HAL
open science

From non parametric statistics to speech denoising

Dominique Pastor, Asmaa Amehraye

To cite this version:

Dominique Pastor, Asmaa Amehraye. From non parametric statistics to speech denoising. ISIVC 2006 : 3d international symposium on image/video communications over fixed and mobile networks, Sep 2006, Hammamet, Tunisia. hal-02136901

HAL Id: hal-02136901

https://hal.science/hal-02136901v1

Submitted on 22 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FROM NON-PARAMETRIC STATISTICS TO SPEECH DENOISING

Dominique Pastor[1], *Asmaa Amehraye*[1],[2]

[1] GET - Ecole Nationale Supérieure des Télécommunications de Bretagne
Technopôle de Brest Iroise, 29238 BREST Cedex, France,
{dominique.pastor,asmaa.amehraye}@enst-bretagne.fr
[2] GSCM-LEESA Faculté des Sciences de Rabat,
4 Avenue Ibn Battouta B.P. 1014 RP, Rabat, Maroc

ABSTRACT

Given some signal additively corrupted by independent white Gaussian noise with unknown standard deviation σ, we present a new estimator of σ. This estimator derives from a theoretical result presented and commented in the paper. Without any preliminary signal detection, the estimate is performed on the basis of the time-frequency components returned by a standard spectrogram where the Discrete Fourier Transform is simply weighted by the square window. No assumption about the signal statistics is made. The signal time-frequency components are assumed to have probabilities of presence less than or equal to one half.

This estimator is suited to speech denoising. It avoids the use of any Voice Activity Detector and is an alternative solution to subspace approaches. Objective performance measurements show that the standard Wiener filtering of speech signals can be tuned with the outcome of this estimator without a significant loss in comparison with the measurements obtained when the noise standard deviation is known.

1. MOTIVATION

Let $s[t]$, $t = 1, \ldots, T$ be the samples of some speech signal and suppose that these T samples are corrupted by additive and independent stationary noise $x[t]$, $t = 1, 2, \ldots, T$ so that the samples of the observed signal are

$$y[t] = s[t] + x[t], t = 1, \ldots, T. \tag{1}$$

We assume that noise is white and Gaussian with null mean and standard deviation σ: for every $t \in \{1, 2, \ldots, T\}$, $x[t] \sim \mathcal{N}(0, \sigma^2)$.

The Wiener filtering of the noisy speech signal y requires prior knowledge of the noise standard deviation σ. A basic and popular solution consists in using a Voice Activity Detector (VAD): the estimate of σ is the square root of the Maximum Likelihood Estimate (MLE) computed on the basis of the samples of the time frames that the VAD has detected as noise alone. Subspace approaches can also be used to estimate σ by computing the smallest eigenvalues of the noisy speech autocorrelation matrix; the model order is difficult to choose and the computation of the eigenvalues may prove unstable.

This paper proposes a new estimator of the noise standard deviation. The theoretical foundation of this estimator is proposition 3.1, stated in section 3 after exposing some preliminary material in section 2. This theoretical result is non-parametric in the sense that it makes no assumption about the probability distributions of the signals and assumes neither that these signals are identically distributed nor that they have equal probabilities of presence.

Section 4 then presents the estimator deriving from proposition 3.1 and a few preliminary experimental results. This estimator is then employed in section 5 to adjust the standard Wiener filtering of the noisy speech signal y introduced above. According to objective performance measurements, the denoised speech signals are not significantly more distorted than those achieved when the filtering is tuned with the exact value of the noise standard deviation.

Conclusions and perspectives are given in section 6.

2. PRELIMINARY MATERIAL

The random vectors and variables are supposed to be defined on the same probability space denoted by $(\Omega, \mathcal{M}, P)$ and for every element $\omega \in \Omega$. As usual, if property $\mathcal{P}$ holds true almost surely, we write $\mathcal{P}$ (a-s).

Given a positive real value σ, a sequence $X = (X_k)_{k \in \mathbf{N}}$ of random complex variables is said to be a *complex white Gaussian noise* (CWGN) with standard deviation σ if the random variables X_k, $k = 1, 2, \ldots$, are complex, mutually independent and identically Gaussian distributed with null mean and variance σ^2. The real and imaginary parts $\Re X_k$ and $\Im X_k$ of each X_k form a two-dimensional random vector such that $(\Re X_k \Im X_k) \sim \mathcal{N}\left(0, (\sigma^2/2)\mathbf{I}\right)$ where $\mathbf{I}$ stands for the 2×2 identity matrix.

The *minimum amplitude* $\mathsf{a}(S)$ of a sequence $S = (S_k)_{k \in \mathbf{N}}$ of random complex variables is defined by

$$\mathsf{a}(S) = \sup\left\{\alpha \in [0, \infty] : \forall k \in \mathbf{N}, |S_k| \geq \alpha \text{ (a-s)}\right\}. \tag{2}$$

If f is some map of the set of all the sequences of complex random variables into $\mathbf{R}$, we say that the limit of f is $\ell \in \mathbf{R}$ when $\mathsf{a}(S)$ tends to ∞ and write that $\lim_{\mathsf{a}(S) \to \infty} f(S) = \ell$ if, for any positive real value η, there exists some $A_0 \in (0, \infty)$ such that, for every $A \geq A_0$ and every S such that $\mathsf{a}(S) \geq A$, $|f(S) - \ell| \leq \eta$.

The set $L^2(\Omega, \mathbf{C})$ stands for the set of those complex random variables Y such that $E[|Y|^2] < \infty$. We then define $\ell^\infty(\mathbf{N}, L^2(\Omega, \mathbf{C}))$ as the set of those sequences $S = (S_k)_{k\in\mathbf{N}}$ of complex random variables such that $S_k \in L^2(\Omega, \mathbf{C})$ for every $k \in \mathbf{N}$ and $\sup_{k\in\mathbf{N}} E[|S_k|^2]$ is finite.

Given a random variable Y and a real number τ, $\mathcal{I}(Y \leq \tau)$ stands for the indicator function of the event $\{Y \leq \tau\}$. As usual, I_0 is the zeroth-order modified Bessel function of the first kind. Throughout the rest of the text, we say 'independent' instead of 'mutually independent' for brevity.

3. A THEORETICAL RESULT

Proposition 3.1 stated below is a corollary of a more general theorem established in [6]. As an introduction to proposition 3.1, we begin with an intuitive approach. It makes the reader understand the main ideas behind proposition 3.1 and the results given in [6].

Consider a sequence $Y = (Y_k)_{k\in\mathbf{N}}$ of complex random variables where each Y_k is either the sum of some signal S_k and noise X_k or noise X_k alone. We assume that $X = (X_k)_{k\in\mathbf{N}}$ is a CWGN with standard deviation σ. For every given $k \in \mathbf{N}$, the presence of noise alone is the null hypothesis whereas the presence of some signal in noise is the alternative one. We assume that, for every $k \in \mathbf{N}$, the index of the true hypothesis is a random variable ε, valued in $\{0, 1\}$ and independent with S_k and X_k. We thus can write that $Y_k = \varepsilon_k S_k + X_k$. The *a priori* probabilities of presence and absence of the signal Λ_k are then $P(\{\varepsilon_k = 1\})$ and $P(\{\varepsilon_k = 0\})$, respectively. Proposition 3.1 significantly reduces the importance of the choice of these probabilities since they will be assumed to be upper-bounded.

At this stage, assume that the random variables ε_k, $k \in \mathbf{N}$, are independent and identically distributed (iid) as well as the random signals S_k, $k \in \mathbf{N}$. It follows that the random vectors Y_k, $k \in \mathbf{N}$, are iid as well. Let p stand for the common value of the probabilities of presence $P(\{\varepsilon_k = 1\})$. We assume that $p \leq 1/2$.

Given some real number T, set

$$A_m(T) = \frac{1}{m}\sum_{k=1}^{m} |Y_k|\mathcal{I}(|Y_k| \leq T)$$

According to Kolmogorov's classical strong limit theorem,

$$\lim_{m\to\infty} A_m(T) = E\left[|Y_k|\mathcal{I}(|Y_k| \leq T)\right] \text{ (a-s)} \qquad (3)$$

where k is any element of $\{1, \ldots, m\}$. An easy computation shows that we can write that

$$\begin{aligned} &E\left[|Y_k|\mathcal{I}(|Y_k| \leq T)\right] \\ &\quad = (1-p)E\left[|X_k|\mathcal{I}(|X_k| \leq T)\right] \times \\ &\quad \left(1 + \frac{p}{1-p}\frac{E\left[|S_k+X_k|\mathcal{I}(|S_k+X_k| \leq T)\right]}{E\left[|X_k|\mathcal{I}(|X_k| \leq T)\right]}\right). \end{aligned} \qquad (4)$$

Let A be a lower bound for the amplitudes of the signals S_k, $k \in \mathbf{N}$. If A is large enough in comparison with σ, we can reasonably expect the existence of some threshold T that makes it possible to distinguish noisy signals from noise alone with a rather small probability of error. As a matter of fact, regarding the choice for T, we can be very specific as follows.

For any given non negative real number h, let $\mathcal{T}_\tau$ stand for the map defined for every complex value z by

$$\mathcal{T}_\tau(z) = \begin{cases} 1 & \text{if } |z| \geq \tau \\ 0 & \text{otherwise.} \end{cases} \qquad (5)$$

Clearly, for every $k \in \mathbf{N}$, $\mathcal{T}_\tau$ is a statistical test for making a decision on the value of ε_k since the composite map $\mathcal{T}_\tau(U_k)$ is measurable and $\mathcal{T}_\tau(U_k) \in \{0, 1\}$. In what follows, $\mathcal{T}_\tau$ is called the thresholding test with threshold height τ. The error probability of this test is then the probability $P\left(\{\mathcal{T}_\tau(U_k) \neq \varepsilon_k\}\right)$ of the event $\{\mathcal{T}_\tau(U_k) \neq \varepsilon_k\}$. This error probability does not depend on k since the observations U_k, $k \in \mathbf{N}$, are assumed to be iid. This is the reason why we simply denote it by $P_e\{\mathcal{T}_\tau\}$, without mentioning the observation under consideration.

Since the probability of presence of any signal S_k, $k \in n\mathbf{N}$, is assumed to be less than or equal to $1/2$ and by taking into account that (see [1, Eq. 9.6.47, p. 377]) $I_0(x) = {}_0F_1(1; x^2/4)$, [4, Theorem VII.1] tells us the following. For every $x \in \mathbf{R}$, set

$$\kappa(x) = I_0^{-1}(e^{x^2})/2x \qquad (6)$$

with $\kappa(0) = 1$; for making a decision on the value of ε_k where k is any natural number, the error probability $P_e\{\mathcal{T}_{\sigma\kappa(A/\sigma)}\}$ of the thresholding test $\mathcal{T}_{\sigma\kappa(A/\sigma)}$ with threshold height $\sigma\kappa(A/\sigma)$ is less than or equal to $\mathcal{Q}(A/\sigma)$ where, for any given non-negative real number x,

$$\mathcal{Q}(x) = e^{-x^2}\int_0^{\kappa(x)} e^{-t^2} t I_0(2xt)dt + \frac{1}{2}e^{-\kappa(x)^2}. \qquad (7)$$

We thus can write that $P_e\{\mathcal{T}_{\sigma\kappa(A/\sigma)}\} \leq \mathcal{Q}(A/\sigma)$. In equation (4), set now $T = \sigma h$ with $h = \kappa(A/\sigma)$. The function $\mathcal{Q}(x)$ decreases very rapidly when x increases. Hence, for large values of A, the probabilities $P(\{|X_k| > \sigma h\})$ and $P(\{|S_k + X_k| \leq \sigma h\})$ are small and, thus, the expectation $E\left[|S_k + X_k|\mathcal{I}(|S_k + X_k| \leq \sigma h)\right]$ can reasonably be expected to be significantly smaller than $E\left[|X_k|\mathcal{I}(|X_k| \leq \sigma h)\right]$. Since p is assumed to be less than or equal to one half, $p/(1-p)$ is less than or equal to 1. Consequently, in a certain sense to specify, we should be able to prove that

$$E\left[|Y_k|\mathcal{I}(|Y_k| \leq \sigma h)\right] \approx (1-p)E\left[|X_k|\mathcal{I}(|X_k| \leq \sigma h)\right].$$

Without caring about mathematical exactness, we combine this approximation to the almost surely convergence of equation (3) to obtain that, in a certain sense,

$$A_m(\sigma h) \approx (1-p)E\left[|X_k|\mathcal{I}(|X_k| \leq \sigma h)\right] \qquad (8)$$

when m and the amplitudes of the signals are both large.

If we now set $B_m(T) = \frac{1}{m}\sum_{k=1}^{m}\mathcal{I}(|Y_k| \leq \sigma h)$, the same type of intuitive approach suggests that

$$B_m(\sigma h) \approx (1-p)E\left[\mathcal{I}(|X_k| \leq \sigma h)\right]. \qquad (9)$$

Consider now the ratio $A_m(\sigma h)/B_m(\sigma h)$. This ratio makes it possible to get rid of the unknown prior p.

Moreover, since $X_k \sim \mathcal{N}_c(0, \sigma^2)$, the distribution of $|X_k|$ is known and its density $f(x)$ is that of the square of a Rayleigh distributed variable. Taking into account that the variance of the real and imaginary parts of X_k both equal $\sigma^2/2$, this density is given by :

$$f(x) = \begin{cases} (2x/\sigma^2)e^{-x^2/\sigma^2} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Therefore, we easily obtain that $E\left[|X_k|\mathcal{I}(|X_k| \leq \sigma h)\right] = 2\sigma \int_0^h t^2 e^{-t^2} dt$ and that $E\left[\mathcal{I}(|X_k| \leq \sigma h)\right] = P(\{X_k \leq \sigma h\}) = 1 - e^{-h^2}$. According to these equalities, equations (3) and (8), we conclude that

$$A_m(\sigma h)/B_m(\sigma h) \approx 2\sigma \int_0^h t^2 e^{-t^2} dt/(1 - e^{-h^2}). \tag{11}$$

Once again, this approximation must be understood with respect to a certain convergence criterion. This one is introduced in proposition 3.1. Its more general form is given in [6]. As a matter of fact, the same type of intuitive approach as that presented above can be used to guess part of the results established in [6]. Proposition 3.1 and its extension not only specify the exact meaning of equation (11) but also significantly extend the conditions of validity of (11) because they state that the convergence holds true even for non iid signals and non iid priors.

Proposition 3.1 *Let $Y = (Y_k)_{k\in\mathbf{N}}$ be a sequence of complex random variables such that, for every $k \in \mathbf{N}$, $Y_k = \varepsilon_k S_k + X_k$ where $S = (S_k) \in \ell^\infty(\mathbf{N}, L^2(\Omega, \mathbf{C}))$, $X = (X_k)_{k\in\mathbf{N}}$ is a CWGN with standard deviation σ and $\varepsilon = (\varepsilon_k)_{k\in\mathbf{N}}$ is a sequence of random variables valued in $\{0, 1\}$ respectively.*

Assume that

(A1) *for every $k \in \mathbf{N}$, S_k, X_k and ε_k are independent;*

(A2) *the random variables Y_k, $k \in \mathbf{N}$, are independent;*

(A3) *the random variables ε_k, $k \in \mathbf{N}$, are independent;*

(A4) *the priors $P(\{\varepsilon_k = 1\}), k \in \mathbf{N}$, are less than or equal to one half.*

Given any natural number m and any pair (x, T) of positive real numbers, define the random variable $D_m(x, T)$ by

$$D_m(x,T) = \left| \frac{\sum_{k=1}^m |Y_k|\mathcal{I}(|Y_k| \leq xT)}{\sum_{k=1}^m \mathcal{I}(|Y_k| \leq xT)} - 2x\frac{\int_0^T u^2 e^{-u^2} du}{1 - e^{-T^2}} \right|.$$

Then, the standard deviation σ is the unique positive real number x such that, for every $\beta_0 \in (0, 1]$,

$$\lim_{\mathsf{a}(S)\to\infty} \left\| \overline{\lim_m} D_m(x, \beta\kappa(\mathsf{a}(S)/x)) \right\|_\infty = 0 \tag{12}$$

uniformly in $\beta \in [\beta_0, 1]$ where, for every $x \in \mathbf{R}$,

$$\kappa(x) = I_0^{-1}(e^{x^2})/2x \tag{13}$$

with $\kappa(0) = 1$.

4. A NEW ALGORITHM FOR ESTIMATING THE NOISE STANDARD DEVIATION

On the basis of proposition 3.1, we start by introducing a discrete cost. A minimum of this discrete cost can be computed and considered as a first estimate of the noise standard deviation. This estimate will be called the Essential Supremum Estimate of type I (ESE-I) because of the crucial role played by the essential supremum norm in its computation. The term ESE-I will also stand for the estimator itself.

Experimental results aimed at assessing the ESE-I suggest another estimate of the noise standard deviation. This new estimate is hereafter called the Essential Supremum Estimate of type II (ESE-II). The term ESE-II will also designate the estimator itself. According to Monte-Carlo experiments of the same type as those mentioned above, the ESE-II performs better than the ESE-I.

4.1. The ESE-I

Let L be some natural number and set $\beta_\ell = \ell/L$, $\ell = 1, 2, \ldots, L$. Suppose that A is some known lower bound for the amplitudes of the signal. We thus have $\mathsf{a}(S) \geq A$. These new notations are kept hereafter with the same meaning.

Consider m observations $Y_1, Y_2, \ldots, Y_m$. If A and m are large enough, proposition 3.1 suggests estimating the noise standard deviation by a possibly local minimum of

$$\sup_{\ell\in\{1,\ldots,L\}} \{D_m(x, \beta_\ell\kappa(A/x))\} \tag{14}$$

when x ranges over a suitable search interval. However, in practice, no lower bound for the amplitudes of the signals is known. Surprisingly enough since 3.1 states that the larger A the better the estimate, the experimental results presented in [5] and [6] suggest that the asymptotic condition on the minimum amplitude of the signals can be relaxed significantly. Therefore, we consider the trivial lower bound $A = 0$ and the discrete cost we minimize is then

$$\sup_{\ell\in\{1,\ldots,L\}} \left\{ \left| \frac{\sum_{k=1}^m |Y_k|\mathcal{I}(|Y_k| \leq x\beta_\ell)}{\sum_{k=1}^m \mathcal{I}(|Y_k| \leq x\beta_\ell)} - 2x\frac{\int_0^{\beta_\ell} u^2 e^{-u^2} du}{1 - e^{-\beta_\ell^2}} \right| \right\}, \tag{15}$$

which straightforwardly derives from (14) with $A = 0$ and seeing that $\kappa(0) = 1$. Any possibly local minimum $\tilde{\sigma}$ of (15) can be considered as an estimate of the noise standard deviation. Because of the crucial role played by the essential supremum norm in proposition 3.1, $\tilde{\sigma}$ will be called the Essential Supremum Estimate of type I (ESE-I).

To compute the ESE-I, we choose $L = m$ as a reasonable trade-off between the expected accuracy of the estimate and the computational load incurred by the minimization routine. However, a better choice can certainly be thought up. This will be made elsewhere.

The search interval used to compute the estimate is $\left[|Y_{[k_{\min}]}|, |Y_{[m]}|\right]$ where $Y_{[k]}, k = 1, 2, \ldots, m$ stands for the sequence $Y_k, k = 1, 2, \ldots, m$ sorted by increasing modulus, $k_{\min} = m/2 - hm$ and $h = 1/\sqrt{4m(1-Q)}$ where Q is some value in $(0, 1)$, close to 1 but less than or equal to $1 - \frac{m}{4(m/2-1)^2}$. The reasons of this choice for the search interval are given in [5] and [7].

4.2. Accuracy of the ESE-I

Let k be some natural number and $\mathcal{L}_k$ stand for the Minimum-Probability-of-Error (MPE) test ([8, section II.B]) for making a decision on the value of ε_k. The null hypothesis is thus $\varepsilon_k = 0$ and the alternative one is $\varepsilon_k = 1$. For the decision problem under consideration, the likelihood ratio test $\mathcal{L}_k$ guarantees the smallest possible probability of error amongst all the possible binary hypothesis tests.

Given $Y_1, Y_2, \ldots, Y_m$, the test $\mathcal{I}(|\cdot| \leq \tilde{\sigma}\kappa(A/\tilde{\sigma}))$ simply consists in substituting the estimate $\tilde{\sigma}$ to the exact value σ in the expression of $\mathcal{T}_{\sigma\kappa(A/\sigma)}$. It assigns the value 1 to any complex value z whose modulus less than or equal to $\tilde{\sigma}\kappa(A/\tilde{\sigma})$ and 0 otherwise. This test is not, strictly speaking, a thresholding test in the sense given above for its "thresholding height" is the random variable $\tilde{\sigma}\kappa(A/\tilde{\sigma})$. However, with a slight abuse of language, we denote it by $\mathcal{T}_{\tilde{\sigma}\kappa(A/\tilde{\sigma})}$.

If $\tilde{\sigma}$ is a reasonably good estimate of σ, the performance of $\mathcal{T}_{\tilde{\sigma}\kappa(A/\tilde{\sigma})}$ can be expected to approach that of the thresholding test $\mathcal{T}_{\sigma\kappa(A/\sigma)}$. In other words, the use of the estimate $\tilde{\sigma}$ instead of the true value σ should not induce a significant performance loss even when the minimum amplitude A is known, provided, of course, that m is large enough. In particular, when the signals S_k, $k \in \mathbf{N}$, are independent, have their probabilities of presence all equal to $1/2$ and are such that $S_k = Ae^{i\Phi_k}$ where Φ_k is uniformly distributed in $[0, 2\pi]$, the error probability of the test $\mathcal{T}_{\sigma\kappa(A/\sigma)}$ equals $\mathcal{Q}(A/\sigma)$ ([4]); therefore, the error probability of the test $\mathcal{T}_{\tilde{\sigma}\kappa(A/\tilde{\sigma})}$ should be close to $\mathcal{Q}(A/\sigma)$ when A and m are both large. Even though the computation of the error probability of the test $\mathcal{T}_{\tilde{\sigma}\kappa(A/\tilde{\sigma})}$ is an open issue, this intuitive claim can easily be verified via Monte-Carlo simulations aimed at comparing the Binary Error Rate (BER) of this test to $\mathcal{Q}(A/\sigma)$. To achieve this simulations, we follow the standard experimental protocol adopted by practitioners in telecommunication systems.

Fix $\sigma = 1$. We carry out independent trials of m observations each by considering a number J of successive independent random copies of the observations $Y_1, \ldots, Y_m$. These copies are henceforth denoted by $Y_{j,1}, Y_{j,2}, \ldots, Y_{j,m}$, $j = 1, 2, \ldots, J$. Of course, they are constructed by using independent random copies $\varepsilon_{j,k}$, $S_{j,k}$ and $X_{j,k}$ of ε_k, S_k and X_k respectively. For every copy j and every given $k \in \{1, \ldots, m\}$, we thus have $Y_{j,k} = \varepsilon_{j,k}S_{j,k} + X_{j,k}$.

For each $j = 1, 2, \ldots, J$, let $\tilde{\sigma}_j$ be the ESE-I of σ obtained during the jth trial, that is on the basis of $Y_{j,k}, k = 1, \ldots, m$; denote by n_j the number of errors made by the test $\mathcal{T}_{\tilde{\sigma}\kappa(A/\tilde{\sigma}_j)} = \mathcal{I}(|\cdot| \leq \tilde{\sigma}_j\kappa(A/\tilde{\sigma}_j)$ applied to the m observations $Y_{j,k}, k = 1, \ldots, m$.

Since the decision is made on the same observations as those used to estimate the noise standard deviation, the accuracy of the estimate affects m decisions at one go. To reduce this effect, we proceed as practitioners in telecommunication systems usually do by fixing a minimum number $J_{\min}$ of trials to achieve and a minimum number $N_{\min}$ of errors to obtain during the experiments.

Trials are thus carried out until the total number J of trials is larger than or equal to $J_{\min}$ and the total number of errors $\sum_{j=1}^{J} n_j$ obtained after these J trials is larger than or equal to $N_{\min}$. The BER of the test $\mathcal{T}_{\tilde{\sigma}\kappa(A/\tilde{\sigma})}$ is then defined as the ratio $\sum_{j=1}^{J} n_j/(J \times m)$.

All the results presented below were achieved with a minimum number of trials equal to $J_{\min} = 150$ and a minimum number of errors equal to $N_{\min} = 400$.

Figure 1 displays the BER of the test $\mathcal{T}_{\sigma\kappa(A/\sigma)}$ for different values of A and m in comparison with the theoretical value $\mathcal{Q}(A/\sigma)$ of the probability of error. Table 1 gives the empirical mean and empirical Mean Square Error (MSE) of the ESE-I obtained during these experiments.

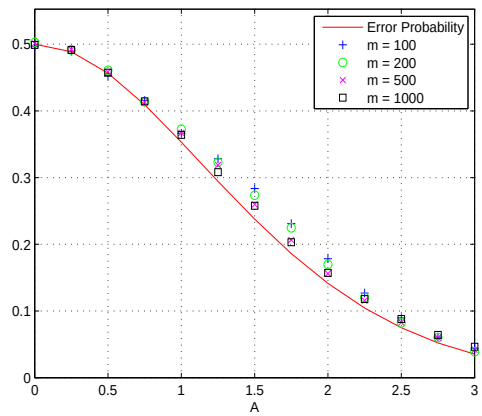

Figure 1. BER of the test $\mathcal{T}_{\tilde{\sigma}\kappa(A/\tilde{\sigma})}$ versus the error probability $\mathcal{Q}(A/\sigma)$ for different values of m and A. The signals S_k, $k \in \mathbf{N}$, are independent, have their probabilities of presence equal to $1/2$ and are such that $S_k = Ae^{i\Phi_k}$ where Φ_k is uniformly distributed in $[0, 2\pi]$.

These results suggest the construction of a new estimator, namely the ESE-II, which basically derives from the ESE-I.

4.3. The ESE-II

With the same notations as those used so far, let Ψ_m be the random variable defined by

$$\Psi_m = \frac{1}{\sigma}\left(\sum_{k=1}^{m} |Y_k|\mathcal{I}(|Y_k| \leq \tilde{\sigma})\right) / \left(\sum_{k=1}^{m} \mathcal{I}(|Y_k| \leq \tilde{\sigma})\right).$$

The empirical mean and standard deviation of Ψ_m were computed during the experiments described in the previous section. The results are those of table 2.

The empirical mean of Ψ_m is rather steady when m varies and the empirical standard deviation of this same

Sample Size	m = 100	m = 200	m = 500	m = 1000
Empirical mean	1.2187	1.2289	1.2094	1.2262
Empirical MSE	0.1275	0.0995	0.0737	0.0756

Table 1. Empirical mean and empirical MSE of the ESE-I for different values of m

Sample Size	m = 100	m = 200	m = 500	m = 1000
Empirical mean	0.7102	0.7120	0.7069	0.7093
Empirical standard deviation	0.0255	0.0152	0.0082	0.0064

Table 2. Empirical mean and standard deviation of Ψ_m for different values of m

random variable decreases with the sample size. Even though we only give the results obtained for $m = 100$, 200, 500 and 1000, the values obtained for other sample sizes less than 1000 are quite the same. The foregoing then suggests defining another estimate $\widehat{\sigma}$ by setting

$$\widehat{\sigma} = \frac{1}{K}\left(\sum_{k=1}^{m} |Y_k| \mathcal{I}(|Y_k| \leq \tilde{\sigma})\right) / \left(\sum_{k=1}^{m} \mathcal{I}(|Y_k| \leq \tilde{\sigma})\right) \tag{16}$$

where $K = 0.7096$ is the average value of the empirical means of Ψ_m for $m = 100, 200, \ldots, 1000$. This new estimate $\widehat{\sigma}$ is called the ESE-II.

We conducted the same type of experiments as those presented in section 4.2. The BERs obtained when the noise standard deviation is estimated by the ESE-II are then those of figure 2. The empirical mean and empirical MSE of this estimate are given in table 3. According to these results, the ESE-II is more accurate than the ESE-I.

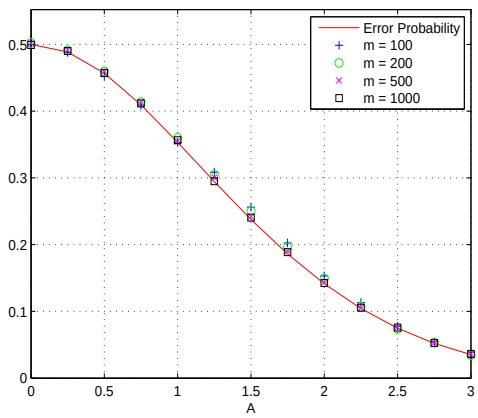

Figure 2. BER of the test $\mathcal{T}_{\widehat{\sigma}\kappa(A/\widehat{\sigma})}$ versus the error probability $\mathcal{Q}(A/\sigma)$ for different values of m and A. These results were obtained with the same signals and the same experimental protocol as those employed to obtain the results of figure 1.

5. APPLICATION TO SPEECH ENHANCEMENT

With the same notations and under the same assumptions as those of section 1, we use the ESE-II to estimate the noise standard deviation and adjust the Wiener filtering of the noisy speech signal y.

5.1. Standard deviation estimation via the ESE-II

We split the T available samples $y[t]$, $t = 1, 2, \ldots, T$, into non-overlapping frames of $N = 2^q$ successive samples each. As usual, q is chosen so that $NF_s \approx 20$ms where F_s is the sampling frequency. Let K stand for the number of frames such constructed. The kth frame is then the finite sequence of samples $y[(k-1)\frac{N}{2}+n]$, $n = 0, 1, \ldots, N-1$. The N-Discrete Fourier Transform (DFT) of this frame is then the sequence $Y_{k,\ell}$, $n = 0, 1, \ldots, N-1$, with

$$Y_{k,\ell} = C \sum_{t=0}^{N-1} y[(k-1)N + t] e^{-i2\pi\ell t/N}, \tag{17}$$

C being some constant, usually chosen in $\{1, 1/N, 1/\sqrt{N}\}$. We thus obtain the matrix $[Y_{k,\ell}]_{k\in\{1,\ldots,K\},\ell\in\{0,\ldots,N-1\}}$. Because of the Hermitian symmetry of the DFT, we can restrict attention to half of this matrix, namely the complex values $Y_{k,\ell}$, $k \in \{1, \ldots, K\}$, $\ell \in \{0, \ldots, N/2-1\}$.

Given a frame k and a bin ℓ, we should write that $Y_{k,\ell} = S_{k,\ell} + X_{k,\ell}$ where, obviously, $S_{k,\ell}$ and $X_{k,\ell}$ stand respectively for the speech and noise time-frequency components for the kth frame and the ℓth bin. Since the frames do not overlap, the complex random variables $X_{k,\ell}$, $k = \{1, \ldots, K\}$, $\ell \in \{0, 1, \ldots, N-1\}$, are iid with $X_{k,\ell} \sim \mathcal{N}_c(0, \gamma^2)$ and $\gamma = \sigma C\sqrt{N}$.

Depending on the type of speech signal present during frame k, some speech time-frequency components $S_{k,\ell}$ can be neglected in comparison with noise and other speech time-frequency components. For instance, high frequency components of voiced speech signals are often negligible in comparison with noise and low-frequency components of the same speech signals; many unvoiced fricative speech signals have low-frequency components significantly smaller than those in high frequency and those due to noise. We model the presence and the absence of the speech time-frequency component $S_{k,\ell}$ by a discrete random variable $\varepsilon_{k,\ell}$ valued in $\{0, 1\}$ and write that the observation is $Y_{k,\ell} = \varepsilon_{k,\ell} S_{k,\ell} + X_{k,\ell}$. With respect to this model, $P(\{\varepsilon_{k,\ell} = 1\})$ is the probability that some speech component be present in bin ℓ during the frame k. This probability of presence may be larger than one half for low frequency components; however, for high frequency components, this probability of presence becomes less than or equal to $1/2$ and even relatively small.

The ESE-II is used as follows to estimate γ. We split the observation set $\{Y_{k,\ell}\}$ where $k \in \{1, \ldots, K\}$, $\ell \in$

Sample Size	m = 100	m = 200	m = 500	m = 1000
Empirical bias	1.0029	1.0103	1.0001	1.0041
Empirical MSE	0.0520	0.0302	0.0159	0.0115

Table 3. Empirical bias and empirical MSE of the ESE-II(m) for different values of m

$\{0, \ldots, N/2-1\}$, into subsets of m observations each; each subset is used to perform an estimate of γ via the ESE-II; we then compute the average value of the $KN/2m$ estimates thus obtained to derive an estimate of γ. Dividing this average by $C\sqrt{N}$ yields an estimate of σ.

In order to deal with m observations that can reasonably be considered as mutually independent, these observations can be chosen randomly amongst the M complex values we have. However, this randomization does not affect significantly the results obtained below.

5.2. The Wiener filtering

The T available samples $y(t)$, $t = 1, 2, \ldots, T$, are still split into frames of $N = 2^q$ samples each but, in contrast with the preceding subsection, the frames overlap now by one half and the samples of each frame are weighted. Despite these differences with the foregoing, the notations used above are kept.

The Wiener filtering of the kth frame consists in seeking the complex values $W_{k,\ell}$, such that, for every bin $\ell \in \{0, 1, \ldots, N-1\}$, $E[|S_{k,\ell} - W_{k,\ell} Y_{k,\ell}|^2]$ is the least value among all the possible quadratic means $E[|S_{k,\ell} - \lambda Y_{k,\ell}|^2]$ when λ ranges over the set of complex values. The well-known solution to this problem is

$$W_{k,\ell} = E[|S_{k,\ell}|^2]/E[|Y_{k,\ell}|^2] = \frac{E[|S_{k,\ell}|^2]}{\gamma^2 + E[|S_{k,\ell}|^2]} \quad (18)$$

since $X_{k,\ell} \sim \mathcal{N}_c(0, \gamma^2)$ and $\gamma = \sigma C\sqrt{N}$. Defining the *a priori* Signal to Noise Ratio (SNR) by

$$\rho_{k,\ell} = E[|S_{k,\ell}|^2]/\gamma^2, \quad (19)$$

equation (18) can be re-written in the form

$$W_{k,\ell} = \rho_{k,\ell}/(1 + \rho_{k,\ell}). \quad (20)$$

The denoised speech signal in the kth frame is then the inverse DFT of the sequence $W_{k,\ell}$, $\ell = 0, 1, \ldots, N-1$.

The main difficulty in performing an estimate of the *a priori* SNR is that speech signals are not stationary. According to the standard recursive filtering procedure originally introduced in [3], we estimate $W_{k,\ell}$ by

$$\widetilde{W}_{k,\ell} = \tilde{\rho}_{k,\ell}/(1 + \tilde{\rho}_{k,\ell}) \quad (21)$$

where

$$\tilde{\rho}_{k,\ell} = (1-\alpha) h\left(\zeta_{k,\ell} - 1\right) + \alpha |\widetilde{W}_{k-1,\ell} Y_{k-1,\ell}|^2/\gamma^2 \quad (22)$$

can be regarded as an estimate of the *a priori* SNR $\rho_{k,\ell}$. In (22), $h(x) = x$ if $x \geq 0$ and $h(x) = 0$ otherwise, α is some weighting factor such that $0 \leq \alpha < 1$ (we chose $\alpha = 0.98$ in our experiments commented below), and

$$\zeta_{k,\ell} = |Y_{k,\ell}|^2/\gamma^2$$

is the so-called *a posteriori* SNR.

When σ is unknown, the value γ can be estimated by proceeding as described in subsection 5.1. Denoting by $\widehat{\gamma}$ the estimate returned for γ by the ESE-II, we modify the recursive filtering approach defined by equations (21) and (22) as follows. The coefficients $W_{k,\ell}$ are now estimated by

$$\widehat{W}_{k,\ell} = \widehat{\rho}_{k,\ell}/(1 + \widehat{\rho}_{k,\ell}), \quad (23)$$

where the estimate $\widehat{\rho}_{k,\ell}$ of the *a priori* SNR is given by

$$\widehat{\rho}_{k,\ell} = (1-\alpha) h\left(\widehat{\zeta}_{k,\ell} - 1\right) + \alpha |\widehat{W}_{k-1,\ell} Y_{k-1,\ell}|^2/\widehat{\gamma}^2, \quad (24)$$

and $\widehat{\zeta}_{k,\ell} = |Y_{k,\ell}|^2/\widehat{\gamma}^2$ is an estimate of $\zeta_{k,\ell}$. The denoised speech signal obtained in frame k is then the inverse DFT of the sequence $\widehat{W}_{k,\ell}$, $\ell = 0, 1, \ldots, N-1$. Since it follows from section 4.3 that the estimate $\widehat{\gamma}$ should approach significantly well the exact unknown value γ, the performance of the recursive procedure defined by equations (23) and (24) can be expected to be close to that obtained by the filtering approach defined by (21) and (22).

5.3. Performance evaluation

We consider twenty sentences of the TIMIT database, downsampled to 8 kHz before adding white Gaussian noise. We estimate the noise standard deviation as described in section 5.1 with frames of $N = 256$ samples each. A frame corresponds to 32ms of noisy speech signals. For estimating the noise standard deviation, these frames do not overlap and are not weighted. As far as the Wiener filtering is concerned, there is a 50% overlap between two adjacent frames and each frame is weighted by a Hanning window before computing the DFT.

We evaluate the quality of the filtered speech signals by means of the standard Segmental Signal to Noise Ratio (SSNR) (see [9]) and the Modified Bark Spectral Distortion (MBSD) (see [10]). The SSNR is the average of the SNR values on short segments. The SSNR is not relevant enough to measure the distortion of the denoised speech signals. This is the reason we use the MBSD. The MBSD proves to be highly correlated with subjective speech quality assessment [10].

The average SNNR and MBSD obtained over the twenty sentences randomly chosen within the TIMIT database are presented in figures 3 and 4. The solid curves are the performance measurements achieved with the filtering defined by equations (23) and (24) where the ESE-II is used

to estimate the noise level. The dashed curves are the results obtained when the filtering is achieved along equations (21) and (22), that is when the noise standard deviation is known. Clearly, the Wiener filtering adjusted with the noise standard deviation estimate yields results that are significantly close to those obtained when the noise standard deviation is known.

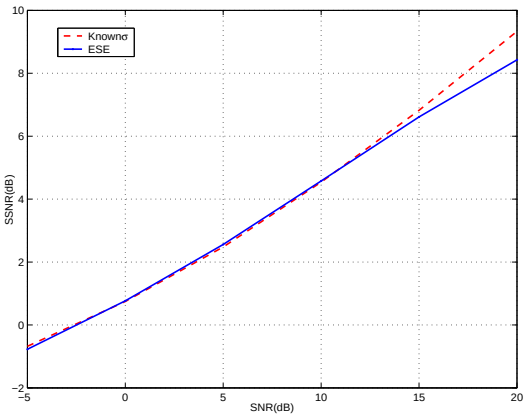

Figure 3. SSNR improvement for speech signals in independent AWGN with various SNRs.

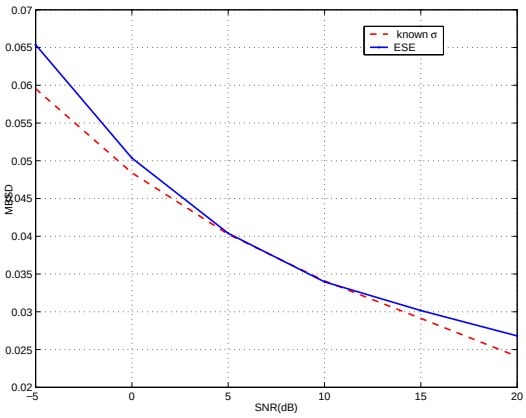

Figure 4. MBSD improvement for speech signals in independent AWGN with various SNRs.

6. CONCLUSION AND PERSPECTIVE

When signals with unknown probability distributions and priors are additively corrupted by independent WGN, the noise standard deviation can be estimated by the ESE-II. This estimator requires no prior knowledge on the signal probability distributions, does not assume that the signals are iid or that the probabilities of presence of these signals are equal. The observations should be independent; the sample size and the signal amplitudes should be large; however, these conditions are seemingly not so constraining in practice.

A direct application of the ESE-II is the estimation of the noise standard deviation when observations are speech signals in additive and independent WGN. The estimate thus performed serves to adjust a standard Wiener filtering without resorting to any VAD or subspace approaches. The SSNR and the MBSD of the denoised speech signals returned by the resulting filtering are very close to those achieved when the noise standard deviation is known.

A rather natural extension of this work is perceptual filtering. For instance, in [2], the same type of estimator is used to adjust some perceptual filtering. Our current work involves the use of the estimator proposed in the present paper to carry out perceptually motivated speech denoising in presence of white and coloured noise.

7. REFERENCES

[1] M. Abramowitz, I. Stegun, *Handbook of Mathematical Functions*, Dover Publications, Inc., New York, Ninth printing, 1972.

[2] A. Amehraye, D. Pastor, S. Ben Jebara, On the Application of Recent Results in Statistical Decision and Estimation Theory to Perceptual Filtering of Noisy Speech Signals, ISCCSP'06, Marrakech, Morocco, 2006.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, 1984, pp. 1109-1121.

[4] D. Pastor, R. Gay, B. Groenenboom, *A Sharp Upper-Bound for the Probability of Error of the Likelihood Ratio Test for Detecting Signals in White Gaussian Noise*, IEEE Transactions on Information Theory, VOL. 48, NO. 1, pp 228-238, January 2002.

[5] D. Pastor, "Un théorème limite et un test pour la détection non paramétrique de signaux dans un bruit blanc gaussien de variance inconnue", GRETSI'05, Louvain-La-Neuve, 2005.

[6] D. Pastor, "On the detection of signals with unknown distributions and priors in white gaussian noise", *Collection des Rapports de Recherche de l'ENST Bretagne*, RR-2006001-SC, 2006.

[7] D. Pastor, Estimating the standard deviation of some additive white Gaussian noise on the basis of non signal-free observations, ICASSP'06, Toulouse, France, 2006.

[8] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag, 2nd Edition, 1994.

[9] S. Quackenbush, T. Barnwell, and M. Clements, "Objective Measures of Speech Quality", *Englewood Cliffs, NJ: Prentice-Hall*, 1988.

[10] W. Yang, M. Dixon and R Yantorno, "Modified bark spectral distortion measure which uses noise masking threshold", *IEEE Speech coding Workshop*, Pocono Manor, 1997, pp. 55-56.