EURASIP Journal on Applied Signal Processing 2003:11, 1125–1134

A Two-Sensor Noise Reduction System: Applications for Hands-Free Car Kit

Alexandre Guérin
Laboratoire Traitement du Signal et de l'Image, Université de Rennes 1, Bât. 22, 35042 Rennes Cedex, France
Email: alexandre.guerin@univ-rennes1.fr

Régine Le Bouquin-Jeannès
Laboratoire Traitement du Signal et de l'Image, Université de Rennes 1, Bât. 22, 35042 Rennes Cedex, France
Email: regine.le-bouquin-jeannes@univ-rennes1.fr

Gérard Faucon
Laboratoire Traitement du Signal et de l'Image, Université de Rennes 1, Bât. 22, 35042 Rennes Cedex, France
Email: gerard.faucon@univ-rennes1.fr

Received 24 September 2002 and in revised form 27 March 2003

This paper presents a two-microphone speech enhancer designed to remove noise in hands-free car kits. The algorithm, based on the magnitude squared coherence, uses speech correlation and noise decorrelation to separate speech from noise. The remaining correlated noise is reduced using cross-spectral subtraction. Particular attention is focused on the estimation of the different spectral densities (noise and noisy signals power spectral densities) which are critical for the quality of the algorithm. We also propose a continuous noise estimation, avoiding the need of vocal activity detector. Results on recorded signals are provided, showing the superiority of the two-sensor approach to single microphone techniques.

Keywords and phrases: two-sensor noise reduction, hands-free telephony, coherence, cross-spectral subtraction, noise estimation, optimization.

1. INTRODUCTION

Hands-freecommunication has undergone huge developments in the past two decades. This technology is considered to have added value in terms of comfort and security for the users. Unfortunately, it is characterized by strong disturbances, namely, echo and ambient noise, which lead to unacceptable communication conditions for the far-end user. In highly adverse conditions, such as the interior of a running automobile (which is under consideration in this paper), the ambient noise—mainly due to the engine, the contact between the tires and road, and the sound of the blowing wind—may be even more powerful than speech and thus has to be reduced.

Since the 1970s, noise reduction has mainly utilized a one-microphone structure, with or without any hypothesis on the noise/speech distribution [1, 2, 3]. These techniques, which are only based on the signal-to-noise ratio (SNR) estimation, use the speech intermittence and noise stationarity hypothesis. These algorithms, and especially spectral subtraction, thanks to its low-computational load, have been investigated with success. Nevertheless, they lead to a compromise between residual noise and speech distortion, especially in the presence of highly energetic noise.

The presence of additional microphones should increase performance, allowing spatial characteristics to be taken into account and the system to get (partially) rid of some hypotheses like noise stationarity. In counterpart, the performance of the algorithms depends highly on speech and noise characteristics.

Microphone array techniques, based on beamformer algorithms like generalized sidelobe canceller (GSC) or superdirective beamformer, have been developed for car noise reduction. These approaches were revealed to be efficient in enhancing the SNR while ensuring no distortion due to time-varying filtering (like spectral subtraction for instance). Nevertheless, the achievable amount of noise reduction is limited by the noise decorrelation. Thus, additional postfiltering is added to cope with decorrelated characteristics: in [4, 5], the beamformer is combined with a Wiener filter in order to remove decorrelated noise. Under more realistic hypothesis, car noise is considered as diffuse, thus presenting a strong correlation in the lower frequencies. Some authors proposed using a spectral subtraction in the lower-frequency

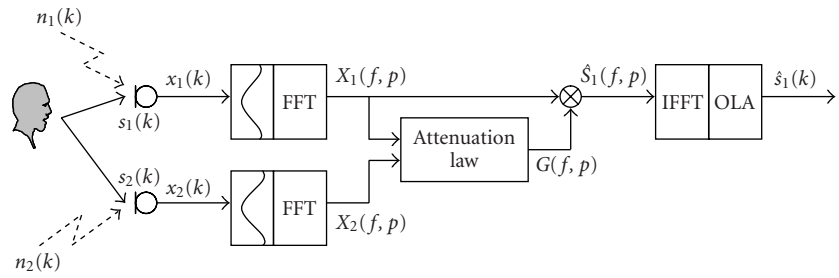

Figure 1: Two-sensor noise reduction system.

bands rather than the Wiener filter [6, 7, 8], or modifying the Wiener filter estimation considering a priori knowledge of the noise spatial statistics [9].

In the GSM context, a two-microphone system, on the contrary to a microphone array, is considered acceptable in terms of cost and ease of installation. The previously described array techniques may be restricted to two-sensor configurations at the expense of reduced performance due to the limited number of microphones. Thus, algorithms specifically dedicated to two-microphone systems have been developed, also depending on signal characteristics. Adaptive noise cancellation has been proposed by Van Compernolle [10], adapted to one point-shaped noise source and linear convolutive mixtures (each microphone picks up noise and speech). A noise reference is formed by linear combination of the two microphone signals, and is then used to remove noise by Wiener filtering. This scheme has recently been adapted to hearing aids with closed microphones [11, 12]. This signal configuration (point-shaped noise sources) is also perfectly suited to source separation under the constraint of less signal sources than sensors [13]. Unfortunately, the speech enhancer usually has to cope with cocktail-party effect (many disturbances with point-shaped sources) and with diffuse noises, which are poorly removed with the previous approach. Maj et al. [14] proposed using generalized singular value decomposition (GSVD) to estimate the Wiener filter. On the contrary to beamformers, this technique is able to remove coherent noise as well as diffuse noise. Though this algorithm provides interesting performance, its huge computational load is not compatible with real-time implementation. In order to reduce the complexity, subband implementation has been investigated, leading to more acceptable complexity, though remaining relatively large [15].

These contributions globally show the advantage of multisensor techniques compared to monosensor. They also demonstrate the difficulty to cope with the real characteristics of signals. The paper, whose concern is a two-sensor noise reduction algorithm, is organized as follows. In the second section, we describe noise and speech signal characteristics. These characteristics then lead into the third section, which discusses a filtering expression based on the coherence function and noise cross-correlation subtraction. We particularly focus on the estimation of the observed signal power spectral densities (psd) as well as those of the noises. Finally, in Section 4, the algorithm is evaluated on real signals and compared to other techniques through objective performance measures.

2. SPATIAL SIGNAL CHARACTERISTICS

Using two microphones, the main question becomes where should we place the microphones inside the car? Indeed, as said in the introduction, the investigated technique depends on their relative position. Obviously, speech has to be picked up as directly as possible to improve the SNR. The position of the second microphone is strictly connected to the noise and speech signal characteristics.

As depicted in Figure 1, we denote by $n_1(k)$ (resp., $n_2(k)$) the noise, by $s_1(k)$ (resp., $s_2(k)$) the speech signal, and by $x_1(k) = n_1(k) + s_1(k)$ (resp., $x_2(k)$) the noisy signals, picked up at the first microphone (resp., at the second microphone). The short-time Fourier transforms (STFT) are denoted by capitals, and indexed by p, the frame number, and f, the frequency (e.g., $N_1(f,p)$ for $n_1(k)$ STFT of pth frame at frequency f). The quantity $G(f,p)$ represents the filtering gain applied to one of the noisy signal in order to remove noise. This gain can be calculated according to the spectral subtraction filter, theEphraim and Malah [3] filter, the coherence, and so forth.

The psd of the noise, speech, and noisy signals are denoted by $\gamma_{n_i}(f)$, $\gamma_{s_i}(f)$, and $\gamma_{x_i}(f)$ on the ith channel ($i = 1, 2$), while $\gamma_{x_1x_2}(f)$ is the observations' cross-power spectral density (cross-psd). The coherence and the magnitude squared coherence (MSC) between the two signals x_1 and x_2 are given respectively by

$$\rho(f) = \frac{\gamma_{x_1x_2}(f)}{\sqrt{\gamma_{x_1}(f)\gamma_{x_2}(f)}}, \qquad \mathrm{MSC}(f) = \left|\rho(f)\right|^2. \tag{1}$$

In a car environment, the signal characteristics are as follows.

(1) Noise is mainly composed of three independent components: the engine, the contact between tires and road, and the wind fluctuations. Their relative importance depends on the car, the road (more or less granular), and the car speed [16]. All these noises can be roughly considered as diffuse. It is well known that the coherence magnitude of diffuse signals is a cardinal

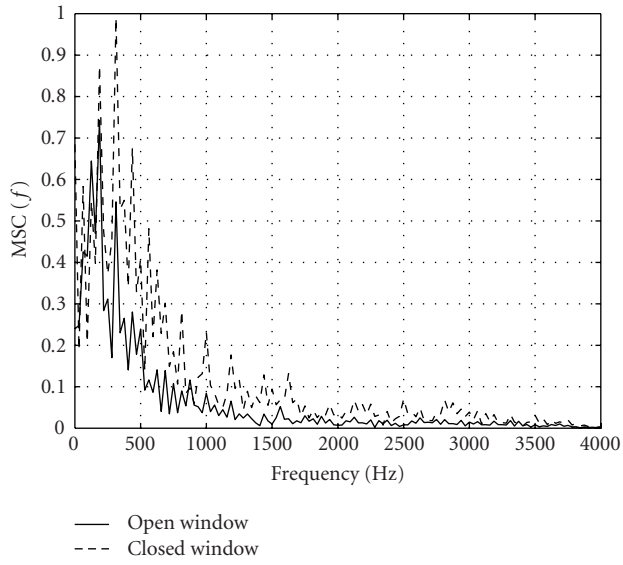

Figure 2: MSC of real car noise signals, with 80-cm spaced microphones, for two conditions: closed (dashed) or open (solid) driver window.

sine modulus function of frequency [17]. This is confirmed by Figure 2, which depicts the MSC of noises corresponding to a car travelling 130 km/h, with either an open or closed driver window. The microphone distance is 80 cm. The MSC profiles show strong correlation in the low part of the spectrum (as predicted by the theory) and decorrelation in the high frequencies. Note that the difference between theoretical and real "cut-off" frequencies is due to noises which are only partially diffuse and also due to microphones characteristics. While the microphones are assumed to be omnidirectional for the theory, they are cardioid in our application.

(2) Speech distribution: speech signals are emitted from a point source. Moreover, the small cockpit size and the interior trim induce no reverberation. Thus, speech signals picked up at different places are highly correlated. A perfect speech correlation is assumed in what follows.

3. TWO-SENSOR ALGORITHM

We first note that it is impossible to create a noise-only reference in the interior of a car. Indeed, speech is strongly reflected in interior car surfaces and is therefore picked up by both microphones wherever they are placed. The main idea is to use the decorrelation of the noises when microphones are sufficiently spaced. With 80 cm-spaced microphones and under diffuse hypothesis, the noises are decorrelated for frequencies above $f = 210$ Hz, that is, above the first minimum of the theoretical MSC function. The lower spectrum, which contains correlated noise, is removed by a bandpass filter in order to respect the telephony requirements (300–3400 Hz). Then, the coherence function is a perfect candidate to operate the filtering of the decorrelated signals and the proposed algorithm is based on it. Indeed, we can show that, under certain hypotheses, the coherence may be equal to the Wiener filter [18]. Hence, applying coherence as a filter to any noisy signal leads to the removal of the decorrelated signals, that is, noise.

Coherence has been widely used in dereverberation techniques. In the car environment, it has been used successfully but with some modifications to cope with low-frequency noise correlation (see [4, 6, 7]). Indeed, in these frequency bands, noises usually exhibit nonnull correlation. Akbari Azirani et al. [18] proposed to estimate noise cross-psd during noise-only periods, and to remove it from the observations' cross-psd during speech activity. The present system is based on this technique named "cross-spectral subtraction." The zero-phase filter $H_{\text{css}}(f, p)$ is given by the following expression:

$$H_{\text{css}}(f) = \frac{|\gamma_{x_1x_2}(f)| - |\gamma_{n_1n_2}(f)|}{\sqrt{\gamma_{x_1}(f)\gamma_{x_2}(f)}}, \tag{2}$$

where $\gamma_{n_1n_2}(f)$ is the noise cross-psd.

The computation of the filter H_{css} needs the estimation of the different psd and cross-psd quantities and is a key point in filtering quality. Concerning spectral subtraction, for instance, many techniques have been developed to remove the well-known problem of musical noise (see [1, 19, 20]). In the MMSE-STSA technique developed by Ephraim and Malah [3], it has been proven that the "decision-directed" approach proposed by the authors to estimate the a priori and a posteriori SNR allows musical noise to be more efficiently controlled [21]. This estimator is still widely used (see, e.g., [18, 22]).

The psd and cross-psd estimation is described in this section. Firstly, we show that the estimation of the observations psd and cross-psd, $\gamma_{x_1}(f)$, $\gamma_{x_2}(f)$, and $\gamma_{x_1x_2}(f)$, should be strictly connected to the signal characteristics, that is, it should respect the long-term noise stationarity and the short-term speech stationarity. This aspect is described in Section 3.1 and the noise cross-psd estimation is considered in Section 3.2. We focus on the noise overestimation and its online estimation, avoiding voice activity detection (VAD).

3.1. Power spectral densities estimation

The noisy signals psd $\gamma_{x_i}(f, p)$ and cross-psd $\gamma_{x_1x_2}(f, p)$ are estimated using a recursive filtering:

$$\begin{aligned}\gamma_{x_i}(f, p) &= \lambda\gamma_{x_i}(f, p-1) \\ &\quad + (1-\lambda)X_i(f, p)X_i^*(f, p), \quad i = 1, 2, \\ \gamma_{x_1x_2}(f, p) &= \lambda\gamma_{x_1x_2}(f, p-1) \\ &\quad + (1-\lambda)X_1(f, p)X_2^*(f, p),\end{aligned} \tag{3}$$

where λ is a forgetting factor usually close to 1. The parameter λ has to cope with two contradictory constraints. On the one hand, the estimation has to respect the short-term speech stationarity, and consequently λ should take low

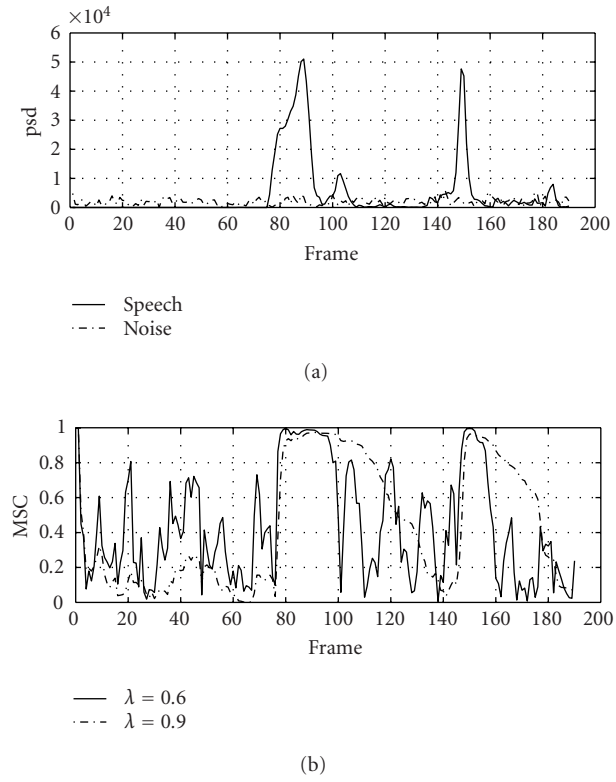

Figure 3: (a) psd of speech (solid) and noise (dash-dotted) in function of the frame index, for $f = 1$ kHz. (b) MSC at $f = 1$ kHz estimated on the observations for two different values, $\lambda = 0.6$ and $\lambda = 0.9$ of the forgetting factor λ.

values; experience shows that for an 8-kHz sampling frequency with 256 sample frames and a 75% overlap, values of λ around 0.6–0.7 are the upper limit. On the other hand, λ has to favor long-term estimation to reduce the estimator variance. The MSC behaviour at 1 kHz is depicted in Figure 3 for two values of λ. The noisy signals used for the MSC computation are composed of correlated speech and decorrelated noise whose psd at 1 kHz, computed for each frame, are displayed at the top figure. For $\lambda = 0.6$, the MSC follows the speech variations, but the estimator variance is high during noise periods. These fluctuations lead to strong filter variations, thus musical noise appears. On the contrary, the variance is highly reduced for $\lambda = 0.9$, but the long-term forgetting factor induces an important reverberant effect especially during speech periods.

Thus, λ has to take small values during speech presence and high values during noise-only periods. To cope with these constraints, we propose the law

$$\lambda(f,p) = 0.98 - 0.3\frac{\mathrm{SNR}(f,p)}{1+\mathrm{SNR}(f,p)}, \tag{4}$$

where $\mathrm{SNR}(f,p)$ is the SNR at the first microphone. The ratio $\mathrm{SNR}(f,p)/(1+\mathrm{SNR}(f,p))$ takes values in the interval $[0,1]$. This type of adaptive coefficient has been proposed by Beaugeant et al. [22] in echo cancellation framework. For low SNR, λ takes high values (close to 0.98), allowing the psd and cross-psd estimations to be smoothed during noise-only periods and thus limits musical noise. On the contrary, for high SNR values, the forgetting factor takes small values (close to 0.68), allowing the estimators to follow the fast speech variations. We propose to approach the ratio $\mathrm{SNR}(f,p)/(1+\mathrm{SNR}(f,p))$ by the previous frame-gain value $H_{\mathrm{css}}(f,p-1)$, assuming that the SNR does not vary too quickly from one frame to another:

$$\frac{\mathrm{SNR}(f,p)}{1+\mathrm{SNR}(f,p)} \simeq H_{\mathrm{css}}(f,p-1). \tag{5}$$

This leads to the following adaptive expression of the forgetting factor λ:

$$\lambda(f,p) = 0.98 - 0.3H_{\mathrm{css}}(f,p-1). \tag{6}$$

The ratio may also be estimated by direct computation of the SNR. Nevertheless, it should not exhibit quick large variations avoiding the rapid fluctuations of $\lambda(f,p)$, thus limiting musical noise. Simulations which we conducted show that we can also use the a priori SNR given by the decision-directed approach [3] (with high time constant). On the contrary, the a posteriori SNR produces overly rapid changes [23].

The proposed law allows the residual noise to be controlled during noise-only periods. Indeed, during speech activity, the adaptive coefficient λ varies quickly with the speech fluctuations, leading to the apparition of musical noise. Although this noise may be partially masked by the speech components, it is still audible and has to be reduced.

3.2. Noise cross-correlation estimation

The musical noise during speech activity is due to two factors:

(1) the long-term estimation of noise cross-psd $\gamma_{n_1n_2}(f)$ during noise-only periods,
(2) the high variance of noise cross-psd included in the term $\gamma_{x_1x_2}(f)$ due to the small forgetting factors.

In addition to its high variance, the short-term estimate $|\gamma_{n_1n_2}(f)|$ also exhibits a mean higher than the long-term one, being more sensitive to instantaneous energetic changes (these ones are less smoothed). Thus, we propose to control musical noise by overestimating the noise cross-psd. First, based on statistical studies, we propose in Section 3.2.1 an overestimation law ensuring the quasiabsence of musical noise. Finally, the noise cross-psd overestimation is achieved in Section 3.2.2 with a novel estimator, giving a long-term estimation without any need for a VAD.

3.2.1. Noise overestimation

Noise overestimation usually consists in multiplying the noise estimate by a constant factor α. For the power spectral subtraction technique, studies show that a 9 dB overestimation factor ($\alpha = 8$) is necessary to remove musical noise [19]; however, this strongly degrades speech. In this section,

we propose to evaluate the overestimation necessary for the cross-correlation spectral subtraction technique, ensuring no musical noise for minimal speech distortion.

To estimate this overestimation, we introduce the cumulative distribution function (cdf) of the short-term noise cross-psd magnitude:

$$F(f,m) = \Pr\left[\,|\gamma_{n_1n_2}(f)| < \mu(f) + m\sigma(f)\right]. \quad (7)$$

In (7), $\mu(f)$ stands for the module of the long-term cross-psd estimate, and $\sigma(f)$ for the short-term cross-psd magnitude standard deviation; the parameter m may take different integer values, $m = 1, 2, 3$. This cdf roughly indicates the probability that the short-term cross-psd module is lower than its long-term estimate plus a positive term depending on its variance. The short-term cross-psd is computed using $\lambda = 0.7$. The cdf curves, computed with real signals, are depicted in Figures 4 (closed window) and 5 (open window). In closed window condition (Figure 4), 95% of the short-term cross-psd are included in the confidence interval $[0; F(f, 2)]$. Note that the profile for $m = 1$ depends highly on the frequency; for $f \leq 500$ Hz, only 80% of the cross-psd are included in the interval $[0; F(f, 1)]$. The explanation is strictly connected to the spatial distribution (diffuse characteristics) but does not come straight forward. Nevertheless, we can conclude that, for closed window condition, $\mu+2\sigma$ is a fairly good overestimation of the short-term noise cross-psd. For an open window, the $F(f, m)$ profiles are similar on the whole spectrum, and the segment $[0; F(f, 1)]$ includes 90% of the short-term cross-psd: $\mu + \sigma$ is a sufficient overestimation ensuring that 90% of the frames do not produce musical noise. The constant profile over the frequency range is due to the noncorrelated characteristics of the noise, whatever the frequency is.

To evaluate the overestimation to be applied, the long-term noise cross-psd module $|\gamma_{n_1n_2}(f)|$ (dashed bottom line) and the $\mu(f) + 2\sigma(f)$ curve (middle solid line) are depicted in Figure 6 (closed window condition). For this condition, the necessary overestimation varies from 2 dB for the low frequencies to 6 dB for the high frequencies. We also displayed the long-term mean psd $\sqrt{\gamma_{n_1}(f)\gamma_{n_2}(f)}$ (top dash-dotted line); this last curve is strictly connected to the $\mu(f) + 2\sigma(f)$ curve. Thus, the long-term estimate $\sqrt{\gamma_{n_1}(f)\gamma_{n_2}(f)}$ is an accurate overestimation of the short-term cross-psd.

The open window condition is considered in Figure 7, with the $\mu(f) + \sigma(f)$ curve (instead of $\mu(f) + 2\sigma(f)$ for closed window), as well as the long-term $|\gamma_{n_1n_2}(f)|$ and $\sqrt{\gamma_{n_1}(f)\gamma_{n_2}(f)}$. The conclusions are exactly the same.

Finally, to limit musical noise, especially during speech periods, we propose to overestimate the noise cross-psd with the mean psd $\sqrt{\gamma_{n_1}(f)\gamma_{n_2}(f)}$. It is important to note that this overestimation does not induce too much speech distortion for the following reasons.

(1) The overestimation is effective for decorrelated noises, that is, especially for high frequencies (see Figures 6 and 7). In this spectrum segment, the SNRs are quite favorable, and the speech components are slightly affected by this overestimation,

(2) In the case of highly correlated noises, that is, for low frequencies, the cross-psd is close to the mean psd. Thus, this slight overestimation for closed-window conditions does not lead to speech distortion (see Figure 6), while the musical noise is controlled. In open window conditions, the overestimation is large (6 dB) because of the noise decorrelation (see Figure 7); more speech distortion is expected.

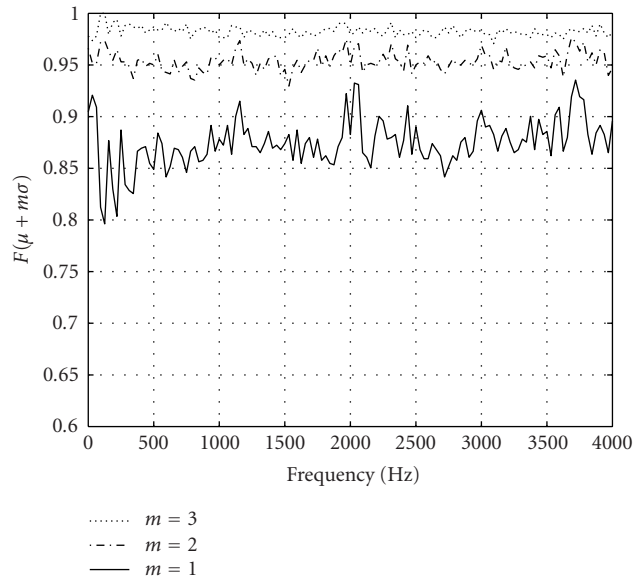

Figure 4: Cumulative distribution functions $F(f, m)$, computed in closed window condition, for three values of m.

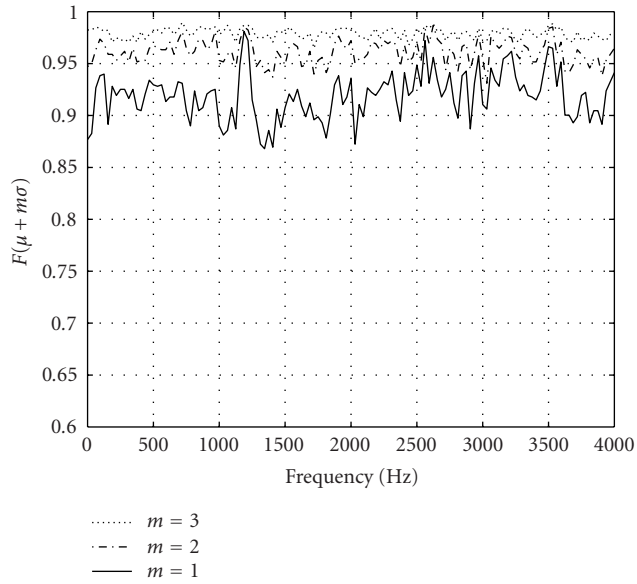

Figure 5: Cumulative distribution functions $F(f, m)$, computed in open window condition, for three values of m.

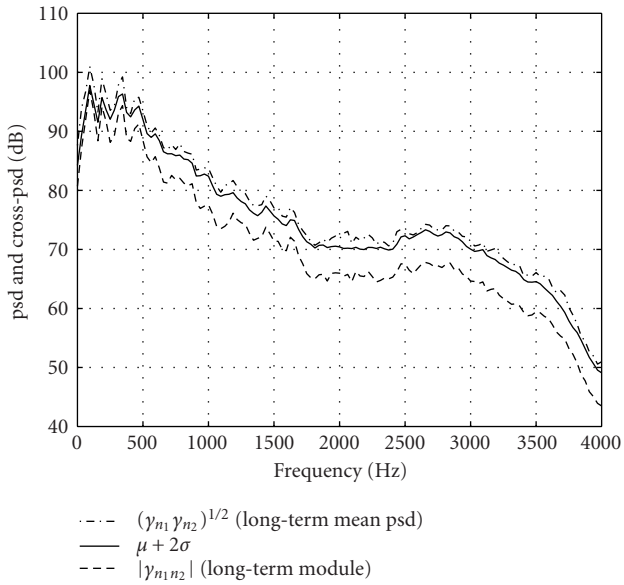

Figure 6: psd and cross-psd module as functions of the frequency in closed window condition: long-term mean psd $\sqrt{\gamma_{n_1}(f)\gamma_{n_2}(f)}$ (dash-dotted), $\mu(f) + 2\sigma(f)$ (solid), and long-term cross-psd module $|\gamma_{n_1 n_2}(f)|$ (dashed).

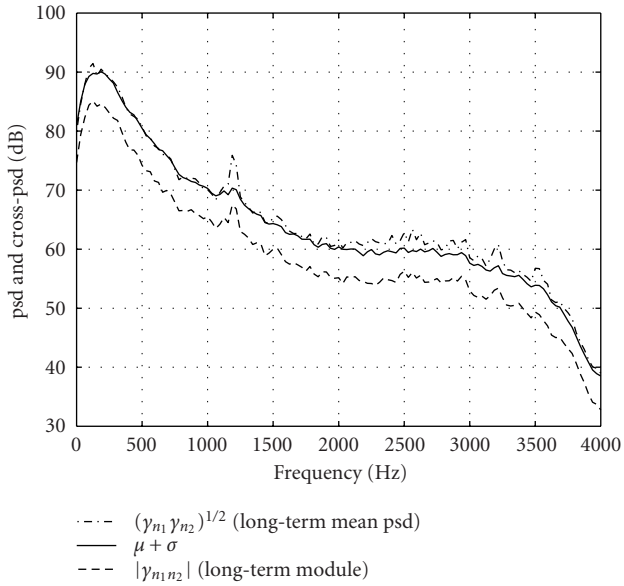

Figure 7: psd and cross-psd module as functions of the frequency in open window condition: long-term mean psd $\sqrt{\gamma_{n_1}(f)\gamma_{n_2}(f)}$ (dash-dotted), $\mu(f)+\sigma(f)$ (solid), and long-term cross-psd module $|\gamma_{n_1 n_2}(f)|$ (dashed).

Experiments on real data show that this overestimation completely removes the musical noise at the cost of a small but acceptable amount of speech distortion.

3.2.2. Continuous noise estimation

Usually, noise psd estimation is achieved during noise-only periods, while being frozen during speech presence. This approach, which is widely used in the literature, needs a robust VAD to help ensure filtering quality. It is especially true for algorithms like spectral subtraction techniques that directly use the noise psd estimation to derive the main signal; a small error in the estimation may lead to musical noise or large amounts of speech distortion. A robust VAD, however, is not as crucial for algorithms using a priori and a posteriori SNR estimation as for the decision-directed approach [3] since the filter estimate also depends on smoothing coefficients.

The cross-spectral technique is strongly affected by the quality of the noise estimate since the filter $H_{\text{css}}(f, p)$ given by (2) depends directly on the noise cross-psd estimate. Experiments show that the filter needs a regularly estimated noise cross-psd to achieve a sufficient denoising with acceptable artefact on speech and noise. In particular, freezing the estimate during a whole sentence is not compatible with the noise stationarity, leading to musical noise emergence. Hence, the VAD has to detect speech pauses or even intersyllabic segments, which may be difficult to achieve with a low-cost stand-alone algorithm. We then propose to use a fuzzy law based on energetic considerations; noise is supposed to be a long-term stationary signal unlike speech. Therefore, a large energy increase between two adjacent frames may be viewed as the presence of speech, whereas small variations only or a decrease in energy corresponds more likely to noise. We propose using the following law, adapted from monosensor algorithm [24]:

$$\begin{aligned}&\sqrt{\gamma_{n_1}(f,p)\gamma_{n_2}(f,p)}\\&\quad= \alpha\Big(\widetilde{\text{SNR}}_{\text{post}}(f,p)\Big)\sqrt{\gamma_{n_1}(f,p-1)\gamma_{n_2}(f,p-1)},\end{aligned} \tag{8}$$

where the function $\alpha(\widetilde{\text{SNR}}_{\text{post}})$ depends on real positive constants b, g, and L:

$$\begin{aligned}\alpha\Big(\widetilde{\text{SNR}}_{\text{post}}\Big) = L + (1-L)\cdot\frac{1}{1+1/(g\cdot\widetilde{\text{SNR}}_{\text{post}})}\\ \cdot\left(1+\frac{1}{1+g\cdot b\cdot\widetilde{\text{SNR}}_{\text{post}}}\right).\end{aligned} \tag{9}$$

The a posteriori modified SNR, $\widetilde{\text{SNR}}_{\text{post}}$, is given by

$$\widetilde{\text{SNR}}_{\text{post}}(f,p) = \frac{|X_1(f,p)X_2(f,p)|}{\sqrt{\gamma_{n_1}(f,p-1)\gamma_{n_2}(f,p-1)}} \tag{10}$$

and takes values in the interval $]0, +\infty[$.

Constants b, g, and L parameterize the $\alpha(\cdot)$ function. Note that, for high values of $\widetilde{\text{SNR}}_{\text{post}}$, indicating an abrupt jump in energy and the emergence of speech, $\alpha(\widetilde{\text{SNR}}_{\text{post}}) \simeq 1$, freezing the noise estimation. The parameter L, comprised in the interval $[0, 1]$, sets the exponential decay of the mean noise psd estimation; for weak values of $\widetilde{\text{SNR}}_{\text{post}}$ (the in-

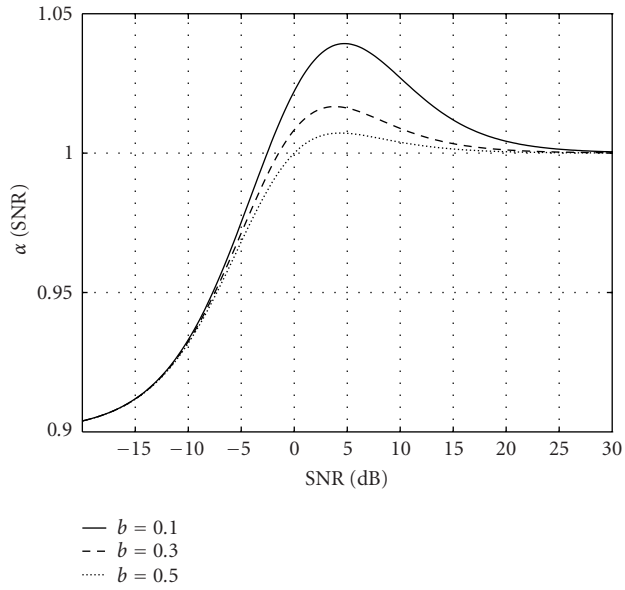

FIGURE 8: Influence of the coefficient b on the α shape, for $g = 2$ and $L = 0.9$.

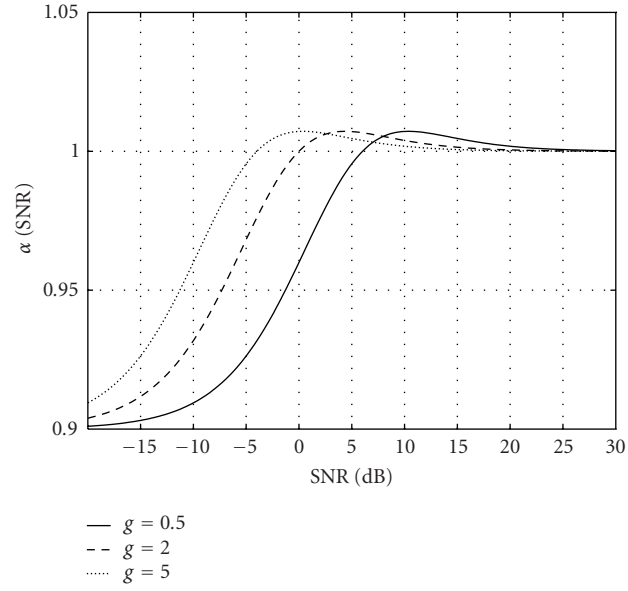

FIGURE 9: Influence of the coefficient g on the α shape, for $b = 0.5$ and $L = 0.9$.

stantaneous amplitudes of the observations are less energetic than those of the previous noise estimate), (9) becomes $\alpha(\widetilde{\mathrm{SNR}}_{\mathrm{post}}) \simeq L$, hence

$$\sqrt{\gamma_{n_1}(f,p)\gamma_{n_2}(f,p)} \simeq L\sqrt{\gamma_{n_1}(f,p-1)\gamma_{n_2}(f,p-1)}. \quad (11)$$

The coefficient b fixes the maximal value reached by α (see Figure 8), while g adjusts this maximum for a given value of $\widetilde{\mathrm{SNR}}_{\mathrm{post}}$ (see Figure 9). Note that L also has an impact on the maximum (the lower L, the higher the maximum). Usually, g is chosen as $g = 1/(1-b)$, fixing the accumulation point $\alpha(1) = 1$; thus, in the case of deterministic noise, the estimator converges towards the true value.

4. SIMULATIONS AND RESULTS

Simulations were conducted on real signals recorded in a driving car. The directional microphones were placed on the left-hand side, upright the windshield and close to the rear view mirror, ensuring a distance of 80 cm. Therefore, the noise decorrelation condition is fulfilled (see Figure 2 for coherence profile). Two different noises are recorded: a quasi-stationary noise, corresponding to a 130 km/h driving car, and a highly nonstationary one at the same speed with open driver window. These two conditions include slow changes in the engine revolution speed caused by accelerations and shifting gears. Artificial files with different SNR from −3 dB to 20 dB were created by adding noise and speech recorded in a quiet environment (stopped car, switched off engine).

The proposed algorithm, called modified cross-spectral subtraction, is also denoted by modified H_{css}. The gain is computed using (2) (as for standard cross-spectral subtraction). The norm of the noise cross-psd $|\gamma_{n_1 n_2}(f)|$ is overestimated by $\sqrt{\gamma_{n_1}(f)\gamma_{n_2}(f)}$, which is computed using (8), (9), and (10). The performances of our algorithm are compared to those of two other techniques, which have been proven to be efficient in those types of environments.

(1) A monosensor technique: the Wiener uncertainty algorithm denoted as WU [25]. The filtering part is achieved by the Wiener filter, with a correcting factor depending on the speech presence probability derived by Ephraim and Malah [3]. Note that this algorithm provides continuous SNR estimation using the decision-directed approach. The noise psd is learned during noise-only periods using a manual VAD.

(2) A two-microphone algorithm: the cross-spectral subtraction denoted by H_{css}. An implementation of this filter is given in [18]. For this algorithm, the noise cross-psd is learned during noise-only periods, then frozen during speech activity using the same manual VAD as the monosensor algorithm. The forgetting factor λ is fixed as 0.7.

In order to compare the performance of the different algorithms, two different measures have been evaluated on processed signals: the cepstral distance (dcep) and the SNR gain which is given by

$$\begin{aligned}&\text{SNR gain (dB)}\\ &\quad = \text{SNR after processing (dB)} - \text{input SNR (dB)}.\end{aligned} \quad (12)$$

The first one evaluates speech distortion while the second shows the noise reduction. These indices are computed on manually segmented speech frames, then averaged on all frames to give a global measure per condition (stationary/nonstationary).

Consider Figures 10 and 11 displaying the results for the quasistationary noise condition. The SNR gain curves

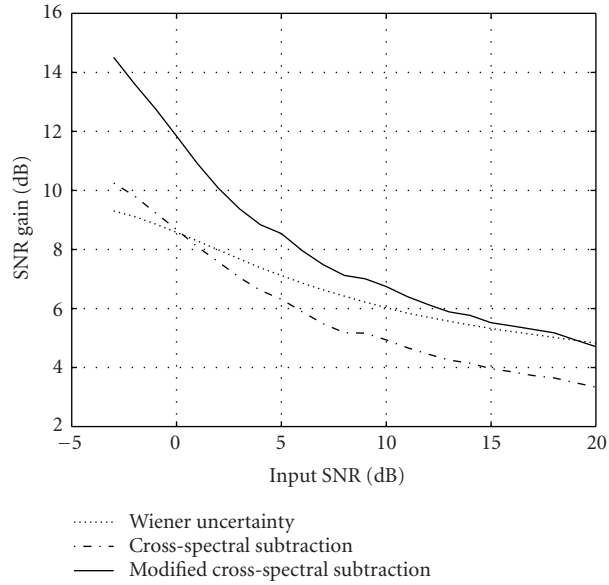

Figure 10: SNR gain as a function of the input SNR for stationary noise condition.

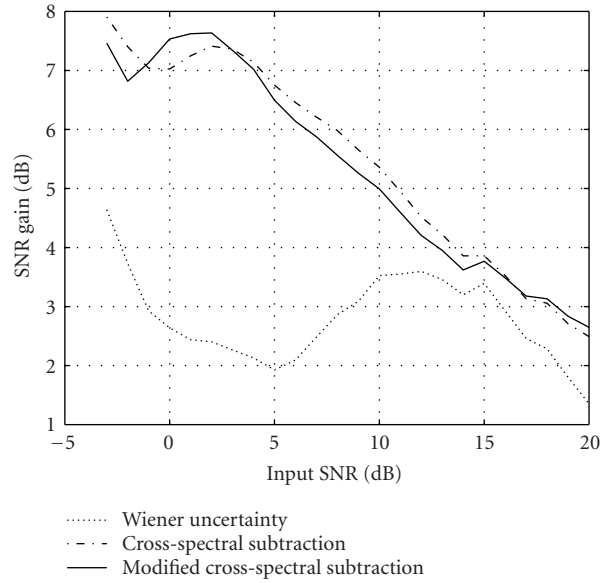

Figure 12: SNR gain as a function of the input SNR for nonstationary noise condition.

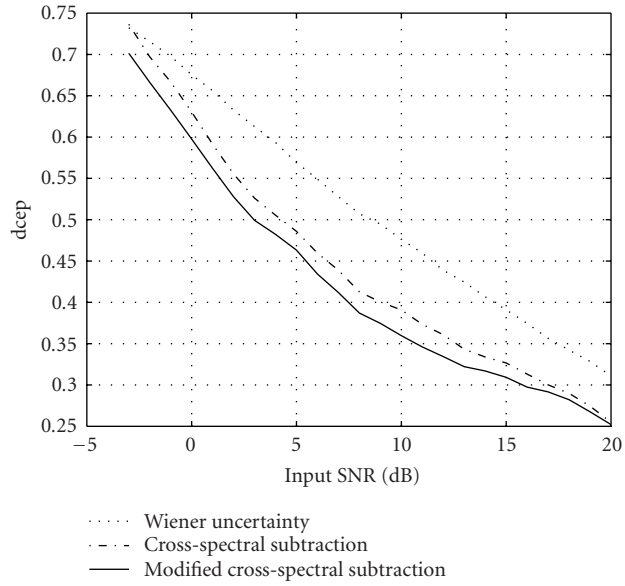

Figure 11: dcep as a function of the input SNR for stationary noise condition.

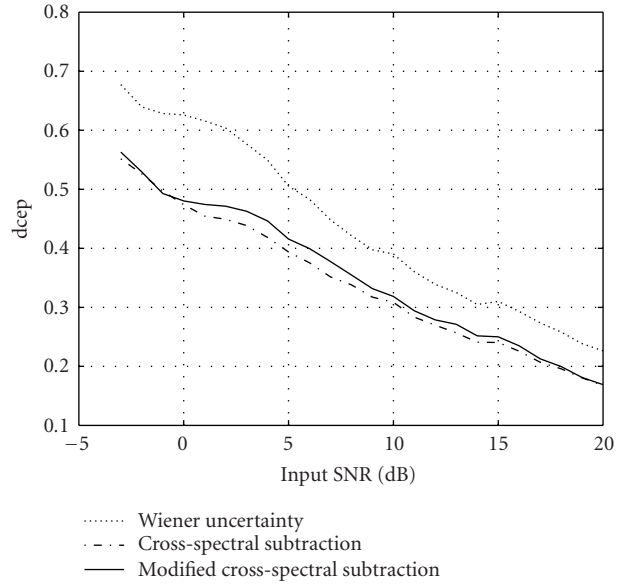

Figure 13: dcep as a function of the input SNR for nonstationary noise condition.

(Figure 10) show the improvement due to noise overestimation and permanent updating; the modified H_{css} performs around 2 to 4 dB better than H_{css}. The monosensor algorithm experiences lower performance than the modified H_{css}, especially for low SNR. In terms of distortion (see Figure 11), the novel technique performs much better than the two others. This result may be explained by the use of the adaptive forgetting factor $\lambda(f, p)$, which prevents overly large smoothing of the psd and cross-psd estimates during speech activity. Note that the monosensor WU algorithm distorts speech much more than the two-microphone techniques, in particular, for high SNR. This confirms the superiority of the modified H_{css} over the WU algorithm for these high SNR despite their equivalent scores in terms of noise reduction.

The results concerning nonstationary noises are depicted in Figures 12 and 13. At a first glance, it is obvious that the two-microphone methods perform much better than the single microphone one in terms of noise reduction as well as speech distortion. It is mainly due to the fact that the two-sensor techniques work particularly well in filtering

these decorrelated noises. Moreover, the fast noise variations prevent the WU from estimating the SNR with accuracy, thus leading to large amounts of speech distortion and residual noise fluctuations. Concerning the two-sensor algorithms, the performance appear to be quite comparable. The reason is that continuous noise updating does not provide any clear advantage; the noise variations, mainly due to the blowing wind, are too rapid to be followed by the estimator. Nevertheless, it should be pointed out that the noise overestimation does not distort the speech signal more than the standard H_{css} filter. Moreover, from a subjective point of view, informal listening tests show that the residual noise appears more natural with the modified H_{css} filter; musical noise and noise level fluctuations, which are audible with standard H_{css} (and monosensor technique), are completely removed. Nevertheless, on very low SNR frames, slight additional speech distortion can be noticed, which is in accordance with the expected behavior of our algorithm. Note also that this distortion is hardly audible due to the energetic noises.

5. CONCLUSION

In this paper, we proposed a two-sensor noise reduction algorithm based on cross-spectral subtraction. The improvement mainly focused on a noise overestimation rule derived from statistical studies, and on spectral densities estimation. With these modifications, simulations showed that the proposed algorithm outperforms proven methods in this environment. With highly nonstationary noises, the new technique is intrinsically better than monosensor ones in terms of speech distortion and noise reduction. In stationary noise conditions, the modified filter outperforms the standard cross-spectral subtraction technique, ensuring much more noise reduction (from 2 to 4 dB) with less speech distortion.

From a computational point of view, this technique is low CPU consuming, about three times the complexity of the spectral subtraction. This allows real-time implementation in GSM mobile phones (e.g., far less CPU consuming than vocoder). The hardware cost caused by the two-microphone approach may be limited by using the terminal microphone, reducing the cost to one additional microphone, like most standard hands-free systems.

REFERENCES

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

[2] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

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

[4] K. U. Simmer, S. Fischer, and A. Wasiljeff, "Suppression of coherent and incoherent noise using a microphone array," *Annales Des Télécommunications*, vol. 49, no. 7-8, pp. 439–446, 1994.

[5] J. Bitzer, K. U. Simmer, and K. D. Kammeyer, "Multi-microphone noise reduction by post-filter and superdirective beamformer," in *Proc. IEEE International Workshop on Acoustic Echo and Noise Control (IWAENC '99)*, pp. 100–103, Pocono Manor, Pa, USA, September 1999.

[6] M. Dorbecker and S. Ernst, "Combination of two-channel spectral subtraction and adaptive Wiener post-filtering for noise reduction and dereverberation," in *Proc. 8th European Signal Processing Conference (EUSIPCO '96)*, pp. 995–998, Trieste, Italy, September 1996.

[7] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '97)*, pp. 1167–1170, Munich, Germany, April 1997.

[8] A. Álvarez, R. Martínez, P. Gómez, and V. Nieto, "A speech enhancement system based on negative beamforming and spectral subtraction," in *Proc. IEEE International Workshop on Acoustic Echo and Noise Control (IWAENC '01)*, Darmstadt, Germany, September 2001.

[9] I. McCowan and H. Bourlard, "Microphone array post-filter for diffuse noise field," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '02)*, vol. 1, pp. 905–908, Orlando, Fla, USA, May 2002.

[10] D. Van Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '90)*, vol. 2, pp. 833–836, Albuquerque, NM, USA, April 1990.

[11] J. Vanden Berghe and J. Wouters, "An adaptive noise canceller for hearing aids using two nearby microphones," *Journal of the Acoustical Society of America*, vol. 103, no. 6, pp. 3621–3626, 1998.

[12] J.-B. Maj, J. Wouters, and M. Moonen, "A two-stage adaptive beamformer for noise reduction in hearing aids," in *Proc. IEEE International Workshop on Acoustic Echo and Noise Control (IWAENC '01)*, Darmstadt, Germany, September 2001.

[13] J. F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[14] J.-B. Maj, M. Moonen, and J. Wouters, "SVD-based optimal filtering technique for noise reduction in hearing aids using two microphones," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 432–443, 2002.

[15] A. Spriet, M. Moonen, and J. Wouters, "A multichannel subband GSVD approach for speech enhancement in hearing aids," in *Proc. IEEE International Workshop on Acoustic Echo and Noise Control (IWAENC '01)*, Darmstadt, Germany, September 2001.

[16] C. Baillargeat, *Contribution à l'amélioration des performances d'un radiotéléphone mains-libres à commande vocale*, Ph.D. thesis, Université de Paris IV, Paris, France, 1991.

[17] N. Dal Degan and C. Prati, "Acoustic noise analysis and speech enhancement techniques for mobile radio applications," *Signal Processing*, vol. 15, no. 4, pp. 43–56, 1988.

[18] A. Akbari Azirani, R. Le Bouquin, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," *IEEE Trans. Speech, and Audio Processing*, vol. 5, no. 5, pp. 484–487, 1997.

[19] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '79)*, pp. 208–211, Washington, DC, USA, April 1979.

[20] R. Le Bouquin, *Traitements pour la réduction du bruit sur la parole. Applications aux communications radio-mobiles*, Ph.D. thesis, Université de Rennes 1, France, 1991.

[21] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech, and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.
[22] C. Beaugeant, V. Turbin, P. Scalart, and A. Gilloire, "New optimal filtering approaches for hands-free telecommunication terminals," *Signal Processing*, vol. 64, no. 1, pp. 33–47, 1998.
[23] A. Guérin, *Rehaussement de la parole pour les communications mains-libres. Réduction de bruit et annulation d'écho non linéaire*, Ph.D. thesis, Université de Rennes 1, France, 2002.
[24] F. Lejay, "Speech enhancement system for Alcatel mobile," Rapport de Description Algorithmique AMP/DTT/SCI/FL0646.95, Alcatel Mobile Phones, 1995.
[25] A. Akbari Azirani, R. Le Bouquin, and G. Faucon, "Speech enhancement using a Wiener filtering under signal presence uncertainty," in *Proc. 8th European Signal Processing Conference (EUSIPCO '96)*, pp. 971–974, Trieste, Italy, September 1996.

Alexandre Guérin was born in Toulouse, France, in 1971. He received the B.S. degree in electrical engineering from the École 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 are concerned with two-sensor noise reduction dedicated to car kit systems and adaptive filtering applied to nonlinear acoustic echo cancellation. He has been an Associate Professor in 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 stereo-electroencephalographic 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 an 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.

Gérard Faucon received the Ph.D. degree in signal processing and telecommunications from the University of Rennes 1, France, in 1975. He is a Professor at the University of Rennes 1 and is a 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 stereo-electroencephalography signals and auditory evoked potentials.