described by (12) is expected to be significantly higher [28].

The analytical study of the MSDs $E[\|D_n\|^2]$ and $E[\|D_n^w\|^2]$ is difficult to carry out as the involved signals are correlated and non-stationary. However, one can see from (9) and (10) that the steady state performance of the WaAEC and the A-WdAEC adaptive filters depends on the instantaneous power of the terms a_n and b_n, respectively. The noise ξ_n is obviously higher (due to the inverse filter H'_N) and less stationary than

[3]PESQ MOS: Perceptual Evaluation of Speech Quality Mean Opinion Score [31].

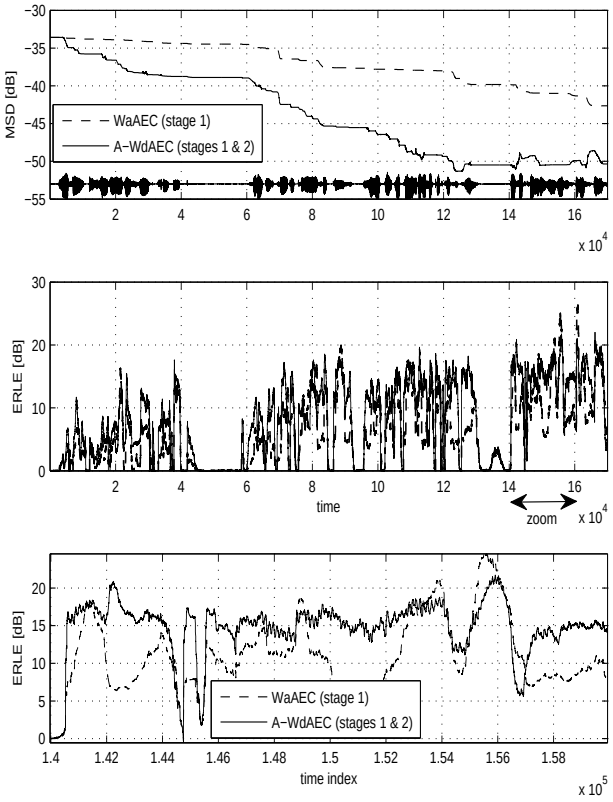

Fig. 6. Top panel: Instantaneous MSD of the A-WdAEC and of the first stage (dotted) only (WaAEC). Middle and bottom panels: Instantaneous ERLE of A-WdAEC and WaAEC (dotted). $\mu = \mu^w = 0.02$, SNR=30 dB, p=200 taps.

ν_n (AWGN). However, $\frac{W_n}{\|W_n\|^2}$ has smoother time variations than $\frac{X_n^w}{\|X_n^w\|^2}$.

Hence, the performance of the NLMS-based second stage is due to the following effects:

- the whiteness of the driving signal w_n enhances the convergence speed;
- the steady state behavior is impaired by the high-variance and the non-stationarity of the noise ξ_n, but this is counterbalanced by the stationarity of the driving signal.

B. Simulation results

The proposed A-WdAEC was simulated to estimate the acoustic IR of a car with $p = 200$ taps, and 16kHz sampled speech with a signal-to-noise ratio SNR=30 dB. The performance of the A-WdAEC is compared with that of the WaAEC (first stage of the A-WdAEC) in Fig. 6, in the case of exact modeling (both AECs are FIR filters with p taps).

To compare the transient state behaviors, the MSD of both stages is depicted in Fig. 6 (top panel), where both A-WdAEC step size values were set to $\mu = \mu^w = 0.02$. As expected theoretically, and since the watermark w_n is less correlated than the watermarked speech x_n^w, the convergence speed of the A-WdAEC is higher than that of the reference WaAEC.

The ERLE -defined by (26)- of both systems is plotted in the middle and zoomed in the bottom panels of Fig. 6. The residual echo of the A-WdAEC is e_n^{tr} of (6) and that of the WaAEC is

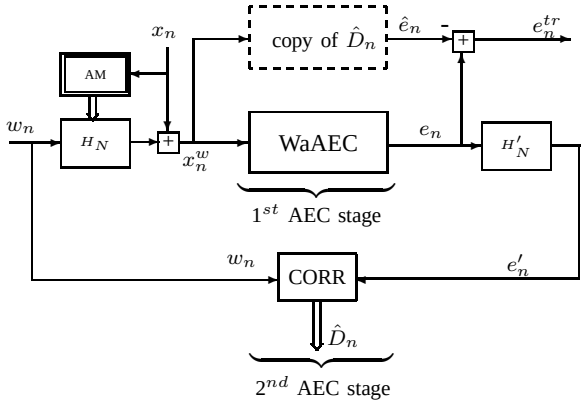

Fig. 7. MLS-WdAEC: block diagram of the two-stage WdAEC with a circular cross-correlator as 2nd AEC stage.

e_n, given by (3). The results show that, in the steady state, the ERLE improvement of the A-WdAEC reached up to 10dB.

C. First conclusions

In this Section, a watermark-driven AEC has been developed, which was designed to be robust against the correlation and non-stationarity of far-end speech. The simulation results provided in section III-B show that the A-WdAEC reached significant enhancement as compared with the reference WaAEC. It is worth noting that despite the high-variance and non-stationary equivalent noise ξ_n, the second stage AEC outperformed the first stage thanks to the appropriate properties of its input w_n.

In the following Section, we present a different implementation of the WdAEC, namely the MLS-WdAEC, the design of which is expected to be even more robust in noisy situations.

IV. MLS WATERMARK-DRIVEN AEC: THE MLS-WDAEC

The main interest of the proposed echo path estimation method lies in the insertion of an L-periodic Maximum Length Sequence (MLS) in the received far-end speech. Indeed, in addition to the required whiteness and stationarity, MLS has advantageous circular correlation properties [32]. The mathematical background behind the use of MLS in the estimation of channels IR is detailed in Appendix B. In particular, when the input to a channel is an MLS, the channel's IR is estimated directly from the circular cross-correlation between the channel's input and output (Fig. 17) [33].

Hence, we embed an MLS watermark w_n in the AEC speech input and we design the 2nd WdAEC stage (Fig.3 and Fig. 4) as a circular cross-correlator. This delivers an estimate $\hat{D}_n$ of the 1st stage misalignment $D_n = F - G_n$ as the cross-correlation between the MLS watermark w_n and the filtered 1st stage residual echo e'_n. We then obtain the new WdAEC design, denoted by MLS-WdAEC and presented in Fig. 7.

As the high identification performance of MLS lies in its circular correlation property and noise immunity [34], it is expected that the MLS-WdAEC will provide a higher robustness to noise, even if the latter is non-stationary and correlated.

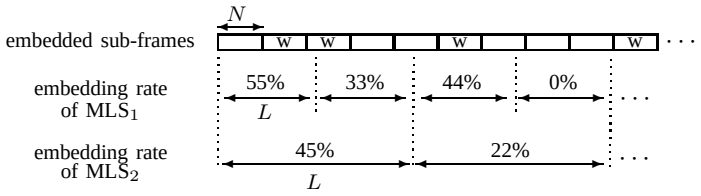

Fig. 8. Diagram of the embedding process of the L-periodic MLS and resulting embedding rate with two different L values. Only sub-frames marked with 'w' are embedded in the received speech due to voice activity threshold.

A. MLS-WdAEC design

1) MLS embedding: The input x_n^w of the MLS-WdAEC system of Fig. 7 is composed of the received far-end speech x_n, in which we embed an L-periodic MLS w_n, spectrally shaped according to Appendix A and subsection II-D. The embedding process of the L-periodic MLS in the speech signal and its piecewise perceptual spectral shaping are outlined in Fig. 8. Each MLS period with length L is segmented into sub-frames with N samples. The sub-frames are spectrally shaped with the perceptual filter H_N if the energy of the host speech frame is higher than the threshold $\mathbf{\Lambda}$ (cf. Section II-D), otherwise no watermark is embedded. Consequently, several MLS periods are only partially embedded in the speech signal. Fig. 8 displays two examples of MLS with different lengths L and the resulting embedding rate per L-period. The latter point is important with respect to identification freezing when the embedding rate is lower than a given threshold. This issue is addressed in Subsection IV-B2.

2) MLS-WdAEC system design: The 1st stage misalignment estimate $\hat{D}_n$ is obtained as the circular cross-correlation between the MLS w_n and the filtered residual echo e'_n given by (4), computed each L-samples frame as

$$\begin{aligned}\hat{d}_l = \phi_{we'}(l) &= \frac{1}{L}\sum_{k=0}^{L-1} w_k e'_{(l+k) \bmod L} \\ &= \sum_{j=0}^{p-1} d_j \phi_w(l-j) + \phi_{w\xi}(l), \end{aligned} \quad (13)$$

where ϕ denotes circular correlation, d_j is the j^{th} component of misalignment vector $D_n = F - G_n$ and $l = 0, \cdots, L-1$. More details about the derivation of (13) are given in Appendix B-B. Using the notation of Appendix B, (13) is rewritten as

$$\hat{d}_l = d_l + \underbrace{P_1(l)}_{=0 \text{ if } p<L} + P_2(l) + \underbrace{P_3(l)}_{=\phi_{w\xi}(l)}. \quad (14)$$

We remind the definitions of $P_1(l)$ and $P_2(l)$ (given in Appendix B) using the notations of this section:

- $P_1(l) = \sum_{j=1}^{\lfloor p/L \rfloor} d_{l+jL}$ is an additive under-modeling noise term when $p > L$, where p is the length of the channel's IR. In case of $p \leq L$, $P_1(l) = 0$.
- $P_2(l) = -\frac{1}{L}\sum_{\substack{j=0 \\ j\neq l+kL}}^{p-1} d_j$ is an additive error term due to $\phi_w(l) \neq 0$ for $l \neq 0$ (cf. (22)).
- $P_3(l) = \frac{1}{L}\sum_{k=0}^{L-1} w_k \xi_{(l+k) \bmod L}$ is an additive term due to the equivalent noise ξ_n given by (5).

Note that the term $P_3(l)$ should be low thanks to the noise immunity of MLS.

We note from (14) and according to the analysis of Appendix B, that the choice of L is a very important issue. Indeed, an optimal estimation is reached under the following conditions:

- $p < L$ and thus $P_1(l) = 0$,
- high L value, and thus $P_2(l) \approx 0$, as $P_2(l)$ is inversely proportional to L,
- high L value implies that $P_3(l)$ is small as it is inversely proportional to L.

Besides, it is possible to further increase the noise immunity of the MLS-based identification by using the preaveraging technique [34]. This technique consists in preaveraging several L periods of the reference signal, e'_n, before computing the cross-correlation (13). Preaveraging reduces the effect of the noise ξ_n, and thus increases the SNR at the cross-correlator input. Both approaches, with and without preaveraging, are evaluated in the following simulations.

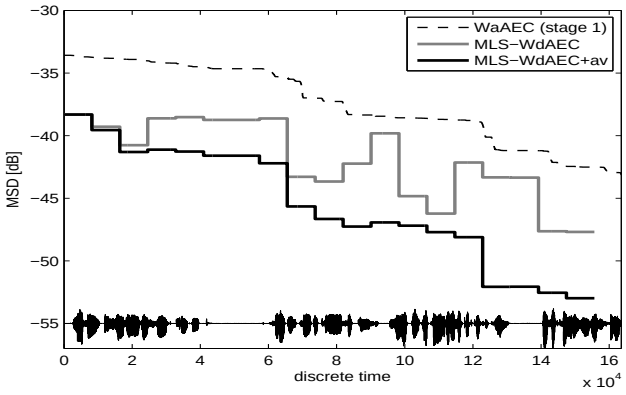

Fig. 9. Convergence speed analysis: Instantaneous MSD of the MLS-WdAEC (full lines) and of the 1st stage WaAEC (dotted line). $\mu = 0.02$, $p = 200$, $L = 8191$, SNR = 30 dB, 44% watermarked.

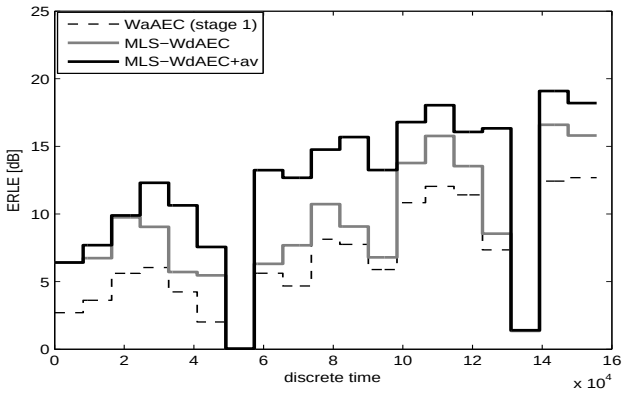

Fig. 10. Instantaneous ERLE of MLS-WdAEC (full line) and WaAEC (dotted line). $\mu = 0.02$, $p = 200$, $L = 8191$, SNR=30 dB, 44% watermarked.

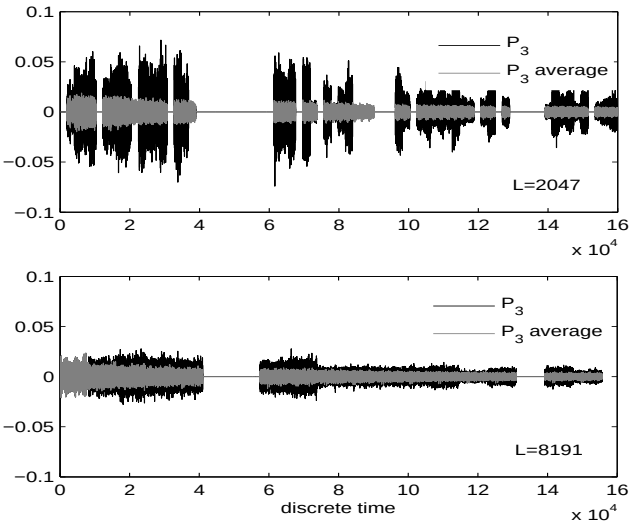

Fig. 11. Comparison of residual noise after cross-correlation for two MLS lengths L (top $L = 2047$, bottom $L = 8191$). The higher L, the higher the noise immunity. ($\mu = 0.02$, $p = 200$, SNR=30dB, 44% watermarked.)

B. General performance analysis

The performance of the MLS-WdAEC of Fig. 7 strongly depends on the values and characteristics of $P_1(l)$, $P_2(l)$, and $\phi_{w\xi}(l)$ given by (14). In this section, we evaluate and discuss the general performance in terms of steady state behavior and convergence speed with respect to the MLS parameters settings. Note that the cross-correlator of the MLS-WdAEC provides one estimate $\hat{D}_n$ per L-period, contrarily to the A-WdAEC which delivers instantaneous estimates.

1) Convergence speed and steady state behavior: The performance analysis was carried out with the same car's IR F ($p = 200$ taps) as in Section III, the additive noise ν_n was set to yield an SNR of 30 dB and the 1st stage AEC has p taps (exact modeling) with a step size $\mu = 0.02$. The overall MLS embedding rate was *ca.* 44% and the embedded MLS has a period of $L = 8191$. Under these conditions, the term $P_1(l)$ of (14) is zero and L is sufficiently large to provide a suitable noise immunity.

In Fig. 9, we compare the MSD of the MLS-WdAEC with that of the WaAEC (1st stage). We observe that the MSD of the MLS-WdAEC is significantly lower, particularly in the case of the preaveraging-based approach. Accordingly, Fig. 10 shows that in the steady state, the MLS-WdAEC achieves a higher ERLE than the reference AEC (up to 10 dB higher for the preaveraging-based method).

Note that MSD and ERLE are here computed once per L-samples frame, yielding a constant value per MLS-period. To facilitate performance visualization, the instantaneous ERLE of the 1st stage WaAEC is averaged over L-periods in Fig. 10.

2) Parameter settings: As mentioned above, the choice of the MLS length is important, in particular regarding the identification immunity against the noise ξ_n. Indeed, in Fig. 11, we show the noisy term P_3 of the cross-correlation (14) for two different L values: P_3 has a lower energy for the largest L value. However, P_3 has nearly the same energy for both L values when preaveraging e'_n prior to cross-correlation.

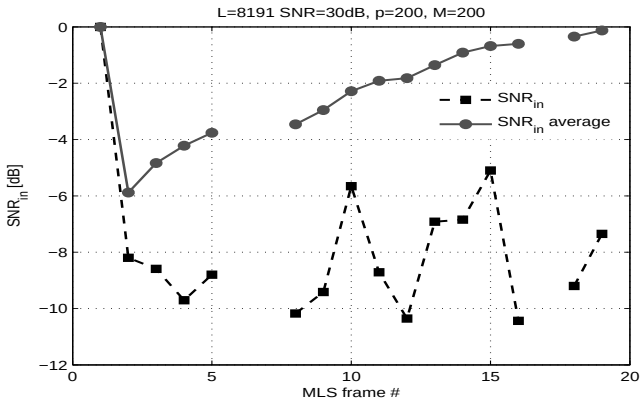

Fig. 12. Improvement of the noise immunity at the cross-correlator input: input SNR (15) with and without preaveraging.($L = 8191$, $\mu = 0.02$, $p = 200$, SNR=30dB, 44% watermarked.)

In fact, for $L = 2047$, the ERLE of the 2nd stage of the MLS-WdAEC without preaveraging is similar to that of the 1st stage (no improvement), while it is nearly the same as the performance of 2nd stage with $L = 8191$ when preaveraging is used.

The increased noise immunity of MLS, thanks to preaveraging, is shown in Fig. 12, where we plot the SNR at the cross-correlator input, computed for each MLS period as:

$$\text{SNR}_{in} = \frac{\text{E}[((F - G_n) * w_n)^2]}{\text{E}[(\xi_n)^2]}. \qquad (15)$$

SNR_{in} is enhanced by *ca.* 8 dB in the steady state, in addition to its stability, compared to the non-averaged approach. Note that in Fig. 12, SNR_{in} is computed only for L-signal periods where the watermark embedding rate is higher than the 20% threshold, below which the 2nd stage identification is frozen. Identification is frozen when the MLS signal is not sufficiently present in e'_n to provide an efficient identification through the cross-correlator.

Besides, from Fig. 8 we observe that for small L, the embedding rate per $L-$period tends to be higher than for large L. Hence, for large L we reach a better noise immunity but a lower MLS embedding rate per L-host period. The simulation results with different L values show that the noise immunity seems to be of more importance for the echo cancellation performance, as it significantly reduces the noise level.

V. Results and Discussion

In this Section, we compare the general performance of the two WdAEC implementations as well as their robustness to noisy and under-modeling conditions. We also emphasize the key points of each of them. It is however worth noting that the performance evaluation and comparison in noisy and under-modeling situations is somewhat delicate, as in the case of NLMS adapted identification, the performance depends on the step size value and its possible regularization. It is possible to improve the identification performance by using noise-dependent step sizes (e.g. Variable Step Size (VSS) adaptation algorithm [9], [10]), or by introducing an SNR dependent regularization term in the adaptation step size, as proposed

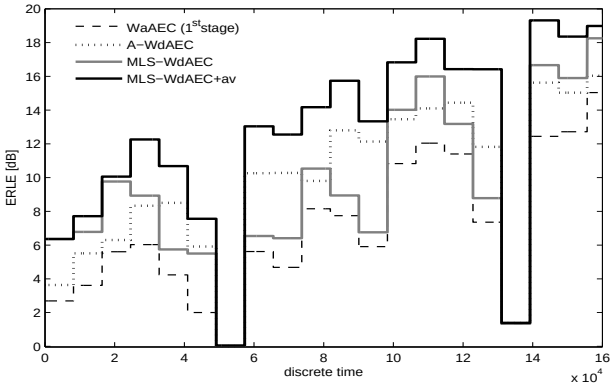

Fig. 13. ERLE comparison of A-WdAEC and MLS-WdAEC ($\mu = \mu^w = 0.02$, $p = 200$, SNR=30dB, $L = 8191$, 44% watermarked signal.)

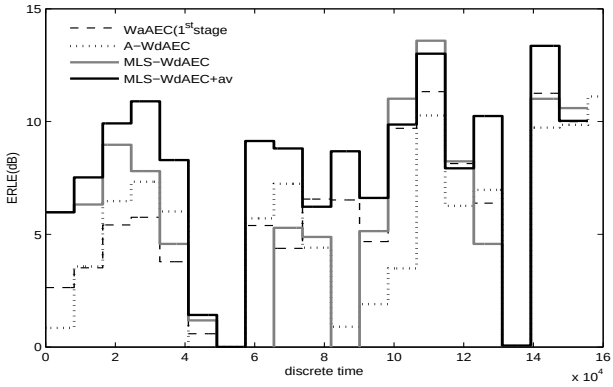

Fig. 14. ERLE comparison of A-WdAEC and MLS-WdAEC for noisy condition SNR=15dB ($\mu = \mu^w = 0.02$, $p = 200$, $L = 8191$, 44% watermarked signal.)

in [35], [36]. These considerations are beyond the focus of this study, which aims at evaluating the general performance of the proposed system, independently of the adaptation algorithm and its very particular settings and variants.

Hence, we are concerned in this Section by providing "objective" results while maintaining the NLMS step sizes at the same values, keeping in mind however that these are probably not the optimal values in noisy and under-modeling conditions. Regardless of the 1st stage adaptation algorithm, the 2nd stage is designed to help the overall system perform better than the 1st stage alone.

A. General performance comparison

For comparison, we report the ERLE of the A-WdAEC and the MLS-WdAEC (with and without preaveraging) in Fig. 13, where the parameters setting is identical to Sections III and IV (i.e. SNR=30dB, $p = M =$200 taps, $\mu = \mu^w = 0.02$, $L = 8191$ and *ca.* 44% embedding rate). Note that, for comparison convenience, the instantaneous ERLE of the A-WdAEC is here averaged per MLS period.

The results show that the MLS-WdAEC with preaveraging has the highest echo cancellation performance, while the other two are nearly equivalent, except in the speech segment between 6.10^4 and 10^5, where the A-WdAEC seems to outperform the MLS-WdAEC. This is mainly due to the low embedding rate in that particular segment, where the signal energy is too low to embed the MLS watermark in each frame. The A-WdAEC computes a channel estimate instantaneously, rather than over L long frames (case of MLS-WdAEC). Consequently, the filter taps are adapted each instant to the presence/absence of watermark and identification is stopped during signal frames where voice activity is too low to embed the watermark. This is not the case of the MLS-WdAEC, which estimates the channel from an L period, containing a minimum of 20% watermark (cf. Fig. 8). Nevertheless, this issue is solved by the preaveraging, which enables the cross-correlator to take advantage of previous signal periods, which may have a higher watermark content.

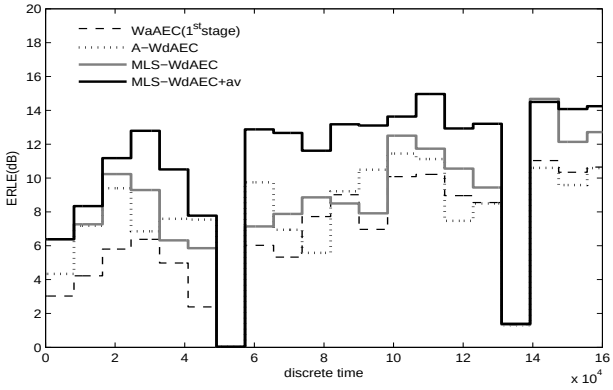

Fig. 15. ERLE comparison of A-WdAEC and MLS-WdAEC with 1st stage under-modeling ($\mu = \mu^w = 0.02$, $p = 200$, $M = 100$, $L = 8191$, SNR=30 dB, 44% watermarked signal).

B. Robustness to noise

As mentioned above, it is a delicate issue to compare the noise robustness of both systems as the noise robustness of NLMS depends on the adaptation step size and its potential regularization. We have chosen in this study to set a fixed step size (μ and μ^w), which provided optimal performance for a moderately high SNR, in order to better focus on the enhancements due to the second AEC stage only. Hence, using the same simulation settings as in Section V-A, we reduce the SNR to 15 dB, and we display in Fig. 14 the ERLE reached by the A-WdAEC and the MLS-WdAEC. These results show the higher robustness of the preaveraging-based MLS-WdAEC as compared with the other two systems, which achieve lower but similar performance.

C. Robustness to first stage under-modeling

In conventional AEC, an under modeling situation occurs when the channel's IR is longer than the IR of the adaptive AEC filter. This usually causes the degradation of the identification quality. The residual echo caused by the part of the echo path that cannot be modeled is equivalent to an additional noise that disturbs the algorithm's performance.

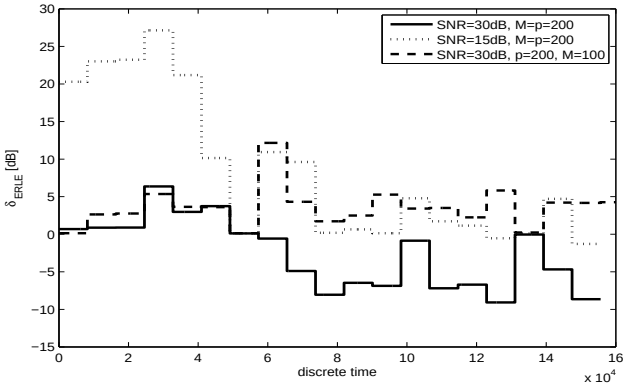

Fig. 16. Performance comparison of the MLS-WdAEC (with preaverage) with the JO-NLMS [8]. $\delta_{\text{ERLE}} = \text{ERLE}_{\text{MLS-WdAEC}} - \text{ERLE}_{\text{JO-NLMS}}$ in different situations. (Referring to the notation of [8], the JO-NLMS parameters are set to ϵ=0.1, $\lambda = 1-1/KM$, $K = 2$, and the noise variance estimation of [37]).

To evaluate the robustness of the proposed A-WdAEC and MLS-WdAEC when the 1st stage is in under-modeling situation, both systems were run with 1st stage AEC having only 100 taps while the echo path IR has $p = 200$ taps. The 2nd stage AEC filter of the A-WdAEC was set in exact modeling mode (200 taps), and the 2nd stage of the MLS-WdAEC used $L = 8191$ for the MLS period.

Fig. 15 displays the steady state behavior of both implementations. As expected, the reached ERLE enhancement is lower than in the case of 1st stage exact modeling. These results confirm the higher robustness of the preaveraging-based MLS-WdAEC to the additional noise due to under-modeling: its ERLE is higher than in the A-WdAEC and MLS-WdAEC cases. The latter two show similar performance and hence equivalent noise robustness properties.

In this context, the 2nd stage of A-WdAEC has to cope with an additional noise (due to under-modeling) with the same step values as in exact modeling, which is not the optimal setting, as already mentioned in Section V-B.

D. Performance comparison with the Joint-Optimized NLMS

The main idea of this study was the following: while most research work on AEC focus on fitting the basic identification algorithms with the inappropriate properties of the speech input (correlation and non-stationarity), we proposed to adapt the input signal to the properties required by a basic algorithm (NLMS and MLS-based correlation). To compare the efficiency of these two opposite approaches, we ran a state-of-the-art algorithm following the first approach, the Joint-Optimized NLMS (JO-NLMS [8]). The JO-NLMS is based on a state-variable model and sets a variable step-size and a regularization term so as to minimize the global system misalignment. The advantage of the JO-NLMS is that it does not rely on a fine parameter setting that could bias the comparison. We ran it in the same simulation conditions as those used in this section (see Fig. 13 to 15) and we compared it to the MLS-WdAEC with preaveraging. The results of the comparison are summarized in Fig. 16. While the JO-NLMS outperforms the MLS-WdAEC in the case of exact modeling and moderately high SNR, the MLS-WdAEC performs better in noisy conditions (SNR=15dB) and in the case of undermodeling. The results for SNR=30dB and $M = p$ should not obliterate the fact that, in these conditions, the MLS-WdAEC reaches an ERLE in the range 15-20 dB, as illustrated in Fig. 13.

VI. Conclusion

Conventional adaptive algorithms for acoustic echo cancellation are sensitive to the non-stationarity and correlation of speech signals. To reduce AEC sensitivity, we proposed in this study a two-stage AEC: the 1st stage is an NLMS based AEC driven by the imperceptibly watermarked far-end speech, and the 2nd stage AEC is driven by the white and stationary watermark signal solely. The role of this 2nd stage is to estimate the 1st stage misalignment in order to further reduce the overall residual echo, as stated by (6). The watermark is used as an auxiliary signal having the adequate properties to drive the 2nd stage identification algorithm.

Although the second stage performs in severe conditions - the misalignment to identify is slowly time-varying; the SNR is low; the noise is non-stationary; and the auxiliary signal is embedded during only a part of the time- the proposed two-stage AEC converges faster than the conventional AEC and achieves a better ERLE.

The two proposed implementations, adaptive (A-WdAEC) or MLS-correlation based (MLS-WdAEC), have similar performance, but the MLS-WdAEC is improved by preaveraging its input over several frames, although it seems counterintuitive as the misalignment to identify is slowly time-varying. This improved implementation is efficient to cope with lower SNR or with under-modeled first stage.

Beyond the proof of concept presented in this paper, other algorithmic refinements could be added to the second stage to take into account its particular operating conditions. For example, a noise-dependent step size could be used in the A-WdAEC [8]–[10], [36] to adapt to the high and variable noise, or a gradient adaptive step size [7] to track the variations of the impulse response to identify.

More fundamentally, watermarking for AEC enhancing is not restricted to the proposed two methods. The convergence speed of the algorithm is related to the condition number of the autocorrelation matrix of the driving signal, which depends, among others, on the signal distribution [2]. While we enhanced this condition number through the whiteness and stationarity of the driving watermark, it could also be enhanced in the first stage AEC by forcing the distribution of the input (doping watermarking, see [12], [22]–[24]) and in the second stage by choosing a watermark with an optimal distribution.

Appendix A
Watermark Embedding

A. Embedding procedure

The considered watermarking technique is based on spread spectrum insertion in the time-domain [38][39]. A frequency masking threshold $M_N(f)$ is computed from the host signal x_n on a frame-by-frame basis. The index N indicates that

$\mathbf{\Lambda}$ / Data	0.001		0.003		0.005	
	PESQ	**%w**	**PESQ**	**%w**	**PESQ**	**%w**
French	3.3	70	3.5	44	3.6	25
AmEnglish	3.5	59	3.6	34.2	3.7	18.9
BrEnglish	3.4	59	3.5	41.7	3.5	25.5
average	**3.4**	**62.67**	**3.53**	**39.96**	**3.6**	**23.13**

TABLE I
LISTENING QUALITY (PESQ MOS) *vs.* EMBEDDING RATE $\%w$ OF THE WATERMARKED SPEECH DATA BASE ITU-T P.50.

the masking threshold is updated each N-samples frame to take into account the non-stationarity of speech. $M_N(f)$ is usually computed from an auditory model, but in the case of speech signals, it can be approximated by the LPC[4] spectral envelope of x_n. The LPC coefficients define a filter with transfer function

$$H_N^0(z) = \frac{b_N}{1 - \sum_{i=1}^{Q} a_N(i) z^{-i}}, \tag{16}$$

whose gain approximates the masking threshold $M_N(f)$. To achieve the inaudibility of the watermark, the white and stationary sequence w_n is spectrally reshaped by the all-pole filter H_N, referred to as perceptual filter and defined by:

$$H_N(z) = \alpha H_N^0\left(\frac{z}{\gamma}\right). \tag{17}$$

The factor α is a constant attenuation to further reduce the power of the filtered watermark to strengthen the inaudibility constraint. The weighting factor $0 < \gamma < 1$ is used to flatten the peaks of the filter's spectral magnitude.
As the coefficient b_N is proportional to the power of the host signal in the processed frame, we define the factor λ_N by:

$$\lambda_N = \alpha . b_N, \tag{18}$$

which indicates the speech energy in each processed frame and allows to control the watermark embedding process, as explained in Section II-D. The spectrally reshaped watermark w_n has a power spectral density fitting $\alpha M_N(f)$ and the watermarked speech signal is given by:

$$x_n^w = x_n + h_N * w_n, \tag{19}$$

where h_N denotes the IR of H_N.
In this study, the filter settings are: $Q = 50$, $\gamma = 0.9$ and α corresponds to an attenuation of -10dB.

B. Embedding rate vs. listening quality

The relationship between embedding rate ($\%w$) and listening quality (assessed by PESQ MOS [31]) of the watermarked speech is given in Tab. I, where we list average PESQ MOS for three different languages (16 sentences male/female voices per language) of the ITU-T speech database [40], and for different voice activity thresholds $\mathbf{\Lambda}$. In the same table we also indicate the resulting average embedding rate $\%w$. We note that the higher the embedding rate, the lower the MOS score.
For the simulations presented in this article, and for comparison purposes, we used the same French speech sample, with $\mathbf{\Lambda} = 0.003$.

[4]LPC: Linear Predictive Coding

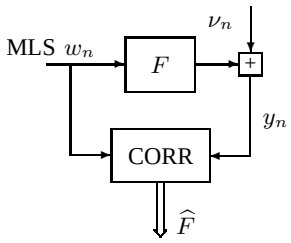

Fig. 17. MLS-based channel identification.

APPENDIX B
ACOUSTIC CHANNEL ESTIMATION USING MLS

In this appendix we emphasize the advantage of introducing maximum length sequences (MLS) as a watermark signal. We first introduce the useful properties of MLS and then describe the functionalities of a didactic AEC driven by an MLS input to display the identification mechanism of the proposed AEC. Finally, we evaluate the robustness of this MLS-driven AEC to additive, correlated and non-stationary noise. This appendix is thought as a complement to Section IV.

A. Correlation characteristic of Maximum Length Sequences

A Maximum Length Sequence is a periodic pseudo-random sequence with period L and values in $\{\pm 1\}$ (symmetrical MLS) [32]. MLS is generated by Galois polynomials of order m, so that the length of the MLS is given by

$$L = 2^m - 1. \tag{20}$$

Note that for a given order m there exist different Galois polynomials, that generate different MLS with the same length L, and each of them generates only one MLS.
The main property of an MLS w_n is its circular autocorrelation, $\phi_w(l)$, which is an L-periodic impulse:

$$\phi_w(l) = \frac{1}{L}\sum_{n=0}^{L-1} w_n w_{(n+l) \bmod_{\mathrm{L}}}, \tag{21}$$

$$= \begin{cases} 1 & \text{if } l \bmod_{\mathrm{L}} = 0, \\ -1/L & \text{otherwise}, \end{cases} \tag{22}$$

where mod denotes the *modulo* function.

B. Acoustic echo cancellation using MLS

The estimation of the IR F of a linear time invariant channel using MLS input is depicted in Fig. 17, where the channel input is an L-periodic MLS w_n and the output y_n is

$$y_n = F * w_n + \nu_n = \sum_{k=0}^{p-1} f_k w_{n-k} + \nu_n. \tag{23}$$

p is the length of the channel's IR $F = [f_0, ..., f_{p-1}]^t$ and ν_n is an additive noise, statistically independent from w_n. The

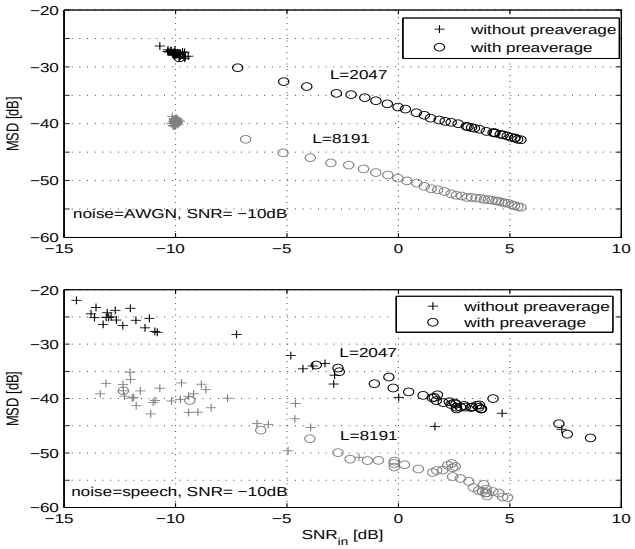

Fig. 18. Robustness to non-stationary (speech) and stationary (AWGN) noise at average SNR=−10dB. MSD *vs.* SNR$_{in}$ per $L-$period for two L values (IR F has $p = 200$ taps).

IR estimate $\widehat{F}$ is obtained as the input output circular cross-correlation, evaluated over L-samples frames as follows

$$\begin{aligned}\phi_{wy}(l) &= \frac{1}{L}\sum_{k=0}^{L-1} w_k y_{(l+k) \bmod_{\mathrm{L}}} \\ &= \frac{1}{L}\left[\sum_{k=0}^{L-1}\sum_{j=0}^{p-1} f_j w_k w_{(l+k-j)\bmod_{\mathrm{L}}} + \sum_{k=0}^{L-1} w_k \nu_{(l+k)\bmod_{\mathrm{L}}}\right] \\ &= \sum_{j=0}^{p-1} f_j \phi_w(l-j) + \phi_{w\nu}(l), \end{aligned} \quad (24)$$

for $l = 0, ..., L-1$. Using the circular correlation property (22), we get

$$\hat{f}_l = \phi_{wy}(l) = f_l + P_1(l) + P_2(l) + P_3(l), \quad (25)$$

for $l = 0, ..., L-1$ and where $\hat{f}_l$ is the estimate of the channel's impulse response taps and

- $P_1(l) = \sum_{j=1}^{\lfloor p/L \rfloor} f_{l+jL}$ is an additive under-modeling noise term when $L < p$,
- $P_2(l) = -\frac{1}{L}\sum_{j=0, j\neq l+kL}^{p-1} f_j$ is an additive error term due to the fact that $\phi_w(l) \neq 0$ for $l \neq 0 \bmod_{\mathrm{L}}$,
- $P_3(l) = \frac{1}{L}\sum_{k=0}^{L-1} w_k \nu_{(l+k) \bmod_{\mathrm{L}}}$ is an additive term due to the presence of noise ν_n.

C. Robustness to non-stationary correlated additive noise

Noise immunity of MLS in linear and non-linear channel identification has been widely studied [33], [41], [42]. In [34], Rife *et al.* demonstrate that the circular cross-correlator of Fig. 17 that recovers F from y_n is equivalent to a matched filter, which is in turn matched to the MLS w_n. Furthermore, MLS has a low crest factor[5] (0 dB), which allows a relatively high SNR for a given peak level. It is also proven in [34] that MLS-based identification is immune to transient noise of different kinds (clicks, coughs, footsteps, etc.).

Thanks to all the properties cited above, the identification system of Fig. 17 should be robust to any additive noise characteristics. This robustness is confirmed by the results plotted in Fig. 18, where we report the MSD per $L-$period *vs.* the SNR in each period (SNR$_{in}$ of (15)). The overall SNR was set to -10dB in both cases: the noise ν_n is AWGN (top panel) or speech. The crosses correspond to the results obtained without preaveraging and the circles correspond to the use of preaveraging (cf. Section IV). We note that MLS-based identification is robust to non-stationary and correlated noise, even if the results, compared to the AWGN case, seem to be less stable (depending on the speech content of each $L-$period). We also note that preaveraging improves the system's robustness for both noise types. Finally, the longer the MLS, the higher the identification performance is.

[5]Signals with low crest factor are desirable as they contain more signal power for a given peak level [34].

APPENDIX C
MATHEMATICAL SYMBOLS AND NOTATION

Time-invariant filter IR taps are denoted by uppercase and single IR taps by lower case letters, thus the notation in case of a time-invariant filter F is:

$$F = [f_0, f_1, \cdots, f_{p-1}]^t,$$

and in case of a time-varying filter F_n:

$$F_n = [f_0(n), f_1(n), \cdots, f_{p-1}(n)]^t,$$

where n is the discrete time index.
Convolution of filter IR F with a signal x_n is written as:

$$F * x_n = \sum_{k=0}^{p-1} f_k x_{n-k}.$$

$\lfloor\cdot\rfloor$ is the floor function.
$a \bmod_{\mathrm{b}}$ denotes the modulo function.
ERLE: Echo Return Loss Enhancement:

$$\mathrm{ERLE} = 10\log_{10}\left(\frac{\mathrm{E}[y_n^2]}{\mathrm{E}[e_n^2]}\right), \quad (26)$$

where y_n is the reference echo and e_n the residual echo. $E[\cdot]$ denotes the mathematical expectation.

REFERENCES

[1] S. Haykin, *Adaptive Filter Theory*. Pearson Education India, 2008.
[2] O. Macchi, *Adaptive processing: The least mean squares approach with applications in transmission*. John Wiley & Sons, 1995.
[3] C. Paleologu, J. Benesty, and S. Ciochina, "A variable step-size affine projection algorithm designed for acoustic echo cancellation," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 16, no. 8, 2008.
[4] M. Mboup, M. Bonnet, and N. Bershad, "LMS coupled adaptive prediction and system identification: A statistical model and transient mean analysis," *IEEE Trans. on Signal Process.*, vol. 42, no. 10, 1994.
[5] P. Scalart and F. Bouteille, "On integrating speech coding functions into echo cancelling filters with decorrelating properties," in *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, 2002.
[6] A. Aicha and S. Ben Jebara, "Decorrelation of input signals for stereophonic acoustic echo cancellation using the class of perceptual equivalence," in *European Signal Process. Conf.*, 2008.

[7] V. Mathews and Z. Xie, "A stochastic gradient adaptive filter with gradient adaptive step size," *IEEE Trans. Signal Process.*, vol. 41, no. 6, 1993.
[8] C. Paleologu, S. Ciochină, J. Benesty, and S. L. Grant, "An overview on optimized NLMS algorithms for acoustic echo cancellation," *EURASIP J. on Advances in Signal Process.*, no. 1, 2015.
[9] L. R. Vega, H. Rey, J. Benesty, and S. Tressens, "A new robust variable step-size NLMS algorithm," *IEEE Trans. on Signal Process.*, vol. 56, no. 5, 2008.
[10] H.-C. Huang and J. Lee, "A new variable step-size NLMS algorithm and its performance analysis," *IEEE Trans. on Signal Process.*, vol. 60, no. 4, 2012.
[11] S. Ben Jebara and H. Besbes, "Variable step size filtered sign algorithm for acoustic echo cancellation," *IET Electron. Lett.*, vol. 39, no. 12, 2003.
[12] G. Mahé, E. Z. Nadalin, R. Suyama, and J. M. Romano, "Perceptually controlled doping for audio source separation," *EURASIP J. on Advances in Signal Process.*, vol. 2014, no. 1, 2014.
[13] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard, "Informed audio source separation: A comparative study," in *European Signal Process. Conf.*, 2012.
[14] M. Parvaix, L. Girin, and J. Brossier, "A watermarking-based method for single-channel audio source separation," in *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, 2009.
[15] Y.-W. Liu, "Sound source segregation assisted by audio watermarking," in *IEEE Int. Conf. Multimedia and Expo*, 2007.
[16] A. Sagi and D. Malah, "Bandwith extension of telephone speech aided by data embedding," *EURASIP J. on Advances in Signal Process.*, 2007.
[17] B. Geiser, P. Jax, and P. Vary, "Artificial bandwidth extension of speech supported by watermark-transmitted side information." in *Interspeech*, 2005.
[18] B. Geiser, F. Mertz, and P. Vary, "Steganographic packet loss concealment for wireless VoIP," in *Sprachkommunikation Conf.*, 2008.
[19] F. Mertz and P. Vary, "Packet loss concealment with side information for voice over IP in cellular networks," in *Sprachkommunikation Conf.*, 2006.
[20] S. Djaziri-Larbi, A. Zaien, and S. Sevestre-Ghalila, "Voicing of animated GIF by data hiding," *Multimedia Tools and Applicat.*, vol. 75, no. 8, 2016.
[21] I. Samaali, G. Mahé, and M. Turki, "Watermark-aided pre-echo reduction in low bit-rate audio coding," *J. of the Audio Eng. Soc.*, vol. 60, no. 6, 2012.
[22] S. Djaziri Larbi, G. Mahé, I. Marrakchi, M. Turki, and M. Jaïdane, "Doping and witness watermarking for audio processing," in *IEEE Int. Workshop on Systems, Signal Process. and their Applicat.*, 2011.
[23] H. Halalchi, G. Mahé, and M. Jaïdane, "Revisiting quantization theorem through audio watermarking," in *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, 2009.
[24] S. Djaziri Larbi and M. Jaïdane, "Audio watermarking: a way to modify audio statistics," *IEEE Trans. on Signal Process.*, vol. 53, no. 2, 2005.
[25] S. Larbi and M. Jaïdane, "Watermarking influence on the stationarity of audio signals," in *Int. Conf. on Acoust., Speech and Signal Process.*, 2003.
[26] F. Mazzenga, "Channel estimation and equalization for M-QAM transmission with a hidden pilot sequence," *IEEE Trans. on Broadcasting*, vol. 46, no. 2, 2000.
[27] I. Mezghani-Marrakchi, G. Mahé, S. Djaziri-Larbi, M. Jaidane, and M. Turki-Hadj Alouane, "Nonlinear audio systems identification through audio input gaussianization," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 22, no. 1, 2014.
[28] 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 *European Signal Process. Conf.*, 2006.
[29] S. Larbi, M. Jaïdane, M. Turki, and M. Bonnet, "On the robustness of an echo canceler robust to speech non stationarities," in *CORESA (COmpression et REpresentation des Signaux Audiovisuels), in french*, 2001.
[30] H. Laurent and C. Doncarli, "Stationarity index for abrupt changes detection in the time-frequency plane," *IEEE Signal process. lett.*, vol. 5, no. 2, 1998.
[31] "Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T, Recommendation P.862, 2001.
[32] D. V. Sarwate and M. B. Pursley, "Crosscorrelation properties of pseudorandom and related sequences," *Proc. of the IEEE*, vol. 68, no. 5, 1980.
[33] J. Borish and J. B. Angell, "An efficient algorithm for measuring the impulse response using pseudorandom noise," *J. of the Audio Eng. Soc.*, vol. 31, no. 7, 1982.
[34] D. D. Rife and J. Vanderkooy, "Transfer-function measurement with maximum-length sequences," *J. of the Audio Eng. Soc.*, vol. 37, no. 6, 1989.
[35] Y.-S. Choi, H.-C. Shin, and W.-J. Song, "Robust regularization for normalized LMS algorithms," *IEEE Trans. on Circuits and Systems II: Express Briefs*, vol. 53, no. 8, 2006.
[36] J. Benesty, C. Paleologu, and S. Ciochina, "On regularization in adaptive filtering," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 19, no. 6, 2011.
[37] M. Asif Iqbal and S. Grant, "Novel variable step size NLMS algorithms for echo cancellation," in *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, 2008.
[38] S. Larbi, M. Jaïdane, and N. Moreau, "A new wiener filtering based detection scheme for time domain perceptual audio watermarking," in *IEEE Int.l Conf. Acoust., Speech, and Signal Process.*, vol. 5, 2004.
[39] L. Boney, A. H. Tewfik, and K. N. Hamdy, "Digital watermarks for audio signals," in *IEEE Int. Conf. on Multimedia Computing and Systems*, 1996.
[40] "Telephone transmission quality, telephone installations, local line networks, objective measuring apparatus, appendix I Test Signals," ITU-T, Recommendation P.50, 1998.
[41] C. Dunn and M. Hawksford, "Distortion immunity of MLS-derived impulse response measurements," *J. of the Audio Eng. Soc.*, vol. 41, no. 5, 1993.
[42] S. K. Olesen, J. Plogsties, P. Minnaar, F. Christensen, and H. Møller, "An improved MLS measurement system for acquiring room impulse responses," in *IEEE Nordic Signal Process. Symp.*, 2000.

Sonia Djaziri-Larbi received the engineering degree in electrical engineering from the Friedrich Alexander Universität of Erlangen-Nürnberg, Germany, in 1996 and the M.Sc.Eng. in digital communications from the National Engineering School of Tunis (ENIT), University of Tunis El Manar, Tunisia, in 1999. She obtained the PhD degree in telecommunications from the Ecole Nationale Supérieure des Télécommunications of Paris and from ENIT in 2005. Since 2001, she has been with the Information and Communications Technologies Department at ENIT, where she is currently assistant Professor. She is a researcher at the Signals and Systems Laboratory at ENIT. Her teaching and research interests are in signal and audio processing.

Gaël Mahé received the engineering degree in telecommunications and the M. Sc. in signal and communications from Télécom Bretagne, France, in 1998. From 1999 to 2002 he has worked at Orange Labs in Lannion, France, where he received the PhD in signal and telecommunications from the University of Rennes 1 in 2002. Since 2003, he has been with the Laboratory of Informatics of Paris Descartes University (LIPADE), where he is currently assistant Professor. His research deals mainly with new uses of watermarking in audio processing.

Imen Mezghani received the engineering degree in electrical engineering in 2002 and the M.Sc.Eng. degree in 2004 both from the National Engineering School of Tunis (ENIT), University of Tunis El Manar, Tunisia. She obtained the PhD degree in telecommunications from the Paris Descartes University, France, and from ENIT in 2010. She is currently assistant Professor at the National Engineering School of Sousse, Tunisia. Her teaching and research interests are in signal and audio processing.

Monia Turki received the Principal Eng., M.Sc. and the PhD degrees from the Department of Electrical Engineering at the National Engineering School of Tunis (ENIT), Tunisia, in 1989, 1991 and 1997 respectively. She joined the Department of Information and Communications Technologies of ENIT since 2001, where she is currently full Professor. Since 2010, she is the Head of the Signals and Systems Lab at ENIT, and since 2014 she is the Head of the ICT Department at ENIT. Her research interest concerns signal processing and adaptive filtering applied to speech, audio and communication systems.

Mériem Jaïdane received the M.Sc. degree in electrical engineering from the National Engineering School of Tunis (ENIT), Tunisia, in 1980. From 1980 to 1987, she has worked as research engineer at the Laboratoire des Signaux et Systèmes, CNRS/Ecole Supérieure d'Electricité, France. She received the Doctorat d'Etat degree in 1987. Since 1987, she has been with ENIT where she is currently a full Professor at the Information and Communications Technologies Department. She is a researcher at the Signals and Systems Lab at ENIT, University of Tunis El Manar. Her teaching and research interests are in adaptive systems for digital communications and audio processing.