SPEECH PROCESSING IN THE WATERMARKED DOMAIN: APPLICATION IN ADAPTIVE ACOUSTIC ECHO CANCELLATION

I. Marrakchi-Mezghani[(1)], *M. Turki-Hadj Alouane*[(1)], *S. Djaziri-Larbi*[(1)], *M. Jaïdane-Saïdane*[(1)], *G. Mahé*[(2)]

[(1)] Unité Signaux et Systèmes, Ecole Nationale d'Ingénieurs de Tunis, Tunisia
email: mezghani_imen@yahoo.fr, {m.turki, sonia.larbi, meriem.jaidane}@enit.rnu.tn
[(2)] UFR Mathématique et Informatique, Université René Descartes, Paris V, France
email: gael.mahe@math-info.univ-paris5.fr

ABSTRACT

Audio watermarking, or embedding information in a host signal was originally used for digital copyright protection purposes. As audio coding, watermarking is progressively brought in audio processing applications.

In this paper, we investigate some benefits of watermarking in signal processing. We focus here on a generic application : adaptive Acoustic Echo Cancellation (AEC). The proposed Watermarked AEC (WAEC) is based on a coupling of two adaptive filters. The first one extracts a rough estimation of the echo response to the known stationary watermark embedded in the speech signal. This extracted estimation constitutes the reference signal for the second adaptive filter. Driven by the known watermark, the second adaptive filter estimates then the actual echo path. The goal here is to drive the estimation of the echo path by the watermark itself, in order to take advantage of its optimal properties (whiteness and stationarity).

As expected, the proposed WAEC exhibits better transient and steady state performance than the classical one. These results of some interest follow from the fact that the second adaptive filter deals with much more stationary signals that the first one.

1. INTRODUCTION

Recently, watermarking, or embedding information in a host signal, originally used for digital copyright protection purposes has arisen the interest of researchers in many other application fields. In particular, watermarking was used to enhance the performance of audio processing systems. Indeed, in [1] watermarking aimed to stationarize the audio signal. The performance of a classical AEC was then improved [2]. In [3], the ill-conditioning of the input covariance matrix in a stereophonic AEC system, was reduced. In [4], Iapatov and Huffman sequences were embedded in speech signal, through perceptual masking, to improve the AEC performance. Note that classical adaptive filters exhibit low performance when dealing with non stationary data.

This paper investigates the benefits of classical signal processing in the watermarked domain, where operating in the watermarked domain refers to processing of watermarked signals. Here, the watermarking is associated to an AEC system according to an original way. Whereas the AEC was driven by the watermarked audio signal in the latter references, the goal here is to drive the estimation of the echo path by the watermark itself, in order to take advantage of the optimal properties of the watermark (whiteness and stationarity).

The paper is organized as follows : section 2 introduces a basic version of the WAEC. To meet the inaudibility constraint the basic version is modified. The retained version of WAEC is then presented in section 3. Performance analysis of the proposed WAEC are discussed in section 4. Simulation results, presented in section 5, demonstrate the benefits of the proposed approach. Comparison with classical AEC is performed.

2. WATERMARKED AEC

The proposed WAEC aims at exploiting the good transient and steady state performance of adaptive filters in presence of an uncorrelated stationary input. Therefore, the structure of a classical AEC is modified in such a way that the actual echo is adaptively estimated using a reference echo signal that includes the convolution between the actual echo path and a known white stationary signal. As a matter of fact, the stationary white sequence, embedded in the received speech, is the carrier of the acoustic impulse response to be identified. The proposed WAEC is depicted on Fig.1. Note that in this basic version, the watermarking process corresponds to a simple addition of the white signal and the received speech.

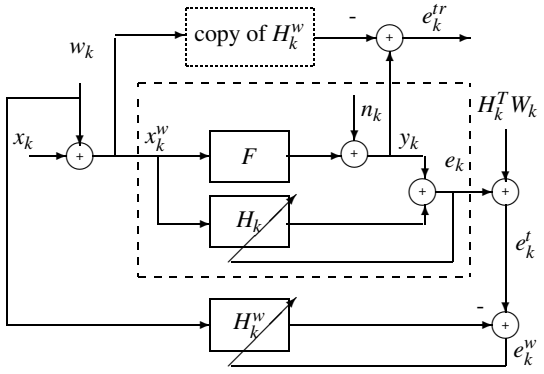

FIG. 1 – Preliminary version of the watermarked AEC.

Fig.1 shows two coupled adaptive filters. The first one H_k is driven by the non stationary watermarked input

$$x_k^w = x_k + w_k,$$

where x_k is the received speech signal and w_k the stationary and white watermark. The generic AEC delimited by a dashed box on Fig.1 provides an estimation of the reference

echo signal y_k used by the second adaptive filter H_k^w. The latter is driven by the watermark w_k only. The estimation of the watermark echo is

$$\begin{aligned} e_k^t &= e_k + (H_k)^T W_k \\ &= F^T W_k + \underbrace{n_k + (V_k)^T X_k}_{n_k'}, \end{aligned} \quad (1)$$

where $V_k = F - H_k$ denotes the deviation vector that describes the behaviour of H_k (^{T} denotes the transpose operator). According to (1), the echo is a sum of the useful information $F^T W_k$ and the non stationary noise n_k'. That is equivalent to a situation where e_k^t is the new reference signal and the echo path F is driven by w_k.

Since H_k^w performs in a more stationary context, it provides a better estimation of the acoustic echo path F than H_k. Consequently, H_k^w is the filter used to compute the echo estimate

$$\hat{y}_k = (H_k^w)^T X_k^w,$$

where $X_k^w = (x_k^w, .., x_{k-p+1}^w)^T$ is the watermarked non stationary tap input vector. Therefore, the transmitted residual echo is given by

$$e_k^{tr} = y_k - \hat{y}_k,$$

where $y_k = F^T X_k^w + n_k$ corresponds to the actual echo corrupted by a noise n_k assumed to be white. Note that an NLMS adaptation is used to adjust the two adaptive filters of length p assumed to be equal to that of the impulse response F.

– The update equations related to H_k are given by

$$e_k = y_k - (H_k)^T X_k^w \quad (2)$$

$$H_{k+1} = H_k + \mu_k \, e_k X_k^w, \quad (3)$$

where $\mu_k = \frac{\mu}{\|X_k^w\|^2}$ is the normalized step size.

– The second adaptive filter, H_k^w, is updated as follows

$$e_k^w = e_k^t - (H_k^w)^T W_k \quad (4)$$

$$H_{k+1}^w = H_k^w + \mu_k^w \, e_k^w W_k, \quad (5)$$

where $W_k = (w_k, .., w_{k-p+1})^T$ is the stationary tap input vector and $\mu_k^w = \frac{\mu^w}{\|X_k^w\|^2}$ the normalized step size.

By substituting in (4) the echo e_k^t given in (1), one can write

$$\begin{aligned} e_k^w &= F^T W_k - (H_k^w)^T W_k + n_k' \\ &= (V_k^w)^T W_k + n_k'. \end{aligned} \quad (6)$$

$V_k^w = F - H_k^w$ denotes the deviation vector related to H_k^w. The analysis of (1) and (6) shows that the steady state performance of the adaptive filter H_k^W and consequently that of the proposed AEC depends on the Signal to Noise Ratio

$$SNR^w = \frac{P_x}{P_w}.$$

The lower the SNR^w is, the better performance of the watermarked AEC can be achieved. But for low values of SNR^w, the watermark w_k becomes audible whereas the inaudibility constraint must be satisfied in practice. Thus, in order to ensure the watermark masking, modifications of this basic version of the WAEC are made.

3. THE PROPOSED WATERMARKED AEC

In general, speech watermarking is performed using a perceptual weighting filter [7, 9]. In our case, the watermark masking is based on a spectral shaping filter derived from the speech spectral envelope. Moreover, the intensity of the spectral shaped output is reduced by about 13 dB in order to ensure an inaudibility quality equivalent to that obtained by a perceptual weighting filter. The spectral shaping filter coefficients are deduced from the speech predictor. The speech predictor is used here to achieve

1. the masking of the white sequence,
2. the whitening of the watermarked input sequence which aims to enhance the convergence speed of the first NLMS adaptive filter. Indeed, in the previous WAEC, the watermarked input is correlated and consequently the convergence speed of H_k is slow. Note that the whitening is already used for system identification purposes [5]. The bloc diagram of the modified WAEC is presented in Fig.2 and detailed in the following.

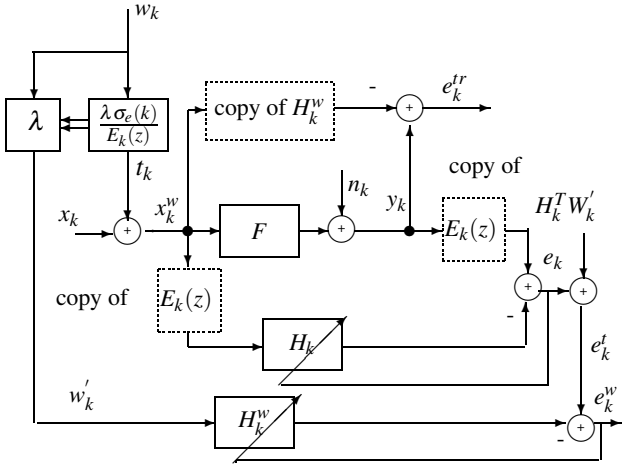

FIG. 2 – The proposed WAEC scheme.

3.1 The whitening filter

The speech is assumed to be an AutoRegressive process of order M (AR (M)). So,

$$x_k = (P_k)^T X_{k-1} + e_k^x,$$

where e_k^x is the prediction error of variance $\sigma_e^2(k)$, $P_k = (p_k^1, ..., p_k^M)^T$ the predictor coefficients vector and $X_{k-1} = (x_{k-1}, ..., x_{k-M})^T$ the input vector. A NLMS adaptation is used to update the predictor coefficients

$$P_k = P_{k-1} + \eta_k \, e_k^x \, X_{k-1}, \quad (7)$$

where $\eta_k = \frac{\eta}{\|X_k\|^2}$ is the normalized step size. The whitening filter transfer function is $E_k(z) = 1 - P_k(z)$, where $P_k(z)$ denotes the predictor transfer function.

3.2 The spectral shaping filter

The transfer function $S_k(z)$ of the all-pole spectral shaping filter is then expressed by

$$S_k(z) = \frac{\lambda\ \sigma_e(k)}{E_k(z)}.$$

The attenuation factor λ is chosen so that $20\log_{10}(\lambda) = -13\ dB$.

Thus, the watermarked signal is obtained according to $x_k^w = x_k + t_k$, as shown on Fig.2, where t_k refers to the spectral shaped watermark signal.

3.3 The echo estimation

Let e_k^x, e_k^{xw} and e_k^n be the through $E_k(z)$ filtered signals x_k, x_k^w and n_k respectively. By applying the whitening filter to y_k, we obtain the reference echo for H_k

$$e_k^y = F^T E_k^x + F^T W_k' + e_k^n,$$

where $E_k^x = (e_k^x, ..., e_{k-p+1}^x)^T$ and

$$W_k' = \lambda \sigma_e(k) W_k.$$

The error signal e_k controlling the update of H_k is

$$e_k = e_k^y - (H_k)^T (E_k^x + W_k'),$$

and the reference echo signal for the second adaptive filter H_k^w becomes

$$\begin{aligned} e_k^t &= e_k + (H_k)^T W_k' \\ &= F^T W_k' + \underbrace{e_k^n + (V_k)^T E_k^{xw}}_{e_k^{n'}}, \end{aligned} \tag{8}$$

where $E_k^{xw} = (e_k^{xw}, ..., e_{k-p+1}^{xw})^T$. The error e_k^t expressed by (8) points out that the useful information for the filter H_k^w is $F^T W_k'$. Therefore, and as shown on Fig.1, the filter H_k^w is driven by w_k', and its coefficients are adapted according to

$$H_{k+1}^w = H_k^w + \mu_k^w e_k^w W_k', \tag{9}$$

where the error e_k^w is given by

$$e_k^w = (V_k^w)^T W_k' + e_k^{n'}. \tag{10}$$

In this section, we have detailed the structure of the WAEC. In the following, we focus on the performance analysis of the proposed system.

4. PERFORMANCE ANALYSIS OF THE PROPOSED WAEC

The transient and steady state performances of the proposed WAEC are analysed in comparison with those of the generic AEC, delimited by a dashed box on Fig.1. For this comparison, the generic AEC must be driven by the watermarked signal x_k^w, because the available reference echo is generated by the watermarked speech. Indeed, it has been proved in [2] that watermarked audio is more stationary than the host signal. Furthermore, in [1] the authors have already shown that

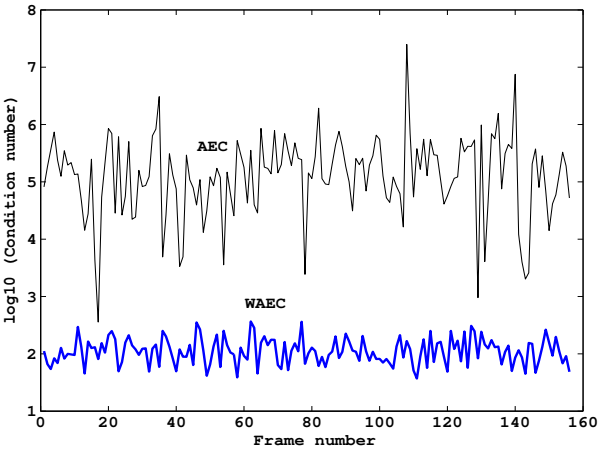

FIG. 3 – Condition number of $R_{w'}(k)$ (WAEC, in blue) and of $R_{x^w}(k)$ (classical AEC with watermarked input in black).

a classical AEC presents higher performance when driven by the watermarked speech instead of the original.

The transient and steady state behaviour of the classical AEC and the WAEC are studied through the time variation of the deviation vectors V_k and V_k^w respectively. Using (1), (4) and (5) it is easy to show that

$$V_{k+1} = \underbrace{(I - \mu_k X_k^w (X_k^w)^T)}_{A} V_k + \underbrace{\mu_k n_k X_k^w}_{C}, \tag{11}$$

where I refers to the identity matrix of rank p. Besides, by combining (8), (9), and (10), it follows that

$$V_{k+1}^w = \underbrace{(I - \mu_k^w W_k' (W_k')^T)}_{B} V_k^w + \underbrace{\mu_k^w e_k^{n'} W_k'}_{D}. \tag{12}$$

We remind that the mean convergence depends on the conditioning of the matrices A and B [8].

Taking the expectation value of both sides of (11), one obtains

$$E[V_{k+1}] = (I - \mu R_{x^w}(k)) E[V_k], \tag{13}$$

where $R_{x^w}(k) = E\left[\frac{X_k^w (X_k^w)^T}{\|X_k^w\|^2}\right]$. This result is obtained assuming the classical hypothesis of statistical independence between V_k and X_k^w. Similarly, assuming that $e_k^{n'}$ is of zero mean and statistically independent from w_k', it follows from (12) that

$$E[V_{k+1}^w] = (I - \mu_w R_{w'}(k)) E[V_k^w], \tag{14}$$

where $R_{w'}(k) = E\left[\frac{W_k' (W_k')^T}{\|W_k'\|^2}\right]$.

From equations (13) and (14) we can notice that in the classical case, the transient behaviour is governed by the matrix $R_{x^w}(k)$, whereas in the case of the proposed WAEC, the transient behaviour depends on the matrix $R_{w'}(k)$. The eigenvalue spread of $R_{w'}(k)$ is expected to be small compared to that of $R_{x^w}(k)$ since w_k' is much more uncorrelated than the watermarked speech x_k^w. This is illustrated by Fig.3, where

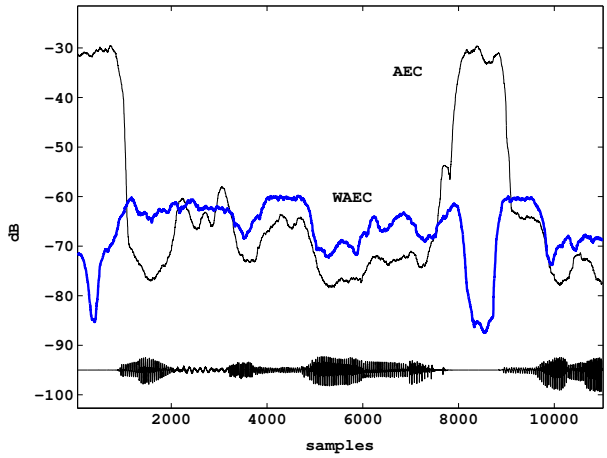

FIG. 4 – Time variation of P_k^D (WAEC, in blue) and P_k^C (classical AEC with watermarked input, in black).

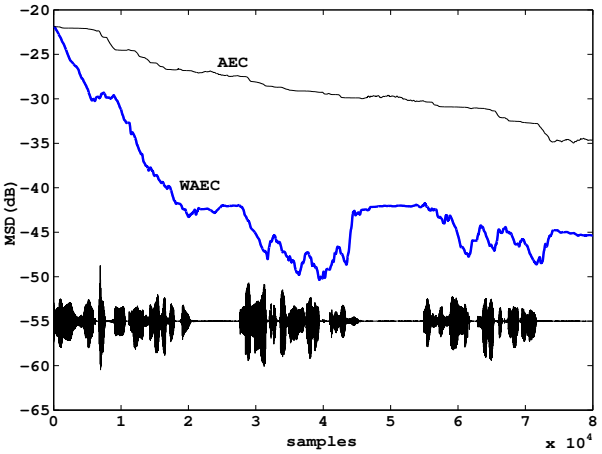

FIG. 5 – *MSD* time variation in the WAEC case (in blue) and in the classical AEC case with watermarked input (in black).

we compare the condition numbers of the covariance matrices $R_{x^w}(k)$ and $R_{w'}(k)$, computed over sample frames of length 20 ms. Knowing that the covariance matrix of a white signal has a logarithmic condition number equal to zero, we conclude from Fig.3 that $R_{w'}(k)$ is better conditioned than $R_{x^w}(k)$. Hence, we may confirm that the convergence rate of the WAEC will be higher than that of the classical AEC.

The analytical study of the mean square convergence is difficult to carry out as the signals are correlated and non stationary. However, one can see from equations (13) and (14) that the steady state performance of the AEC and the WAEC adaptive filters depends crucially on the instantaneous values of $P_k^C = E\left[(C_k)^T C_k\right]$ and $P_k^D = E\left[(D_k)^T D_k\right]$ respectively. As H_k^w performs in a more stationary context, P_k^D is expected to have smoother variations and lower values than P_k^C. This aspect will improve the WAEC performance in the steady state.

5. SIMULATION RESULTS

The parameter settings chosen for all reported simulations are :

- a speech signal sampled at 8 kHz and a car acoustic impulse response truncated to 100 taps,
- $SNR = 40$ dB,
- $\eta = 0.0025$,
- $M = 4$ and $p = 100$.

For the transient behaviour analysis, the same step size value 0.01 is used for the two coupled adaptive filters of the proposed WAEC. For the classical AEC with watermarked input, the step size is also fixed to 0.01. To demonstrate the enhancement of the convergence rate achieved by the proposed WAEC, we plot on Fig.5 the time variations of the *MSD* related to the WAEC $\left(H_k^w\right)$ and the *MSD* obtained by the classical AEC with a watermarked input (H_k). The *MSD* measure is defined here for a deviation vector V_k as

$$MSD_{dB} = 10\log_{10}\left(E\left[V_k^T V_k\right]\right).$$

As expected theoretically, since the watermark is much more white than the watermarked signal, the convergence rate of the proposed WAEC is much higher than that of the classical AEC.

To have comparative results in steady state, the choice of the step size values must ensure the same convergence rate for both AEC systems (the classical and the proposed one). To meet this condition, the same step size value 0.01 is used for both coupled adaptive filters of the proposed WAEC, whereas, for the classical AEC with watermarked input, the step size is fixed to 0.1.

The measure of the effect of an echo cancellation filter is the so-called Echo Return Loss Enhancement (ERLE)

$$\mathrm{ERLE}(k) = 10\log_{10}\left(\frac{E[y_k^2]}{E[e_k^2]}\right),$$

where y_k is the echo signal (it is the same in both cases) and e_k is the residual error in the generic AEC case. The equivalent residual error in the WAEC case is denoted e_k^{tr} on Fig.2.

To confirm the steady state analysis presented in Section 4, we depict on Fig.4 the time variation of P_k^C and P_k^D. As expected, this figure shows that the variations of P_k^D are much more smoother than those of P_k^C. In particular, P_k^C takes very large values during the low dynamics of the watermarked speech. This results in considerable disturbance of the MSD time evolution as shown on Fig.6. Consequently, the ERLE related to the generic AEC is impaired as displayed by Fig.7. Indeed, we notice that the WAEC achieved a 10 dB gain in the ERLE.

The reported results presented on Fig.5 and Fig.7 show that both the transient and the steady state performances are considerably improved by the WAEC.

6. CONCLUSION

This paper investigated the advantages of speech processing in watermarked domain. We have presented an original way to exploit watermarking for the improvement of a generic AEC performance. The new proposed WAEC benefits from the fact that white and stationary signals are the ideal inputs for classical adaptive filters. Compared to the classical AEC, the proposed WAEC exhibits much higher

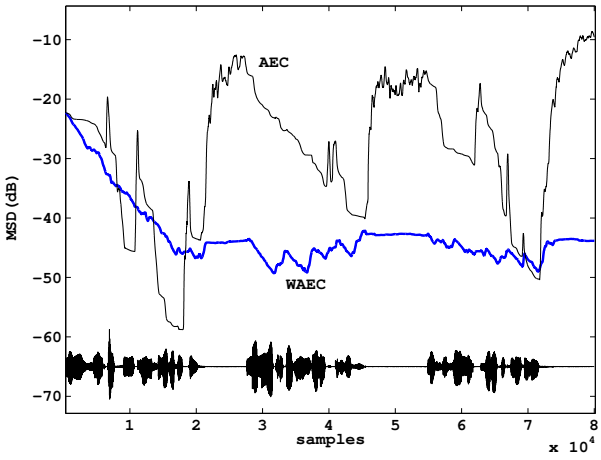

FIG. 6 – MSD time variation in the steady state and in both cases : WAEC (in blue) and classical AEC with watermarked input (in black).

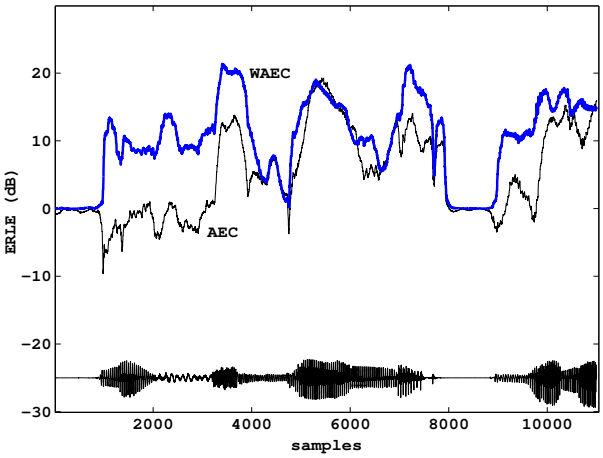

FIG. 7 – ERLE evolution for the WAEC (in blue) and the classical AEC with watermarked input (in black).

convergence rate and lower MSD. Moreover, an ERLE enhancement of up to 10 dB was achieved. These reported results are very promising and the slightly higher complexity of the WAEC is not a serious drawback considering the actual high-performance digital signal processors (DSP). However, a deeper theoretical and experimental analysis needs to be carried out to further improve the WAEC system.

REFERENCES

[1] S. Larbi, M. Jaïdane, M.Turki and M. Bonnet, "Réflexion sur un annuleur d'écho robuste aux non stationarités de la parole", in *CORESA*, Compression et représentation des signaux audio visuels, Dijon, France, 2001.

[2] S. Larbi and M. Jaïdane, "Audio watermarking : a way to modify audio statistics", *IEEE Trans. on Signal Processing*, vol. 53(2), 2005.

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

[4] M. Peters, "Psychoacoustical excitation of the (N)LMS algorithm for acoustical system identification", in *ISSPA* (fifth Int. Symp. on Signal Processing and its Applications), Brisbane, Australia, 1999.

[5] 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 Processing*, vol. 42(10), 1994.

[6] P. Scalart, F. Bouteille, "On integrating speech coding functions into echo cancelling filters with decorrelating properties ", in *ICASSP*, Orlando, USA, 2002.

[7] B. Paillard, P. Mabilleau, S. Morisette, "PERCEVAL : perceptual evaluation of the quality audio signals ", *JASA*, Vol. 40(12), 1992.

[8] S. Haykin, "Adaptive Filter Theory", Prentice-Hall, 1991.

[9] Wai C. Chu, "Speech coding algorithms", Wiley, 2003.