Nonlinear Acoustic Echo Cancellation Based on Volterra Filters

Article *in* IEEE Transactions on Speech and Audio Processing · December 2003

DOI: 10.1109/TSA.2003.818077 · Source: IEEE Xplore

Alexandre Guerin
Orange Labs
40 PUBLICATIONS **431** CITATIONS
SEE PROFILE

Régine Le Bouquin Jeannès
Université de Rennes 1
162 PUBLICATIONS **1,578** CITATIONS
SEE PROFILE

Nonlinear Acoustic Echo Cancellation Based on Volterra Filters

Alexandre Guérin, Gérard Faucon, and Régine Le Bouquin-Jeannès

***Abstract*—This paper describes a nonlinear acoustic echo cancellation algorithm, mainly focused on loudspeaker distortions. The proposed system is composed of two distinct modules organized in a cascaded structure: a nonlinear module based on polynomial Volterra filters models the loudspeaker, and a second module of standard linear filtering identifies the impulse response of the acoustic path. The tracking of the overall system model is achieved by a modified Normalized-Least Mean Square algorithm for which equations are derived. Stability conditions are given, and particular attention is placed on the transient behavior of cascaded filters. Finally, results of real data recorded with Alcatel GSM material are presented.**

***Index Terms*—Acoustic echo cancellation, cascaded structure, nonlinear adaptive filtering, transient behavior analysis, Volterra filters.**

I. Introduction

ACOUSTIC echo cancellation (AEC) is a major concern in telecommunications, and especially in the GSM domain (global system for mobile communication) where echo delay is particularly annoying for speakers. Historically, under the assumption of a completely linear acoustic chain (including amplifier, loudspeaker, acoustic path and microphone), a number of adaptive algorithms based on the gradient theory were developed to remove echo while keeping full-duplex communication characteristics. The most famous of these is the least mean square (LMS) algorithm [1], [2]. Several techniques based on affine projection algorithm (APA) or recursive least-squares (RLS) algorithm have been developed to cope with an ill-conditioned input autocorrelation matrix that degrades LMS performance [1], [3]. Nevertheless, as indicated before, all these algorithms assume that the acoustic echo is linearly dependent on the loudspeaker input signal.

Recently, in order to improve customer comfort and security, particularly in vehicle communication, the GSM telecommunications area has seen important developments for new hands-free functions in each terminal. These functions imply higher power-amplification and powerful loudspeakers, which are unfortunately not compatible with the miniaturization trend. In fact, the loudspeaker is commonly driven with maximum voltages, resulting in nonlinearities. A statistical study of the LMS algorithm performed by Costa *et al.* showed that a nonsignificant saturation could degrade the achievable performance in case of linear active noise control [4]. We have demonstrated in [5], that small but inaudible nonlinearities containing memory have a dramatic influence on classical linear AEC.

These theoretical results led to some nonlinear AEC algorithm developments: for example Stenger *et al.* studied the gradient identification of a memoryless polynomial-type nonlinearity before the linear acoustic path, modeling an amplifier nonlinearity [6]. Other studies also focused on the problem of identification of loudspeaker memory nonlinearities. Different contributions have shown the efficiency of second-order Volterra filters (SOVF) and their superiority compared to linear algorithms (see [7] and [8]). The proposed systems are based on a parallel structure composed of first and second-order Volterra filters. However, the real difficulty is related to algorithm complexity and convergence rate. Indeed, the number of coefficients grows with the square of the memory size for the SOVF. Ill-conditioning of the input signal auto-correlation matrix in addition to the high number of coefficients lead to low convergence rate. These characteristics are all the more restricting since the nonlinear part varies rapidly with the acoustic path due to the effect of their convolution (see right-hand part of Fig. 1). Fermo proposed using an APA2 algorithm [7] to speed-up convergence. Stenger *et al.* observed that the response of the SOVF is concentrated around the main section of the linear response, which corresponds to the delay introduced by the acoustic path [8]. Thus, they proposed reducing the complexity, while increasing at the same time the convergence speed, by truncating these first nonsignificant coefficients. Other models were developed to reduce complexity based on the decomposition of the higher order Volterra kernels into products of first-order kernels: let us refer to the "parallel-cascade" structure of Panicker *et al.* [9], and the MMD (multi-memory decomposition) filters of Frank [10]. All these contributions show the growing importance of nonlinear filtering in the AEC field.

In the context of hands-free communications, some of our studies on real databases showed that loudspeaker nonlinearities are characterized by powerful harmonics whose energy depends on the excitation frequency. A new echo cancellation algorithm has thus been developed, so as to take into account the loudspeaker nonlinearities. In order to separate the identification of the nonlinear loudspeaker parameters and the tracking of the linear acoustic path changes, the proposed adaptive nonlinear AEC system is based on the model displayed by the left-hand part of Fig. 1.

Manuscript received September 27, 2001; revised April 30, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dirk van Compernolle.

The authors are with the Laboratoire Traitement du Signal et de l'Image, INSERM-Université de Rennes 1, 35042 Rennes Cedex, France (e-mail: alexandre.guerin@univ-rennes1.fr).

Digital Object Identifier 10.1109/TSA.2003.818077

1063-6676/03$17.00 © 2003 IEEE

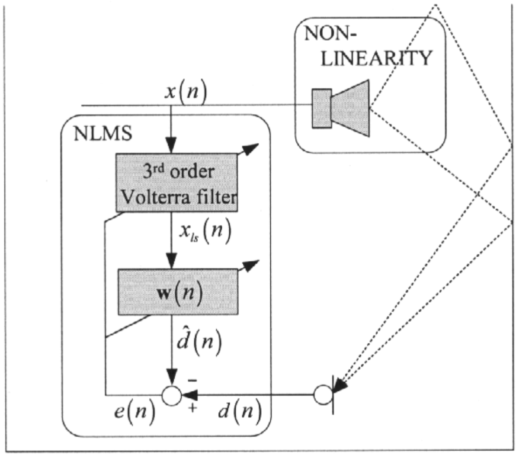


Fig. 1. Adaptive cascaded system dedicated to nonlinear acoustic echo cancellation: the loudspeaker is modeled by third-order Volterra filters while the acoustic path is modeled by a standard linear filter.

1) The first module, based on polynomial Volterra filters, identifies the loudspeaker parameters. This identification is achieved by the Normalized-LMS algorithm (NLMS), using the linear property of Volterra operators.
2) The output signal of the Volterra filter feeds a classical linear filter which identifies the linear acoustic impulse response. The identification is also achieved by the NLMS algorithm.

Intuitively, the loudspeaker model may be assumed as fixed and learnt off-line. Unfortunately, the nonnegligible component dispersion requires some parameter learning that depends upon the loudspeaker itself. Moreover, the characteristics of a given loudspeaker depend on different parameters like aging, temperature, or even the duration of the communication. All these considerations lead to an adaptive loudspeaker identification (and obviously the acoustic path) as depicted in Fig. 1.

The overall system model is described in the second section. The LMS identification equations used for the Volterra parameters and the linear acoustic path are derived in the third section. These equations for the cascaded system are more complicated than those of the classical parallel approach, especially for the upstream module. The stability problem will be considered in particular, leading to the NLMS formulation. We also focus on the problem of a permanent oscillating system, due to the combined effect of the cascaded structure and filters misadjustments, especially during the transient phase. A particular strategy, which we will discuss, was developed to avoid this phenomenon. The last section presents simulations using real data recorded on Alcatel material which compare the performance of the linear NLMS algorithm, the standard parallel Volterra filters, and that of the proposed system.

II. Model

As presented in the introduction, previously used models based on parallel structure suffer from low convergence speed and lead us to consider the cascaded approach. The loudspeaker model uses Volterra filters, based on the fact that loudspeakers usually have memory imbedded in their nonlinear part, creating harmonic distortions depending on the frequency. Then, we propose in Section II-B a new global model of the system.

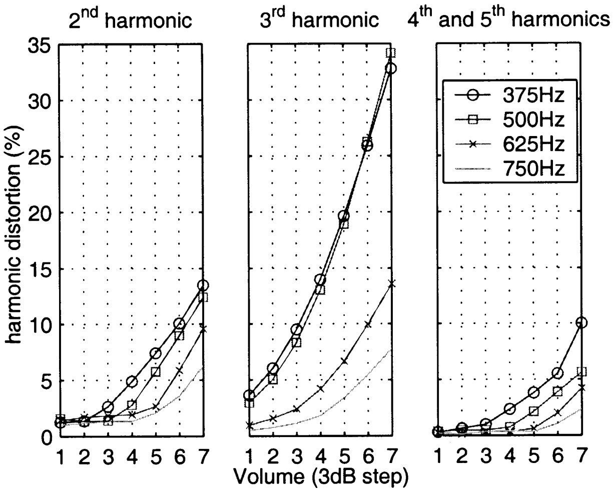


Fig. 2. Harmonic distortion values computed for different harmonics and for different frequencies: 375 Hz (o), 500 Hz (x), 625 Hz (square), and 750 Hz (no marker). Left box: second harmonic distortion. Center box: third harmonic distortion. Right box: fourth and fifth harmonics distortion.

A. Loudspeaker Modeling

Polynomial Volterra filters can model any nonlinear function with or without memory due to their structure: they constitute the memory counterpart of the Taylor series [11]. If we denote the time sample index by n and the loudspeaker input by $x(n)$, the loudspeaker may be modeled by the following causal p-order Volterra filter:

$$\begin{aligned} x_{ls}(n) = & \sum_{\alpha_1=0}^{+\infty} g_1(\alpha_1)\, x(n-\alpha_1) \\ & + \sum_{\alpha_1=0}^{+\infty} \sum_{\alpha_2=\alpha_1}^{+\infty} g_2(\alpha_1,\alpha_2)\, x(n-\alpha_1)\, x(n-\alpha_2) \\ & \vdots \\ & + \sum_{\alpha_1=0}^{+\infty} \cdots \sum_{\alpha_p=\alpha_{p-1}}^{+\infty} g_p(\alpha_1,\ldots,\alpha_p)\, x(n-\alpha_1) \\ & \ldots x(n-\alpha_p) \end{aligned}$$

where $x_{ls}(n)$ stands for the loudspeaker signal (see Fig. 1), the kernel g_1 for linear impulse response and the higher orders $g_i, i \geqslant 2$, for the nonlinear harmonic impulse responses.

The question is the choice of the model order. Fig. 2 depicts the harmonic distortion for different frequencies from 375 Hz to 750 Hz, with 125 Hz step, depending on the power of the input signal (3 dB step): the maximal volume "7"corresponds to the maximal power applied to the loudspeaker imposed by the manufacturer. The left box shows the second harmonic distortion, the center box the third harmonic distortion, and the right box the combined distortion created by the fourth and fifth harmonics. First of all, these curves show different harmonic distortions depending on the frequency, which confirms the presence of nonlinearity with memory. Secondly, these curves show very

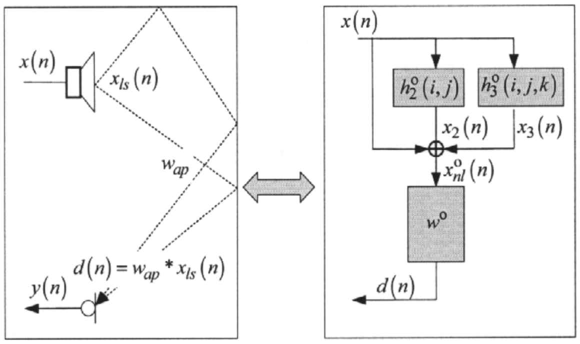


Fig. 3. Modified cascaded model of the hands-free system: the linear parts of the loudspeaker and the acoustic path are grouped into one single filter.

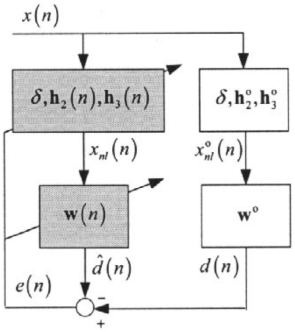


Fig. 4. Cascaded model of the system (right hand part) and mirror adaptive system (left hand part).

important distortions up to 30% (center box), which are highly irritating for pure tones; whereas for more complex sounds like speech, the distortion is less audible because of the presence of harmonics in the signal itself. Thirdly, the third harmonic distortion (center box) has to be taken into account, due to its predominance over the other distortions. The second harmonic will also be modeled, considering that it contains up to 15% distortion, especially for high volumes and low frequencies. We decided to leave out the higher harmonics due to their relative weakness compared to the others (10% maximum for very low frequencies), and also due to their complexity when modeled by Volterra filters.

According to these observations, the Volterra model of the loudspeaker, limited up to the third order, may be expressed as

$$\begin{aligned} x_{ls}(n) = &\sum_{i=0}^{L-1} g_1(i)x(n-i) \\ &+ \sum_{i=0}^{L-1}\sum_{j=i}^{L-1} g_2(i,j)x(n-i)x(n-j) \\ &+ \sum_{i=0}^{L-1}\sum_{j=i}^{L-1}\sum_{k=j}^{L-1} g_3(i,j,k) \\ &\times x(n-i)x(n-j)x(n-k). \end{aligned} \tag{1}$$

The upper limit, L, derives from the damping nature of real systems: over a given index L, the response of the loudspeaker may be considered as negligible compared to its response over the period $[0; L-1]$.

B. System Model

Classically, the acoustic path is modeled by a linear tap delay filter w_{ap} (see Fig. 3). The identification of the cascaded model, including loudspeaker and acoustic path suffers from nonunique solutions, especially for the linear part (zeros of g_1 may be included in w_{ap} and vice-versa). To circumvent this problem, we grouped the linear parts g_1 and w_{ap} in one single filter $w^o = g_1 * w_{ap}$ as depicted on Fig. 3. The key point of such modeling is the existence of causal second and third-order kernels h_2^o and h_3^o that verify the equivalence of the systems: $\begin{pmatrix} g_1, g_2, g_3 \\ w_{ap} \end{pmatrix} \Leftrightarrow \begin{pmatrix} \delta, h_2^o, h_3^o \\ w^o \end{pmatrix}$, where δ is the kroneker symbol. Intuitively, their existence is related to the invertibility of the linear part g_1 of the loudspeaker. We can show that h_2^o (resp. h_3^o) must be a type of convolution between g_2 (resp. g_3) and the inverse of g_1. A priori, it cannot be proved that g_1 is a minimal phase filter, but it is often assumed in linearization techniques. Klippel's works on loudspeaker linearization showed that this hypothesis seems valid in most of cases [12]–[14]. In addition, many other authors such as Frank [15] and Nomura [16], also assume this invertibility. In any case, it is important to keep in mind that the nonlinear filters h_2^o and h_3^o in the modified model only depend on the loudspeaker characteristics, but not on the acoustic path.

Assuming invertibility, the global system loudspeaker/acoustic path may be modeled, without any loss of generality, by the cascaded system (nonlinear/linear) depicted on Fig. 3. The upstream nonlinear module, whose output is denoted by $x_{nl}^o(n)$, is then modeled by a Dirac in parallel with second- and third-order kernels

$$\begin{aligned} x_{nl}^o(n) = x(n) &+ \sum_{i=0}^{L-1}\sum_{j=i}^{L-1} h_2^o(i,j)\,x(n-i)\,x(n-j) \\ &+ \sum_{i=0}^{L-1}\sum_{j=i}^{L-1}\sum_{k=j}^{L-1} h_3^o(i,j,k)\,x(n-i)\,x(n-j)\,x(n-k) \end{aligned} \tag{2}$$

where the constant L stands for the finite memory of the Volterra filters (L is different from the nonlinear part memory in (1), since it includes the linear part inverse).

III. Nonlinear Acoustic Echo Cancellation

The adaptive system, depicted on Fig. 4, is based on the model defined by (2). The equations defining the different quantities of the model (input, output and intermediate variables) are introduced in Section III-A. Then, the update equations of the adaptive system are derived in Section III-B, using the LMS algorithm, and stability conditions are made explicit in Section III-C. The transient behavior of the adaptive filters is analyzed in Section III-D, under a quasistationarity hypothesis, and leads us to formulate a strategy of adaptation in Section III-E avoiding permanent oscillations.

A. Notations

We define by the bold symbols $\mathbf{w}^o$, $\mathbf{h}_2^o$ and $\mathbf{h}_3^o$ on Fig. 4 the vector notation of the impulse responses w^o, h_2^o and h_3^o. Let $\mathbf{w}(n) = [w_0(n)\, w_1(n) \ldots w_{N-1}(n)]^T$ denote the linear adaptive filter used to identify the global impulse response $\mathbf{w}^o$.

The input signal of the loudspeaker is $x(n)$, and $\mathbf{x}^N(n) = [x(n)\,x(n-1)\ldots x(n-N+1)]^T$ is the input data vector of length N. In the following for convenience reasons, the N index will be omitted.

We denote by $x_{nl}(n)$ the output of the nonlinear upstream module expressed as

$$x_{nl}(n) = x(n) + x_2(n) + x_3(n), \tag{3}$$

where

$$x_2(n) = \sum_{i=0}^{L-1}\sum_{j=i}^{L-1} h_2(i,j;n)x(n-i)\,x(n-j), \tag{4}$$

$$x_3(n) = \sum_{i=0}^{L-1}\sum_{j=i}^{L-1}\sum_{j=k}^{L-1} h_3(i,j,k;n) \times x(n-i)\,x(n-j)\,x(n-k), \tag{5}$$

and $h_2(.,.;n)$ and $h_3(.,.,.;n)$ are the nonlinear second- and third-order Volterra kernels.

We denote by $\mathbf{x}_{nl}(n)$, $\mathbf{x}_2(n)$ and $\mathbf{x}_3(n)$ the N length nonlinear vectors corresponding to $x_{nl}(n)$, $x_2(n)$ and $x_3(n)$. Then, the estimated main signal $\hat{d}(n)$ is given by the convolution

$$\hat{d}(n) = \mathbf{x}_{nl}^T(n)\,\mathbf{w}(n). \tag{6}$$

Using (6), the error signal defined by $e(n) = d(n) - \hat{d}(n)$ (where $d(n)$ is the real echo) is given by the formula

$$e(n) = d(n) - \mathbf{x}_{nl}^T(n)\,\mathbf{w}(n). \tag{7}$$

The mean-squared error $J(n)$ is defined by the equation

$$J(n) = E\left\{e^2(n)\right\}, \tag{8}$$

where $E\{.\}$ stands for the expectation operator.

If $\mathbf{x}$ and $\mathbf{y}$ are two column vectors of same length, we denote by $\mathbf{\Gamma}_{\mathbf{x},\mathbf{y}} = E\left\{\mathbf{x}\mathbf{y}^T\right\}$ the cross-correlation matrix ($\mathbf{\Gamma}_{\mathbf{x}} = E\left\{\mathbf{x}\mathbf{x}^T\right\}$ is the auto-correlation matrix of vector $\mathbf{x}$).

For convenience in later computations, the vectors $\mathbf{x}_2$ and $\mathbf{x}_3$ may be expressed as products of matrices. We define first the filter $\mathbf{h}_2$, corresponding to the second-order nonlinearity as the vector of dimension $L_2 \times 1$ (where $L_2 = L(L+1)/2$)

$$\mathbf{h}_2(n) = \begin{bmatrix} h_2(0,0;n) \\ h_2(0,1;n) \\ \vdots \\ h_2(0,L-1;n) \\ h_2(1,1;n) \\ \vdots \\ h_2(L-1,L-1;n) \end{bmatrix}. \tag{9}$$

We associate to the vector $\mathbf{h}_2$ the $N \times L_2$ matrix $\mathbf{U}_2$, whose expression is given by (10), shown at the bottom of the page, so that their product yields $\mathbf{x}_2$

$$\mathbf{x}_2(n) = \mathbf{U}_2(n)\,\mathbf{h}_2(n). \tag{11}$$

Using the same conventions, the nonlinear vector of order 3 of dimension $L_3 = (L(L+1)(L+2)/6)$ may be expressed as

$$\mathbf{x}_3(n) = \mathbf{U}_3(n)\,\mathbf{h}_3(n). \tag{12}$$

B. LMS Adaptation

By developing the convolution $\mathbf{x}_{nl}^T(n)\,\mathbf{w}(n)$, it may be easily shown that the error is a bilinear function of the adaptive coefficients (products of $w_i(n)$ and $h_2(i,j;n)$ or $h_3(i,j,k;n)$). Therefore, the update equations can be easily derived using the Least Mean-Square algorithm [1]. The update equation of a given vector $\mathbf{v}(n)$ ($\mathbf{w}(n)$, $\mathbf{h}_2(n)$ or $\mathbf{h}_3(n)$), using the LMS algorithm is given by

$$\mathbf{v}(n+1) = \mathbf{v}(n) - \frac{1}{2}\mu\nabla_{\mathbf{v}}\left\{e^2(n)\right\} \tag{13}$$

where μ is the step size and $\nabla_{\mathbf{v}}$ is the gradient with respect to $\mathbf{v}$.

Substituting the expression of the squared error given by (7) leads to

$$\mathbf{v}(n+1) = \mathbf{v}(n) + \mu e(n)\,\nabla_{\mathbf{v}}\left\{\mathbf{x}_{nl}^T(n)\,\mathbf{w}(n)\right\}. \tag{14}$$

Equation (14) will be used to derive the different update equations.

$$\mathbf{U}_2^T(n) = \begin{bmatrix} x(n)^2 & x(n-1)^2 & \ldots & x(n-N+1)^2 \\ x(n)x(n-1) & x(n-1)x(n-2) & & \vdots \\ \vdots & \vdots & & \vdots \\ x(n)x(n-L+1) & \vdots & & \vdots \\ x(n-1)^2 & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ x(n-L+1)^2 & \vdots & & x(n-L-N+2)^2 \end{bmatrix} \tag{10}$$

1) Linear Filter $\mathbf{w}(n)$*:* Replacing $\mathbf{v}$ by $\mathbf{w}$ (and μ by μ_1) in (14) yields

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu_1 e(n) \nabla_{\mathbf{w}} \left\{ \mathbf{x}_{nl}^T(n) \mathbf{w}(n) \right\}. \quad (15)$$

The computation of the partial derivative in the above equation is well-known and leads to

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu_1 e(n) \mathbf{x}_{nl}(n). \quad (16)$$

This equation is similar to the classical LMS equation with $x_{nl}(n)$ standing for the input signal of the adaptive linear filter.

2) Nonlinear Filters $\mathbf{h}_2(n)$ *and* $\mathbf{h}_3(n)$*:* The explicit gradient computations will be detailed only for the second-order kernel, the extension to the third-order being obvious. Replacing $\mathbf{v}$ by $\mathbf{h}_2$ (and μ by μ_2) in (14) yields

$$\mathbf{h}_2(n+1) = \mathbf{h}_2(n) + \mu_2 e(n) \nabla_{\mathbf{h}_2} \left\{ \mathbf{x}_{nl}^T(n) \mathbf{w}(n) \right\}. \quad (17)$$

Using (11) and the decomposition of $\mathbf{x}_{nl}(n)$ given by (3), it easily becomes

$$\begin{aligned} \nabla_{\mathbf{h}_2} \left\{ \mathbf{x}_{nl}^T(n) \mathbf{w}(n) \right\} =& \nabla_{\mathbf{h}_2} \left\{ \mathbf{w}^T(n) (\mathbf{x}(n) + \mathbf{x}_2(n) \right. \\ & \left. + \mathbf{x}_3(n)) \right\} \\ =& \nabla_{\mathbf{h}_2} \left\{ \mathbf{w}^T(n) \mathbf{U}_2(n) \mathbf{h}_2(n) \right\} \\ =& \mathbf{U}_2^T(n) \mathbf{w}(n). \end{aligned} \quad (18)$$

Substituting (18) in (17) yields

$$\mathbf{h}_2(n+1) = \mathbf{h}_2(n) + \mu_2 e(n) \mathbf{U}_2^T(n) \mathbf{w}(n). \quad (19)$$

Using the same arguments as those developed for the second-order kernel, the update equation of the third-order coefficients vector $\mathbf{h}_3(n)$ is

$$\mathbf{h}_3(n+1) = \mathbf{h}_3(n) + \mu_3 e(n) \mathbf{U}_3^T(n) \mathbf{w}(n). \quad (20)$$

C. Stability Condition: Normalized-LMS

The update equations (16), (19) and (20) do not ensure stability unless a strong condition is imposed on the step sizes μ_1, μ_2 and μ_3. The optimum step size for the LMS algorithm which guarantees stability and fast convergence leads to the so-called NLMS algorithm. The variable step size depends upon the ambient noise and residual echo power [17]. Here, the computation of the normalization is extended for the cascaded system, neglecting ambient noise.

For demonstration purposes, another expression for the squared error is

$$\begin{aligned} e^2(n) =& d^2(n) - 2d(n) \mathbf{x}_{nl}^T(n) \mathbf{w}(n) \\ & + \mathbf{w}^T(n) \mathbf{x}_{nl}(n) \mathbf{x}_{nl}^T(n) \mathbf{w}(n). \end{aligned} \quad (21)$$

This expression will be differentiated with respect to the different adaptive filters. This allows us to express the bounds for the step sizes $\mu_i, \quad i \in \{1,2,3\}$, which guarantee the tradeoff between convergence speed and stability.

1) Linear Filter $\mathbf{w}(n)$*:* Differentiating (21) relatively to $\mathbf{w}(n)$ yields

$$\nabla_{\mathbf{w}} \left\{ e^2(n) \right\} = -2d(n) \mathbf{x}_{nl}(n) + 2\mathbf{x}_{nl}(n) \mathbf{x}_{nl}^T(n) \mathbf{w}(n). \quad (22)$$

Substituting the previous result in (13) leads to

$$\begin{aligned} \mathbf{w}(n+1) =& \mathbf{w}(n) + \mu_1 d(n) \mathbf{x}_{nl}(n) - \mu_1 \mathbf{x}_{nl}(n) \mathbf{x}_{nl}^T(n) \\ & \times \mathbf{w}(n) \\ =& \left[\mathbf{I}_N - \mu_1 \mathbf{x}_{nl}(n) \mathbf{x}_{nl}^T(n) \right] \mathbf{w}(n) + \mu_1 d(n) \\ & \times \mathbf{x}_{nl}(n), \end{aligned} \quad (23)$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix.

If we suppose that the nonlinear filters $\mathbf{h}_2$ and $\mathbf{h}_3$ are time-invariant, the adaptation of $\mathbf{w}$ is similar to the classical case (no nonlinear filter). Therefore, the NLMS equation driving the update of the linear filter finally becomes (see [1], [18])

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu_1 e(n) \frac{\mathbf{x}_{nl}(n)}{\|\mathbf{x}_{nl}(n)\|_2^2}, \quad (24)$$

where $\|.\|_2$ is the L_2 norm. Stable convergence of the average tap weight vector $E\{\mathbf{w}(n)\}$ can be proven for $0 < \mu_1 < 2$, on the basis of the "independence assumption" [1]. Although the independence is not guaranteed (because of the memory of the nonlinear filters), it can be proved that deviation from the theory introduced by these nonlinearities may be neglected due to their relative weakness compared to the main signal $x(n)$.

2) Nonlinear Filters $\mathbf{h}_2(n)$ *and* $\mathbf{h}_3(n)$*:* Computations are derived for the second-order filter and directly extended to the third-order filter. Equation (21) is once more used to analytically express the convergence condition

$$\begin{aligned} \nabla_{\mathbf{h}_2} \left\{ e^2(n) \right\} =& -2\nabla_{\mathbf{h}_2} \left\{ d(n) \mathbf{x}_{nl}^T(n) \mathbf{w}(n) \right\} \\ & + \nabla_{\mathbf{h}_2} \left\{ \mathbf{w}^T(n) \mathbf{x}_{nl}(n) \mathbf{x}_{nl}^T(n) \mathbf{w}(n) \right\}. \end{aligned} \quad (25)$$

Using (18), the first derivative of the right-hand term of (25) becomes

$$\nabla_{\mathbf{h}_2} \left\{ d(n) \mathbf{x}_{nl}^T(n) \mathbf{w}(n) \right\} = d(n) \mathbf{U}_2^T(n) \mathbf{w}(n). \quad (26)$$

This term is independent of vector $\mathbf{h}_2(n)$.

Using (18), the computation of the second derivative term leads to

$$\begin{aligned} \nabla_{\mathbf{h}_2} & \left\{ \mathbf{w}^T(n) \mathbf{x}_{nl}(n) \mathbf{x}_{nl}^T(n) \mathbf{w}(n) \right\} \\ & = 2\mathbf{x}_{nl}^T(n) \mathbf{w}(n) \mathbf{U}_2^T \mathbf{w}(n). \end{aligned} \quad (27)$$

Using the decomposition of the nonlinear vector $\mathbf{x}_{nl}(n)$, defined by (3), (27) becomes

$$\begin{aligned} \mathbf{x}_{nl}^T(n) \mathbf{w}(n) \mathbf{U}_2^T \mathbf{w}(n) =& (\mathbf{x}(n) + \mathbf{x}_3(n))^T \mathbf{w}(n) \\ & \times \mathbf{U}_2^T \mathbf{w}(n) \\ & + \mathbf{x}_2^T(n) \mathbf{w}(n) \mathbf{U}_2^T \mathbf{w}(n) \\ =& (\mathbf{x}(n) + \mathbf{x}_3(n))^T \mathbf{w}(n) \\ & \times \mathbf{U}_2^T \mathbf{w}(n) \\ & + \mathbf{U}_2^T \mathbf{w}(n) \mathbf{w}^T(n) \mathbf{U}_2(n) \\ & \times \mathbf{h}_2(n). \end{aligned} \quad (28)$$

Replacing the different results in (25) yields

$$\begin{aligned} \nabla_{\mathbf{h}_2} \left\{ e^2(n) \right\} =& -2d(n) \mathbf{U}_2^T(n) \mathbf{w}(n) \\ & + 2(\mathbf{x}(n) + \mathbf{x}_3(n))^T \mathbf{w}(n) \mathbf{U}_2^T \mathbf{w}(n) \\ & + 2\mathbf{U}_2^T \mathbf{w}(n) \mathbf{w}^T(n) \mathbf{U}_2(n) \mathbf{h}_2(n). \end{aligned} \quad (29)$$

Substituting this form of the gradient in (13) leads to

$$\begin{aligned}\mathbf{h}_2(n+1) =& \mathbf{h}_2(n) - \mu_2 \mathbf{U}_2^T(n)\mathbf{w}(n)\mathbf{w}^T(n)\mathbf{U}_2(n)\mathbf{h}_2(n)\\ &+ \mu_2 d(n)\mathbf{U}_2^T(n)\mathbf{w}(n)\\ &- \mu_2(\mathbf{x}(n)+\mathbf{x}_3(n))^T\mathbf{w}(n)\mathbf{U}_2^T(n)\mathbf{w}(n)\end{aligned} \tag{30}$$

which may be also expressed as

$$\mathbf{h}_2(n+1) = [\mathbf{I}_{L_2} - \mu_2\mathbf{M}]\mathbf{h}_2(n) + \mathbf{c} \tag{31}$$

where $\mathbf{M} = \mathbf{U}_2^T(n)\mathbf{w}(n)\mathbf{w}^T(n)\mathbf{U}_2(n)$. This ensures that $E\{\mathbf{M}\}$ is a positive semi-definite matrix, and that its eigenvalues are all positive or null.

If we suppose now that the different filters $\mathbf{w}(n)$, $\mathbf{w}^o(n)$, and $\mathbf{h}_3(n)$, and the statistical characteristics of the input signal are constant during the adaptation of the filter $\mathbf{h}_2(n)$, the matrix $E\{\mathbf{M}\}$ and the vector $E\{\mathbf{c}\}$ are constant. Thus, the NLMS update equation of $\mathbf{h}_2(n)$ may be formulated as

$$\mathbf{h}_2(n+1) = \mathbf{h}_2(n) + \mu_2 e(n)\frac{\mathbf{U}_2^T(n)\mathbf{w}(n)}{\left\|\mathbf{U}_2^T(n)\mathbf{w}(n)\right\|_2^2}, \tag{32}$$

and that of the third-order filter as

$$\mathbf{h}_3(n+1) = \mathbf{h}_3(n) + \mu_3 e(n)\frac{\mathbf{U}_3^T(n)\mathbf{w}(n)}{\left\|\mathbf{U}_3^T(n)\mathbf{w}(n)\right\|_2^2}. \tag{33}$$

These are classical equations and may be compared to the equation of the linear adaptive filter. The conditions on the step sizes $0 < \mu_i < 2, \quad i \in \{2,3\}$ ensure convergence in the mean sense. Nevertheless, the independence assumptions do not hold in this particular case and this result has to be considered carefully (see [18]).

D. Transient Behavior Analysis

The Section III-C ensures the stability of the whole system, but does not guarantee the convergence toward the optimal parameters in the system sense (the parameters to identify). Indeed, the nonquadratic form of the MSE J produces some local minima, implying the convergence toward incorrect parameters depending on the initialization. However, the experiments we conducted exhibited no convergence toward local minima, with respect to "reasonable" initialization. The MSE surfaces we computed confirmed that the local minima were placed far away from the system parameters. Similar results were noticed by Bershad *et al.* with the identification of Wiener-Hammerstein models [19]. Let us point out that the convergence speed may be very small due to the nearly null gradients in some regions of the parameters.

In any case, the influence of the cascaded structure is more complex than simply the effects of local minima. In [20], using an adaptive Wiener-Hammerstein cascaded model, Nollet *et al.* noted that problems of convergence occured when simultaneously adapting the postfilter and the nonlinear one: the system was not able to remove any echo at all. They particularly point out the influence of filter initialization on the final convergence. To circumvent this problem, they adopted a strategy consisting in the adaptation of the postfilter before the nonlinear one. This phenomenon is also mentioned by Stenger *et al.* in [18] for an adaptive Hammerstein system. To ensure convergence, the authors used a smaller step size for the upstream nonlinear filter. They also recommend not to adapt the nonlinear filter until the linear one has "sufficiently" converged.

In this section, we describe theoretically this phenomenon, and explain why the adaptive system may oscillate without converging. For demonstration purposes, the unknown system in Fig. 4 consists of two cascaded filters: a nonlinear one composed of second- and third-order Volterra filters $\mathbf{h}_2^o$ and $\mathbf{h}_3^o$ plus a Dirac, and a linear filter $\mathbf{w}^o$. Thus $d(n)$ may be expressed as the convolution

$$d(n) = \mathbf{x}_{nl}^{o\,T}(n)\mathbf{w}^o, \tag{34}$$

where $\mathbf{x}_{nl}^o(n) = \mathbf{x}(n) + \mathbf{U}_2(n)\mathbf{h}_2^o + \mathbf{U}_3(n)\mathbf{h}_3^o$ is the nonlinear module output.

The system is identified by adaptive filters of the same length, $\mathbf{h}_2(n)$, $\mathbf{h}_3(n)$ and $\mathbf{w}(n)$.

The squared error given by (22) can be used to calculate the optimal parameter set in the minimum mean-squared error sense (MMSE). These filters, corresponding respectively to $\mathbf{w}$ and $\mathbf{h}_2$ are denoted by $\mathbf{w}_{MMSE}$ and $\mathbf{h}_{2,MMSE}$. To give an analytical expression of these filters, we suppose that the situation is pseudo-stationary, *i.e.*, the other filters are quasiconstant. Although this assumption does not exactly correspond to the real situation (permanent adaptation of the whole system), it is necessary because the optimal filters actually depend on transient states, as a result of the cascaded structure. This may be interpreted as the evaluation of the optimal filters at a given time n, the state of the other filters being a "mean state" at this time.

1) Linear Filter $\mathbf{w}$*:* The gradient $\nabla_{\mathbf{w}}\{J(n)\}$ is given by

$$\begin{aligned}\nabla_{\mathbf{w}}\{J(n)\} = &-2E\{d(n)\mathbf{x}_{nl}(n)\}\\ &+2E\left\{\mathbf{x}_{nl}(n)\mathbf{x}_{nl}^T(n)\right\}\mathbf{w}.\end{aligned} \tag{35}$$

Substituting the expression (34) in the previous equation yields

$$\begin{aligned}\nabla_{\mathbf{w}}\{J(n)\} = &-2E\left\{\mathbf{x}_{nl}(n)\mathbf{x}_{nl}^{o^T}(n)\right\}\mathbf{w}^o\\ &+2E\left\{\mathbf{x}_{nl}(n)\mathbf{x}_{nl}^T(n)\right\}\mathbf{w}.\end{aligned} \tag{36}$$

Equation (36) shows that $\mathbf{w}_{MMSE}$, which satisfies the equality $\nabla_{\mathbf{w}}\{J(\mathbf{w})\}|_{\mathbf{w}=\mathbf{w}_{MMSE}} = \mathbf{0}$, depends on the correlation matrices $\mathbf{\Gamma}_{\mathbf{x}_{nl}}$ and $\mathbf{\Gamma}_{\mathbf{x}_{nl},\mathbf{x}_{nl}^o}$, and thus on the state of the adaptive nonlinear filters. Using now the quasistationarity hypothesis, the optimal filter is then defined by the following expression

$$\mathbf{w}_{MMSE} = \mathbf{\Gamma}_{\mathbf{x}_{nl}}^{-1}\mathbf{\Gamma}_{\mathbf{x}_{nl},\mathbf{x}_{nl}^o}\mathbf{w}^o. \tag{37}$$

In (37), the Wiener filter should also be indexed by time n since it depends on $\mathbf{h}_2(n)$ and $\mathbf{h}_3(n)$.

Equation (37) shows that if the upstream nonlinear module has not yet converged

1) $\mathbf{w}(n)$ will not converge (in the mean) toward the filter $\mathbf{w}^o$ but to a linear transformed version, depending on the nonlinear filters. The transient behavior of $\mathbf{w}(n)$ may be chaotic,
2) the Wiener filter also depends on input signal statistics (variance and higher orders). Hence, for a given state of

the nonlinear filters, $\mathbf{w}_{MMSE}$ varies over the time with the signal statistics;

3) if we suppose now that the nonlinear filters adapt continuously, the statistics of the correlation matrices $\mathbf{\Gamma}_{\mathbf{x}_{nl},\mathbf{x}_{nl}^{o}}$ and $\mathbf{\Gamma}_{\mathbf{x}_{nl}}$ change rapidly. Thus, even with a time invariant system to be identified, the adaptive system has to cope with a rapidly moving target.

However, the impact of the misadjustment of the nonlinear filters in real situations has to be downgraded. With wide band signals like speech, loudspeakers' nonlinearities are much less energetic than the linear part (at least 15 dB lower). Thus, if we assume that, during their adaptation, filters $\mathbf{h}_2(n)$ and $\mathbf{h}_3(n)$ remain comparable to those of the real system ($\mathbf{h}_2^o$ and $\mathbf{h}_3^o$), the whole nonlinearity may be neglected in the correlation matrices $\mathbf{\Gamma}_{\mathbf{x}_{nl}}$ and $\mathbf{\Gamma}_{\mathbf{x}_{nl},\mathbf{x}_{nl}^{o}}$, which become quite comparable and equal to $\mathbf{\Gamma}_{\mathbf{x}}$. Using (37) leads to the following approximation

$$\mathbf{w}_{MMSE} \simeq \mathbf{w}^{o}, \quad \text{if } \|x_2\| \ll \|x\| \, \text{and} \, \|x_3\| \ll \|x\|. \qquad (38)$$

Hence, under the hypothesis of small nonlinearities, the linear filter should converge toward the filter to be identified. However, let us specify that if we can neglect the influence of the nonlinearities on the optimal linear filter, it is obviously not the case for the residual error.

2) Nonlinear Filter $\mathbf{h}_2(n)$*:* Combining gradient equation (29) and (34) leads to

$$\begin{aligned}\nabla_{\mathbf{h}_2}\{J(n)\} =& -2E\left\{\mathbf{U}_2^T(n)\,\mathbf{w}(n)\,{\mathbf{x}_{nl}^{o}}^{T}(n)\,\mathbf{w}^{o}\right\}\\ &+2E\left\{(\mathbf{x}(n)+\mathbf{x}_3(n))^T\,\mathbf{w}(n)\right.\\ &\quad\left.\times\mathbf{U}_2^T(n)\mathbf{w}(n)\right\}\\ &+2E\left\{\mathbf{U}_2^T(n)\mathbf{w}(n)\,\mathbf{w}^T(n)\,\mathbf{U}_2(n)\right\}\\ &\times\mathbf{h}_2(n).\end{aligned} \qquad (39)$$

Replacing $\mathbf{x}_{nl}^{o}$ by its decomposed form yields

$$\begin{aligned}\nabla_{\mathbf{h}_2}\{J(n)\} =& -2E\left\{\mathbf{U}_2^T(n)\,\mathbf{w}(n)\,\mathbf{w}^{oT}\mathbf{x}_2^o(n)\right\}\\ &-2E\left\{\mathbf{U}_2^T(n)\,\mathbf{w}(n)\,\mathbf{w}^{oT}(\mathbf{x}(n)+\mathbf{x}_3^o(n))\right\}\\ &+2E\left\{(\mathbf{x}(n)+\mathbf{x}_3(n))^T\,\mathbf{w}(n)\,\mathbf{U}_2^T(n)\right.\\ &\quad\left.\times\mathbf{w}(n)\right\}\\ &+2E\left\{\mathbf{U}_2^T(n)\,\mathbf{w}(n)\,\mathbf{w}^T(n)\,\mathbf{U}_2(n)\right\}\\ &\times\mathbf{h}_2(n).\end{aligned} \qquad (40)$$

Finally, using (11) with optimal parameters $\mathbf{h}_2^o$ and putting $\nabla_{\mathbf{h}_2}\{J(n)\} = \mathbf{0}$ produces

$$\begin{aligned}\mathbf{h}_{2,MMSE} =& \mathbf{h}_2^o + \mathbf{\Gamma}_{\mathbf{U}_2^T\mathbf{w}}^{-1}\left[E\left\{\mathbf{U}_2\mathbf{w}\mathbf{e}_{\mathbf{w}}^T\mathbf{x}\right\}\right.\\ &+E\left\{\mathbf{U}_2^T\mathbf{w}\mathbf{e}_{\mathbf{w}}^T(\mathbf{x}_2^o+\mathbf{x}_3^o)\right\}\\ &\left.+E\left\{\mathbf{U}_2^T\mathbf{w}\mathbf{w}^T(\mathbf{x}_3^o-\mathbf{x}_3)\right\}\right]\end{aligned} \qquad (41)$$

where $\mathbf{e}_{\mathbf{w}}(n)$ is the linear filter misadjustment

$$\mathbf{e}_{\mathbf{w}}(n) = \mathbf{w}^o - \mathbf{w}(n), \qquad (42)$$

and $\mathbf{e}_{\mathbf{h}_3}(n)$ is the third-order filter misadjustment

$$\mathbf{e}_{\mathbf{h}_3}(n) = \mathbf{h}_3^o - \mathbf{h}_3(n). \qquad (43)$$

The (41) shows that, at a given time n, $\mathbf{h}_{2,MMSE}$ is equal to the desired filter $\mathbf{h}_2^o$ plus other terms depending linearly on instantaneous weight-error $\mathbf{e}_{\mathbf{w}}(n)$ and $\mathbf{e}_{\mathbf{h}_3}(n)$, noticing that $\mathbf{x}_3^o - \mathbf{x}_3 = \mathbf{U}_3\mathbf{e}_{\mathbf{h}_3}$. Then, if any of the filters $\mathbf{w}(n)$ or $\mathbf{h}_3(n)$ have not converged toward the desired parameters, $\mathbf{h}_2(n)$ will not converge toward its optimal values.

Experiments have shown that the nonlinearities generated by loudspeakers are relatively weak compared to the linear component (even if the impact on the performance is not negligible). Therefore, if we assume that $\mathbf{e}_{\mathbf{w}}$ is of the same importance as $\mathbf{w}$, the expectations of lines 2 and 3 of (41) may be neglected compared to $E\left\{\mathbf{U}_2^T\mathbf{w}\mathbf{e}_{\mathbf{w}}^T\mathbf{x}\right\}$ and lead to the approximation

$$\mathbf{h}_{2,MMSE} \simeq \mathbf{h}_2^o + \mathbf{\Gamma}_{\mathbf{U}_2^T\mathbf{w}}^{-1}E\left\{\mathbf{U}_2^T\mathbf{w}\mathbf{x}^T\right\}\mathbf{e}_{\mathbf{w}}. \qquad (44)$$

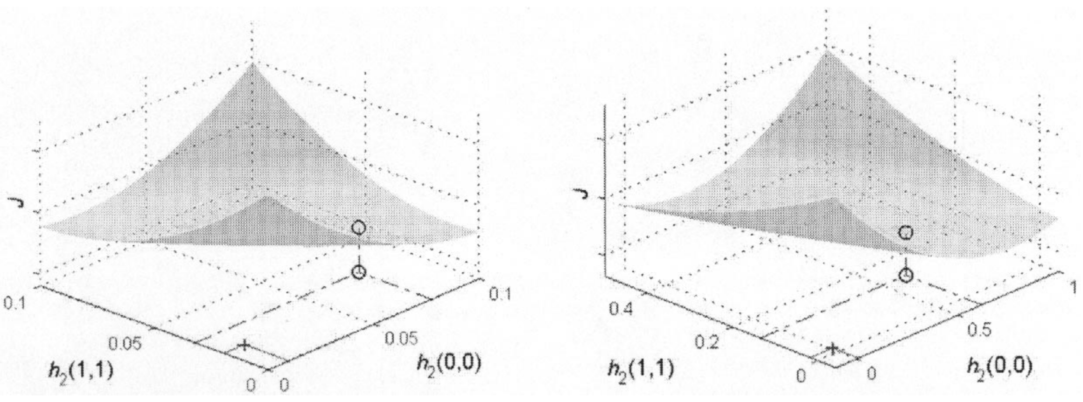
Note that this approximation is valid for colored signals like speech, but does not hold for zero-mean white Gaussian noise because the expectancy of the third-order powers is null. It can be seen in (44) that the error vector $\mathbf{e}_{\mathbf{w}}$ has a major influence on the transient behavior of $\mathbf{h}_2$, and thus on convergence. Indeed, simulations on speech signals show that the second term, depending on the error linear filter $\mathbf{e}_{\mathbf{w}}$, may be much higher than the filter $\mathbf{h}_2^o$, even for small misadjustments of the linear filter.

To illustrate this situation, we conducted simulations with a simplified system to be identified. The upstream module consists of a Dirac plus a second-order filter $\mathbf{h}_2^o$

$$\begin{cases} h_2^o(0,0) = 0.01\\ h_2^o(0,1) = 0\\ h_2^o(1,1) = 0.02\end{cases}$$

which corresponds to a nonlinear component $x_2^o(n) = 0.01x^2(n) + 0.02x^2(n-1)$, and $\mathbf{w}^o = [1\ 0.5]^T$ is the linear downstream module. The adaptive system contains the same number of coefficients: two for the linear and two for the nonlinear filters.

Fig. 5(a) and (b), respectively, depict the MSE surface as a function of the coefficients $h_2(0,0)$ and $h_2(1,1)$ for 10% and 50% linear filter misadjustment values. These surfaces have been evaluated with a quasistationary voiced speech segment "a". Note the quadratic form of the surfaces, since the error is linear in $h_2(0,0)$ and $h_2(1,1)$ once the linear coefficients are fixed. For a 10% misadjustment [Fig. 5(a)], the optimal filter (circle) takes completely different values than those of $\mathbf{h}_2^o$ (cross). It is all the more disturbing since the linear filter may vary in an important manner during a communication, changes in the acoustic path impulse response being relatively drastic. For a 50% misadjustment [Fig. 5(b)], the optimal filter takes values 10 times larger than those of $\mathbf{h}_2^o$. With these values of $\mathbf{h}_2$, the nonlinearities may be no longer considered as negligible compared to the linear echo, leading to important misadjustments of the linear filter $\mathbf{w}$ (see (37)). The transient states of linear and nonlinear filters are at the origin of convergence problems encountered in cascaded systems, since each filter has to track a continuously varying target. In fact, each filter behaves



(a) (b)

Fig. 5. MSE J as a function of $h_2\,(0,0)$ and $h_2\,(1,1)$ for two values of misadjustment of the linear filter w: (a) 10%, (b) 50%. The filter to be identified $\mathbf{h}_2^o$ is plotted as a cross, while the optimal filter (in the MMSE sense) is plotted as a circle.

to compensate the other filters' misadjustment, leading to a perpetual oscillating system.

E. Adaptation Strategy

As seen in (37) and (44), the adaptive system is not able to identify the optimal set of parameters, unless one of the modules is known (upstream or downstream). Simulations showed that, continuously adapting the filters can lead to oscillations, because the optimal filters in the MMSE sense always vary in time. Nevertheless, three remarks are worth being pointed out:

1) The linear filter has to adapt continuously so as to react to any change in the acoustic path.
2) Equation (38) points out that, assuming the nonlinear adaptive filters $\mathbf{h}_2$ and $\mathbf{h}_3$ are of the same order of magnitude as those of the real system $\mathbf{h}_2^o$ and $\mathbf{h}_3^o$, $\mathbf{w}\,(n)$ converges (in the mean sense) toward a slightly degraded version of $\mathbf{w}^o$.
3) The acoustic path may vary in an important manner during a communication: in the event of acoustic path changing, the error vector $\mathbf{e}_\mathbf{w}\,(n)$ is not negligible compared to $\mathbf{w}\,(n)$. Therefore, the nonlinear filters must not adapt until the linear filter has "sufficiently" converged and is stable.

From these observations, it comes out that the state of the linear filter $\mathbf{w}\,(n)$ has to be detected and included in the adaptation strategy: in case of transient phase (change in the acoustic path) the nonlinear filters must not adapt, whereas, in steady-state situation, the Volterra filters can adapt. Breining *et al.* proposed to evaluate the weight error norm using delay coefficients [17]. This solution works correctly for time-invariant impulse response, which is not the case here. We propose achieving the detection of the linear filter steady-state by comparing its standard deviation with a set of thresholds. It can be proved that the variance of the linear filter reaches its minimum at convergence. This minimum depends upon the input signal and noise variances, and also on the step size μ_1. Let us denote by $\overline{\mathbf{w}}\,(n)$, the $\mathbf{w}\,(n)$ time mean estimate by first-order recursive filtering

$$\overline{\mathbf{w}}\,(n+1) = \alpha\overline{\mathbf{w}}\,(n) + (1-\alpha)\,\mathbf{w}\,(n) \qquad (45)$$

where α is a scalar close to 1. The standard deviation $\sigma_\mathbf{w}\,(n)$ of the adaptive filter is also achieved by first-order IIR filtering (see (46) at the bottom of the page) where β_{up} and β_{dn} are scalars close to 1, with $\beta_{up} \leq \beta_{dn}$. The use of the $L1$ norm $\|.\|_1$ is not critical: the $L2$ norm may be equivalently used, the only difference being in the values taken by the parameter $\sigma_\mathbf{w}\,(n)$. The order relation between the forgetting factors ensures that (*i*) in case of transient phase of $\mathbf{w}\,(n)$, the standard deviation does not decrease too rapidly so as to avoid adapting the nonlinear filters until $\mathbf{w}\,(n)$ has converged, (*ii*) the standard deviation is reactive to any change in the acoustic path, so it can be employed to stop the adaptation of $\mathbf{h}_2\,(n)$ and $\mathbf{h}_3\,(n)$ at the beginning of the adaptation of $\mathbf{w}\,(n)$.

Moreover, the nonlinear module only adapts in the presence of nonlinearities, avoiding misadjustment. The detection of the nonlinearities is based on the *a priori* knowledge that they occur when the power of the loudspeaker signal is relatively high. Therefore, a simple detector based on the power of the loudspeaker signal is implemented and is compared to a fixed threshold. Nevertheless, we have to be very careful concerning the delay introduced by the acoustic path. Because of it, detection based on loudspeaker power may be in advance with respect to the presence of the nonlinearity at the microphone, especially at the beginning of speech periods. Moreover, due to the low power of the signal preceding the distortion phase, the normalization term $\left\|\mathbf{U}_2^T\,(n)\,\mathbf{w}\,(n)\right\|_2^2$ in the update equation of $\mathbf{h}_2$ is relatively weak and leads to large step sizes and possible instability. These rapid adaptations may lead to nonnegligible nonlinear terms, compromising the convergence of the linear filter $\mathbf{w}$. In [18], Stenger *et al.* overcome this problem by adding a constant δ_u to the normalization term which prevents the

$$\sigma_\mathbf{w}(n) = \begin{cases} \beta_{up}\sigma_\mathbf{w}(n) + (1-\beta_{up})\|\overline{\mathbf{w}}(n) - \mathbf{w}(n)\|_1, & \text{if } \|\overline{\mathbf{w}}(n) - \mathbf{w}(n)\|_1 \geq \sigma_\mathbf{w}(n) \\ \beta_{dn}\sigma_\mathbf{w}(n) + (1-\beta_{dn})\|\overline{\mathbf{w}}(n) - \mathbf{w}(n)\|_1, & \text{elsewhere} \end{cases} \qquad (46)$$

denominator from being too small. We propose here to use the normalization term $\left\|\mathbf{U}_2^T(n)\,\mathbf{w}(n)\right\|_2^2$ as the detector parameter because it corresponds to the delayed squared auto-correlation norm of the loudspeaker signal. In order to be independent of $\mathbf{w}(n)$, this moment is normalized by the squared norm of $\mathbf{w}(n)$. Finally, let $\gamma(n)$ be given by

$$\gamma(n) = \frac{\left\|\mathbf{U}_2^T(n)\,\mathbf{w}(n)\right\|_2^2}{\|\mathbf{w}(n)\|_2^2}. \tag{47}$$

$\gamma(n)$ is compared to a fixed threshold determined as a function of the range of the signal $x(n)$. This parameter detects the large delayed amplitudes of the input signal and, at the same time, prevents large step sizes and incorrect adaptations just before the beginning of speech periods.

F. Complexity Reduction

As explained in the introduction, the complexity of Volterra filters is of key importance, especially for real-time implementations. For this type of algorithm, it may be evaluated in terms of multiplications per sample (MPS), since each addition is always combined with a multiplication (an instruction multiply/accumulate can be processed in one DSP cycle). The algorithm may be divided into two predominant tasks: the computation of the nonlinear vector $\mathbf{x}_{nl}(n)$ and the update of the nonlinear filters $\mathbf{h}_2(n)$ and $\mathbf{h}_3(n)$, each task being processed for every sample.

The computation of each sample of $\mathbf{x}_{nl}(n)$ requires $3L_3+L_2$ MPS. Nevertheless, let us point out that, due to small step sizes μ_2 and μ_3, the nonlinear filters $\mathbf{h}_2(n)$ and $\mathbf{h}_3(n)$ do not vary too much over N samples. Thus, we only compute sample $x_{nl}(n)$ with the recent filters $\mathbf{h}_2(n-1)$ and $\mathbf{h}_3(n-1)$, while the old samples $x_{nl}(n-k)\,(0<k<N)$ of $\mathbf{x}_{nl}(n)$ are not computed again; $(N-1)(3L_3+L_2)$ MPS may be saved.

Concerning the adaptation, the main load is expressed in the computation of the convolutions $\mathbf{U}_2^T(n)\,\mathbf{w}(n)$ and $\mathbf{U}_3^T(n)\,\mathbf{w}(n)$. These require approximately NL_3 MPS. If we consider typical memories like $N=128$ for the linear adaptive filter, and $L=10$ for the nonlinear filters (*i.e.*, $L_3=220$), this may lead to a huge computational load, not compatible with real-time constraints. This load may be drastically reduced by observing that most of the information in the convolution stands around the main energetic part of the linear filter $\mathbf{w}(n)$, corresponding to the delay of the acoustic path. Thus, the convolutions $\mathbf{U}_i^T(n)\,\mathbf{w}(n)$, $i\in\{2,3\}$, may be evaluated using a truncated section of $\mathbf{w}(n)$ containing its most relevant coefficients. Simulations have shown that using roughly ten coefficients does not degrade the performance of the algorithm. Using these values, the nonlinear filter computation load reduces approximately to $13L_3$ which corresponds, with $L_3=220$, to the load of a linear NLMS filter which has a length of 1300. This complexity is acceptable for real-time implementation.

IV. Simulations

This section validates the algorithm on data recorded with real material. The example consists of a sentence pronounced twice, at the sampling frequency of 8 kHz, represented on Fig. 6. This signal drives the loudspeaker at two different volumes, maximal and maximal minus 15 dB. An advance of four coefficients at the microphone is artificially introduced in the second sentence. This allows us to have perfect knowledge of the position of the acoustic echo impulse response change. For the linear part of the system, the parameters are $N=128$ and $\mu_1=0.5$.

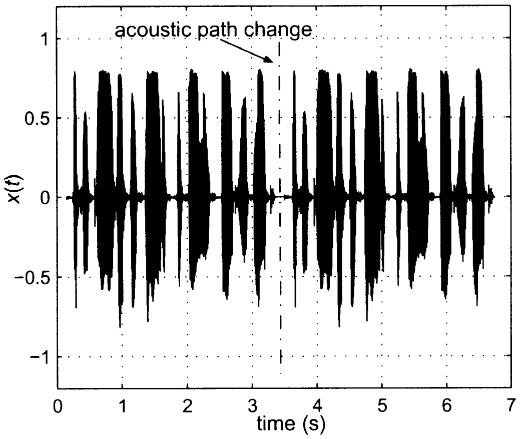


Fig. 6. Loudspeaker speech signal $x(t)$: a 4 coefficients delay is introduced in the acoustic path at approximatively $t = 3.5$ s.

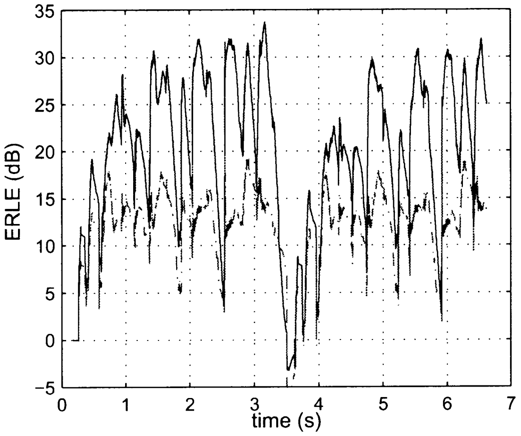


Fig. 7. ERLE for maximal volume minus 15 dB (solid line) versus maximal volume (dash-dotted line).

Fig. 7 shows the Echo Return Loss Enhancement (ERLE) resulting from NLMS linear filtering for the two different volumes. First, note that the impact of the nonlinearities on the loudspeaker signal is not audible: there is no noticeable distortion. Nevertheless, the gap in performance is obvious: the mean difference in ERLE is about 12 dB with some peaks around 15 dB for the voiced segments of speech. This shows the drastic influence of the loudspeaker nonlinear behavior on the achievable performance of linear NLMS.

Now consider the nonlinear part of the system. The choice of L is the center point and is determined by the linear and nonlinear characteristics of the loudspeaker. Identifications obtained using the described algorithm for different values of L showed that the coefficients corresponding to delays of more than 10 samples were negligible and had no influence on the results. Therefore, the length L was fixed to $L=10$. This results in 55 coefficients for $\mathbf{h}_2$ and 220 coefficients for $\mathbf{h}_3$. The length of the truncated linear filter used in the adaptation of $\mathbf{h}_2$ and $\mathbf{h}_3$ is equal to 20.

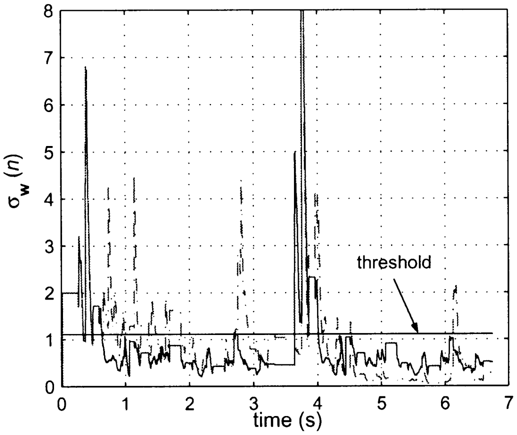


Fig. 8. Behavior of $\sigma_{\mathbf{w}}(n)$ for two different values for the couple (μ_2, μ_3): (0.05, 0.05) (solid line) versus (0.5, 0.5) (dash-dotted line).

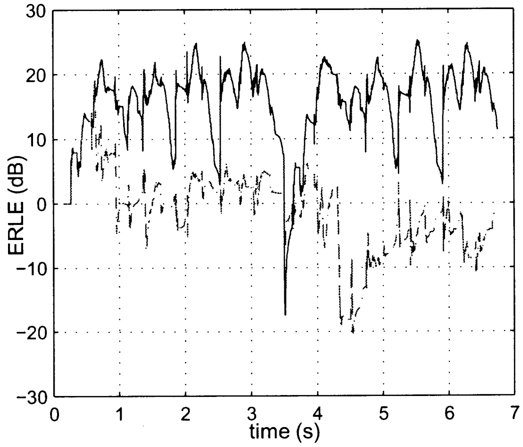


Fig. 9. ERLE for two different values for the couple (μ_2, μ_3): (0.05, 0.05) (solid line) versus (0.5, 0.5) (dash-dotted line).

The threshold used for linear filter steady-state detection is set to $\sigma_{thresh} = 1.1$. Although this value is not critical, it should not be too large: a small value prevents adaptation of the nonlinear filters when the echo to noise ratio at the microphone is weak and when the nonlinearities are masked by the surrounding noise.

The values of the fixed step sizes μ_2 and μ_3 are not critical. They can even be relatively small so as to avoid important fluctuations which lead to permanent oscillations of the system. Figs. 8 and 9 display the influence of the couple (μ_2, μ_3) on convergence and the performance of the nonlinear algorithm. For large values, the echo added by the nonlinear filters is no longer negligible and influence the convergence of the linear filter (see Fig. 8), leading to erratic behavior in the system. Thus, the adaptive system does not fulfill its echo cancellation function (Fig. 9). Simulations show that the values $\mu_2 = \mu_3 = 0.05$ are a good compromise between convergence and stability. Note that the relatively small values for μ_2 and μ_3, and the ill-conditioning of $\mathbf{\Gamma}_{\mathbf{x}_2}$ and $\mathbf{\Gamma}_{\mathbf{x}_3}$ (due to high correlations properties of powers of the input signal $x(n)$), result in a very low convergence speed. In [18], Stenger *et al.* propose to circumvent this problem by either pre-whitening the input (assuming speech as a Laplacian distribution) or using an RLS algorithm, leading to robust algorithms with rapid convergence. Nevertheless, the goal of the nonlinear filters in this situation is to identify slowly varying loudspeaker characteristics. Thus, their convergence speed is not of key importance. This is confirmed by the comparable results of adaptive RLS and off-line LS identification in [18].

Fig. 10 displays, for maximal volume, the performance of the linear NLMS algorithm compared to the cascaded and parallel nonlinear ones, using the previous parameters values and the delayed nonlinearity detector. For the parallel algorithm, we used a memory of 40 coefficients for the nonlinear filters, resulting in 820 coefficients for the second-order filter, and 22 140 for the third order. The filters are updated using the NLMS algorithm. The steady-state detector $\sigma_{\mathbf{w}}(n)$ is depicted on Fig. 8. Regarding the mean performance (during steady-state phases), the ERLE of the cascaded filter reaches 17 dB, compared to 15 dB for the parallel one and 12 dB for the linear case, the improvement being higher for speech peaks. Surprisingly, the cascaded filter is more efficient than the parallel one. The explanation stands in the number of coefficients: the parallel has to cope with a very huge number of coefficients (22 000!) whose variance may produce some additional noise. When considering transient behavior, the difference in the performance of the filters is even more defined. At the beginning of the adaptation $(t = 0)$, all filters are initialized to zero. Firstly, the ERLE of the nonlinear filters are comparable to that obtained with the linear algorithm: at this stage, the ERLE is mainly due to suppression of the linear echo, due to its predominance compared to the nonlinear echo. It is important to note that, for the parallel nonlinear algorithm, the linear filter adapts more quickly than the nonlinear ones (which are allowed to adapt all the time), explaining the equivalent results. Secondly, once the linear filter has converged, the cascaded Volterra filters are allowed to adapt and the ERLE is higher than that of the linear case. Let us note that the nonlinear parallel algorithm converges much more slowly than the cascaded one, due to the number of coefficients. The situation at the beginning of the second sentence (in the input signal) is much more instructive: the acoustic path has changed (about $t = 3.5$ seconds). Until $t = 3.8$ seconds approximately, the ERLE is still the same for the 3 algorithms, since the echo is for the most part linear. After $t = 4$ seconds however, the difference between cascaded and parallel algorithms becomes obvious: the latter has to adapt since the nonlinear impulse response has changed with the acoustic path, whereas the former has already adapted. This is shown by the ERLE curve of the cascaded structure which immediately reaches its maximum (see performance of the first sentence) whereas the parallel algorithm has to re-adapt its nonlinear filters. Indeed, the higher-order filters of the cascaded structure do not depend on the acoustic path.

Finally, Fig. 11 shows the impact of nonlinearity detection on algorithm performance. In the present situation, the acoustic path introduces a delay of about 20 coefficients. As expected, the system based only on the energy of the loudspeaker signal $x(n)$ does not converge properly and does not offer a better performance than the linear algorithm, due to the advance of the nonlinearity detection and the resulting high step sizes.

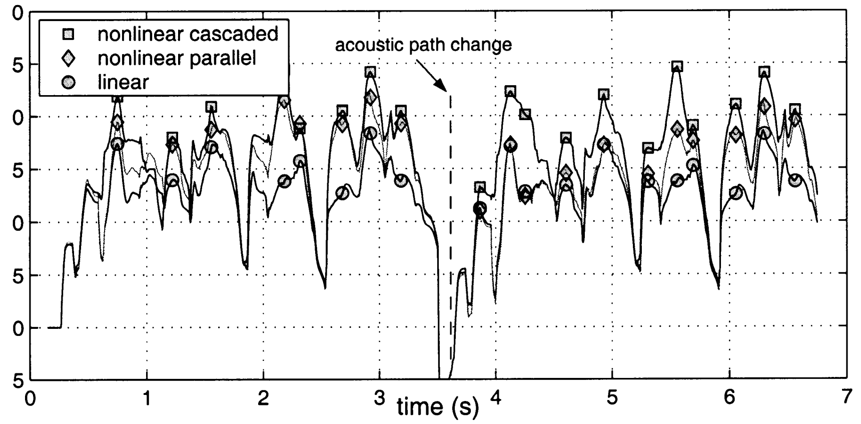


Fig. 10. ERLE at maximal volume for nonlinear cascaded algorithm (square), nonlinear parallel algorithm (diamond-gray line) and linear algorithm (circle).

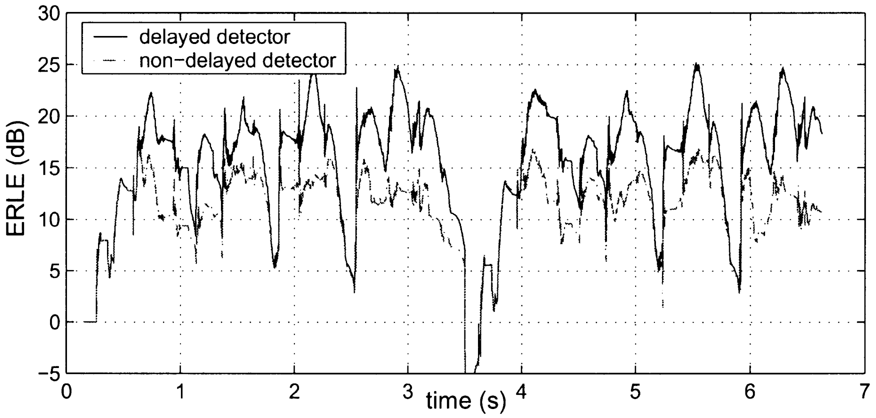


Fig. 11. Comparison of the ERLE for two different methods detecting the nonlinearity: delayed version (solid line) versus nondelayed version (dash-dotted line).

V. Conclusion

This paper has proposed a nonlinear acoustic echo canceler based on a two-module cascaded system: the upstream module, based on second and third-order Volterra filters, identifies the loudspeaker impulse response while the downstream module, consisting of a classical linear filter, identifies the global linear response. Using gradient formalism, the update equations of the linear and nonlinear adaptive filters were derived, as well as the normalization. This leads to the formulation of the NLMS equations. We focused in particular on the problem of convergence inherent in the cascaded systems: a theoretic study was performed to explain the fundamentals of this phenomenon. An adaptation strategy was derived from this study which ensures the convergence of the system while keeping the ability to track the acoustic path changes. The results computed from the recorded signal using the GSM loudspeaker show the advantage in terms of ERLE of our system compared to the standard linear filter (5 dB better) and parallel Volterra filters (2 dB better). They also demonstrate the advantage in terms of convergence speed of the cascaded structure over the parallel one, mainly because the nonlinear filters remain adapted even after an acoustic path change.

Nevertheless, even if the computation load has been lowered (the complexity is lower than that of a parallel structure), it remains relatively high due to the number of coefficients still in use [in $O(L^3)$] and the convolutions involved in the update equations. Further studies are under consideration to reduce the number of coefficients and to optimize computation of the update equations.

Acknowledgment

The authors would like to thank S. Dufossé for providing the material to record the database. They also acknowledge the valuable discussions with I. Bourmeyster, engineer at STMicroelectronics. The authors finally wish to express their gratitude to the anonymous reviewers for their careful reading and help in improving the paper's quality.

References

[1] S. Haykin, "Adaptive filter theory," in *Prentice-Hall Information and System Sciences*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[2] O. Macchi, *Adaptive Processing: The Least Mean Squares Approach With Applications in Transmission*. New York: Wiley, 1995.

[3] S. L. Gay and J. Benesty, "Acoustic signal processing for telecommunication," in *Acoustic Signal Processing for Telecommunication*. Norwell, MA: Kluwer, 2000, sec. 551.

[4] M. Costa, J. Bermudez, and N. Bershad, "Statistical analysis of the LMS algorithm with a zero-memory nonlinearity after the adaptive filter," in *Int. Conf on Acoustics. Speech and Signal Processing*, Phoenix, AZ, 1999.

[5] A. Guérin, G. Faucon, and R. Le Bouquin-Jeannès, "Influence de nonlinéarités d'ordre 2 de volterra sur l'algorithme lms," in *Gretsi*, vol. 2, Toulouse, France, 2001, pp. 293–296.

[6] A. Stenger, W. Kellermann, and R. Rabenstein, "Adaptation of acoustic echo cancellers incorporating a memoryless nonlinearity," in *Int. Workshop Acoustic Echo and Noise Control*, 1999.

[7] A. Fermo, A. Carini, and G. L. Sicuranza, "Simplified volterra filters for acoustic echo cancellation in GSM receivers," in *European Signal Processing Conf.*, Tampere, Finland, 2000.

[8] A. Stenger and R. Rahenstein, "Adaptive volterra filters for acoustic echo cancellation," in *Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, vol. 2, Antalya, Turkey, 1999, pp. 679–683.

[9] T. M. Panicker and V. J. Mathews, "Parallel-cascade realizations and approximations of truncated volterra systems," *IEEE Trans. Signal Processing*, vol. 46, pp. 2829–2832, Oct. 1998.

[10] W. Frank, "An efficient approximation to the quadratic volterra filter and its application in real-time loudspeaker linearization," *Signal Process.*, vol. 45, no. 1, pp. 97–113.

[11] M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems*. New York: Krieger, 1989.

[12] W. J. Klippel, "The mirror filter—A new basis for reducing nonlinear distorsions and equalizing response in woofer systems," *J. Audio Eng. Soc.*, vol. 32, no. 9, pp. 675–691, 1992.

[13] W. Klippel, "Identification of nonlinear sm-systems," in *IEEE Workshop on Nonlinear Signal and Image Processing*, Neos Marmaras, Greece, pp. 230–233.

[14] ——, "Adaptive inverse control of weakly nonlinear systems," in *Proc. ICASSP*, vol. 1, Munich, Germany, 1997, pp. 355–358.

[15] W. Frank, "On the compensation of nonlinear distortions," in *Proc. Signal and Image Processing.*, Orlando, FL, 1996, pp. 195–197.

[16] Y. Nomura and Y. Kajikawa, "An elimination method of the nonlinear distortion in frequency domain by the volterra filter," in *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, 1997.

[17] C. Breining, P. Dreiseitel, E. lIansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control—An application of very-high-order adaptive filters," *IEEE Signal Processing Mag.*, vol. 16, no. 4, pp. 42–69, 1999.

[18] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Process.*, vol. 80, no. 9, pp. 1747–1760.

[19] N. J. Bershad, S. Bouchired, and F. Castanie, "Stochastic analysis of adaptive gradient identification of wiener-hammerstein systems for Gaussian inputs," *IEEE Trans. Signal Processing*, vol. 48, pp. 557–560, Feb. 2000.

[20] B. Nollet and D. Jones, "Nonlinear echo cancellation for hands-free spcakerphones," in *Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, 1997.

Alexandre Guérin was born in Toulouse, France, in 1971. He received the B.S. degree in electrical engineering from the Ecole Nationale Supérieure des Télécommunications de Bretagne, France, in 1995, and the Ph.D. degree from the University of Rennes, France, in 2002.

From 1997 to 2001, he was with Alcatel Mobile Phones, where he was involved in the development and study of speech enhancement algorithms for GSM hands-free systems. His research activities concerned two-sensor noise reduction dedicated to car-kit systems and adaptive filtering applied to nonlinear acoustic echo cancellation. He has been Associate Professor with the Laboratory of Signal and Image Processing, University of Rennes 1, since September 2002. His research interests are in the area of biomedical engineering, more particularly on the auditory cortex modeling through the analysis of stereoelectroencephalographic signals and auditory evoked potentials.

Gérard Faucon received the Ph.D. degree in signal processing from the University of Rennes, France, in 1975.

He is Professor at University of Rennes and is Member of the Laboratory of Signal and Image Processing. He worked on adaptive filtering, speech and near-end speech detection, noise reduction, and acoustic echo cancellation for hands-free telecommunications. His research interests are analysis of stereo-electroencephalography signals and auditory evoked potentials.

Régine Le Bouquin-Jeannès was born in 1965. She received the Ph.D. degree in Signal Processing and Telecommunications from the University of Rennes 1, France, in 1991.

Her research focused on speech enhancement for hands-free telecommunications (noise reduction and acoustic echo cancellation) until 2002. She is currently Associate Professor in the Laboratory of Signal and Image Processing, University of Rennes 1, and her research activities are essentially centered on biomedical signals processing and more particularly on human auditory cortex modeling through the analysis of auditory evoked potentials recorded on depth electrodes.