Robust Estimation of Non-Stationary Noise Power Spectrum for Speech Enhancement

van Khanh Mai, Dominique Pastor, Abdeldjalil Aissa El Bey, Raphaël Le Bidan

Robust Estimation of Non-Stationary Noise Power Spectrum for Speech Enhancement

Van-Khanh Mai, *Student Member, IEEE*, Dominique Pastor, *Member, IEEE*,
Abdeldjalil Aïssa-El-Bey, *Senior Member, IEEE*, and Raphaël Le-Bidan, *Member, IEEE*
Institut Télécom; Télécom Bretagne; UMR CNRS 6285 Lab-STIC, Technopôle Brest-Iroise CS 83818, 29238 Brest, France
Université européenne de Bretagne

***Abstract*—We propose a novel method for noise power spectrum estimation in speech enhancement. This method called extended-DATE (E-DATE) extends the d-dimensional amplitude trimmed estimator (DATE), originally introduced for additive white gaussian noise power spectrum estimation in [1], to the more challenging scenario of non-stationary noise. The key idea is that, in each frequency bin and within a sufficiently short time period, the noise instantaneous power spectrum can be considered as approximately constant and estimated as the variance of a complex gaussian noise process possibly observed in the presence of the signal of interest. The proposed method relies on the fact that the Short-Time Fourier Transform (STFT) of noisy speech signals is sparse in the sense that transformed speech signals can be represented by a relatively small number of coefficients with large amplitudes in the time-frequency domain. The E-DATE estimator is robust in that it does not require prior information about the signal probability distribution except for the weak-sparseness property. In comparison to other state-of-the-art methods, the E-DATE is found to require the smallest number of parameters (only two). The performance of the proposed estimator has been evaluated in combination with noise reduction and compared to alternative methods. This evaluation involves objective as well as pseudo-subjective criteria.**

***Index Terms*—Speech enhancement, noise power spectrum estimation, noise reduction, robust statistics.**

I. Introduction

NOWADAYS communication electronic support in general and telephone conversation in particular often take place in noisy and non-stationary environments such as the inside of a car, in the street or inside an airport for example. Hence many research efforts have aimed at improving not only the quality but also the intelligibility of speech. Noise power spectrum estimation is a key issue in designing robust noise reduction methods for speech enhancement. Most noise power spectrum estimation algorithms found in the literature can be classified into four main categories [2], namely histogram-based methods, minimal-tracking algorithms, time-recursive averaging algorithms, and other techniques derived from Maximum-Likelihood (ML) or Bayesian estimation principles, *e.g.* minimum mean square error (MMSE) methods.

In the first category of algorithm, the noise power spectrum is estimated from the maximum of the histogram in the time-frequency domain of the observed signal power spectrum, the latter being determined by using a first-order smoothing recursion [3]. An improvement of this method involves updating the noise power spectrum uniquely on the frames detected as noise-only by a chi-square test [4]. However, most of the histogram-based algorithms have the drawback of being relatively complex in terms of computational cost and memory resources [5].

In the second family of methods, the noise power spectrum is tracked via minimum statistics, according to the reasonable hypothesis that the noise power spectrum level is below that of noisy speech [6], [7]. First, the smoothed noisy speech power spectrum is evaluated by a first-order recursive operation. Then, the noise variance is computed as the statistical minimum of the smoothed power spectrum with a factor of correction. The main difference between the two methods in [6] and [7] lies in the computation of the smoothing parameter used in the first order recursion. In [6], the smoothing parameter is chosen empirically, whereas this parameter is derived by minimizing the mean square error between the noise and the smoothed noisy speech power spectrum in [7]. Minimum-statistics methods require observing the noisy signals on a sufficiently long time interval so as to track speech power instead of noise power. On the other hand, a long time interval is detrimental to the quality of the estimate in case of non stationary noise. A trade-off is thus necessary, leading to a typical time-delay of 1 to 3 seconds in practice. This causes underestimation which decreases in turn the performance of noise reduction algorithms.

Famous methods of the third category include the Minima-Controlled Recursive-Averaging (MCRA) algorithm [8] and its many modifications such as the Improved-MCRA (IMCRA) [5] or the MCRA2 [9] methods. In this class of algorithms, the noise power spectrum in a given frequency bin is estimated by first-order recursive operations where smoothing parameters depend on the conditional speech presence probability in the bin. The main difference between MCRA, MCRA2 and IMCRA lies in the way the speech-presence probability is estimated. MCRA and MCRA2 directly estimate the speech-presence probability frame-by-frame via a smoothing operation whereby, for a given frame, the probability of speech presence is increased when this frame is detected as noisy speech and decreased otherwise. A frame is detected as noisy speech if the ratio of the smoothed noisy speech power spectrum to its local minimum is above a certain threshold, the local minimum being computed by using the minimum-statistics technique proposed in [7]. Fixed and frequency-dependent thresholds are used in MCRA and MCRA2, respectively. On the other

hand, IMCRA derives the speech-presence probability in each bin by a two-step estimation of the speech-absence probability. The first iteration aims at detecting the absence of speech in a given frame, while the second iteration actually estimates the speech-absence probability from the power spectral components in the speech-absence frame. The main disadvantage of these methods is the estimation delay in case of sudden rising noise, this delay being mainly due to the use of the minimum-statistics methods of [7].

Techniques derived from ML or Bayesian estimation principles overcome the problem of sudden rising noise by estimating the noise power spectrum from the noise periodogram via a statistical criterion. In [10], [11], the noise instantaneous power is evaluated by MMSE and then incorporated in a recursive noise power spectrum estimation technique. [10] proposes a simple bias compensation of the noise instantaneous power before estimating the noise power spectrum via the same recursive smoothing and under the same hypotheses as in [11]. However, the noise instantaneous power estimate in [10] remains biased. In contrast, an unbiased estimator for the noise instantaneous spectrum is obtained in [11] by soft-weighting the noisy speech instantaneous power and the previous noise power spectrum estimate by the conditional probabilities of speech-absence and speech-presence, respectively. The noise power spectrum estimation can also be carried out by recursive ML-Expectation-Maximization [12], similar to MCRA and IMCRA. This approach allows for rapid noise power spectrum estimation and tracking by avoiding the use of minimum-statistics methods.

In this paper, we propose a new approach for noise power spectrum estimation, without requiring any model or any prior knowledge for the probability distributions of the speech signals. Fundamentally, we do not even take into consideration the fact that the signal of interest here is speech. The approach is henceforth called extended-DATE (E-DATE) since it basically extends the d-dimensional amplitude trimmed estimator (DATE), initially proposed in [1] for white gaussian noise (WGN), to colored stationary and non-stationary noise. The main principle at the heart of the E-DATE algorithm is the weak-sparseness property of the STFT of noisy signals, according to which the sequence of complex values returned by the STFT in a given time-frequency bin can be modeled as a complex random signal with unknown distribution and whose unknown probability of occurrence in noise does not exceed one half. Noise in each bin is assumed to follow a zero-mean complex gaussian distribution [2, p. 210], so that estimating the noise power spectrum amounts to estimating the noise variance in each bin, the latter being provided by the DATE. The DATE trims the amplitudes in each given bin, after having sorted them by increasing norm. Noise power spectrum estimation by E-DATE is thus similar to and actually extends the quantile-based approach of [13], which relies on assumptions that the weak-sparseness model embraces. More generally, the reader will notice similarities between the proposed method and the state-of-the-art techniques mentioned above. A main difference between the E-DATE approach and standard ones is actually the mathematical justification of the former via the weak-sparseness model, which formalizes more or less standard heuristics in speech processing and yields a reduced number of parameters for more robustness. Although the E-DATE does not rely on minimum-statistics principles or methods, it does however require a time buffer having the same length — typically 80 frames for a frequency sampling rate of 8 kHz — as other popular algorithms.

The paper is organized as follows. In Section II, the main features of the DATE are reviewed. Section III develops the weak-sparseness model for noisy speech. The E-DATE is then introduced in Section IV, following a step-by-step methodology where we successively deal with WGN, stationary noise and non-stationary noise. Two practical implementations of the E-DATE algorithm are then described. The performance of the E-DATE algorithm is evaluated in Section V and compared to state-of-the-art methods in terms of number of parameters and estimation errors. Speech enhancement experimental comparisons using objective as well as pseudo-subjective criteria are also conducted by combining the noise power spectrum estimation methods with a noise reduction system. Conclusions are finally given in Section VI.

II. The DATE

For the sake of self-completeness, this section presents the DATE in its full generality. Given d-dimensional observations of random signals randomly absent or present in independent and additive WGN, the purpose of the DATE is to estimate the noise standard deviation. Such an estimation may serve to detect the signals or estimate them as in speech denoising. As in [14], the DATE addresses the frequently-encountered case where 1) most observations follow the same zero-mean normal distribution with unknown variance, 2) signals of interest have unknown distributions and occurrences in noise. Standard robust scale estimators such as the very popular median absolute deviation (MAD) estimator and the trimmed estimator (T-estimator) have performance that degrades significantly when the proportion of signal increases. In contrast, the DATE can still estimate the noise standard deviation when possible signals occur with a probability too large for usual scale estimators to perform well. As indicated by its name, the DATE basically trims the norms of the d-dimensional observations. However, in contrast to the conventional T-estimator, which applies to one-dimensional data and fixes the number of outliers to remove, the DATE applies to any dimension and chooses adaptively the number of outliers to discard. It performs the trimming by assuming that the signal norms are above some known lower-bound and that the signal probabilities of occurrence are less than one half. These assumptions bound our lack of prior knowledge about the signals and make it possible to separate signals from noise. Moreover, these assumptions are suitable for signal processing applications where noisy signals are considered as outliers with respect to the noise distribution. They are particularly suitable for observations obtained

after sparse transforms capable of representing signals by coefficients that are mostly small except a few ones whose norms are relatively big. In particular, the sequel will exhaustively use the fact that the Fourier transform of speech signals is sparse in a weak sense detailed hereafter.

The DATE basically relies on [1, Theorem 1], which is asymptotic and can be viewed as a method of moments. A detailed presentation of the theoretical background of the DATE is beyond the scope of this paper and the reader is referred to [1] for details. However, the following brief heuristic presentation may be convenient for the reader. This heuristic exposure departs from that proposed in [1, Theorem 1], so as to shed different light on the theory behind the DATE.

Notation: In what follows, $\|\cdot\|$ is the usual euclidean norm in the space of all d-dimensional real vectors, $\mathbf{I}_d$ stands for the $d \times d$ identity matrix, $\mathcal{N}(0, \sigma_0^2 \mathbf{I}_d)$ designates the d-dimensional gaussian distribution with null mean and covariance matrix $\sigma^2 \mathbf{I}_d$ and $\mathbb{1}[X \in B]$ stands for the indicator function of the event $[U \in B]$, where U is any random variable and B is any borel set of the real line: $\mathbb{1}[U \in B] = 1$ if $U \in B$ and $\mathbb{1}[U \in B] = 0$, otherwise. In addition, Γ is the standard Gamma function and ${}_0\mathrm{F}_1$ is the generalized hypergeometric function [15, p. 275]. All the random vectors and variables are henceforth assumed to be defined on the same probability space $(\Omega, \mathbb{P}, \mathbb{E})$.

Let $(Y_n)_{n \in \mathbb{N}}$ be a sequence of d-dimensional random observations such that:

(A0) The observations $Y_1, Y_2, \ldots, Y_n, \ldots$ are mutually independent, $Y_n = \varepsilon_n \Lambda_n + X_n$ where $X_n \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}_d)$ and ε_n is Bernoulli distributed with values in $\{0, 1\}$ for each $n \in \mathbb{N}$.

In this model, each observation is either noise alone or the sum of some signal and noise. The probability distributions of the signals Λ_n are supposed to be unknown. Our purpose is then to estimate σ_0.

If all the ratios $\|\Lambda_n\|/\sigma_0$ are known to be above some sufficiently large signal to noise ratio (SNR) ρ, it can be expected that some threshold height $\sigma_0 \xi(\rho)$ can suitably be chosen to decide with small error probability that Λ_n is present (resp. absent) whenever $\|Y_n\|$ is above (resp. less) $\sigma_0 \xi(\rho)$. Therefore, most of the non-zero terms in the sum $\sum_{n=1}^N \|Y_n\| \mathbb{1}[\|Y_n\| \leqslant \sigma_0 \xi(\rho)]$ should pertain to noise alone. If the number $\sum_{n=1}^N \mathbb{1}[\|Y_n\| \leqslant \sigma_0 \xi(\rho)]$ of these non-zero terms is itself large enough, we should have an approximation of the form $\frac{\sum_{n=1}^N \|Y_n\| \mathbb{1}[\|Y_n\| \leqslant \sigma_0 \xi(\rho)]}{\sum_{n=1}^N \mathbb{1}[\|Y_n\| \leqslant \sigma_0 \xi(\rho)]} \approx \lambda \sigma_0$. Such an approximation can actually be proved asymptotically with the help of some additional assumptions. More precisely, suppose that:

(A1) Λ_n, X_n and ε_n are independent for every $n \in \mathbb{N}$;

(A2) the set of priors $\{\mathbb{P}[\varepsilon_n = 1] : n \in \mathbb{N}\}$ is upper-bounded by $1/2$ and the random variables ε_n, $n \in \mathbb{N}$, are independent;

(A3) $\sup_{n \in \mathbb{N}} \mathbb{E}\left[\|\Lambda_n\|^2\right] < \infty$.

These assumptions including **(A0)** deserve some comments. To begin with, the independence assumption in **(A0)** is mainly technical to prove the results stated in [1]. In fact, our experimental results below suggest that this assumption is not so constraining in speech processing, where we deal with non-overlapping but not necessarily independent time frames. Assumption **(A1)** simply means that the two hypotheses for the observation occur independently and that the noise and signal are independent. The model thus assumes prior probabilities of presence and absence through the random variables ε_n. However, the impact of these priors is reduced by assuming that the probabilities of presence and absence are actually unknown. The role of Assumption **(A2)** is then to bound this lack of prior knowledge about the occurrences of the two possible hypotheses that any Y_n is supposed to satisfy. Assumption **(A3)** simply means that the signals Λ_n have finite energy.

Under assumptions **(A0)**-**(A3)** and with the help of [16, Theorem 1], [1, Theorem 1] then guarantees that σ_0 is the unique positive real number σ such that:

$$\lim_{\rho \to \infty} \left\| \limsup_{N \to \infty} \left| \frac{\sum_{n=1}^N \|Y_n\| \mathbb{1}[\|Y_n\| \leqslant \sigma \xi(\rho)]}{\sum_{n=1}^N \mathbb{1}[\|Y_n\| \leqslant \sigma \xi(\rho)]} - \lambda \sigma \right| \right\|_\infty = 0 \quad (1)$$

where $\lambda = \sqrt{2}\,\Gamma\left(\frac{d+1}{2}\right)/\Gamma\left(\frac{d}{2}\right)$ and $\xi(\rho)$ is the unique positive solution in x to the equality ${}_0\mathrm{F}_1(d/2; \rho^2 x^2/4) = e^{\rho^2/2}$. It is thus natural to estimate the noise standard deviation σ_0 by seeking a possibly local minimum of:

$$\left| \frac{\sum_{n=1}^N \|Y_n\| \mathbb{1}[\|Y_n\| \leqslant \sigma \xi(\rho)]}{\sum_{n=1}^N \mathbb{1}[\|Y_n\| \leqslant \sigma \xi(\rho)]} - \lambda \sigma \right|, \quad (2)$$

when σ ranges over some search interval $[\sigma_{\min}, \sigma_{\max}]$. Given a lower bound ρ for the ratios $\|\Lambda_n\|/\sigma_0$, the DATE computes the solution in σ to the equality:

$$\frac{\sum_{n=1}^N \|Y_n\| \mathbb{1}[\|Y_n\| \leqslant \sigma \xi(\rho)]}{\sum_{n=1}^N \mathbb{1}[\|Y_n\| \leqslant \sigma \xi(\rho)]} = \lambda \sigma. \quad (3)$$

Indeed, such a solution trivially minimizes (2).

In addition, an application of Bienaymé-Chebyshev's inequality makes it possible to determine the value $n_{\min} \in \{1, 2, \ldots, N\}$ such that the probability that the number of observations due to noise alone be above $n_{\min}$ is larger than or equal to some given probability value Q. The main steps of the DATE are summarized in Algorithm 1, where $Y_{(1)}, Y_{(2)}, \ldots, Y_{(N)}$ is the sequence $Y_1, Y_2, \ldots, Y_N$ sorted by increasing norm so that $\|Y_{(1)}\| \leqslant \|Y_{(2)}\| \leqslant \ldots \|Y_{(N)}\|$, and where we have defined

$$\mathrm{M}^*_{\{\|Y_1\|, \|Y_2\|, \ldots, \|Y_N\|\}}(n) = \begin{cases} \frac{1}{n} \sum_{k=1}^n \|Y_{(k)}\| & \text{if } n \neq 0 \\ 0 & \text{if } n = 0, \end{cases} \quad (4)$$

The parameters on which the DATE relies are thus: the dimension d of the observations, the number N of observations and the lower bound ρ for the possible SNRs. The two parameters that directly influence the DATE performance are N and ρ. As recommended in [1, Remark 4], we can use $\rho = 4$ in practice. Theoretically, N should be large since the theoretical result on which the DATE relies is asymptotic by nature. However, experimental results show that the DATE performance is acceptable when N is above 200. This will

Algorithm 1 DATE algorithm for estimation of noise standard deviation

Input:

- A finite subsequence $\{Y_1, Y_2, \ldots, Y_N\}$ of a sequence $Y = (Y_n)_{n\in\mathbb{N}}$ of d-dimensional real random vectors satisfying assumptions **(A0-A3)** above
- A lower bound ρ for the SNRs $\|\Lambda_n\|/\sigma_0,\ n \in \mathbb{N}$
- A probability value $Q \leqslant 1 - \frac{N}{4(N/2-1)^2}$

Constants: $n_{\min} = N/2 - \sqrt{N/4(1-Q)}$, $\xi(\rho)$, λ
Output: The estimate $\sigma^*_{\{Y_1,Y_2,\ldots,Y_N\}}$ of σ_0

Computation of $\sigma^*_{\{Y_1,Y_2,\ldots,Y_N\}}$:

Sort $Y_1, Y_2, \ldots, Y_N$ by increasing norm so that $\|Y_{(1)}\| \leqslant \|Y_{(2)}\| \leqslant \ldots \leqslant \|Y_{(N)}\|$
if there exists a smallest integer n in $\{n_{\min}, \ldots, N\}$ such that: $\|Y_{(n)}\| \leqslant \left(\mathrm{M}^*_{\{\|Y_1\|,\|Y_2\|,\ldots,\|Y_N\|\}}(n)/\lambda\right)\xi(\rho) < \|Y_{(n+1)}\|$
    $n^* = n$
else
    $n^* = n_{\min}$
end if
$\sigma^*_{\{Y_1,Y_2,\ldots,Y_N\}} = \mathrm{M}^*_{\{\|Y_1\|,\|Y_2\|,\ldots,\|Y_N\|\}}(n^*)/\lambda$

be confirmed by the application to speech processing of Sections IV and V.

Another means to choose the minimal SNR required by the DATE is to resort to the notion of universal threshold [17], as proposed in [18]. Indeed, the coordinates of all the N observations $Y_1, Y_2, \ldots, Y_N$ form a set of $N \times d$ random variables. If no signals were present, these $N \times d$ random variables would be i.i.d (independent and identically distributed) gaussian with null mean and variance equal to σ_0^2. According to [19, Eqs. (9.2.1), (9.2.2), Section 9.2, p. 187] [20, p. 454] [21, Section 2.4.4, p. 91], the universal threshold $\lambda_u(N \times d) = \sigma_0\sqrt{2\ln(N \times d)}$ could then be regarded as the maximum absolute value of these gaussian random variables when $N \times d$ is large. Instead of proceeding as in wavelet shrinkage [17] where the universal threshold is utilized to discriminate noisy signal wavelet coefficients from wavelet coefficients of noise alone, the trick proposed in [22] and [18] is to consider $\lambda_u(N \times d)$ as the minimum amplitude that a signal must have to be distinguishable from noise. The minimal SNR can then be defined as $\rho = \rho(N \times d) = \lambda_u(N \times d)/\sigma_0 = \sqrt{2\ln(N \times d)}$. It is an interesting fact that the value of $\rho(N \times d)$ grows rapidly to 4 with $N \times d$.

In the sequel, we will consider values returned by STFT. The DATE will therefore be applied to sequences of real and complex values, that is, one- and two-dimensional data since complex values can be regarded as 2-dimensional real vectors. It is thus worth recalling the specific values of $\xi(\rho)$ and λ for $d = 1$ and $d = 2$. If $d = 1$, $\xi(\rho) = \cosh^{-1}(e^{\rho^2/2}) = \frac{1}{2}\rho + \frac{1}{\rho}\log(1 + \sqrt{1 - e^{-\rho^2}})$ and $\lambda = 0.7979$. If $d = 2$, $\xi(\rho) = \mathrm{I}_0^{-1}(e^{\rho^2/2})/\rho$ where I_0 is the zeroth order modified Bessel function of the first kind and $\lambda = 1.2533$. Note that $1/\lambda$ can be regarded as a bias correction factor, similar to those employed by minimum-statistics approaches.

III. Weak-sparseness model of noisy speech

The main motivation for utilizing the DATE is that noisy speech signals in the time-frequency domain after STFT reasonably satisfy the same type of weak-sparseness model as used to establish [1, Theorem 1]. This weak-sparseness model essentially assumes that the noisy speech signal can be represented by a relatively small number of coefficients with large amplitudes. Indeed, let us consider the spectrograms of Figure 1 obtained by STFT of typical examples of clean and noisy speech signals. In the time-frequency domain, speech is composed of a set of time-frequency components or atoms. Most atoms with small amplitudes are masked in the presence of noise. Only the few atoms whose amplitude is above some minimum value remain visible in noise. Clearly, the proportion of these significant atoms does not exceed one half. These remarks lead to the following model for noisy speech STFTs. In the time domain, the observed signal is given by

$$y(t) = s(t) + x(t) \tag{5}$$

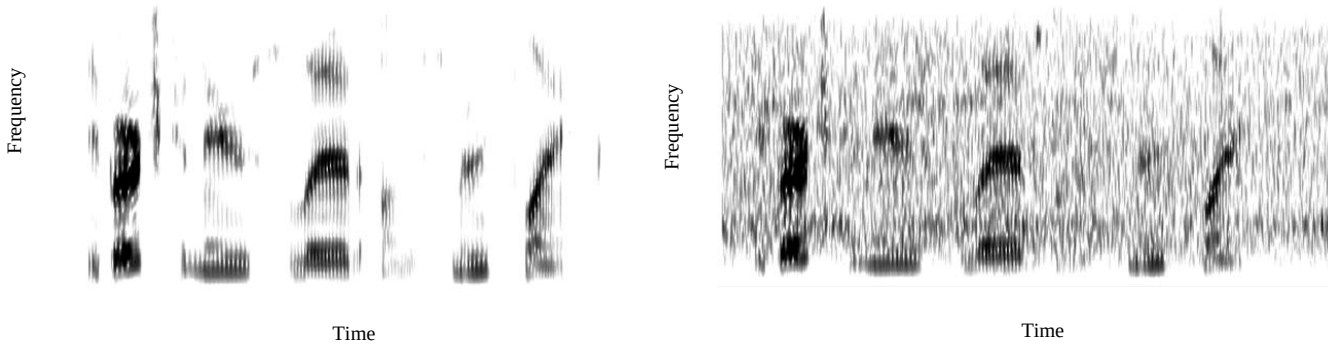
where $s(t)$ and $x(t)$ denotes clean speech and independent additive noise. Note that both are real-valued signals. The signal in the time domain is transformed into the time-frequency domain by STFT since most noise reduction systems operate in this particular transform domain. Hence, all processing is frame-based. Let K be the frame length, or equivalently, the STFT length. The corresponding system model in the time-frequency domain then reads:

$$Y(m,k) = S(m,k) + X(m,k) \tag{6}$$

in which m denotes the frame index, k is the frequency-bin index, and $S(m,k)$ (resp. $X(m,k)$) stands for the STFT component of the speech signal (resp. noise) at time-frequency point (m,k). Following [2, page 210], we model each $X(m,k)$ as a complex Gaussian random variable. Complex values $Y(m,k)$ are manipulated as 2-dimensional real vectors.

According to the empirical remarks above, the weak-sparseness model first assumes that an atomic speech audio source is either present or absent at any given time-frequency point (m,k). The presence or the absence of this source is modeled by a Bernoulli random variable $\varepsilon(m,k)$. This Bernoulli model is tantamount to and justified by the concept of ideal binary masking in the time-frequency domain, as used in audio source separation [18], [23]. The probability of presence is assumed to be less than or equal to 1/2. Thus $\mathbb{P}\left[\varepsilon(m,k) = 1\right] \leq 1/2$. Second, the atomic audio source must have significant amplitude so as to contribute effectively to the mixture that composes the speech signal. The minimum amplitude that such a source must have will hereafter be denoted by ρ. Let us further denote by $\Theta(m,k)$ the underlying atomic audio source. Then, under the previous assumptions, the noisy speech signal at time-frequency point (m,k) can be modeled as:

$$Y(m,k) = \varepsilon(m,k)\Theta(m,k) + X(m,k) \tag{7}$$



(a) Clean speech (b) Noisy speech

Fig. 1: Spectrograms of clean and noisy speech signals from the NOIZEUS database. The noise source is car noise. No weighting function was used to calculate the STFT.

We recognize here the weak-sparseness model [24] applied to speech processing, in the continuation of [18].

In summary, our model essentially assumes that the STFT of noisy speech signals satisfies the following three key properties in each time-frequency bin (m,k):
(A'1): the presence/absence of speech $\varepsilon(m,k)$ and the atomic speech audio source $\Theta(m,k)$ are independent,
(A'2): the speech-presence probability does not exceed 1/2,
(A'3): the instantaneous power of the random clean speech signal is upper-bounded by a finite value.

Assumptions (**A'1-A'3**) are adaptations of (**A1-A3**) to the particular case of noisy speech signals. Regarding (**A0**), its equivalent form for noisy speech signals is simply Eq. (7).

Our purpose is then to estimate the noise power spectrum $\sigma_X^2(m,k) = \mathbb{E}\left[|X(m,k)|^2\right]$ at any given time-frequency point (m,k). This problem is similar to that addressed in [18], where the signal of interest was a mixture of audio signals, possibly including speech signals, and where additive noise was stationary, gaussian and white. The DATE was used to estimate the noise power spectrum in [18] because this estimator does not make prior assumption on the statistical nature of the signals of interest. In the present paper and in contrast to [18], we do not restrict our attention to WGN and generalize the approach of [18] to the estimation of colored and possibly non-stationary noise in the presence of speech.

IV. Noise power spectrum estimation by E-DATE

In this section, we derive the E-DATE algorithm that will be used for noise power spectrum estimation in all the experiments conducted in Section V. The derivation follows a three-step process, which aims at gradually introducing the modifications required to evolve from the academic WGN model to the much more realistic, but also more challenging, practical case of non-stationary noise. More precisely, we first describe the application of the DATE algorithm to noise power spectrum estimation of noisy speech signals in the time-frequency domain. We extend the DATE to the case of colored stationary gaussian noise, and then discuss the estimation of non-stationary noise. This leads to the E-DATE algorithm, which is specifically designed for noise power spectrum estimation in non-stationary noisy environments, but can be used with stationary noise as well.

In the following, we suppose to be given M noisy speech frames of K samples. The frames are assumed to be non-overlapping so as to satisfy assumption (**A0**). The STFTs are normalized by $1/\sqrt{K}$.

A. Stationary WGN

In this case, the noise power spectrum is constant and equals σ_X^2 over the whole time-frequency plane. Accordingly, and by properties of the (normalized) STFT, each noise sample $X(m,k)$ in the time-frequency domain is a zero-mean circularly-symmetric gaussian complex random variable with variance σ_X^2:

$$X(m,k) \sim \mathcal{N}_c(0,\sigma_X^2).$$

Equivalently, $X(m,k)$ may be viewed as a zero-mean two-dimensional real gaussian random vector with covariance matrix $(\sigma_X^2/2)\boldsymbol{I}_2$:

$$X(m,k) \sim \mathcal{N}\left(\boldsymbol{0},(\sigma_X^2/2)\boldsymbol{I}_2\right).$$

Since the STFT of noisy speech signals is weakly-sparse in the sense of Section III, the $M\times(K/2-1)$ values $Y(m,k)$ for $m\in\{1,2,\ldots,M\}$ and $k\in\{1,2,\ldots,K/2-1\}$ can be used as inputs of the two-dimensional ($d=2$) version of the DATE to provide an estimate $\widehat{\sigma}_X^2$ of σ_X^2. Note that, in principle, another estimate of σ_X^2 could be obtained by applying a one-dimensional ($d=1$) DATE on the $2\times M$ real dataset $Y(1,0),Y(2,0),\ldots,Y(M,0),Y(1,K/2),Y(2,K/2),\ldots,Y(M,K/2)$. However, the size of this second dataset is usually much smaller than that of the first one. Thus only the first option is used in practice as it leads to a more reliable estimate. Note also that, due to the Hermitian property of the STFT of real input signals, $|Y(m,k)| = |Y(m,K-k)|$. Therefore the frequency bins $K/2+1$ to K are not used in the estimation process as they do not bring additional information.

B. Colored stationary noise

For colored stationary noise, the noise power spectrum is no longer constant over the whole time-frequency plane but may vary as a function of frequency. Consequently, each noise sample $X(m,k)$ in a given frequency bin k will now be modeled as a zero-mean complex gaussian random variable with variance $\sigma_X^2(k)$:

$$X(m,k) \sim \mathcal{N}_c\big(0,\sigma_X^2(k)\big).$$

Here again, the STFT output sequence $Y(m,k)$ for $m = 1,2,\ldots,M$ is assumed to be weakly-sparse in the sense of Section III so that in each frequency bin k, only a few of these values will have an SNR above ρ and in a proportion that does not exceed $1/2$. As a result and as illustrated in Figure 2, the extension to colored stationary noise involves running concurrently $K/2+1$ independent instances of the DATE to estimate $\sigma_X^2(k)$ in each frequency bin $k = 0,1,2,\ldots,K/2$. As discussed earlier, we do not use

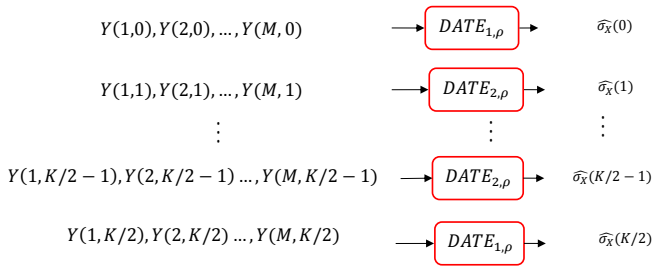


Fig. 2: Principle of noise power spectrum estimation based on the DATE in colored stationary noise

the DATE to estimate $\sigma_X^2(k)$ for $k > K/2$ because of the Hermitian symmetry. For $k \in \{1,2,K/2-1\}$, the estimate of $\sigma_X^2(k)$ is computed by the two-dimensional ($d=2$) DATE whereas the one dimensional ($d=1$) DATE is used for bins 0 and $K/2$. For colored noise, assumption (**A'1**) may not always rigorously hold, especially at low frequencies. However, as supported by the experimental results of Section V, this deviation with respect to the underlying theoretical model turns out to be no real issue in practice, thanks to the robust behavior of the DATE, even when the signal presence probability may exceed $1/2$ (see [1, Figure 2]).

In contrast to WGN for which the whole time-frequency plane ($\approx MK/2$ observations) is used to estimate the noise variance σ_X^2, M frames only are available here to estimate $\sigma_X^2(k)$ in each frequency bin. Clearly a more reliable estimate can be obtained by increasing M, but this increases in return the overall computational cost and may also entail some time-delay. A possible solution is to begin with a first estimate $\widehat{\sigma}_X^2(k)$ computed over the first M frames, and then to periodically update this estimate as new frames are acquired. For stationary noise, the initial number of frames M need not be very high. Even if the first estimate is not very accurate, it is expected to improve rapidly as new frames enter the estimation process.

C. Extension to non-stationary noise: The E-DATE algorithm

Most practical applications including speech denoising usually face a mix of stationary as well as non-stationary noise. Unlike white or colored stationary noise, the power spectrum of non-stationary noise varies over time and frequency, and, as such, proves to be much more challenging to estimate. Interestingly, non-stationary noise models including car noise, babble noise, exhibition noise and others, usually exhibit some form of local stationarity in time and frequency. In such cases, non-stationary noise can be considered as approximately stationary within short time periods of D consecutive frames, where parameter D has to be defined appropriately for each noise model. This amounts to assuming the existence of a noise power spectrum in this time interval, which is a function of frequency only. The DATE algorithm for colored stationary noise introduced in Section IV-B can then be used to estimate the noise power spectrum within this time window of D frames. This is the basis of the E-DATE algorithm.

Parameter D can be preset once for all or could be optimized for applications where prior knowledge about noise is available. The choice for duration D results from a trade-off between estimation accuracy, stationarity and practical constraints such as computational cost and time-delay. A large value for D may violate the local stationary property. On the other hand, the number of frames D should be large enough to produce reliable estimates $\sigma_X^2(k)$. In case D is too small to provide the DATE with a sufficient number of input data, a possible solution consists in grouping several consecutive frequency bins. This is tantamount to assuming that the noise power spectrum is approximately constant over those frequencies. Such a procedure however requires prior knowledge on the noise spectrum properties, which can be irrelevant in practical applications where noise has often unknown type and may evolve across time. For this reason, this solution will not be further studied below.

In summary, the E-DATE algorithm consists in carrying noise power spectrum estimation by running a per-bin instance of the DATE (see Figure 2) on periods of D consecutive non-overlapping frames, where D is chosen so that noise can be considered as approximately stationary within this time interval. Once an estimate of the noise power spectrum has been obtained, it can be used for denoising purpose for instance, but will not be taken into account in the computation of future estimates, as the local power spectrum of non-stationary noise may change significantly from one period of D frames to the next.

Although the E-DATE algorithm was specifically designed for power spectrum estimation of non-stationary noise, it can be used without modification for power spectrum estimation of WGN or colored stationary noise, thereby offering a robust and universal noise power spectrum estimator whose parameters are fixed once for all types of noise considered above. Let us now discuss the practical implementation of the E-DATE algorithm.

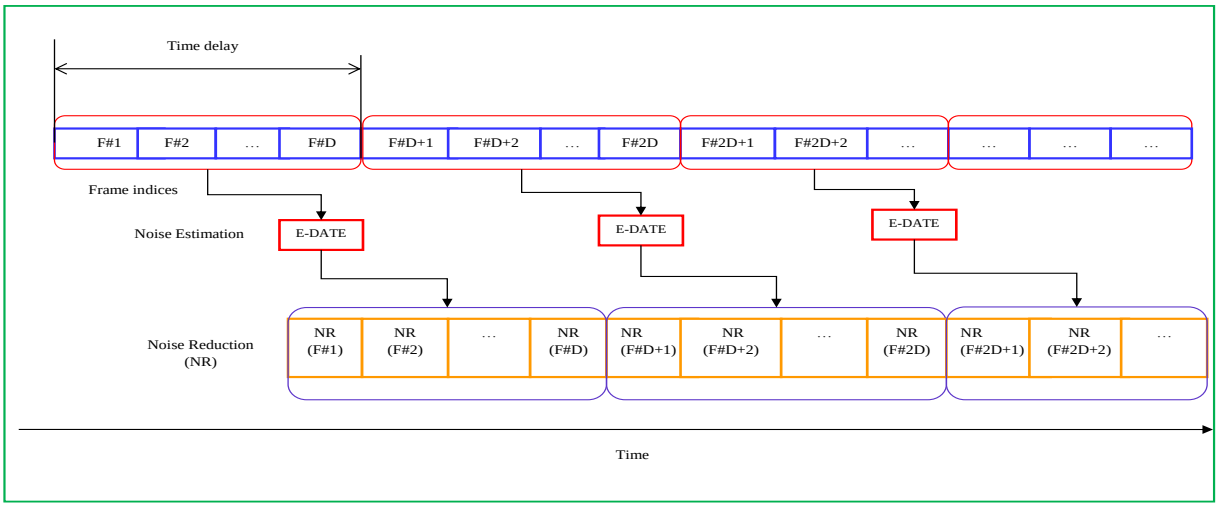


Fig. 3: Block E-DATE (B-E-DATE) combined with noise reduction (NR). A single noise power spectrum estimate is calculated every D non-overlapping frames and used to denoise each of these D frames.

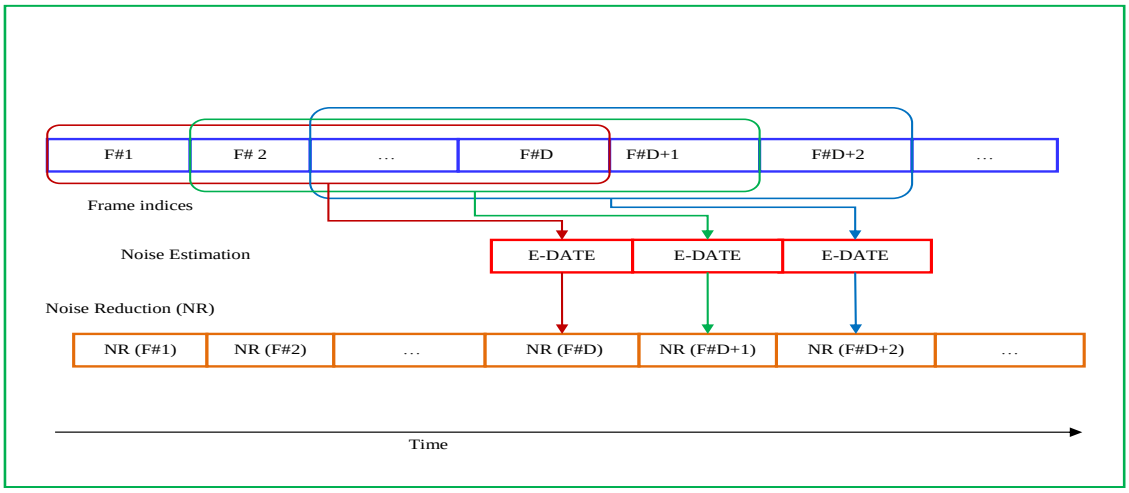


Fig. 4: Sliding-Window E-DATE (SW-E-DATE) combined with noise reduction. For the first $D-1$ frames, a surrogate method for noise power spectrum estimation is used in combination with noise reduction. Once D frames are available and upon reception of frame $D+\ell$, $\ell \geq 0$, the SW-E-DATE algorithm provides the NR system with a new estimate of the noise power spectrum computed using the last D frames $F_{\ell+1},\ldots,F_{\ell+D}$ for denoising of the current frame.

D. Practical implementation of the E-DATE algorithm

Two different implementations of the E-DATE algorithm are proposed here.

The first approach is a straightforward block-based implementation of the algorithm described in Section IV-C. It involves estimating the noise power spectrum on each period of D successive non-overlapping frames. This requires storing D frames, calculating the $K/2+1$ estimates $\widehat{\sigma}_X^2(k)$ using the observations in these D frames, and then waiting for D new non-overlapping frames. The resulting algorithm is called Block-E-DATE (B-E-DATE) and summarized in Algorithm 2, where $\widehat{\sigma} = \mathrm{DATE}_{d,\rho}(y_1, y_2, \ldots, y_n)$ denotes the standard deviation estimate $\widehat{\sigma}$ returned by the d-dimensional DATE with minimal SNR ρ and n real d-dimensional inputs $y_1, y_2, \ldots, y_n$.

Estimation of the noise power spectrum over separate periods of D non-overlapping frames reduces the overall algorithm complexity. However, this entails a time-delay of D frames, which must be considered in applications. Consider the particular example of speech denoising illustrated in Figure 3. Noise reduction is performed on a frame-by-frame basis. A new noise power spectrum estimate is provided to the noise reduction system by the B-E-DATE algorithm once every D non-overlapping frames, and then used to denoise each of those D frames. Clearly, denoising cannot start before the first D non-overlapping frames have been recorded. This results in an overall latency of about 1 or 2 seconds for typical sampling rates of 8 and 16 kHz. This delay can then have some impact for speech applications embedded in current mobile devices. It will naturally be

lesser in applications such as Active Noise Cancellation (ANC) where frequency rates are much higher.

The delay limitation can be bypassed as follows. First, a standard noise power spectrum tracking method is used to estimate the noise power spectrum during the first $D-1$ non-overlapping frames. Any of the methods mentioned in the introduction can be used for this purpose. Afterwards, starting from the D^{th} frame onwards, a sliding-window version of the E-DATE algorithm is used to estimate the noise spectrum on a per-frame basis, using the latest recorded D non-overlapping frames. This alternative implementation called Sliding-Window E-DATE (SW-E-DATE) is summarized in Algorithm 3. Its application to speech denoising is illustrated in Figure 4.

The B-E-DATE and the SW-E-DATE algorithm may be viewed as two particular instances of a more general buffer-based algorithm. More precisely, the B-E-DATE algorithm corresponds to the extreme case where the buffer is totally flushed and updated once every D non-overlapping frames. In contrast, the SW-E-DATE algorithm corresponds to the other extreme case where only the oldest frame is discarded in order to store the current one, in a First-In First-Out (FIFO) mode. Clearly, a more general approach between these two extremes consists in partially updating the buffer by renewing only L frames among D. This point has not been further investigated in the present work.

Note finally that the proposed implementations of the E-DATE algorithm are not limited to speech denoising but could find use in any application involving signals corrupted by additive and independent non-stationary noise, and to which the weak-sparseness model locally applies.

Algorithm 2 Block-Extended-DATE (B-E-DATE) algorithm for noise power spectrum estimation

for $m \geq D$ **do**
  if $\mod(m,D) = 0$
    $m^* = m$
    $\widehat{\sigma}_X(m^*,0) =$
    $\text{DATE}_{1,\rho}\big(Y(m-D+1,0), Y(m-D+2,0), \ldots, Y(m,0)\big)$
    $\widehat{\sigma}_X(m^*,K/2) =$
    $\text{DATE}_{1,\rho}\big(Y(m-D+1,K/2), Y(m-D+2,K/2), \ldots, Y(m,K/2)\big)$
    for $k := 1$ to $\frac{N}{2}-1$ **do**
      $\widehat{\sigma}_X(m^*,k) =$
      $\text{DATE}_{2,\rho}\big(Y(m-D+1,k), Y(m-D+2,k), \ldots, Y(m,k)\big)$
      $\widehat{\sigma}_X(m^*,K-k) = \widehat{\sigma_X}(m^*,k)$
    end for
  else
    $\widehat{\sigma}_X(m-D,k) = \widehat{\sigma}^*_X(m^*,k)$
  end if
end for

V. Performance Evaluation

Several comparisons and experiments were conducted in order to assess the performance and benefits of the E-DATE noise power spectrum estimator in comparison with other state-of-the-art algorithms. Both the B-E-DATE and the SW-E-DATE implementations were considered in two different benchmarks. In subsection V-A, we first compare the number of parameters required by the E-DATE and several classical or more recent noise power spectrum estimators. Then, we compare in subsection V-B the estimation quality of the different algorithms in several distinct noise environments. The combination of the noise power spectrum estimation algorithms with a noise reduction system based on the Log-MMSE algorithm is investigated using the NOIZEUS speech corpus in subsection V-C. Finally, the time-complexity of the E-DATE algorithm is analyzed in subsection V-D.

Algorithm 3 Sliding-Window Extended-DATE (SW-E-DATE) algorithm for noise power spectrum estimation

for $m = 1$ to the end of signal **do**
  if $m < D$
    Calculate $\widehat{\sigma}_X$ by an alternative method
  else
    $\widehat{\sigma}_X(m,0)$
    $= \text{DATE}_{1,\rho}\big(Y(m-D+1,0), Y(m-D+2,0), \ldots, Y(m,0)\big)$
    $\widehat{\sigma}_X(m,K/2)$
    $= \text{DATE}_{1,\rho}\big(Y(m-D+1,K/2), Y(m-D+2,K/2)), \ldots, Y(m,K/2)\big)$
    for $k := 1$ to $\frac{K}{2}+1$ **do**
      $\widehat{\sigma}_X(m,k)$
      $= \text{DATE}_{2,\rho}\big(Y(m-D+1,k), Y(m-D+2,k), \ldots, Y(m,k)\big)$
      $\widehat{\sigma}_X(m,K-k) = \widehat{\sigma_X}(m,k)$
    end for
  end if
end for

A. Number of parameters

Table I gives the number of parameters required by the E-DATE as well as by the state-of-the-art noise power spectrum estimation algorithms mentioned in the introduction. Derived from robust statistical signal processing concepts, the E-DATE is the simplest algorithm to configure, with only two parameters to specify, namely the SNR lower bound ρ and the number of frames D. This stands in sharp contrast with other popular approaches such as Minimum Statistics [7], which involves 7 parameters. In practice, the minimal SNR ρ can be set as explained at the end of Section II so that the only crucial parameter is D. Working with $D = 80$ non-overlapping frames of $K = 256$ samples was found to yield good performance in all the experiments reported here.

B. Noise Estimation Quality

The estimation quality of the noise power spectrum estimation algorithms listed in Table I was evaluated on several noise models using the symmetric segmental logarithmic estimation error measure defined in [25]. The difference between the estimated noise power spectrum $\widehat{\sigma}^2_X(m,k)$ and reference noise power spectrum $\sigma^2_X(m,k)$ is evaluated by

$$LogErr = \frac{1}{MK}\sum_{m=0}^{M-1}\sum_{k=0}^{K-1}\left|10\log_{10}\frac{\widehat{\sigma}^2_X(m,k)}{\sigma^2_X(m,k)}\right| \qquad (8)$$

where M denotes the total the number of available frames. For WGN, the theoretical reference noise power spectrum

TABLE I: Number of parameters required by different noise power spectrum estimation algorithms

Method	MARTIN[7]	IMCRA[5]	MCRA2[9]	MMSE1[10]	MMSE2[11]	ML-ME[12]	E-DATE
Parameters number	7	10	7	3	5	3	2

is known and can be substituted to $\sigma_X^2(m,k)$ in (8). This is no longer the case for non-stationary noise involved in the NOIZEUS database. For non stationary noise, the reference noise power spectrum $\sigma_X^2(m,k)$ is estimated as follows [25]:

$$\sigma_X^2(m,k) = \alpha\sigma_X^2(m-1,k) + (1-\alpha)|X(m,k)|^2, \text{with } \alpha = 0.9.$$

Both the B-E-DATE and the SW-E-DATE implementations of the E-DATE algorithm were evaluated and compared. The SW-E-DATE uses the recently-introduced MMSE2 method [11] as a surrogate algorithm to provide an estimate for the first $D-1$ frames since, as shown below, this algorithm turns out to offer excellent performance among state-of-the-art noise estimators.

The $LogErr$ measures obtained with the different noise power spectrum estimators are given in Figure 5. All algorithms have been benchmarked at four SNR levels and against various noise models, namely WGN, auto-regressive (AR) colored stationary noise, and 6 typical non-stationary noise environments.

The results for white and colored stationary noise are given in Figures. 5(a) and 5(b), respectively. The B-E-DATE and SW-E-DATE methods yield the lowest $LogErr$ error, the best performance being achieved by the B-E-DATE algorithm in WGN. This is no surprise since the underlying DATE algorithm was originally developed for estimating the standard deviation of additive WGN.

For non-stationary noise with slowly-varying noise spectrum like exhibition, car, station or train noise, and depending on the noise level, the B-E-DATE algorithm uniformly obtains either the best score, or comes very close to the best score, as shown in Figures 5(c), 5(d) and 5(e), respectively.

Figures 5(f), 5(g) and 5(h) present the results obtained with the least favorable types of non-stationary noise. In the case of modulate WGN (resp. babble noise), the SW-E-DATE (resp. B-E-DATE) algorithm yields the smallest $LogErr$ error. As illustrated in Figure 5(h), the two proposed algorithms are among the best in estimating the very challenging airport noise environment. Their performance closely match those obtained with the state-of-the-art MMSE2 and ML-EM estimators.

C. Performance Evaluation in Speech Enhancement

In complement to the previous study, the performance of the noise power spectrum estimation algorithms listed in Table I have also been evaluated and compared in combination with a noise reduction system. The speech denoising experiments are based on the NOIZEUS database [2], which contains IEEE sentences corrupted by eight types of noise coming from the AURORA noise database, at four SNR levels, namely 0, 5, 10 and 15 dB. The noise reduction algorithm retained for our experiments is the Log-MMSE estimator [26]. This method is a standard reference in speech denoising. It can easily be implemented and is known to reduce residual noise without distorting too much the speech signal [2, p. 230, Sec. 7.7].

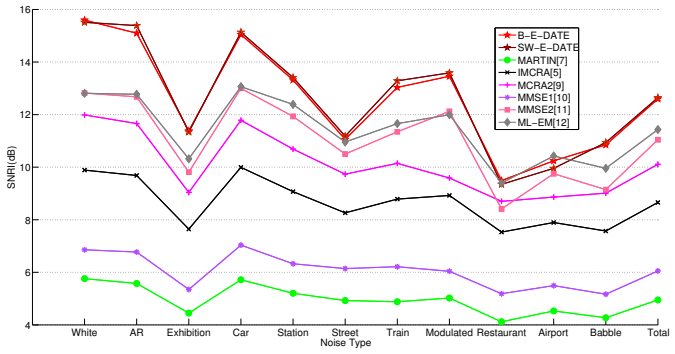


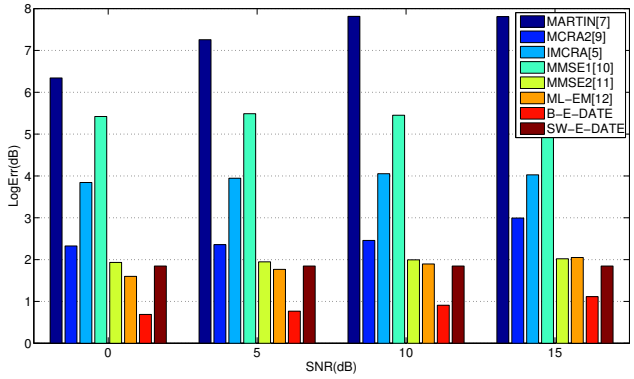
Fig. 6: SNRI with various noise types

Two different criteria have been used to compare the different algorithms. The first one is the Signal-to-Noise Ratio Improvement (SNRI) objective criterion standardized in the ITU-T G.160 recommendation for evaluating noise reduction systems [27]. The SNRI performance obtained with the Log-MMSE combined with the noise power spectrum estimators of Table I are shown in Figure 6 for various noise environments. Note that 4 noise levels were used for each noise type, the final SNRI score being computed as the average score over these 4 levels. We observe that the B-E-DATE and SW-E-DATE yield similar performance measurements and that they outperform all other methods for each type of noise except airport noise. The average SNRI score computed over the 11 noise types and labeled *Total* at the right of Figure 6 clearly emphasizes the SNRI gain brought by the E-DATE in comparison to other methods.

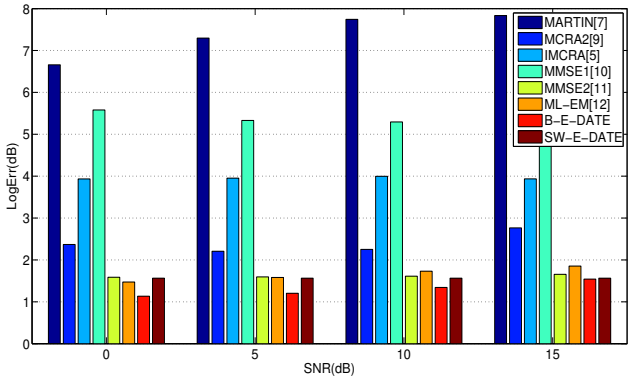
The second criterion used to assess noise power spectrum estimation in speech enhancement is the composite objective measures proposed in [28] (see also [2]). This criterion introduces three measures C_{sig}, C_{bak} and C_{ovl} that are linear combination of some widely used measures like segmental SNR (segSNR), weighted-slope spectral (WSS), log likelihood ratio (LLR), and perceptual evaluation of speech quality (PESQ):

$$\begin{cases} C_{sig} = 3.093 - 1.029\,\text{LLR} 0.603\,\text{PESQ} - 0.009\,\text{WSS} \\ C_{bak} = 1.634 + 0.478\,\text{PESQ} - 0.00\,\text{WSS} + 0.063\,\text{segSNR} \\ C_{vol} = 1.594 + 0.805\,\text{PESQ} - 0.512\,\text{LRR} - 0.007.\text{WSS} \end{cases}$$

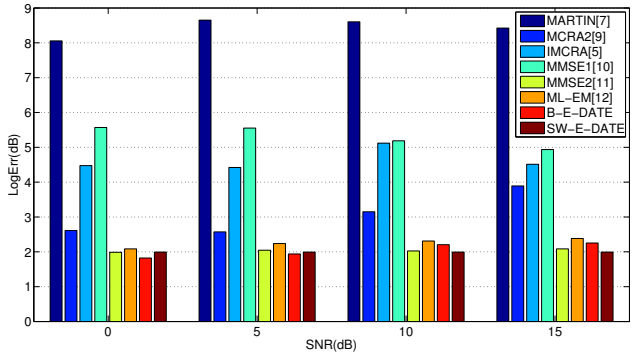
The three measures C_{sig}, C_{bak} and C_{ovl} are designed so as to provide a high correlation with the three usual corresponding subjective measures that are signal distortion (SIG), background intrusiveness (BAK) and Mean Opinion Score (OVRL). We focus here on the C_{ovl} criterion since it has the highest correlation with the real subjective tests.



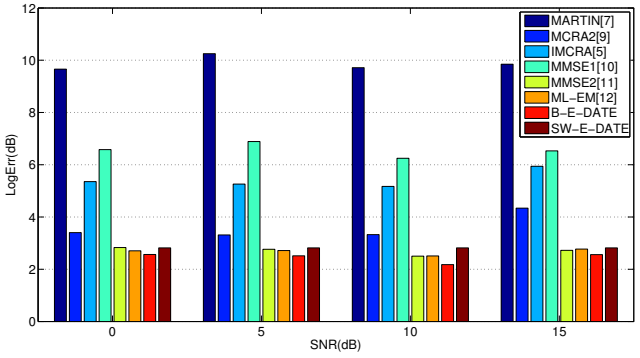
(a) WGN



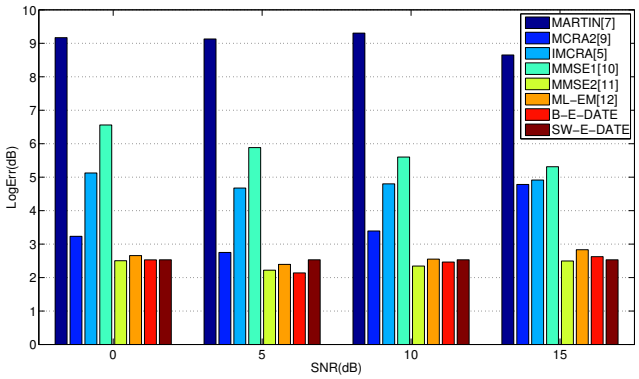
(b) AR noise



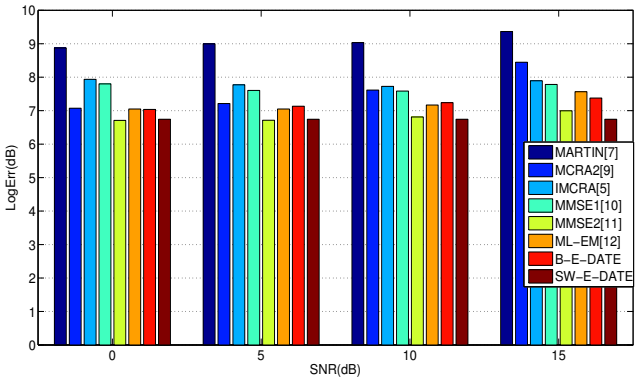
(c) car noise



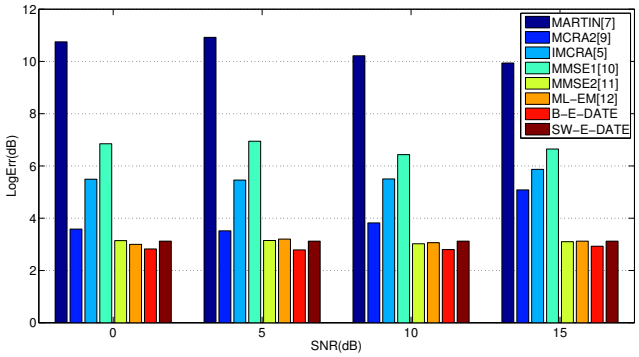
(d) train noise



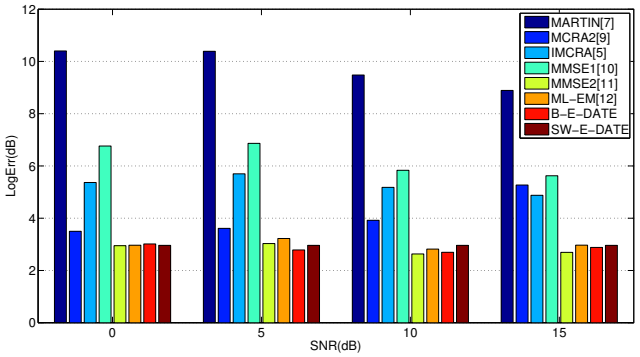
(e) station noise



(f) modulated WGN

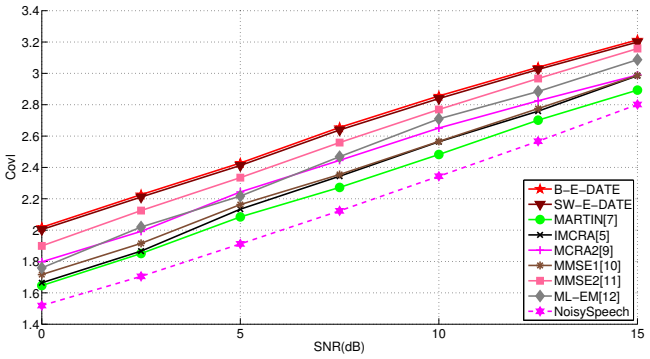


(g) babble noise

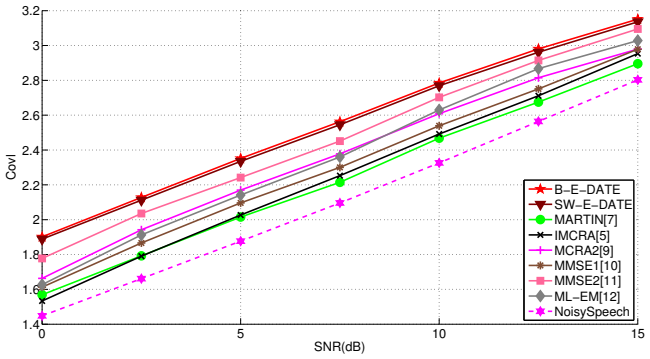


(h) airport noise

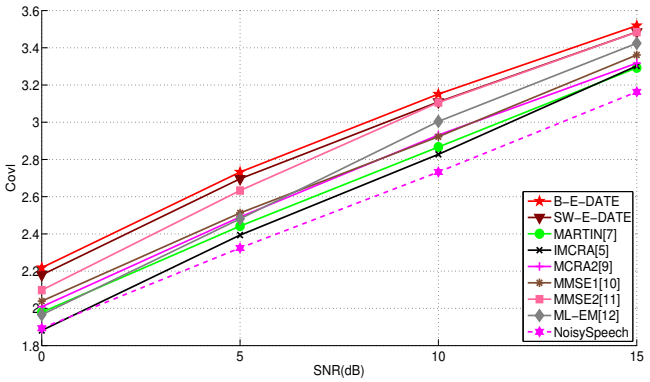
Fig. 5: Noise estimation quality comparison of several noise power spectrum estimators at different SNR levels and with different kind of noise.



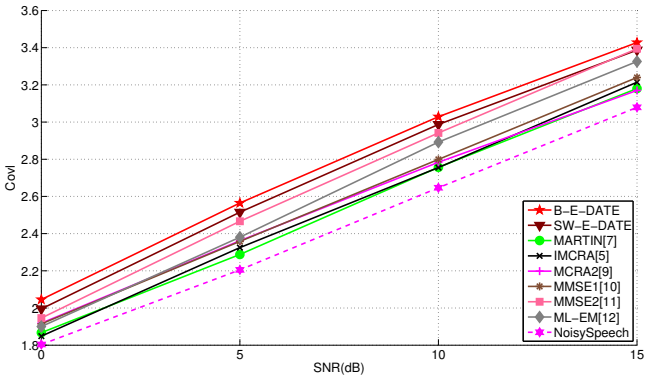
(a) WGN



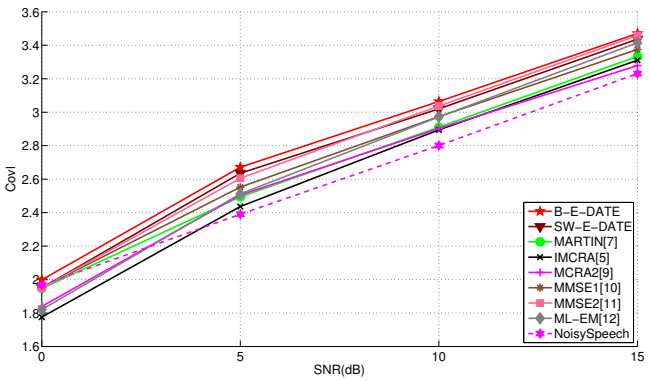
(b) AR noise



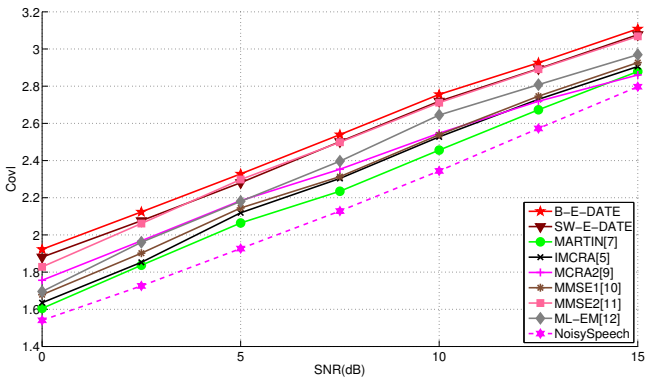
(c) car noise



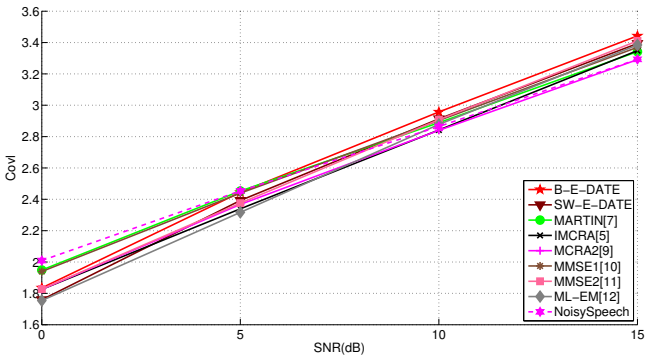
(d) train noise



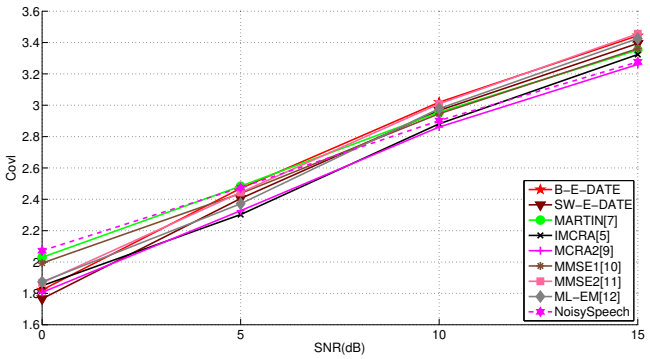
(e) station noise



(f) modulated WGN

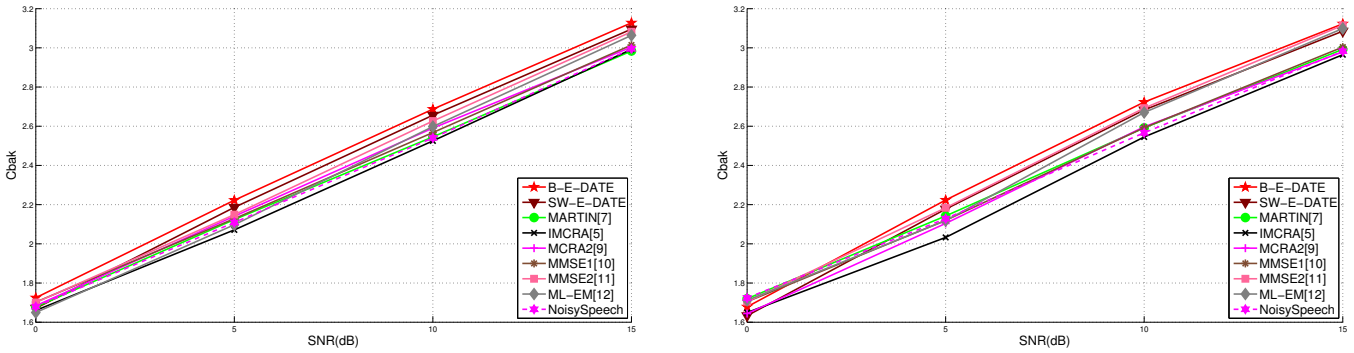


(g) babble noise



(h) airport noise

Fig. 7: Speech quality evaluation after speech denoising (C_{ovl} composite criterion).



(a) babble noise

(b) airport noise

Fig. 8: Speech quality evaluation after speech denoising (C_{bak} composite criterion).

Figure 7 shows the C_{ovl} scores obtained with the different noise power spectrum estimators and noise environments. For reference purpose, the C_{ovl} score obtained with noisy speech but without noise reduction is shown in dashed lines in each sub-figure. The good performance of the B-E-DATE and SW-E-DATE are confirmed by the C_{ovl} measures obtained in the case of WGN, AR noise, car noise, station noise and train noise. These results allow us to conclude that the E-DATE approach is well-suited for stationary or slowly varying non-stationary noise. Although not shown here for space limitation, we hasten to mention that very similar trends were observed for the other two criteria C_{sig} and C_{bak}. In the challenging case of airport noise, all the methods in this paper introduce a large signal distortion at 0dB and 5 dB. At 10 and 15 dB, the E-DATE C_{ovl} scores are similar to that obtained by the other methods (see Figure 7(h)). A detailed analysis of the C_{bak} scores in babble and airport noise (see Figure 8) nevertheless reveals that the E-DATE algorithms perform best in terms of background noise reduction. Two final remarks are in order here. First, the B-E-DATE algorithm generally performs better than the SW-E-DATE algorithm. This is particularly evident in Figure 7 and can also be noticed in the other experimental results. This is mainly due to the fact that our implementation of the SW-E-DATE initially resorts to a surrogate algorithm to estimate noise power spectrum during the first $D = 80$ frames, which has inferior performance to the B-E-DATE. Since these D frames represent a significant part of the total duration of many of the tested utterances, the performance loss incurred by the use of a worse estimator significantly impacts the overall score. Second, in the previous section was evoked the possibility to partially update the buffer by renewing only L frames among D instead of flushing it completely (B-E-DATE), or renewing it only one frame at a time in a FIFO manner (SW-E-DATE). The difference in performance between these two E-DATE implementations suggests that such a partial renewal should not dramatically modify the results. This means that buffer optimization can be performed in practice whenever required by practical constraints, and without significantly impacting the denoising performance. For instance, additional experimental results with airport, babble, station, car and train noises suggest that D can be chosen in the range $[50, 80]$ without really affecting C_{ovl} for SNR $>$ 0dB.

TABLE IV: Computational cost of MMSE2 per new frame and per frequency bin

Addition	Multiplication	Division	Exponent
12	10	2	1

D. Complexity analysis

Tables II and III compare the computational costs of the B-E-DATE and SW-E-DATE implementations, respectively. Each table gives the number of real additions, multiplications, divisions and square roots required to perform the estimate. Both the B-E-DATE and the SW-E-DATE use D frames to compute the noise power spectrum estimate. However computation is performed only once every D frames for the B-E-DATE algorithm, whereas it is performed once per frame in the SW-E-DATE implementation. Hence the number of operations in Table II should be divided by D to allow for a fair per-frame computational cost comparison between the two implementations. For reference purpose, Table IV lists the number of operations required by the MMSE2 estimator of [11]. Inspection of Tables II and IV shows that the B-E-DATE and MMSE2 estimators have similar computational complexity. This is confirmed by execution times of Matlab implementations of these algorithms where the B-E-DATE algorithm is found to have a processing time about 1.53 times that of the MMSE2 algorithm. We also note from Tables II and III that SW-E-DATE requires approximately $D/3$ times more operations that B-E-DATE. Indeed, B-E-DATE requires $3D$ multiplications to process D frames at once, whereas SW-E-DATE requires $D+2$ multiplications per frame. Execution times of Matlab implementations of these algorithms also confirm this ratio.

TABLE II: Computational cost of B-E-DATE per group of D frames and per frequency bin

	Addition	Multiplication	Division	Square root
Norm	D	$2D$	0	D
Sorting	$D\log D$	0	0	0
Search n^*(worst case)	$D(D-1)/2$	D	D	0
Total	$D\left(\log D+(D+1)/2\right)$	$3D$	D	D

TABLE III: Computational cost of SW-E-DATE per new frame and per frequency bin

	Addition	Multiplication	Division	Square root
Norm	1	2	0	1
Sorting	$\log D$	0	0	0
Search n^*(worst case)	$D(D-1)/2$	D	D	0
Total	$1+\log D+D(D-1)/2$	$D+2$	D	1

VI. Conclusion

In this paper, we have proposed a novel method to estimate the power spectrum of some non-stationary noise, in applications where a weak-sparse transform makes it possible to represent the signal of interest by a relatively small number of coefficients with significantly large amplitude. The resulting estimator called Extended-DATE (E-DATE) is robust in that it does not use prior knowledge about the signal or the noise except for the weak-spareness property. Compared to other methods in the literature, the E-DATE algorithm has the remarkable advantage of requiring only two parameters to specify. A straightforward block-based implementation of the E-DATE, called B-E-DATE, has first been introduced. This implementation entails an estimation delay, which diminishes as the frequency rate increases. This delay could be reduced by grouping frequency bins. Another solution to shorten this delay involves resorting to a sliding-window implementation called SW-E-DATE, but at the price of a higher computational cost. The B-E-DATE and SW-E-DATE have been benchmarked against various classical and recent noise power spectrum estimation methods in two situations: with and without noise reduction. The experimental results show that the E-DATE estimator generally provides the most accurate noise estimate, and that it outperforms other methods for speech denoising in the presence of various noise types and levels. For its good performance and low complexity, the B-E-DATE should be preferred in practice when frequency rates are high enough to induce acceptable or even negligible time-delay.

Although the present paper focused on noise reduction in speech enhancement systems, it must be emphasized that the E-DATE estimator is not restricted to speech signals and could find other applications in any scenario where noisy signals have a weakly-sparse representation. For many signals of interest, not limited to speech, such a weakly-sparse representation can be provided by an appropriate wavelet transform. In this respect, the application of the E-DATE algorithm to audio separation could be considered in continuation of [18], [29], [30], [31].

The E-DATE estimator fundamentally relies on the DATE estimator which, as emphasized in [1], can be regarded as an outlier detector. Consequently the E-DATE can also be used as an outlier detector in each frequency bin. This opens interesting perspectives in voice activity detection based on frequency analysis as well as in the detection and estimation of chirp signals in various types of noise.

References

[1] D. Pastor and F. Socheleau, "Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurrences," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1545–1555, Apr. 2012.

[2] P. C. Loizou, *Speech enhancement: theory and practice*. New York: CRC Press, 2013.

[3] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Detroit, Michigan, USA, May 1995, pp. 153–156.

[4] B. Ahmed and W. H. Holmes, "A voice activity detector using the chi-square test," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Montreal, Quebec, Canada, 2004, pp. I–625.

[5] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[6] R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedins of the European Signal Processing Conference (EUSIPCO)*, 1994, pp. 1182–1185.

[7] ——, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[8] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[9] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *ELSEVIER Speech communications*, vol. 48, no. 2, pp. 220–231, Feb. 2006.

[10] R. Yu, "A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction," in *IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 4421–4424.

[11] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[12] M. Souden, M. Delcroix, K. Kinoshita, T. Yoshioka, and T. Nakatani, "Noise power spectral density tracking: A maximum likelihood perspective," *IEEE Signal Processing Letters*, vol. 19, no. 8, pp. 495–498, Aug. 2012.

[13] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and wiener filtering," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2000, pp. 1875–1878 vol.3.

[14] P. Davies and U. Gather, "The identification of multiple outliers (with discussion)," *J. Amer. Statist. Assoc.*, no. 423, pp. 782 – 801, 1993.

[15] N. N. Lebedev, *Special Functions and their Applications*. Prentice-Hall, Englewood Cliffs, 1965.

[16] D. Pastor, "A theoretical result for processing signals that have unknown distributions and priors in white gaussian noise," *Computational Statistics & Data Analysis, CSDA*, vol. 52, no. 6, pp. 3167 – 3186, 2008.
[17] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
[18] S. M. Aziz Sbai, A. Aïssa-El-Bey, and D. Pastor, "Contribution of statistical tests to sparseness-based blind source separation," *EURASIP journal on applied signal processing*, Jul. 2012.
[19] S. M. Berman, *Sojourns and extremes of stochastic processes*. Wadsworth, Reading, MA, January 1992.
[20] S. Mallat, *A wavelet tour of signal processing, second edition*. Academic Press, 1999.
[21] R. J. Serfling, *Approximations theorems of mathematical statistics*. Wiley, 1980.
[22] A. M. Atto, D. Pastor, and G. Mercier, "Detection thresholds for non-parametric estimation," *Signal, Image and Video processing*, vol. 2, no. 3, pp. 207–223, February 2008.
[23] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.
[24] D. Pastor and A. M. Atto, *Wavelet shrinkage: from sparsity and robust testing to smooth adaptation; In Fractals and Related Fields, Eds: J. Barral & S. Seuret*. Birkhaüser, 2010.
[25] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 541–553, Mar. 2008.
[26] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
[27] "ITU recommendation, G. 160," *Voice Enhancement Devices for Mobile Networks*, 2005.
[28] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement." in *Proc. Interspeech*, 2006, pp. 1447–1450.
[29] F.-X. Socheleau, D. Pastor, and A. Aïssa-El-Bey, "Robust statistics based noise variance estimation: Application to wideband interception of noncooperative communications," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 47, no. 1, pp. 746–755, January 2011.
[30] S. M. Aziz Sbai, A. Aïssa-El-Bey, and D. Pastor, "Robust underdetermined blind audio source separation of sparse signals in the time-frequency domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 3716–3719.
[31] A. Aïssa-El-Bey and K. Abed-Meraim, "Blind identification of sparse simo channels using maximum a posteriori approach," in *16th European Signal Processing Conference EUSIPCO*, August 2008, pp. 1–5.

PLACE PHOTO HERE

Van-Khanh Mai was born in Vietnam in 1987. He received the engineer degree in electronic and information from Hanoi University of Technology, Hanoi,Vietnam and the Research Master degree in electronics and telecommunications from the Rennes I University, Rennes, France, in 2013. He is currently a Ph.D. student in signal and communication at the Signal & Communications department of Telecom Bretagne. His research interests include audio signal processing, noise reduction and speech enhancement.

Dominique Pastor was born in Cahors, France, in 1963. He graduated from Telecom Bretagne (Brest, France) in 1986 and from the University of Rennes (France) in 1997 (Ph.D.). From 1987 until 2000, he was with Thales. In particular, between 1990 and 1998, he was with Thales Avionics where his research concerned speech processing for applications to speech recognition systems embedded in military fast jet cockpits and, from 1998 to 2000, he was with Thales Nederland where he worked on the detection of radar targets in sea clutter. In September 2000, he joined Altran Technologies Nederland as a senior consultant. Since September 2002, he is with Institut Telecom, where he is currently Professor at Telecom Bretagne. His current research interests focus on statistical signal processing and sparse transforms with applications to physiological signals including speech.

Abdeldjalil Aïssa-El-Bey (M'07, SM'12) was born in Algiers, Algeria, in 1981. He received the State Engineering degree from École Nationale Polytechnique (ENP), Algiers, Algeria, in 2003, the M.S. Degree in signal processing from Supélec and Paris XI University, Orsay, France, in 2004 and the Ph.D. degree in signal and image processing from Telecom ParisTech Paris, France in 2007. He is currently and since 2007 Associate Professor at Signal & Communications department of Telecom Bretagne. His research interests are blind source separation, blind system identification and equalization, statistical signal processing, wireless communications, and adaptive filtering.

Raphaël Le-Bidan (M'03) was born in Fontenay-Le-Comte, France, in 1977. He received the Eng. Degree in Telecommunications and the M. Sc. Degree in Electrical Eng. from the Institut National des Sciences Appliquées (INSA), Rennes, France, in June 2000, and the Ph. D. degree in Electrical Eng. from the INSA, Rennes, in November 2003. Since December 2003, he is working as an Associate Professor at Telecom Bretagne, in the Signal & Communication department. His research interests are in the area of Communication Theory and Information Theory, with an emphasis on coding theory, sparse graph codes and iterative decoding algorithms, energy-efficient communications, and digital transmission systems design. Recent research interests also include advanced noise cancellation and speech processing techniques for mobile voice communications.