Combined Noise and Echo Reduction in Hands-Free Systems: A Survey

Régine Le Bouquin Jeannès, Pascal Scalart, Gérard Faucon, and Christophe Beaugeant

***Abstract*—The modern telecommunications field is concerned with freedom and, in this context, hands-free systems offer subscribers the possibility of talking more naturally, without using a handset. This new type of use leads to new problems which were negligible in traditional telephony, namely the superposition of noise and echo on the speech signal. To solve these problems and provide a quality that is sufficient for telecommunications, combined reduction of these disturbances is required. This paper presents a summary of the solutions retained for this dual reduction in the context of mono-channel and two-channel sound pick-ups.**

***Index Terms*—Combined noise and echo reduction, echo cancellation, hands-free applications, noise reduction.**

I. Introduction

RECENT developments in telecommunications and more particularly in mobile communications makes the problems inherent to sound pick-up important. The many papers that have dealt with noise reduction and/or echo control over the last ten years bear witness to the amount of scientific activity surrounding the problem of improving the quality of speech signals, which remain the principal telecommunications commodity. The large majority of papers, however, only consider one of these problems, either echo cancellation or noise reduction. More recent studies attempt to combine these two approaches in order to propose solutions combining noise reduction and echo cancellation. This paper offers a bibliographical synthesis of these solutions.

As a result of their practical aspects and the reduction in user constraints, hands-free sound pick-up systems have become the norm in a number of telephone applications. Amongst these we can cite teleconferencing, audioconferencing on telephone sets, a range of hands-free applications for multimedia services, and hands-free sets for mobile radiotelephony services, particularly for use in vehicles. All these systems offer sound pick-up and sound reproduction away from the user, which means that the speaker is no longer restricted by having to hold the telephone in his hand.

The use of these hands-free sound pick-up terminals does indeed render the "traditional" telephone set obsolete, but also gives rise to new problems previously negligible because of the close sound pick-up inherent to handsets: reverberation and the effect of noise and echo. Because of its importance, only noise and echo reduction will be studied here.

The term ambient noise designates all of the sound waves, apart from those emitted by the speaker and by the hands-free system loudspeaker, which are superimposed on the useful signal to be transmitted.

The term "echo" represents the transmission of the signal back to the transmitter. It results from various couplings, that is various interactions between two physical phenomena all along the speech transmission chain. Three types of coupling are usually distinguished (Fig. 1): electrical coupling due to interactions in the network (changeover from two wires to four wires), solid-carried coupling generated by the mechanical interactions (vibrations) which may exist between loudspeaker and microphone(s) (propagation of sound within the telephone set), and finally acoustic coupling resulting from acoustic interactions (sound propagation within the immediate room) between loudspeaker and microphone(s). In contrast to the last two couplings, the microphone signal cannot be used to solve the problem of electrical coupling; this problem needs specific solutions. Moreover, noise is not as dominant in the electric line as in a mobile acoustic environment. Therefore, combined echo and noise reduction is not really an issue in line echo cancellation and this case will not be treated here.

The presence of echo and noise can prove annoying to the remote speaker, with fatigue and problems of comprehension because of the noise and the unsettling effect of hearing one's own speech delayed by echo return (delay due to the propagation time across the network). Likewise, the performance of speech recognition systems (multimedia applications and dialogue with an "intelligent" carrier) or those of speech encoders (typical for GSM) placed upstream from a microphone in a hands-free system is reduced by the presence of these disturbances. In fact, in many hands-free telecommunications applications, an improvement in the quality of captured sound is an absolute necessity. This improvement is achieved by a reduction of the two disturbances, echo and noise.

Faced with this double problem, the presence of echo and the presence of noise, historically, the first solution has consisted in dealing with these two disturbances independently. There is thus a great deal of literature published concerning firstly noise

Manuscript received May 18, 2000; revised June 30, 2001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael S. Brandstein.

R. Le Bouquin Jeannès and G. Faucon are with the Laboratoire de Traitement du Signal et de l'Image, Université de Rennes 1, 35042 Rennes Cedex, France (e-mail: regine.le-bouquin-jeannes@univ-rennes1.fr; gerard.faucon@univ-rennes1.fr).

P. Scalart is with the France Télécom R&D, France Telecom, DIH/DIPS, Technopole Anticipa, 22307 Lannion Cedex, France and also with the Laboratoire d'analyse des Systèmes de Traitement de l'Information, Ecole Nationale Supérieure de Sciences Appliquées et de Technologie, Technopôle Anticipa, 22305 Lannion Cedex, France (e-mail: pascal.scalart@rd.francetelecom.com).

C. Beaugeant is with the France Télécom R&D, France Telecom, DIH/DIPS, Technopole Anticipa, 22307 Lannion Cedex, France.

Publisher Item Identifier S 1063-6676(01)09670-5.

1063–6676/01$10.00 © 2001 IEEE

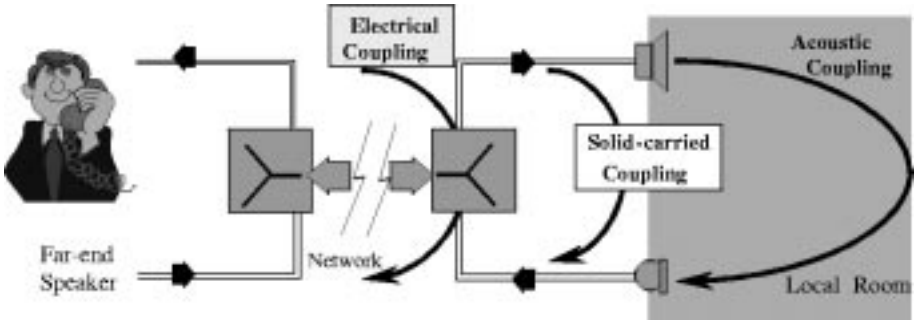

Fig. 1. Coupling effects.

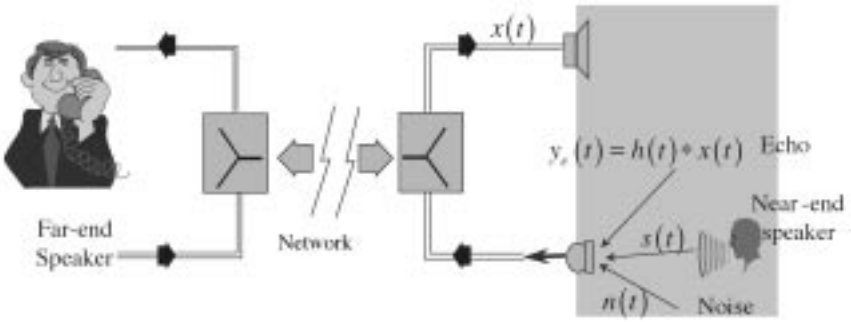

Fig. 2. Notation.

reduction [1]–[3] and, secondly, echo cancellation [4], [5]. More recently, researchers have concentrated on finding a global solution for both problems, noise reduction and echo cancellation, by proposing solutions that aim at reducing both types of disturbance together. After a brief look at the techniques of noise reduction and acoustic echo cancellation, the foundations for combined solutions, we shall concentrate on the papers proposing combined noise and echo reduction solutions.

II. Notation

Throughout this document we maintain the following notational conventions to describe hands-free sound pick-up mathematically. The diagram in Fig. 2 summarizes this notation. The person located "at the other end of the line" will be called the remote speaker (or far-end speaker). The term local speaker (or near-end speaker) will designate the person using the hands-free terminal.

For a mono-channel sound pick-up system, the signals captured by the hands-free system microphone are as follows.

- Speech uttered by the local speaker, henceforth called useful signal and denoted as the time signal $s(t)$, where t is the time variable.
- Echo, denoted as $y_e(t)$, produced by the coupling between the loudspeaker and the microphone in the terminal. In many communication systems, the acoustic channel is characterized by nonlinearities mainly due to the loudspeaker and amplifier. Nevertheless, many papers assume that this acoustic channel is linear. So, the coupling is characterized by the impulse response $h(t)$ between loudspeaker and microphone, so that the echo signal received on the microphone is the result of the convolution between the signal present on the loudspeaker $x(t)$ and $h(t)$, that is

$$y_e(t) = h(t) * x(t). \tag{1}$$

- Noise, denoted as $n(t)$, corresponds to all of the sound sources captured by the microphone apart from the useful signal and the echo. The term disturbance, $d(t)$, designates all of the signals, apart from the useful signal, captured by the microphone: $d(t) = y_e(t) + n(t)$.

Finally, the microphone signal $y(t)$ is written as the sum of the terms previously described

$$\begin{aligned} y(t) &= s(t) + d(t) \\ &= s(t) + y_e(t) + n(t). \end{aligned} \tag{2}$$

The environmental noise is assumed to be independent from speech signals, i.e., the useful signal uttered by the local speaker and the signal emitted by the loudspeaker. Moreover, the useful signal and the echo are mutually independent.

Throughout this paper, signals and physical phenomena satisfy the assumption of short-term stationarity, which constitutes a theoretical concept enabling the optimal filters to be defined with a view toward combined reduction of noise and echo.

In practice, filters in the frequency domain are implemented in accordance with the short-term spectral attenuation principle, described in detail in [6]. The transformation between time domain and frequency domain, and vice versa (analysis/synthesis), is performed by the short-time Fourier transform and inverse short-time Fourier transform, a complete analysis of which can be found in [7]. This technique uses the local stationarity property of speech signals over the length of an analysis frame of

about 20 to 60 ms. Thus, the assumption of short-term stationarity implies a theoretical context which, in practice, is assumed on each analysis frame.

In addition and in general, the following notation is used in this paper.

- Discrete time index is denoted as k.
- For a stationary time signal $u(k)$, the Fourier transform will be denoted $U(f)$.
- Estimation of a value v (whether time or frequency) will be denoted $\hat{v}$.
- Cross-power spectral density between two signals $u(k)$ and $v(k)$ will be denoted $\gamma_{uv}(f)$.
- Notation $E[.]$ will designate the expectation.
- With a two-channel sound pick-up, the notation will be similar to that mentioned previously and will be indexed by 1 and 2. This gives $y_i(k) = s_i(k) + y_{e_i}(k) + n_i(k), i \in \{1, 2\}$ with $y_i(k)$, $s_i(k)$, $y_{e_i}(k)$, and $n_i(k)$, respectively, the microphone signal, the useful signal, the echo, and the noise captured by the microphone i. Index i shall likewise be applied to the notation for the Fourier transforms and to the various estimators. The useful speech signals are coming from the same source and are correlated, while noises are decorrelated as long as the distance between microphones is on the order of few times the wavelength. The echo signals $y_{e_1}(k)$ and $y_{e_2}(k)$ are correlated since they are generated from the same loudspeaker signal.

Two terms that will be used below are the *double-talk* (DT) mode, which corresponds to the simultaneous presence of local speech and echo (local and remote speakers speaking simultaneously), and *single-talk* (ST) mode, which corresponds to the presence of echo only. For both of these modes, we assume that noise is present in the microphone observation(s).

III. Noise and Echo Reduction Techniques

In this section we present a general survey of noise reduction (Section III-A) and echo canceling techniques (Section III-B). These descriptions only explore the solutions that may be used in the methods combining echo and noise reduction described below (Sections IV and V). They do not, therefore, provide an exhaustive review, but simply the concepts required to understand this paper.

A. Principles of Noise Reduction

Single-channel noise reduction is a quite difficult challenge, since the speech and the noise are mixed in the same channel $y(k) = s(k) + n(k)$ [$y_e(k) = 0$ in this section]. Most of these techniques are based on a well-known family of speech enhancement algorithms: *short-time spectral attenuation* algorithms. They attempt to estimate the short-time spectral magnitude of the speech by applying an attenuation to each one of the short-time Fourier transform coefficients of the noisy speech $y(k)$ in relation to the estimation of the signal-to-noise ratio (SNR) for each component. The phase of the noisy speech is not processed, based on the assumption that phase distortion is not perceived by the human ear [8].

These techniques can be classified into three types:

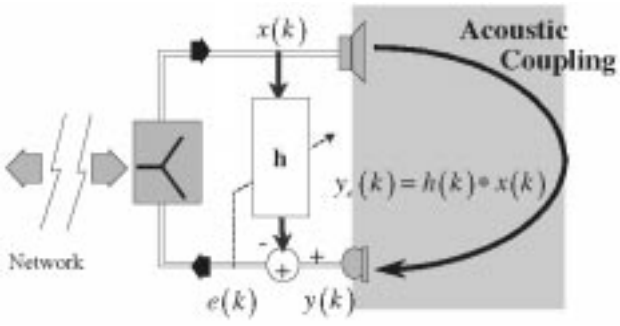

Fig. 3. Principle of echo cancellation by channel identification.

- power spectral subtraction [6] consists in subtracting an estimate of the noise power spectral density from the power spectral density of the microphone signal;
- spectral amplitude subtraction [9], which consists of subtracting from amplitude $Y(f)$ the estimate of the amplitude of the spectral noise component $\hat{N}(f)$;
- direct implementation of the Wiener solution by an open-loop filter of the microphone signal, which consists of minimizing the root-mean-square (RMS) error [2].

However, it has been widely reported that the noise remaining after the processing has a very unnatural disturbing quality, generated by sinusoidal components with random frequencies that come and go in each short-time frame. This artifact is known as the "musical noise" phenomenon. Various modifications of the basic suppression rules have been proposed to overcome this problem: magnitude averaging [9], overestimation of the noise power and introduction of a spectral floor [10], soft-decision noise suppression filtering [11], optimal MMSE estimation of the short-time spectral amplitude [12], a decision- directed approach [12], [13], nonlinear spectral subtraction [14], introduction of masking properties of the human auditory system [15], and morphological-based spectral constraints [3], [16].

B. Principles of Echo Cancellation

The usual techniques for echo cancellation are mainly based on the identification of the acoustic channel $h(k)$ (Fig. 3). This channel is generally modeled by a finite impulse response (FIR) filter with length L, $\mathbf{h}_{\text{opt}}(k)$. This linear modeling can be justified by the fact that the channel is, at first approximation, composed essentially of delay and attenuation. The longer the useful time support of the impulse response, the greater is the length L required for modeling. In practice, this time support may vary from a few dozen milliseconds (vehicle compartment) to several hundred milliseconds (conference room).

The echo cancellation algorithm enables filter $\mathbf{h}_{\text{opt}}(k)$ to be estimated by the L-size vector $\mathbf{h}(k)$ using a criterion based on the *a priori* estimation error. This estimation error, called residual echo, is written, for each sample k

$$e(k) = y(k) - \mathbf{h}^T(k)\mathbf{x}(k) \tag{3}$$

where $\mathbf{x}(k) = [x(k), x(k-1), \ldots, x(k-L+1)]^T$ represents the L last samples of the loudspeaker signal. The filter is updated

at each instant by feedback of the estimation error proportional to the adaptation gain, denoted as $\mathbf{c}(k)$, and according to

$$\mathbf{h}(k+1) = \mathbf{h}(k) - \mathbf{c}(k)e(k). \tag{4}$$

The different echo cancellation algorithms are distinguished by the gain calculation $\mathbf{c}(k)$. These algorithms can be classified as follows.

- Algorithms derived from the gradient (LMS: least mean squares) [17], for which the optimization criterion corresponds to a minimization of the mean-square error. In this case, $\mathbf{c}(k) = \mu\mathbf{x}(k)$, where μ is the step size parameter. Versions in blocks [18] minimize the error criterion on a block of samples. The frequency versions, multidelay filter (MDF) and generalized MDF (GMDF) [19], are the result of using block gradient algorithms in the frequency domain [20].
- Recursive least squares (RLS) algorithms are based on a minimization of the criterion of the least squares with exponential forgetting

$$J[\mathbf{h}(k)] = \sum_{i=0}^{k} \lambda^{k-i} \left[y(i) - \mathbf{h}^T(i)\mathbf{x}(i)\right]^2 \tag{5}$$

  where $\lambda \in [0, 1]$ is a forgetting factor. Fast versions of these algorithms, namely, the fast Kalman [21], the fast *a posteriori* error sequential technique (FAEST) [22], and fast transversal filter (FTF) [23] algorithms, are derived from the RLS by the introduction of forward and backward predictors in the calculation of the "Kalman gain" $\mathbf{c}(k)$. Moreover, the predictable part of the input signal can be extracted with predictors of lower order than the filter size L, leading to a class of Newton-type algorithms known as fast Newton transversal filters (FNTF) [24].
- Affine projection (AP) algorithms [25] are based on a projection that is no longer co-linear with the observation vector $\mathbf{x}(k)$ of the loudspeaker signal, as is the case for LMS-type algorithms, but on a projection that is orthogonal to the intersection of several hyperplanes [defined as all of the vectors $\boldsymbol{\nu}(k)$ such that $\boldsymbol{\nu}(k)^T\mathbf{x}(k) = y(k)$]. A fast version of this algorithm, i.e., the FAP algorithm, relies on a sliding windowed fast RLS algorithm to generate forward and backward prediction vectors and expected prediction error energies [26], [27]. Block algorithms that have exactly the same convergence rate as the original sample-by-sample AP algorithm have also been proposed [28], which have even smaller computational complexity than the FAP algorithm.

C. Echo Control

We have seen that the usual acoustic echo canceling techniques rely on the implementation of a finite impulse response filter of length L. In the asymptotic phase, the coefficients of the adaptive filter generally converge toward those of the Wiener filter, which minimizes the mean value of the filtering error power. Given the cost restrictions imposed by market laws, the number of coefficients L is limited to a value compatible with the characteristics (memory and computation) imposed by the target processor. Consequently, in most applications a residual echo will remain, which may be audible. It is therefore essential to insert a device guarding against this effect in the transmission chain, which leads traditionally to gain variation techniques [29].

The general principle of these techniques consists of determining the active channel (transmission or reception), then applying an attenuation value to the passive channel. Attenuation control is very sensitive since it must respond to a double objective: reducing the residual echo sufficiently while minimizing the effects introduced by local speech and background noise. This function thus plays a critical role within the entire echo canceling device, and dictates, in many cases, the speech quality in the same way as the convergence properties of the adaptive algorithm.

The approaches presented so far propose solutions specific to a given situation, namely noise reduction or echo cancellation. In a situation requiring combined reduction of both disturbances, the approaches retain these solutions as "foundations," but combined reduction leads to a number of considerations, not the least to which is the order in which the two operations must be performed. The following sections present the practical solutions proposed in this field.

IV. Combined Methods for Mono-Channel Pick-Up

This part deals with the techniques developed in situations where only one microphone and loudspeaker are involved. Two situations are considered, one where filtering is applied to both microphone and loudspeaker observations, and one where filtering is applied only to the microphone observation.

A. Filtering Applied to Both Observations

1) Optimal Filter: Let the vector $\mathbf{z}(k)$ be constituted by observations $y(k)$ and $x(k)$

$$\mathbf{z}(k) = [\,y(k) \quad x(k)\,]^T. \tag{6}$$

If we suppose that estimator $\hat{s}(k)$ of $s(k)$ is a linear function of $\mathbf{z}(k)$, the MSE is written

$$E\left[\left|S(f) - \hat{S}(f)\right|^2\right] = E\left[\left|S(f) - \mathbf{W}^T(f)\mathbf{Z}(f)\right|^2\right] \tag{7}$$

where $\mathbf{W}(f)$ is the filter applied to the two observations. Minimizing this error in relation to $\mathbf{W}(f)$ leads to the estimator [30]

$$\hat{S}(f) = \left[\boldsymbol{\Gamma}_{\mathbf{zz}}^{-1}(f)\boldsymbol{\Gamma}_{\mathbf{z}s}(f)\right]^H \mathbf{Z}(f) \tag{8}$$

where

$\boldsymbol{\Gamma}_{\mathbf{zz}}(f)$ power spectral density matrix for the vector $\mathbf{z}(k)$;

$\boldsymbol{\Gamma}_{\mathbf{z}s}(f)$ cross-power spectral density vector between $\mathbf{z}(k)$ and $s(k)$;

superscript H Hermitian (complex conjugate) transpose.

After substitution, (8) becomes

$$\hat{S}(f) = \left[\begin{pmatrix} \gamma_{yy}(f) & \gamma_{yx}(f) \\ \gamma_{xy}(f) & \gamma_{xx}(f) \end{pmatrix}^{-1} \begin{pmatrix} \gamma_{ys}(f) \\ 0 \end{pmatrix}\right]^H \begin{pmatrix} Y(f) \\ X(f) \end{pmatrix}. \tag{9}$$

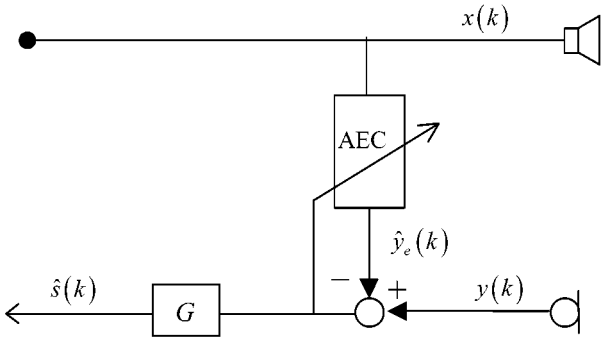

Fig. 4. Cascaded structure derived from optimal filtering.

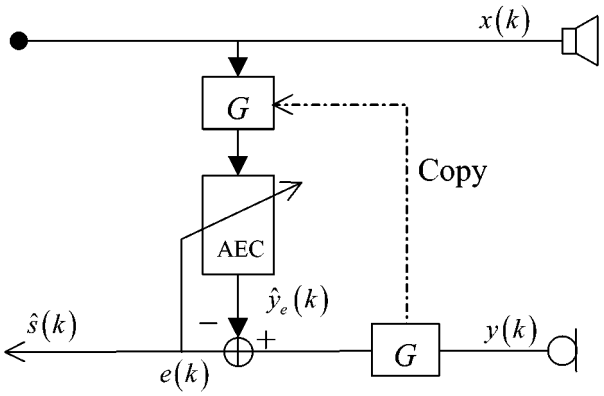

Fig. 5. "Dual" structure of optimal filtering.

After simplification, (9) is reduced to [31]

$$\hat{S}(f) = \left[Y(f) - \frac{\gamma_{yx}(f)}{\gamma_{xx}(f)} X(f)\right] \frac{\gamma_{ss}(f)}{\gamma_{ss}(f) + \gamma_{nn}(f)}. \quad (10)$$

Formula (10) translates the order in which the two operations performed succeed each other. First there is an echo canceling stage which identifies the channel [calculation of the ratio $\gamma_{yx}(f)/\gamma_{xx}(f)$] followed by a noise reduction stage performed through a Wiener filter. We emphasize the fact that the optimal filter is written as an ordered chain of two optimal filters relating to each operation. With an optimal echo canceler, the echo is completely suppressed by the first filter, leaving the useful signal and noise unchanged. The output from the acoustic echo canceler (AEC) is ideally $s(k) + n(k)$. The second stage consists in reducing noise through the Wiener gain filter $\gamma_{ss}(f)/[\gamma_{ss}(f) + \gamma_{nn}(f)]$.

2) Cascaded Structure Derived From Optimal Filtering: One of the first structures appearing in the available literature corresponds, naturally, to the implementation of the optimal filter for which the AEC precedes the noise canceling filter (Fig. 4) [31]–[33]. This structure has been evaluated using different filters. For example, channel identification has been performed using the normalized LMS (NLMS) algorithm and the second-order soft decision AP algorithm (SDAPA2) in [34] as well as the GMDF algorithm in [35]. In these papers, the noise reduction algorithms are often derived from the estimator proposed in [12] and [36].

Implementing a finite-length AEC leads to the presence of a residual echo on its output. The estimator given in (10) therefore cannot be obtained and other structures have been studied.

3) "Dual" Structure of Optimal Filtering: In the structure presented in Fig. 4 it appears that the echo canceling system is disturbed by the continual presence of noise and the intermittent presence of the useful signal. Therefore, in order to minimize the effect of the noise on the AEC, it has been proposed to place a noise canceler, denoted as G, upstream from this system [32], [35] (Fig. 5). If the noise reduction operation enables the signal-to-noise ratio to be improved, it may introduce nonlinear distortion to the echo, which disturbs the identification operation. Copying filter G on the identification branch aims at reducing this potential disturbance [37]. The algorithms used here are those presented in the previous section.

4) Structure With Preprocessing: The structure presented above enables the effect of the noise on the AEC to be reduced. Despite the distortion introduced by filter G, an experimental study [35] has shown that it is preferable to follow this procedure in order to obtain a more precise estimate of the echo. Thus, in the structure shown in Fig. 6, the effect of the noise is initially lessened by the introduction of a preprocessing filter, called G_1. The estimated echo, $\hat{y}_e(k)$, is then subtracted from the observation $y(k)$ to give the value $v(k) = s(k) + n(k) + y_e(k) - \hat{y}_e(k)$. A second noise reduction filter, G_2, is then applied to the signal $v(k)$ to give a final estimate. This is in fact a structure performing echo cancellation followed by noise reduction as in Fig. 4 but including noise reduction preprocessing.

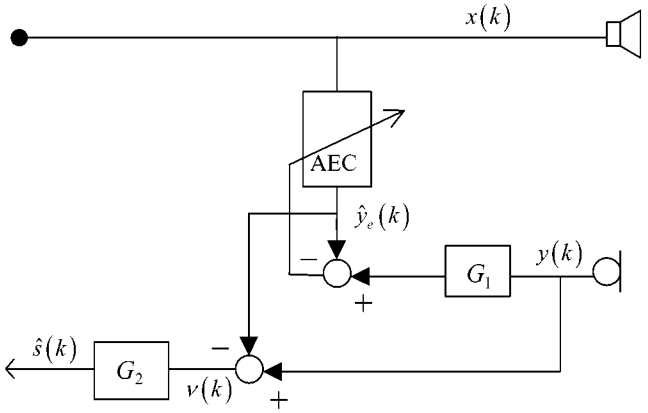

Fig. 6. Cascaded structure with preprocessing.

5) Parallel Structure: This structure follows the organization imposed by the optimal filter. The designation "parallel" given in [31] can be explained by the fact that, to determine the noise reduction filter, the analysis is performed using the microphone observation and not the signal coming from the AEC (Fig. 7). The echo and noise canceling filters are thus estimated using the microphone channel. This structure was proposed in order to reduce the distortion introduced to the useful signal by the noise reducer when it is calculated from the AEC output. In terms of performance, the structure derived from the optimal filter (Fig. 4) leads to good echo cancellation in the ST mode whereas the parallel structure improves the gain of the signal-to-disturbance ratio in the DT mode [38].

6) Improvement of AEC Adaptation: The presence of noise on the echo canceler output disturbs its adaptation. In [39], noise

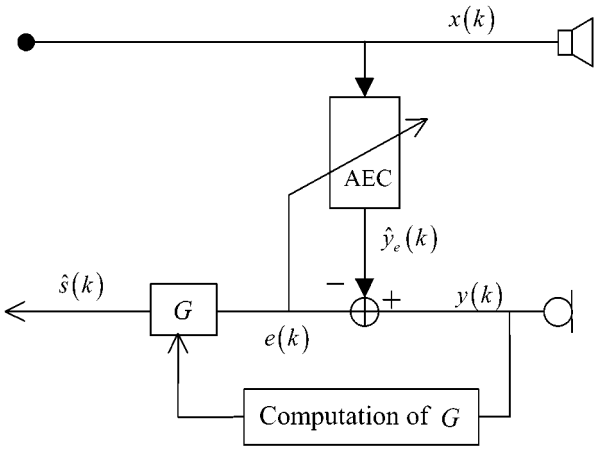

Fig. 7. Parallel structure.

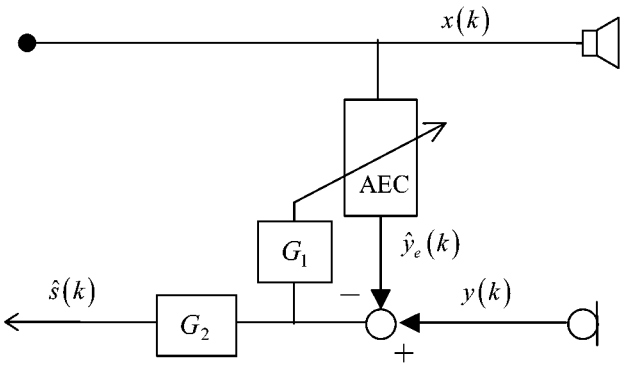

Fig. 8. Noise reduction for AEC adaptation.

reduction symbolized by G_1 is performed on the AEC output to reduce the effect of the noise on the echo canceler adaptation (Fig. 8). A second filter noted G_2 enables a structure comparable to the structure derived from the optimal filter to be obtained. For echo canceling, the multidelay frequency domain algorithm with overlap (MDFO) is used and processes the incoming sample blocks with an overlap of over half the size of the fast Fourier transform (FFT) in order to improve performance and reduce the delay. The nonlinear spectral subtraction (NSS) algorithm is used to reduce noise [14].

B. Filtering Applied to the Microphone Observation Only

Previously, optimal filtering was applied to both observations, microphone and loudspeaker. The solutions presented next lead to the estimation of a filter which, applied only to the microphone signal, globally reduce the components constituting the disturbance. The information coming from the loudspeaker remains useful in calculating the filter.

1) Optimal Filtering: The mean square error in the frequency domain is written

$$E\left[\left|S(f)-\hat{S}(f)\right|^2\right]=E\left[\left|S(f)-W(f)Y(f)\right|^2\right] \tag{11}$$

where $W(f)$ is the filter applied to the microphone observation. Minimizing this error in relation to $W(f)$ gives

$$\hat{S}(f)=\left[\gamma_{yy}^{-1}(f)\quad \gamma_{ys}(f)\right]^{*}Y(f) \tag{12}$$

where $*$ represents complex conjugate.

The filter $W(f)$ can be expressed

$$W(f)=\frac{\gamma_{sy}(f)}{\gamma_{yy}(f)}=\frac{SDR(f)}{1+SDR(f)} \tag{13}$$

where $SDR(f)=\gamma_{ss}(f)/\gamma_{dd}(f)$ denotes the signal-to-disturbance ratio.

Different expressions can be given for the filter which reveal the spectral densities of various signals leading to the expression of $W(f)$ in the form of a single filter or of several cascaded filters.

2) Implementing the Optimal Filter: Various modifications of the basic approach were proposed in [40]

$$W(f)=\frac{1}{1+[SER(f)]^{-1}+[SNR(f)]^{-1}} \tag{14}$$

$$W(f)=\frac{1}{1+[SNR(f)]^{-1}[1+ENR(f)]} \tag{15}$$

where $SER(f)=\gamma_{ss}(f)/\gamma_{y_ey_e}(f)$ is the signal-to-echo ratio; $SNR(f)=\gamma_{ss}(f)/\gamma_{nn}(f)$ is the signal-to-noise ratio; and $ENR(f)=\gamma_{y_ey_e}(f)/\gamma_{nn}(f)$ echo-to-noise ratio.

In contrast, other studies [41] realize $W(f)$ by cascading several filters, one relating to noise reduction, the other to echo reduction. Thus, filter $W(f)$ can be expressed in the form

$$W(f)=\frac{SNER(f)}{1+SNER(f)}\,\frac{SNR(f)}{1+SNR(f)} \tag{16}$$

which can also be written

$$W(f)=[1-MSC_{yx}(f)]\,\frac{SNR(f)}{1+SNR(f)} \tag{17}$$

where $SNER(f)=\gamma_{(s+n)(s+n)}(f)/\gamma_{y_ey_e}(f)$ denotes the signal-plus-noise-to-echo ratio and $MSC_{yx}(f)=|\gamma_{yx}(f)|^2/\gamma_{yy}(f)\gamma_{xx}(f)$ denotes the magnitude-squared coherence between the transmission and reception channels.

The calculation of the filter transfer function can be reduced to the problem of estimating the different ratios between the spectral densities of the various signals. These estimators lead in practice to different behaviors of filter $W(f)$ because of the properties of the estimators $SER(f)$, $SNR(f)$, $ENR(f)$, $SNER(f)$, and $MSC_{yx}(f)$.

In [40] and [41], it is proposed to estimate the power spectral density of the useful signal by using a "decision directed" approach, initially introduced in [12]. The power spectral density of the echo signal is estimated as follows:

$$\gamma_{y_ey_e}(f)=\frac{|\gamma_{yx}(f)|^2}{\gamma_{xx}(f)}. \tag{18}$$

Filtering the microphone observation opens very vast perspectives. Indeed, the traditional estimation of the echo signal by adaptive identification of the channel between the loudspeaker and the microphone supplies information which is too rich in relation to its final use in the implementation of global processing. Indeed, the latter only requires, in this particular case, an estimation of the spectral density of the disturbing signal.

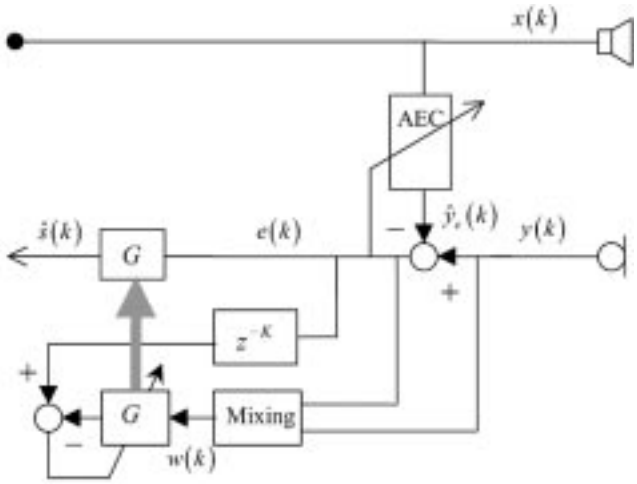

Fig. 9. AEC followed by a closed-loop post-filter.

Such perspectives are interesting as they lead in the long term to suppressing the adaptive echo canceling device, inherent to the traditional solutions presented in Section III-B, and enable the development of future systems at lower cost to be envisaged. This aspect seems particularly important in the global system for mobile (GSM) radio context where cost restrictions play a primordial role in the development of future hands-free terminals and new mobile services.

C. Echo Canceling and Post-Filtering

If we refer to the optimal filter applied to both observations (Section IV-A), we considered the ideal case where the echo was completely suppressed at the end of the first AEC stage. In practice, a residual echo exists and the second stage (initially a noise reduction filter) must be modified in order to reduce both the effect of the noise and that of the residual echo. Several authors have proposed using, for this second stage, the filters described in Section IV-B. Two classes can be distinguished, depending on whether filtering in the second stage is implemented in closed loop or open loop.

1) Estimation of the Closed-Loop Post-Filter: In the approach developed in [42], a post-filter is applied on the AEC output to provide approximately the same echo attenuation as the one provided by the AEC. This second stage is implemented as a closed-loop structure fed by a signal (Fig. 9) generated by a linear combination of the microphone signal and the output $e(k)$ of the AEC

$$w(k) = a(k)y(k) + [1 - a(k)]e(k) \tag{19}$$

where $a(k)$ is an adaptive coefficient in the range [0, 1]. This signal serves as reference for an NLMS adaptive filter whose main channel is the output of the AEC delayed by K samples to obtain the decorrelated noise components. This filter, G, implemented in time domain, is copied in the second filter which is applied at the output of the AEC to obtain the estimated signal. An analysis in the frequency domain was proposed in [43], where it was shown that the optimal post-filter gain is written

$$G(f) = \frac{SNER(f) + \beta[\beta + a(1-\beta)]}{SNER(f) + [\beta + a(1-\beta)]^2} \tag{20}$$

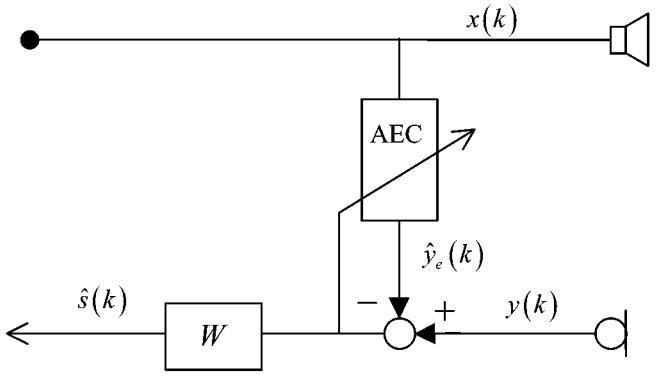

Fig. 10. Echo canceling followed by an open-loop post-filter.

where β represents the scalar attenuation of the echo signal so that

$$y_e(k) - \hat{y}_e(k) = \beta y_e(k) \tag{21}$$

and $SNER(f)$ is the ratio defined in Section IV-B2.

It is thus possible to discover the attenuation provided by the filter in relation to the $SNER(f)$ ratio, for different values in the mixing factor $a(k)$ and the attenuation factor β. The post-filter provides a maximum attenuation for $a(k) = 1$ and for the echo-only situation. In this case, for $SNER(f) \ll \beta$, the attenuation provided by the post-filter is equal to that obtained by the canceler. For $SNER(f) \gg 1$, which corresponds to local speech only, the filter gain is close to 1, which signifies that the post-filter does not modify the useful signal.

In the DT mode, the choice of an intermediary value for the mixing factor, for example $a(k) = 0.3$, enables the post-filter attenuation to be reduced and thus the distortion of the local speech signal is limited. This asymptotic study shows the importance of the choice of the adaptation factor depending on the conditions. An adaptation rule is thus proposed in [42] where the mixing factor depends on measurements of the speech activity of the local speaker and the remote speaker. In practice, the convergence speed and the identification error of the post-filter are determined by its adaptation step. The nonstationary nature and the correlation of speech signals do not enable the maximum attenuation to be attained. Simulations performed in [42] show, nevertheless, the merit of this combined system which enables a reduction of one third the number of AEC coefficients for the same echo attenuation in comparison with the solution developed in Section IV-A2.

2) Estimation of an Open-Loop Post-Filter: In the approach developed by [44]–[46], the second stage is considered as an open-loop post-filter (Fig. 10) implemented in the frequency domain. This filter $W(f)$ is computed according to (13) where the disturbance to be considered is the sum of the noise and the residual echo, in contrast to Figs. 4 and 7 where filter $G(f)$ was only a noise reduction filter.

The performance of such a system depends on the ratio of the residual echo energy to the noise energy, and so, on the considered application. In the case where the residual echo is very energetic at the AEC output in relation to the ambient noise [$|e(k)| \gg |n(k)|$, which is typically the case in teleconferencing], filter $W(f)$ may be considered as a generalization in

TABLE I
MEAN-SCORES OF THE DIFFERENT SOLUTIONS (MONO-CHANNEL PICK-UP) RELATIVE TO THE OPEN-LOOP POST-FILTER (SECTION IV-C2)

Type	Relative mean-score
Cascaded structure derived from optimal filtering (Section IV-A2)	-0.33
"Dual" structure of optimal filtering (Section IV-A3)	-0.25
Structure with preprocessing (Section IV-A4)	-0.15
Filtering applied to the microphone observation (Section IV-B2)	-0.20

the frequency domain of the principle of gain variation set out in Section III-C [44], [47]. For a length L of the coupling impulse response $\mathbf{h}_{\rm opt}(k)$, the AEC/post-filter association takes into account the following analysis:

$$\mathbf{h}_{\rm opt}(k) = \mathbf{h}_{\rm dir}(k) + \mathbf{h}_{\rm res}(k) \tag{22}$$

where $\mathbf{h}_{\rm dir}(k)$ designates the (direct) first N coefficients of $\mathbf{h}_{\rm opt}(k)$ ($N < L$), and $\mathbf{h}_{\rm res}(k)$ the (residual) last ($L - N$). The first stage (AEC) produces an estimate $\hat{\mathbf{h}}_{\rm dir}(k)$ of $\mathbf{h}_{\rm dir}(k)$, and the post-filter $W(f)$ reduces the residual echo $e_{\rm res}(k) = x(k) * \{\mathbf{h}_{\rm res}(k) + [\mathbf{h}_{\rm dir}(k) - \hat{\mathbf{h}}_{\rm dir}(k)]\}$. When $N \ll L$, this solution enables the complexity to be reduced by using a short AEC instead of a single filter of size L trying to identify $\mathbf{h}_{\rm opt}(k)$. In the context of teleconferencing [40], this solution enables the length of the AEC to be reduced by as much as 75%.

When ambient noise is no longer negligible (mainly the case in mobile telephony), filter $W(f)$ in Fig. 10 reduces both noise and residual echo [48]–[50]. This dual reduction is produced using the analysis expounded in Section IV-B. A comparative study of several techniques is presented in [49] and shows that the performance of the AEC/post-filter association is better in comparison with those given by the filters produced by the relations (10) and (13).

D. Evaluation and Complexity

It is difficult to find in the literature comparative studies of the proposed solutions since the evaluation is usually performed on various recordings with different properties (SNR, noise power spectral density …). Nevertheless, it seems judicious to select an algorithm on the basis of its performance and real-time constraints (delay, complexity) for a given application.

Experimental conditions are the following:

- for hands-free telephony in vehicles, noise is recorded in a moving car at different speeds, loudspeaker-to-microphone distance is approximately 1 m (measured in a middle-size car), and a 256 taps echo canceler is used to obtain approximately 40 dB echo return loss;
- for hands-free teleconferencing, the length of the acoustic echo canceler is increased to 512 taps.

Evaluation of speech quality at the processing output is difficult. The subjective impression results from a trade-off between reduction of the disturbances and distortion of the speech signal. First of all, we give some subjective results concerning our own methods. We only conducted a comparison category rating (CCR) test. An absolute category rating (ACR) test would not allow comparison of our results to those found in the literature since the databases are different.

We present test results obtained for hands-free telephony in cars. Our test is made on the same database covering a wide range of realistic situations, including single-talk and double-talk modes as well as more or less noisy environments [51]. This test consists in a comparison between two signals A and B: the degradation scale is composed of seven discrete values in the interval $[-3; +3]$ ($+3$ corresponds to a better quality of the second signal and -3 corresponds to a lower quality), the number of listeners was 16. The open-loop post filter (Section IV-C2) is chosen as the reference structure since informal listening tests indicate good performance. Compared to this structure, Table I gives the relative mean scores of the other methods.

From results (objective and subjective) presented in [51], we can conclude that, for the considered conditions, the structure combining echo canceling and post-filtering described in (Section IV-C2) gives better speech quality than the solutions given in Sections IV-A and IV-B. The double echo cancellation (obtained by adaptive echo canceler and post-filtering) provides an efficient echo cancellation even in a noisy environment, allowing compensation for the convergence difficulties of the adaptive echo canceler. Now, we would like to give some results obtained with the optimal filtering. In low noise level conditions, the structure derived from optimal filtering (Section IV-A2) is equivalent to the structure chosen as reference, i.e., the open-loop filter. When noise level increases, the first structure becomes less attractive than "dual" structure of optimal filtering, structure with preprocessing, and filtering applied to the microphone observation. Its performance also decreases when the signal-to-echo ratio decreases. In this case, the open-loop post-filter performs best, since it allows reduction of the residual echo. Applying optimal filtering only to the microphone observation seems interesting. In noisy conditions, this solution is always better than the cascaded structure (Section IV-A2). In many noise and echo conditions, its performance is equivalent to that of the algorithm given in Section IV-C2 as long as the echo-to-noise ratio is below about 5 dB.

Moreover, the complexity of the different solutions has to be considered. Let the cascaded structure derived from optimal filtering (Section IV-A2) be chosen now as the reference structure, whereby echo cancellation and noise reduction are implemented in the frequency domain. Regarding this reference

TABLE II
RELATIVE COMPLEXITY OF THE DIFFERENT SOLUTIONS (MONO-CHANNEL PICK-UP), (MUL: REAL MULTIPLICATION OPERATOR; ADD: REAL ADDITION OPERATOR)

Type	Algorithmic complexity
Cascaded structure derived from optimal filtering (Section IV-A2)	Cmplx_0 =Cmplx[AEC]+Cmplx[G]+Cmplx[2 FFT,1 IFFT]
"Dual" structure of optimal filtering (Section IV-A3)	$\text{Cmplx}_0 + \left(2 + N_{\text{FFT}}\right)$ Mul
Structure with preprocessing (Section IV-A4)	$\text{Cmplx}_0 + \left(2 + N_{\text{FFT}}\right)$ Mul + $\left(2 + N_{\text{FFT}}\right)$ Add + Cmplx[G]
Parallel structure (Section IV-A5)	Cmplx_0
Improvement of AEC adaptation (Section IV-A6)	$\text{Cmplx}_0 + \left(2 + N_{\text{FFT}}\right)$ Mul + Cmplx[G]
Filtering applied to the microphone observation (Section IV-B)	Cmplx[W] + Cmplx[2 FFT,1 IFFT]
Estimation of an open-loop post-filter (Section IV-C2)	Cmplx_0 - Cmplx[G] + Cmplx[W]

structure, it obvious that the "dual" structure of optimal filtering (Section IV-A3), the structure with preprocessing (Section IV-A4), and the structure with improvement of AEC adaptation (Section IV-A6) are more complex due to the introduction of a second noise reduction filter. The comparison of the algorithmic solutions presented in this paper is difficult because various and specific implementations can be realized for a given structure. Table II gives the complexity of different solutions where the complexity of the cascaded structure derived from optimal filtering is denoted as Cmplx_0.

As an example for a specific implementation choice, let us compare the computational complexity of the two following solutions: the cascaded structure derived from optimal filtering (Section IV-A2) and the filtering applied to the microphone signal [Section IV-B2, (17)]. For the first solution, the acoustic echo cancellation is realized by a generalized multidelay adaptive filter (GMDFα) algorithm with $L = 512$ coefficients as required in teleconferencing applications. With the following parameters: overlapping factor $\alpha = 4$, block size $N = 2^b = 128$, and division of the impulse response into $K = L/N = 4$ segments, this algorithm requires $\alpha[(8K+12)b+2K-13]$ operations for the constrained adaptation procedure [19]. The computational complexity of a 256-point FFT of a real sequence was evaluated assuming a real split-radix algorithm. In addition, the noise reduction filter corresponds to an open-loop Wiener filter implemented in the frequency domain, where the power spectrum of the enhanced speech is estimated with a power spectral subtraction procedure. We assume that the overlap between successive frames is equal to 50% and that the FFT (or inverse FFT) length, N_{FFT}, is equal to twice the block length N of the GMDFα algorithm.

On the other hand, the global filter applied to the microphone signal is implemented in the frequency domain according to relation (17). We assume that the overlap between successive frames is equal to 50% and that the power spectral densities are evaluated on $(1 + N_{\text{FFT}}/2)$ frequency bins in accordance to the Hermitian symmetry of real signals. For this algorithm, we have evaluated the complexity of the optimal filter given by (17), which requires approximately $21(1 + N_{\text{FFT}}/2) + 3$ Cmplx(FFT) to provide $N_{\text{FFT}}/2$ samples at the output [Here, Cmplx(.) denotes the complexity of the operator within the parenthesis]. Noting that the complexity of the cascaded structure (Section IV-A2) and the echo canceling and post-filtering (Section IV-C2) are similar, they are approximately 18 times higher than the complexity of the structure described in Section IV-B2.

V. COMBINED METHODS FOR TWO-CHANNEL PICK-UP

We shall now deal with the case where two microphones and a loudspeaker are used. It is supposed that the signal for estimation is the signal in the first channel, that is $s_1(k)$. As previously for single-channel pick-up, two situations are envisaged: firstly when the filter is applied to the loudspeaker and two microphone observations, and secondly when it is applied only to the two microphone observations.

A. Filter Applied to the Loudspeaker and Two Microphone Observations

Let the vector $\mathbf{z}(k)$ constituted by observations $y_1(k)$, $y_2(k)$, and $x(k)$ denoted as

$$\mathbf{z}(k) = [y_1(k) \quad y_2(k) \quad x(k)]^T. \tag{23}$$

The mean-square error in the frequency domain is written

$$E\left[\left|S_1(f) - \hat{S}_1(f)\right|^2\right] = E\left[\left|S_1(f) - \mathbf{W}^T(f)\mathbf{Z}(f)\right|^2\right]. \tag{24}$$

By minimizing this error in relation to filter $\mathbf{W}(f)$, we obtain the estimator

$$\hat{S}_1(f) = \left[\mathbf{\Gamma}_{\mathbf{zz}}^{-1}(f) \quad \mathbf{\Gamma}_{\mathbf{z}s_1}(f)\right]^H \mathbf{Z}(f). \tag{25}$$

Thus [52]

$$\hat{S}_1(f) = \left[\begin{pmatrix} \gamma_{y_1y_1}(f) & \gamma_{y_1y_2}(f) & \gamma_{y_1x}(f) \\ \gamma_{y_2y_1}(f) & \gamma_{y_2y_2}(f) & \gamma_{y_2x}(f) \\ \gamma_{xy_1}(f) & \gamma_{xy_2}(f) & \gamma_{xx}(f) \end{pmatrix}^{-1} \begin{pmatrix} \gamma_{y_1s_1}(f) \\ \gamma_{y_2s_1}(f) \\ 0 \end{pmatrix}\right]^H \cdot \begin{pmatrix} Y_1(f) \\ Y_2(f) \\ X(f) \end{pmatrix}. \tag{26}$$

Assuming that the noises are decorrelated with the signals as well as each other, (26) becomes

$$\hat{S}_1(f) = \left(Y_1(f) - \frac{\gamma_{y_1x}(f)}{\gamma_{xx}(f)} X(f)\right) \frac{\gamma_{s_1s_1}(f)\gamma_{n_2n_2}(f)}{\Delta(f)} + \left(Y_2(f) - \frac{\gamma_{y_2x}(f)}{\gamma_{xx}(f)} X(f)\right) \frac{\gamma_{s_1s_2}(f)\gamma_{n_1n_1}(f)}{\Delta(f)} \tag{27}$$

where

$$\Delta(f) = \gamma_{s_1s_1}(f)\gamma_{n_2n_2}(f) + \gamma_{s_2s_2}(f)\gamma_{n_1n_1}(f) + \gamma_{n_1n_1}(f)\gamma_{n_2n_2}(f). \tag{28}$$

Equation (27) corresponds to echo cancellation in each channel followed by noise reduction performed by a vectorial Wiener filter. Indeed, the output of each ideal echo canceler only contains the ambient noise and the signal produced by the near-end speaker, echoes being completely deleted. The noise reduction system has as ideal inputs $s_1(k) + n_1(k)$ and $s_2(k) + n_2(k)$, and estimates $s_1(k)$ from these two inputs. The vectorial Wiener filter for noise reduction, $\mathbf{G}(f)$, is given by:

$$\mathbf{G}(f) = \left(\frac{\gamma_{s_1s_1}(f)\gamma_{n_2n_2}(f)}{\Delta(f)}, \frac{\gamma_{s_1s_2}(f)\gamma_{n_1n_1}(f)}{\Delta(f)}\right)^T. \tag{29}$$

As with a single microphone, two operations are distinguished: echo cancellation followed by noise reduction. The structure given in [52] corresponds to the implementation of this optimal filter where the echo cancellation is performed by the GMDF algorithm, whereas noise reduction is obtained using the preprocessing and signal identification (PSI) technique. This involves three stages. The first performs Wiener filtering on each channel, taking into account the uncertainty in the presence of the signal applied to each channel. The second stage consists of identifying the signal present in channel 1 from channel 2 to give another estimate of $s_1(k)$. Finally, in the third stage, the average of the two estimates of $s_1(k)$ is taken and the time signal is obtained by inverse Fourier transform and overlap [53].

As with the single-channel situation, the echo cancellation system is disturbed by the presence of noise. Using the approach developed for the single-channel case, a noise reducer is placed on each of the observations upstream from the preceding structure (Fig. 11) [52]. Despite the distortion introduced by the noise reduction filter, an experimental study shows that the echo is better estimated, as the effect of the noise is reduced by the preprocessing. After echo cancellation on each channel, the vectorial noise reducer is applied to the two new values obtained. In the preprocessing, which performs noise reduction on each observation, a compromise has to be found between distortion and noise reduction; Wiener filtering is calculated taking into account the uncertainty in the presence of the signal, and noise reduction is rendered more or less severe by raising the filter to a given power.

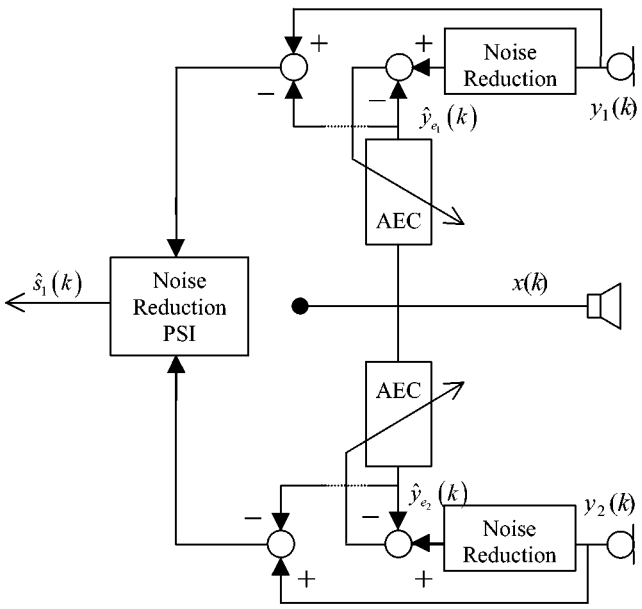

Fig. 11. Two-channel processing with preprocessing.

B. Filtering Applied to the Two Microphone Observations Only

Considering the case where the observations to be filtered are the two microphone channels, [41], we write

$$\mathbf{z}(k) = [y_1(k) \quad y_2(k)]^T. \tag{30}$$

The estimator $\hat{s}_1(k)$ of $s_1(k)$, minimizing the mean-square error in the frequency domain, is given by

$$E\left[\left|S_1(f) - \hat{S}_1(f)\right|^2\right] = E\left[\left|S_1(f) - \mathbf{W}^T(f)\mathbf{Z}(f)\right|^2\right]. \tag{31}$$

The optimal estimator is still given by (25), that is

$$\hat{S}_1(f) = \left[\begin{pmatrix} \gamma_{y_1y_1}(f) & \gamma_{y_1y_2}(f) \\ \gamma_{y_2y_1}(f) & \gamma_{y_2y_2}(f) \end{pmatrix}^{-1} \begin{pmatrix} \gamma_{y_1s_1}(f) \\ \gamma_{y_2s_1}(f) \end{pmatrix}\right]^H \begin{pmatrix} Y_1(f) \\ Y_2(f) \end{pmatrix} \tag{32}$$

or, equivalently

$$\hat{S}_1(f) = \left[\left(Y_1(f) - \frac{\gamma_{y_1y_2}(f)}{\gamma_{y_2y_2}(f)} Y_2(f)\right) \frac{\gamma_{s_1s_1}(f)}{\gamma_{y_1y_1}(f)} - \left(MSC_{y_1y_2}(f)Y_1(f) - \frac{\gamma_{y_1y_2}(f)}{\gamma_{y_2y_2}(f)} Y_2(f)\right) \cdot \frac{\gamma_{s_1y_2}(f)}{\gamma_{y_1y_2}(f)}\right] \frac{1}{1 - MSC_{y_1y_2}(f)} \tag{33}$$

where $MSC_{y_1y_2}(f)$ is the magnitude-squared coherence between $y_1(k)$ and $y_2(k)$.

The philosophy behind this approach is identical to that given for the single-channel case in Section IV-B2. The complexity is decreased even further with two channels.

C. Echo Canceling and Post-Filtering

The concept explained in Section IV-C1 is applied to the two-channel situation in [54] and [55] (Fig. 12). Echo canceling is performed on each channel. Each microphone observation and

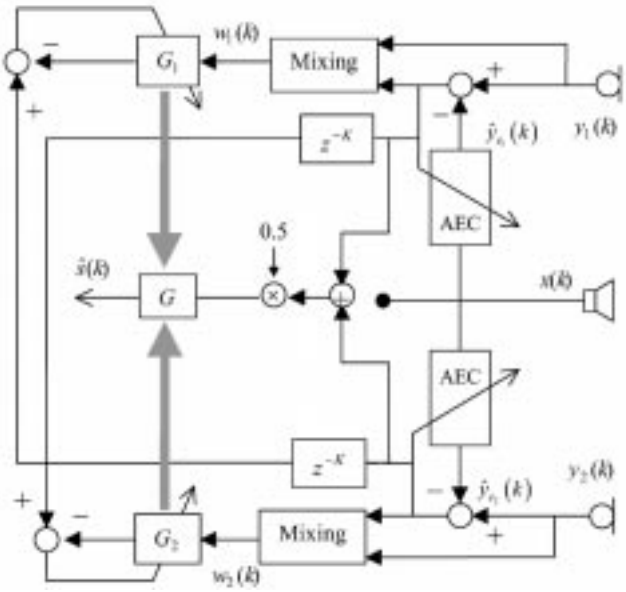

Fig. 12. Two-channel echo canceler and post-filter.

the output of the corresponding echo canceler are mixed as in (19) to supply two signals $w_1(k)$ and $w_2(k)$. Each of these signals serves as a reference for NLMS-type adaptive filtering, with the main channel being the output of the echo canceler from the other channel delayed by K samples. Two filters are thus obtained, G_1 and G_2. A copy of these filters enables a filter $G = (G_1 + G_2)/2$ to be applied to the average of the echo canceler outputs. The main difference in relation to the structure given for the single-channel case is that each filter G_1 and G_2 uses the delayed AEC output in the opposite channel as the desired signal. Local speech is strongly correlated and transmitted by these filters without modification. Time compensation is necessary if the local speaker is not in a symmetrical position in relation to the microphones. This compensation (not shown in Fig. 12) is performed by using a cross-correlation estimator.

D. Comparison

In the two-channel pick-up, it is difficult to give objective results on the speech quality: the signal to be estimated is $s_1(k)$ and every influence of $s_2(k)$ on the final output is not necessarily prejudicious to output signal quality. Subjective tests show the utility of the approach, whereby the processing is applied to microphonic observations, providing performance similar to that obtained when the processing is applied to all observations [41], with lower complexity. As a consequence, we draw the same conclusions as those previously mentioned for the mono-channel case.

VI. Introduction of Psychoacoustic Properties

An understanding of the functioning of the human ear and of the relationship between hearing sensations and acoustic stimuli constitutes one of the keys to the problem of sound pick-up and restoration. In fact, in all systems where the speech signal represents the essential message for transmission, the subscriber's ear constitutes the information receiver. It is a good idea to understand the main psychoacoustic characteristics in order to optimize speech signal processing and generate a true message that is pleasant to listen to. This concern may be evidenced by taking into account restrictions linked to the properties of the human ear in processing noise reduction and echo cancellation.

The domain of psychoacoustics raises the issue of the masking effect, which is the fact that one sound can render another sound partially or completely inaudible. In the frequency domain, the concept corresponds to the notion of simultaneous masking [56]. When the useful signal masks the disturbance(s), the noise reduction and/or echo canceling processing becomes useless. In this case, not performing the processing limits the degradation to the useful signal. Exploiting this property constitutes the essence of the contribution of psychoacoustics to the processing presented above.

An important concept concerns the useful signal masking curve, $Mask(f)$, which enables the determination of the level from which the frequency components of the disturbing signal can be considered to be audible (nonmasked components) or inaudible (masked components). Supposing that the masking curve, $Mask(f)$, is known, the general principle of filtering with psychoacoustic constraints can be expressed as follows:

$$\hat{S}(f) = \begin{cases} Y(f), & \text{if } \gamma_{dd}(f) \leq Mask(f) \\ W(f)Y(f), & \text{otherwise.} \end{cases} \tag{34}$$

This relation expresses simply the fact that it is not necessary to apply the disturbance reduction filter $W(f)$ when the power spectral density of the disturbing signal $\gamma_{dd}(f)$ is lower than the masking threshold $Mask(f)$.

The relevance of this type of procedure is linked to the validity of the model used to estimate the masking threshold. The different methods proposed in the literature to calculate the masking threshold are mainly used for audio coding applications. Some of these methods have been studied in the noise reduction context [15], [53], [57]. In relation to these studies, the noise and echo reduction introduces the original masking property of a speech signal (acoustic echo) by another speech signal (local speech). The calculation of $Mask(f)$ must therefore be modified in relation to traditional techniques, which consider only noise masking by a speech signal [58], [59]. One solution proposed in [60] consists of using a "hybrid" technique which enables the optimization of the masking threshold calculation, particularly at low frequencies where distortion proves more audible.

VII. Conclusion

Noise and echo are phenomena that are inherent in the development of hands-free terminals. This article gives a survey of the research activities conducted on the problem of combined reduction of these two interfering signals for a mono or two-channel type sound pick-up. The algorithmic solutions presented are divided into two main families depending on whether the optimal filter is derived using microphone signals only, or else all of the observation signals (i.e., including loudspeaker channel). The first family of algorithms offers solutions whose arithmetic complexity is greatly reduced, but at the price of distortion being introduced to the speech for transmission.

It remains difficult to prefer one or another of these approaches, as they are very rarely compared in the literature. In reality, the choice must come from a complex compromise between the constraints linked to the terminal's acoustic environment, those linked to cost, and those concerned with speech quality. For applications where speech quality is paramount, it is preferable to look to solutions where the filter is applied to all of the observation signals. In contrast, when cost restrictions take precedence, techniques based on filtering the microphone channels only are more suitable.

Even if current predictions lead to favoring data transmission, there is no doubt that speech messages will continue to play a significant role in the future. Therefore, improving speech quality remains a major preoccupation, and even more so as it constitutes a tool for differentiation in a competitive field such as telecommunications. In order to master this speech quality, several studies have been directed recently toward the search for an optimized solution to a wider problem. In the domain of mobile terminals for cellular networks, this leads for example to optimizing global noise reduction processing and source encoding operations [61], [62], and thus to a point where these two problems are no longer considered as independent.

Improving the quality of speech transmission implies the possibility of measuring this quality. In this domain, there are still few objective measurements and evaluation methodologies, and they are often not very suitable for evaluating speech processing. Thus, the problem engendered by noise and echoes is a real one. Techniques for reducing these disturbances exist, but it is difficult to specify the correspondence between perception and filtering. Further reflection is required to compare techniques effectively, to harmonize the results, and evaluate the speech quality of a system.

REFERENCES

[1] J. S. Lim, *Speech Enhancement*, A. V. Oppenheim, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[2] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*. New York: Wiley-Teubner, 1996.

[3] J. H. L. Hansen and J. R. Deller, "Speech enhancement and quality assessment with applications to robust recognition and coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1995.

[4] E. Hänsler, "The hands-free telephone problem: An annotated bibliography update," *Signal Process.*, vol. 27, pp. 259–271, 1992.

[5] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1991.

[6] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, 1979.

[7] R. E. Crochière and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[8] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, no. 4, pp. 679–681, 1982.

[9] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[10] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1979, pp. 208–211.

[11] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 137–145, 1980.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.

[13] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 345–349, Apr. 1994.

[14] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and projection, for robust speech recognition in cars," *Speech Commun.*, vol. 11, no. 2–3, pp. 215–228, 1992.

[15] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 126–137, Mar. 1999.

[16] J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 598–614, Oct. 1994.

[17] B. Widrow and S. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[18] G. A. Clark, S. K. Mitra, and S. R. Parker, "Block implementation of adaptive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, no. 3, pp. 744–752, 1981.

[19] E. Moulines, O. Ait Amrane, and Y. Grenier, "The generalized multidelay adaptive filter: Structure and convergence analysis," *IEEE Trans. Signal Processing*, vol. 43, pp. 14–28, Jan. 1995.

[20] J. S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 373–376, Feb. 1990.

[21] L. Ljiung, M. Morf, and D. Falconer, "Fast calculations of gains matrices for recursive estimation schemes," *Int. J. Contr.*, vol. 27, pp. 1–19, 1978.

[22] G. Carayannis, D. Manolakis, and N. Kalouptsidis, "A fast sequential algorithm for least-squares filtering and prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, no. 6, pp. 1394–1402, 1983.

[23] J. Cioffi and T. Kailath, "Fast recursive LS transversal filters for adaptive processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 304–337, 1984.

[24] G. V. Moustakides and S. Theodoridis, "Fast Newton transversal filters—A new class of adaptive estimation algorithms," *IEEE Trans. Signal Processing*, vol. 39, pp. 2184–2193, Oct. 1991.

[25] K. Ozeki and T. Umeda, "An adaptive algorithm using an orthogonal projection to an affine subspace and its properties," *Electron. Commun. Jpn.*, vol. 67-A, no. 5, pp. 19–25, 1984.

[26] S. Gay and S. Tavathia, "The fast affine projection algorithm," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1995, pp. 3023–3027.

[27] M. Tanaka, Y. Kaneda, S. Makino, and J. Kojima, "Fast projection algorithm and its step-size control," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1995, pp. 945–948.

[28] M. Tanaka, S. Makino, and J. Kojima, "A block exact fast affine projection algorithm," *IEEE Trans. Signal Processing*, vol. 7, pp. 79–86, Jan. 1999.

[29] A. Gilloire and J. F. Zurcher, "Achieving the control of the acoustic echo in audio terminals," in *Proc. EUSIPCO*, 1988, pp. 491–494.

[30] M. Feder, A. V. Oppenheim, and E. Weinstein, "Maximum likelihood noise cancellation using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, no. 2, pp. 204–216, 1989.

[31] B. Ayad and G. Faucon, "Acoustic echo and noise canceling for hands-free communication systems," in *Proc. Int. Workshop Acoustic Echo Control*, 1995, pp. 91–94.

[32] Y. Guélou, A. Benamar, and P. Scalart, "Analysis of two structures for combined acoustic echo cancellation and noise reduction," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1996, pp. 637–640.

[33] P. Dreiseitel and H. Puder, "A combination of noise reduction and improved echo cancellation," in *Proc. Int. Workshop Acoustic Echo Control*, 1997, pp. 180–183.

[34] P. Scalart and A. Benamar, "A system for speech enhancement in the context of hands-free radiotelephony with combined noise reduction and acoustic echo cancellation," *Speech Commun.*, vol. 20, no. 3–4, pp. 203–214, 1996.

[35] G. Faucon and R. Le Bouquin-Jeannès, "Joint system for acoustic echo cancellation and noise reduction," in *Proc. EUROSPEECH*, 1995, pp. 1525–1528.

[36] A. Akbari, R. Le Bouquin-Jeannès, and G. Faucon, "Speech enhancement using a Wiener filtering under signal presence uncertainty," in *Proc. EUSIPCO*, 1996, pp. 971–974.

[37] A. Benamar, "Etude et implémentation de la fonction de contrôle de l'écho acoustique pour la radiotéléphonie mains-libres," Ph.D. dissertation, Univ. Paris-Sud, France, 1996.

[38] R. Le Bouquin-Jeannès, G. Faucon, and B. Ayad, "How to improve acoustic echo and noise canceling using a single talk detector," *Speech Commun.*, vol. 20, pp. 191–202, 1996.
[39] F. Capman, J. Boudy, and P. Lockwood, "Acoustic echo cancellation and noise reduction in the frequency domain: A global optimization," in *Proc. EUSIPCO*, 1996, pp. 29–32.
[40] C. Beaugeant, V. Turbin, P. Scalart, and A. Gilloire, "New optimal filtering approaches for hands-free telecommunication terminals," *Signal Process.*, vol. 64, pp. 33–47, 1998.
[41] B. Ayad, "Systèmes combinés d'annulation d'écho acoustique et de réduction de bruit pour les terminaux mains-libres," Ph.D. dissertation, Univ. de Rennes 1, France, 1997.
[42] R. Martin and J. Altenhöner, "Coupled adaptive filters for acoustic echo control and noise reduction," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1995, pp. 3043–3046.
[43] R. Martin and S. Gustafsson, "The echo shaping approach to acoustic echo control," *Speech Commun.*, vol. 20, pp. 181–190, 1996.
[44] V. Turbin, "Combinaison du filtrage adaptatif et du filtrage optimal pour réaliser l'annulation de l'écho acoustique dans un contexte de téléconférence," Ph.D. dissertation, Univ. de Rennes 1, France, 1998.
[45] S. Gustafsson and P. Jax, "Combined residual echo and noise reduction: a novel psychoacoustically motivated algorithm," in *Proc. EUSIPCO*, 1998, pp. 961–964.
[46] S. Gustafsson and R. Martin, "Combined acoustic echo control and noise reduction based on residual echo estimation," in *Proc. Int. Workshop Acoustic Echo Control*, 1997, pp. 160–163.
[47] V. Turbin, A. Gilloire, and P. Scalart, "Comparison of three post-filtering algorithms for residual acoustic echo reduction," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1997, pp. 307–310.
[48] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Process.*, vol. 64, pp. 21–32, 1998.
[49] C. Beaugeant and P. Scalart, "Combined systems for noise reduction and echo cancellation," in *Proc. EUSIPCO*, 1998, pp. 957–960.
[50] S. Gustafsson, P. Jax, A. Kamphausen, and P. Vary, "A postfilter for echo and noise reduction avoiding the problem of musical tones," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. # 1281, 1999.
[51] C. Beaugeant, "Réduction de bruit et contrôle d'écho pour les applications radiomobiles," Ph.D. dissertation, Univ. Rennes 1, France, 1999.
[52] R. Le Bouquin-Jeannès, G. Faucon, and B. Ayad, "A two-microphone approach for speech enhancement in hands-free communications," in *Proc. Int. Conf. Commun. Technology*, 1996, pp. 424–427.
[53] A. Akbari, "Rehaussement de la parole en ambiance bruitée. Application aux télécommunications mains-libres," Ph.D. dissertation, Univ. Rennes 1, France, 1995.
[54] R. Martin, "Combined acoustic echo cancellation, spectral echo shaping, and noise reduction," in *Proc. Int. Workshop on Acoustic Echo Control*, 1995, pp. 48–51.
[55] ——, "Design and optimization of a two microphone speech enhancement system," in *Proc. EUROSPEECH*, 1995, pp. 2009–2012.
[56] R. Zwicker and R. Feldkeller, *Das Ohr als Narichtenempfänger ou Psychoacoustique. L'oreille Récepteur d'Information*. Stuttgart, Germany: Hirzler Verlag, 1967. French translation by C. Sorin, Masson, 1981.
[57] D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos, "Speech enhancement using psychoacoustic criteria," in *Proc IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1993, pp. 359–362.
[58] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.
[59] ISO, Int. Recommend. ISO 11 172-3 MPEG audio, London, U.K., 1992.
[60] V. Turbin, A. Gilloire, P. Scalart, and C. Beaugeant, "Using psychoacoustic criteria in acoustic echo cancellation algorithms," in *Proc. Int. Workshop Acoustic Echo Control*, 1997, pp. 53–56.
[61] R. Martin and R. V. Cox, "New speech enhancement techniques for low bit rate speech coding," in *Proc. IEEE Workshop Speech Coding*, 1999, pp. 165–167.
[62] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. # 1761, 1999.

Régine Le Bouquin Jeannès received the Ph.D. degree in signal processing and telecommunications from the University of Rennes 1, France, in 1991, where her research focused on speech enhancement for hands-free telecommunications (noise reduction and acoustic echo cancellation).

She is currently Assistant Professor with the University of Rennes 1 and her research interests include biomedical signals analysis and processing (stereoelectroencephalographic signals and auditory evoked potentials).

Pascal Scalart received the Ph.D. degree from the University of Rennes, France, in 1992.

In 1993, he held a postdoctoral position with Laval University, Quebec, PQ, Canada, engaging in research on digital signal processing for communications, particularly multicarrier modulation systems. Since 1994, he has been with France Télécom R&D, Lannion, France, where he has been involved in research on speech signal processing for multimedia applications. He is currently Assistant Professor with the Ecole Nationale Supérieure de Sciences Appliquées et de Technologie, Lannion, France. His current research interests are in the field of speech enhancement and adaptive filtering techniques for echo cancellation.

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 now analysis of stereoelectroencephalographic signals and auditory evoked potentials.

Christophe Beaugeant was born in Saint-Etienne, France, in 1972. He received the engineering degree from the Institut National des Télécommunications, France, in 1996, and the Ph.D. degree in signal processing and telecommunications from the University of Rennes, France, in 1999.

Since 1999, he has been with Siemens, Munich, Germany, as a Research Engineer in acoustic and speech signal processing for ICM mobile phones. His current research interests include signal processing for hands-free communication in mobile and multimedia teleconferencing contexts.