

Audio Denoising by Time-Frequency Block Thresholding

Guoshen Yu*, Stéphane Mallat and Emmanuel Bacry¹

Submitted to IEEE TRANSACTIONS ON SIGNAL PROCESSING

EDICS: SSP-REST

Abstract— Removing noise from audio signals requires a non-diagonal processing of time-frequency coefficients to avoid producing “musical noise”. State of the art algorithms perform a parameterized filtering of spectrogram coefficients with empirically fixed parameters. A block thresholding estimation procedure is introduced, which adjusts all parameters adaptively to signal property by minimizing a Stein estimation of the risk. Numerical experiments demonstrate the performance and robustness of this procedure through objective and subjective evaluations.

Index Terms— Audio denoising, Block thresholding, Ephraim and Malah, Power spectrum, Power subtraction, Thresholding

I. INTRODUCTION

Audio signals are often contaminated by background environment noise and buzzing or humming noise from audio equipments. Audio denoising aims at attenuating the noise while retaining the underlying signals. Applications such as music and speech restoration are numerous.

Diagonal time-frequency audio denoising algorithms attenuate the noise by processing each window Fourier or wavelet coefficient independently, with empirical Wiener [48], power subtraction [2], [3], [38] or thresholding operators [20]. These algorithms create isolated time-frequency structures that are perceived as a “musical noise” [7], [60]. Ephraim and Malah [21], [22] showed that this musical noise is strongly attenuated with non-diagonal time-frequency estimators that regularize the estimation by recursively aggregating time-frequency coefficients. This approach has further been improved by optimizing the SNR estimation with parameterized filters [10] that rely on stochastic audio models. However, these parameters should be adjusted to the nature of the audio signal, which often varies and is unknown. In practice, they are empirically fixed [7], [10], [21], [22].

This paper introduces a new non-diagonal audio denoising algorithm through adaptive time-frequency block thresholding [60]. Block thresholding has been introduced by Cai and Silverman in mathematical statistics [4], [5], [6] to improve the asymptotic decay of diagonal thresholding estimators. For audio

¹The authors are with CMAP, Ecole Polytechnique, 91128 Palaiseau, France. Tel: + 33 (0)1 69 33 45 91. Fax: +33 (0)1 69 33 46 46. Email: yu@cmaph.polytechnique.fr, mallat@cmaph.polytechnique.fr, emmanuel.bacry@polytechnique.fr.

time-frequency denoising, we show that block thresholding regularizes the estimate and is thus effective in musical noise reduction. Block parameters are automatically adjusted by minimizing a Stein estimator of the risk [55], which is calculated analytically from the noisy signal values. Numerical experiments show that this new adaptive estimator is robust to signal type variations and improves the SNR and the perceived quality with respect to state of the art audio denoising algorithms.

The paper first reviews the state of the art time-frequency audio denoising algorithms by emphasizing the difference between diagonal and non-diagonal methods. Section III introduces time-frequency block thresholding and computes a Stein unbiased estimate of the resulting risk to adjust automatically the block parameters. Numerical experiments and comparisons are presented in Section IV, with objective and subjective measures.

II. STATE OF THE ART

A. Time-frequency Audio Denoising

Time-frequency audio-denoising procedures compute a short-time Fourier transform or a wavelet transform or a wavelet packet transform of the noisy signal, and processes the resulting coefficients to attenuate the noise. These representations reveal the time-frequency signal structures that can be discriminated from the noise. We concentrate on the coefficient processing as opposed to the choice of representations. Numerical experiments are performed with short-time Fourier transforms that are most commonly used in audio processing.

The audio signal f is contaminated by a noise ϵ that is often modeled as a zero mean Gaussian process independent of f :

$$y[n] = f[n] + \epsilon[n], \quad n = 0, 1, \dots, N-1. \quad (1)$$

A time-frequency transform decomposes the audio signal y over a family of time-frequency atoms $\{g_{l,k}\}_{l,k}$ where l and k are the time and frequency (or scale) localization indices. The resulting coefficients shall be written:

$$Y[l, k] = \langle y, g_{l,k} \rangle = \sum_{n=0}^{N-1} y[n] g_{l,k}^*[n].$$

where $*$ denotes the conjugate. These transforms define a complete and often redundant signal representation. In this paper we shall suppose that these time-frequency atoms define a tight frame [18], [43], which means that there exists $A > 0$ such that

$$\|y\|^2 = \frac{1}{A} \sum_{l,k} |\langle y, g_{l,k} \rangle|^2.$$

This implies a simple reconstruction formula

$$y[n] = \frac{1}{A} \sum_{l,k} Y[l, k] g_{l,k}[n].$$

The constant A is a redundancy factor and if $A = 1$ then a tight frame is an orthogonal basis. A tight frame behaves like a union of A orthogonal bases.

A frame representation provides an energy control. The redundancy implies that a signal f has a non-unique way to be reconstructed from a tight frame representation: $f[n] = \frac{1}{A} \sum_{l,k} C[l, k] g_{l,k}[n]$, but all such reconstructions satisfy

$$\|f\|^2 \leq \frac{1}{A} \sum_{l,k} |C[l, k]|^2, \quad (2)$$

with an equality if $C[l, k] = \langle f, g_{l,k} \rangle, \forall l, k$.

Short-time Fourier atoms can be written: $g_{l,k}[n] = w[n-lu] \exp\left(\frac{i2\pi kn}{K}\right)$, where $w[n]$ is a time window of support size K , which is shifted with a step $u \leq K$. l and k are respectively the integer time and frequency indices with $0 \leq l < N/u$ and $0 \leq k < K$. In this paper, $w[n]$ is the square root of a Hanning window and $u = K/2$ so one can verify that the resulting window Fourier atoms $\{g_{l,k}\}_{l,k}$ define a tight frame with $A = 2$.

A denoising algorithm modifies time-frequency coefficients by multiplying each of them by an attenuation factor $a[l, k]$ to attenuate the noise component. The resulting “denoised” signal estimator is:

$$\hat{f}[n] = \frac{1}{A} \sum_{l,k} \hat{F}[l, k] g_{l,k}[n] = \frac{1}{A} \sum_{l,k} a[l, k] Y[l, k] g_{l,k}[n]. \quad (3)$$

Time-frequency denoising algorithms differ through the calculation of the attenuation factors $a[l, k]$. The noise coefficient variance

$$\sigma^2[l, k] = E\{|\langle \epsilon, g_{l,k} \rangle|^2\}$$

is supposed to be known or estimated with methods such as [16], [20], [45]. If the noise is stationary, which is often the case, then the noise variance does not depend upon time: $\sigma^2[l, k] = \sigma^2[k]$.

B. Diagonal Estimation

Simple time-frequency denoising algorithms compute each attenuation factor $a[l, k]$ only from the corresponding noisy coefficient $Y[l, k]$ and are thus called *diagonal estimators*. These algorithms have a limited performance and produce a musical noise.

To minimize an upper bound of the quadratic estimation risk

$$r = E\{\|f - \hat{f}\|^2\} \leq \frac{1}{A} \sum_{l,k} E\{|F[l, k] - \hat{F}[l, k]|^2\}, \quad (4)$$

(4) being a consequence of (2), one can verify [20] that the optimal attenuation factor is

$$a[l, k] = 1 - \frac{1}{\xi[l, k] + 1} \quad (5)$$

where $\xi[l, k] = F^2[l, k]/\sigma^2[l, k]$ is the *a priori* SNR. The resulting risk lower bound, also called oracle risk r_o , is

$$r_0 \leq \frac{1}{A} R_o \quad \text{where} \quad R_o = \sum_{l, k} \frac{|F[l, k]|^2 \sigma^2[l, k]}{|F[l, k]|^2 + \sigma^2[l, k]}. \quad (6)$$

This lower bound cannot be reached because the “oracle” attenuation factor (5) depends upon the *a priori* SNR $\xi[l, k]$ which is unknown. It is thus necessary to estimate this SNR.

Diagonal estimators of the SNR $\xi[l, k]$ are computed from the *a posteriori* SNR defined by $\gamma[l, k] = |Y[l, k]|^2/\sigma^2[l, k]$. One can verify that

$$\hat{\xi}[l, k] = \gamma[l, k] - 1 \quad (7)$$

is an unbiased estimator. Inserting this estimator in the oracle formula (5) defines the empirical Wiener estimator [38], [48]

$$a[l, k] = \left(1 - \frac{1}{\hat{\xi}[l, k] + 1} \right)_+ \quad (8)$$

with the notation $(z)_+ = \max(z, 0)$. Variants of this empirical Wiener are obtained by minimizing a sum of signal distortion and residual noise energy [23], [25], [30], [41] or by computing a maximum likelihood estimate [38], [48], [59].

Power subtraction estimators [2], [3], [38], [51], [53] generalize the empirical Wiener attenuation rule:

$$a[l, k] = \left(1 - \lambda \left[\frac{1}{\hat{\xi}[l, k] + 1} \right]^{\beta_1} \right)^{\beta_2}_+ \quad (9)$$

where $\beta_1, \beta_2 \geq 0$ and $\lambda \geq 1$ is an over-subtraction factor to compensate variation of noise amplitude.

Following the statistical work of Donoho and Johnstone [20], thresholding estimators have also been studied for audio noise removal. A hard thresholding [26], [35], [39], [57] either retains or sets to zero each noisy coefficient with

$$a[l, k] = 1_{\{\hat{\xi}[l, k] + 1 > \lambda^2\}}. \quad (10)$$

Soft-thresholding estimator [8], [34], [37], [50] is a special case of power subtraction (9) with $\beta_1 = 1/2$, $\beta_2 = 1$. Donoho and Johnstone have proved that for Gaussian white noises, the quadratic risk of thresholding estimators is close to the oracle lower bound [20].

The attenuation factor $a[l, k]$ of these diagonal estimators only depends upon $Y[l, k]$ with no time-frequency regularization. The resulting attenuated coefficients $a[l, k]Y[l, k]$ thus lack of time-frequency

regularity. It produces isolated time-frequency coefficients which restore isolated time-frequency structures that are perceived as a musical noise. Fig. 1 shows the denoising of a short recording of the Mozart oboe concerto with an additive Gaussian white noise. Fig. 1(a) and 1(b) show respectively the log spectrograms $\log |F[l, k]|$ and $\log |Y[l, k]|$ of the original signal f and its noisy version y . Fig. 1(c) displays a power subtraction attenuation map $a[l, k]$, with black points corresponding to values close to 1. The zoom in Fig. 1(c') shows that this attenuation map contains many isolated coefficients close to 1 (black points). These isolated coefficients restore isolated windowed Fourier vectors $g_{l,k}[n]$ that produce a musical noise.

C. Non-diagonal Estimation

To reduce musical noise as well as the estimation risk, several authors have proposed to estimate the *a priori* SNR $\xi[l, k]$ with a time-frequency regularization of the *a posteriori* SNR $\gamma[l, k]$. The resulting attenuation factors $a[l, k]$ thus depend upon the data values $Y[l', k']$ for (l', k') in a whole neighborhood of (l, k) and the resulting estimator $\hat{f}[n] = \frac{1}{A} \sum_{l,k} a[l, k] Y[l, k] g_{l,k}[n]$ is said to be *non-diagonal*.

In their pioneer paper Ephraim and Malah [21] have introduced a *decision-directed* SNR estimator obtained with a first order recursive time filtering:

$$\hat{\xi}[l, k] = \alpha \tilde{\xi}[l - 1, k] + (1 - \alpha) (\gamma[l, k] - 1)_+, \quad (11)$$

where $\alpha \in [0, 1]$ is a recursive filter parameter and $\tilde{\xi}[l - 1, k] = |\hat{F}[l - 1, k]|^2 / \sigma^2[l, k]$ is an empirical SNR estimate of $F[l - 1, k]$ based on the previously computed estimate. This decision-directed SNR estimator has been applied with various attenuation rules such as empirical Wiener estimator (8) [11], Ephraim and Malah's minimum mean-square error spectral amplitude (MMSE-SA) [21], log spectral amplitude estimator (MMSE-LSA) [22] and Wolfe and Godsill's minimum mean-square error spectral power estimator (MMSE-SP) [58] that are derived from a Bayesian formulation using a Gaussian speech model [13], [15], [17], [21], [22], [35], [42], as well as Martin's MMSE estimators using a Gamma speech model [44]. These work clearly showed that the regularization of the SNR estimation reduces musical noise as well as the estimation risk $r = E\{\|\hat{f} - f\|^2\}$.

Cohen [10] improved the decision-directed SNR estimator by combining a causal recursive temporal filter with a noncausal compactly supported time-frequency filter to get a first SNR estimation. He then refines this estimation in a Bayesian formulation by computing a new SNR estimation using the MMSE-SP attenuation rule [58] from the first SNR estimate. This noncausal *a priori* SNR estimator has been combined with attenuation rules derived from Gaussian [10], [13], Gamma and Laplacian speech models [12]. Other SNR estimators have been proposed by Cohen [14] with generalized autoregressive

conditional heteroscedasticity (GARCH), applied with MMSE-LSA attenuation rules of Gamma and Laplacian speech models [14].

Matz and Hlawatsch have also proposed to estimate the SNR with a rectangular time-frequency filter and to use it together with the empirical Wiener estimator (8) [46]. In one example, they showed a noticeable performance gain with respect to a diagonal SNR estimation. The same non-diagonal SNR estimation has been applied in [47] where the authors automatically adapted the size of the short-time Fourier windows to the signal properties.

Thresholding estimators [20] have also been studied with time-regularized thresholds [29], [40], which are indirectly based on non-diagonal SNR estimations $\hat{\xi}[l, k]$. Such thresholds can further be adapted to a detection of speech presence [1], [9], [56].

Non-diagonal estimators clearly outperform diagonal estimators but depend upon regularization filtering parameters. Large regularization filters reduce the noise energy but introduce more signal distortion [7], [13], [21], [24]. It is desirable that filter parameters are adjusted depending upon the nature of audio signals. In practice, however, they are selected empirically [7], [10], [13], [21], [22]. Moreover, the attenuation rules and the *a priori* SNR estimators that are derived with a Bayesian approach [10], [12], [13], [14], [15], [17], [21], [22], [35], [42] model audio signals with Gaussian, Gamma or Laplacian processes. Although such models are often appropriate for speech, they do not take into account the complexity of other audio signals such as music, that include strong attacks.

III. TIME-FREQUENCY BLOCK THRESHOLDING

Block thresholding was introduced in statistics by Cai and Silverman [4], [5], [6] and studied by Hall et al. [31], [32], [33] to obtain nearly minimax signal estimators. The “p-point uncertainty model” proposed by Matz and Hlawatsch [46] also led to a block thresholding estimator with fixed parameters that are chosen empirically. For audio signal denoising, we describe an adaptive block thresholding non-diagonal estimator that automatically adjusts all parameters. It relies on the ability to compute an estimate of the risk, with no prior stochastic audio signal model, which makes this approach particularly robust.

A. Block Thresholding Algorithm

A time-frequency block thresholding estimator regularizes power subtraction estimation (9) by calculating a single attenuation factor over time-frequency blocks. The time-frequency plane $\{l, k\}$ is segmented in I blocks B_i whose shape may be chosen arbitrarily. The signal estimator \hat{f} is calculated from the

noisy data y with a constant attenuation factor a_i over each block B_i

$$\hat{f}[n] = \sum_{i=1}^I \sum_{(l,k) \in B_i} a_i Y[l, k] g_{l,k}[n]. \quad (12)$$

To understand how to compute each a_i , one relates the risk $r = E\{\|f - \hat{f}\|^2\}$ to the frame energy conservation (2) and obtains

$$r = E\{\|f - \hat{f}\|^2\} \leq \frac{1}{A} \sum_{i=1}^I \sum_{(l,k) \in B_i} E\{|a_i Y[l, k] - F[l, k]|^2\}. \quad (13)$$

Since $Y[l, k] = F[l, k] + \epsilon[l, k]$ one can verify that the upper bound of (13) is minimized by choosing

$$a_i = 1 - \frac{1}{\xi_i + 1} \quad (14)$$

where $\xi_i = \overline{F_i^2} / \overline{\sigma_i^2}$ is the average *a priori* SNR in B_i . It is calculated from

$$\overline{F_i^2} = \frac{1}{B_i^\#} \sum_{(l,k) \in B_i} |F[l, k]|^2 \quad \text{and} \quad \overline{\sigma_i^2} = \frac{1}{B_i^\#} \sum_{(l,k) \in B_i} \sigma^2[l, k],$$

which are the average signal energy and noise energy in B_i , and $B_i^\#$ is the number of coefficients $(l, k) \in B_i$. The resulting oracle block risk r_{bo} satisfies

$$r_{bo} \leq \frac{1}{A} R_{bo} \quad \text{where} \quad R_{bo} = \sum_{i=1}^I \frac{\overline{F_i^2} \overline{\sigma_i^2}}{\overline{F_i^2} + \overline{\sigma_i^2}}. \quad (15)$$

The oracle block attenuation coefficients a_i in (14) can not be calculated because the *a priori* SNR ξ_i is unknown. Cai and Silverman [4] introduced block thresholding estimators that estimate the SNR over each B_i by averaging the noisy signal energy:

$$\hat{\xi}_i = \frac{\overline{Y_i^2}}{\overline{\sigma_i^2}} - 1 \quad (16)$$

where

$$\overline{Y_i^2} = \frac{1}{B_i^\#} \sum_{(l,k) \in B_i} |Y[l, k]|^2.$$

Observe that if $\sigma[l, k] = \overline{\sigma_i}$ for all $(l, k) \in B_i$ then $\hat{\xi}_i$ is an unbiased estimator of ξ_i . The resulting attenuation factor a_i is computed with a power subtraction estimator (9)

$$a_i = \left(1 - \frac{\lambda}{\hat{\xi}_i + 1} \right)_+. \quad (17)$$

A block thresholding estimator can thus be interpreted as a non-diagonal estimator derived from averaged SNR estimations over blocks. Each attenuation factor is calculated from all coefficients in each block, which regularizes the time-frequency coefficient estimation. Fig 1(d) displays a block thresholding

attenuation map a_i with black points corresponding to values close to 1. The zoom in Fig 1(d') shows that non-diagonal block thresholding attenuation factors are much more regular than the diagonal power subtraction attenuation factors in Fig 1(c') and they do not keep isolated points responsible for musical noise.

B. Block Thresholding Risk and Choice of λ

An upper bound of the risk of the block thresholding estimator is computed by analyzing separately the bias and variance terms. Observe that the upper bound of the oracle risk r_{bo} in (15) with blocks is always larger than that of the oracle risk r_o in (6) without blocks, because the former is obtained through the same minimization but with less parameters as attenuation factors remain constant over each block. A direct calculation shows that

$$R_{bo} - R_o = \sum_{i=1}^I \sum_{(l,k) \in B_i} \frac{\bar{\xi}_i \xi[l,k] (\bar{\sigma}_i^2 - \sigma^2[l,k]) + (\bar{F}_i^2 - |F[l,k]|^2)}{(\bar{\xi}_i + 1)(\xi[l,k] + 1)} \geq 0. \quad (18)$$

R_{bo} is close to R_o if both the noise and the signal coefficients have little variation in each block. This bias term is thus reduced by choosing the blocks so that in each block B_i either (i) $F[l,k]$ and $\sigma[l,k]$ vary little; or (ii) $\xi[l,k] \gg 1$ and $\sigma[l,k]$ varies little; or (iii) $\xi[l,k] \ll 1$ and $F[l,k]$ varies little.

Block thresholding (17) approximates the oracle block attenuation (14) by replacing ξ_i with an estimate $\hat{\xi}_i$ in (16) and by setting an over-subtraction factor $\lambda \geq 1$ to control the variance term of risk due to the noise variation. If the noise ϵ is a Gaussian white noise, then the resulting risk $r = E\{\|f - \hat{f}\|^2\}$ can be shown to be close to the oracle risk (15). The average noise energy over a block B_i is

$$\bar{\epsilon}_i^2 = \frac{1}{B_i^\#} \sum_{(l,k) \in B_i} |\epsilon[l,k]|^2. \quad (19)$$

If the frame is an orthogonal basis, in the particular case where all blocks B_i have the same size $B^\#$ and the noise is Gaussian white noise with variance σ^2 (hence $\bar{\epsilon}_i^2 = \bar{\epsilon}^2$) then then Cai [4] proved that

$$r = E\{\|\hat{f} - f\|^2\} \leq 2\lambda R_{bo} + 4N\sigma^2 \text{Prob}\{\bar{\epsilon}^2 > \lambda\sigma^2\}, \quad (20)$$

where $\text{Prob}\{\}$ is the probability measure. We have mentioned that a tight frame behaves very similarly to a union of A orthogonal bases. Therefore the oracle inequality with a frame representation holds as well:

$$r = E\{\|\hat{f} - f\|^2\} \leq \frac{2\lambda}{A} R_{bo} + \frac{4M}{A} \sigma^2 \text{Prob}\{\bar{\epsilon}^2 > \lambda\sigma^2\}, \quad (21)$$

where $M \geq N$ is the number of vectors $g_{l,k}$ in the frame. For the window Fourier frame used in this paper, $M = 2N$ and $A = 2$.

The second term $4M\sigma^2 \text{Prob}\{\bar{\epsilon}^2 > \lambda\sigma^2\}$ is a variance term corresponding to a probability of keeping pure noise coefficients, i.e., f is zero ($y = \epsilon$) and $a_i \neq 0$ (c.f. (17)). $\text{Prob}\{\bar{\epsilon}^2 > \lambda\sigma^2\}$ is the probability to keep a residual noise. The oracle risk and the variance terms in (21) are competing. When λ increases the first term increases and the variance term decreases. Similarly, when the block size $B^\#$ increases the oracle risk R_{bo} increases whereas the variance decreases. Adjusting λ and the block sizes $B^\#$ can be interpreted as an optimization between the bias and the variance of our block thresholding estimator. The parameter λ is set depending upon $B^\#$ by adjusting the residual noise probability

$$\text{Prob}\{\bar{\epsilon}^2 > \lambda\sigma^2\} = \delta. \quad (22)$$

The probability δ is a perceptual parameter. We set $\delta = 0.1\%$ in (22) as our psychoacoustic experiments show that with a residual noise probability $\delta \approx 0.1\%$, musical noise is hardly perceptible.

Let $B_i^\# = L_i \times W_i$ be a rectangular block size, where $L_i \geq 2$ and $W_i \geq 2$ are respectively the block length in time and the block width in frequency (the unit being the time-frequency index in the window Fourier transform). One can verify that with half overlapping Hanning windows the average noise energy $\bar{\epsilon}^2$ follows approximatively a χ^2 distribution degrees with $B_i^\#$ degree of freedom. Thus solving λ in (22) amounts to looking up a χ^2 table. Table I gives values for a frequency width $W_i \geq 2$. Due to discretization effects, λ takes nearly the same values for $W_i = 1$ and $W_i = 2$. We thus compute λ for $W_i = 1$ by multiplying $B_i^\#$ by 2 and looking at Table I. That (22) holds with λ shown in Table I can also be verified by Monte Carlo simulation.

C. Adaptive Block Thresholding

A block thresholding segments the time-frequency plane in disjoint rectangular blocks of length L_i in time and width W_i in frequency. In the following by “block size” we mean a choice of block shapes and sizes among a collection of possibilities. The adaptive block thresholding chooses the sizes by minimizing an estimate of the risk.

The risk $E\{\|f - \hat{f}\|^2\}$ cannot be calculated since f is unknown, but it can be estimated with a Stein risk estimate [55]. Best block sizes are computed by minimizing this estimated risk. We saw in (13) that the block thresholding risk satisfies

$$r = E\{\|f - \hat{f}\|^2\} \leq \frac{1}{A} \sum_{i=1}^I \sum_{(l,k) \in B_i} E\{|a_i Y[l,k] - F[l,k]|^2\}. \quad (23)$$

Since $Y[l,k] = F[l,k] + \epsilon[l,k]$ and $\epsilon[l,k]$ has a zero mean, $F[l,k]$ is the mean of $Y[l,k]$. To estimate the block thresholding risk Cai [6] uses the Stein estimator of the risk when computing the mean of a random vector, which is given by Stein theorem [55].

Theorem (Stein Unbiased Risk Estimate SURE). *Let $\mathbf{Y} = (Y_1, \dots, Y_p)$ be a normal random vector with the identity as covariance matrix and mean $\mathbf{F} = (F_1, \dots, F_p)$. Let $\mathbf{Y} + \mathbf{h}(\mathbf{Y})$ be an estimator of \mathbf{F} , where $\mathbf{h} = (h_1, \dots, h_p) : \mathbb{R}^p \rightarrow \mathbb{R}^p$ almost differentiable ($h_j : \mathbb{R}^p \rightarrow \mathbb{R}^1, \forall j$). Define $\nabla \cdot \mathbf{h} = \sum_{j=1}^p \frac{\partial}{\partial Y_j} h_j$. If $E \left\{ \sum_{j=1}^p \left| \frac{\partial}{\partial Y_j} h_j(\mathbf{Y}) \right| \right\} < \infty$, then*

$$R = E \|\mathbf{Y} + \mathbf{h}(\mathbf{Y}) - \mathbf{F}\|^2 = p + E \{ \|\mathbf{h}(\mathbf{Y})\|^2 + 2\nabla \cdot \mathbf{h}(\mathbf{Y}) \}. \quad (24)$$

So

$$\hat{R} = p + \|\mathbf{h}(\mathbf{Y})\|_2^2 + 2\nabla \cdot \mathbf{h}(\mathbf{Y}) \quad (25)$$

is an unbiased estimator of the risk R of $\mathbf{Y} + \mathbf{h}(\mathbf{Y})$, called Stein Unbiased Risk Estimator [55].

An estimation of the risk $E\{\|\hat{f} - f\|^2\}$ upper bound (23) is derived from this theorem by computing an estimator \hat{R}_i of the risk in each block B_i : $R_i = \sum_{(l,k) \in B_i} E\{|F[l, k] - a_i Y[l, k]|^2\}$. Over a block B_i , the mean vector $\mathbf{F}_i = (F[l, k])_{(l,k) \in B_i}$ of $\mathbf{Y}_i = (Y[l, k])_{(l,k) \in B_i}$ is estimated by $\hat{\mathbf{F}}_i = (\hat{F}[l, k])_{(l,k) \in B_i}$ with $\hat{\mathbf{F}}_i = a_i \mathbf{Y}_i = \mathbf{Y}_i + \mathbf{h}(\mathbf{Y}_i)$. From the expression (17) of a_i we derive that

$$\mathbf{h}(\mathbf{Y}_i) = -\mathbf{Y}_i \left(\lambda \frac{\bar{\sigma}_i^2}{\bar{Y}_i^2} \mathbf{1}_{\bar{Y}_i^2 \geq \lambda \bar{\sigma}_i^2} + \mathbf{1}_{\bar{Y}_i^2 < \lambda \bar{\sigma}_i^2} \right).$$

Under the hypothesis that the noise variance remains constant on each block, $\sigma^2[l, k] = \bar{\sigma}_i^2$ for $(l, k) \in B_i$, the resulting Stein estimator of the risk $R_i = \sum_{l,k \in B_i} E\{|F[l, k] - a_i Y[l, k]|^2\}$ is

$$\hat{R}_i = \bar{\sigma}_i^2 \left(B_i^\# + E \{ \|\mathbf{h}(\mathbf{Y}_i / \bar{\sigma}_i)\|^2 + 2\nabla \cdot \mathbf{h}(\mathbf{Y}_i / \bar{\sigma}_i) \} \right) \quad (26)$$

and a direct calculation shows that

$$\hat{R}_i = \bar{\sigma}_i^2 \left(B_i^\# + \frac{\lambda^2 B_i^\# - 2\lambda(B_i^\# - 2)}{\bar{Y}_i^2 / \bar{\sigma}_i^2} \mathbf{1}_{\bar{Y}_i^2 \geq \lambda \bar{\sigma}_i^2} + B_i^\# (\bar{Y}_i^2 / \bar{\sigma}_i^2 - 2) \mathbf{1}_{\bar{Y}_i^2 < \lambda \bar{\sigma}_i^2} \right). \quad (27)$$

If the noise is Gaussian white and the frame is an orthogonal basis then the noise coefficients are uncorrelated with same variance and Stein theorem proves that \hat{R}_i is an unbiased risk estimator of the risk R_i . If the noise is not white but stationary then the noise variance does not change in time. If the blocks B_i are sufficiently narrow in frequency then the noise variance still remains constant over each block so the risk estimator remains unbiased. We mentioned that a tight frame behaves very similarly to a union of A orthogonal bases. As a consequence, the theorem result applies approximately and the resulting estimator mains nearly unbiased.

The adaptive block thresholding groups coefficients in blocks whose sizes are adjusted to minimize the Stein risk estimate and it attenuates coefficients in those blocks. To regularize the adaptive segmentation in blocks, the time-frequency plane is first decomposed in macroblocks M_j , $j = 1, 2, \dots, J$, as illustrated

in Fig. 2. Each macroblock M_j is segmented in blocks B_i of same size which means that $B_i^\# = P_j$ is constant over a macroblock M_j . The Stein risk estimation over M_j is $\frac{1}{A} \sum_{i \in M_j} \hat{R}_i$. Several such segmentations are possible and we want to choose the one that leads to the smallest risk estimation. The optimal block size and hence P_j is calculated by choosing the block shape that minimizes $\sum_{i \in M_j} \hat{R}_i$. Once the block sizes are computed, coefficients in each B_i are attenuated with (17), where λ is calculated with (22).

In numerical experiments, each macroblock is segmented with 15 possible block sizes $L \times W$ with a combination of block length $L = 8, 4, 2$ and block width $W = 16, 8, 4, 2, 1$. The size of macroblocks is set to be equal to the maximum block size 8×16 . Fig. 2 illustrates different segmentations of these macroblocks into time-frequency blocks of same size. Minimizing the estimated risk adapts the blocks to the signal time-frequency properties. In particular, it eliminates “pre-echo” artifacts on signal onsets and results in less distortion on signal transients.

Fig.3(a) zooms on the onset of “Mozart” signal whose log-spectrogram is illustrated in Fig 1(b). The attenuation factors of block thresholding with a fixed block size $L = 8$ and $W = 1$ are displayed in Fig.3(b). At the beginning of the harmonics, blocks of large attenuation factors spread beyond the onset of the signal. Fig3(b’) illustrates the horizontal blocks at the onsets marked in Figs 3(a) and (b). In the time interval where the blocks exceed the signal onset, moderate attenuation is performed, and since the noise is not eliminated a transient noise component is heard before the signal beginning. This can be called as a “pre-echo” artifact. On the other hand, this moderate attenuation in the blocks that exceeds signal onsets muffles the onsets as well.

In Figs 3(c)(c’), the adaptive block method chooses blocks of shorter length L in the first part of “Mozart”, which hardly exceed the onset of the signal. This reduces considerably the “pre-echo” artifact. After the onset, the adaptive block method chooses narrow horizontal blocks, to better capture the harmonic signal structures.

D. Non-Diagonal Wiener Post-Processing and Masking Noise

Similarly to the bootstrapping algorithm of Cohen [10] which performs a second SNR estimation from the signal obtained after a first denoising, the block thresholding estimation is improved by applying a second thresholding estimation. A block-thresholding algorithm regularizes the time-frequency estimation as compared to a diagonal thresholding, but it outputs a time-frequency estimation with some block structures as shown in Fig. 4(b). This first estimation is used as an input to compute a Wiener time-frequency estimation that takes advantage of the time-frequency regularization provided by the block

thresholding estimation.

Let \hat{f} be the block thresholding estimation from the noisy data y . Similarly to the post-processing proposed by Baraniuk for images denoising [28], this first estimation is post-processed by computing a new attenuation factor using the oracle formula (5) calculated from its time-frequency coefficients $\hat{F}[l, k] = \langle \hat{f}, g_{l,k} \rangle$:

$$\tilde{a}[l, k] = \frac{|\hat{F}[l, k]|^2}{|\hat{F}[l, k]|^2 + \sigma^2[l, k]}. \quad (28)$$

This new attenuation factor is applied on the noisy time-frequency coefficients to reconstruct a second estimator.

$$\tilde{f}[n] = \frac{1}{A} \sum_{l,k} \tilde{a}[l, k] Y[l, k] g_{l,k}[n].$$

This Wiener estimator is non-diagonal since the attenuation coefficients $\tilde{a}[l, k]$ depend upon values of $Y[l', k']$ in a time-frequency neighborhood of (l, k) . Comparing with Fig 4(b), Fig 4(c) shows that the amplitude of the non-diagonal Wiener attenuation factors $\tilde{a}[l, k]$ is more regular than the block thresholding attenuation factors and is closer to the oracle attenuation (5) displayed in Fig. 4(d). Experiments show that this post-processing increases the SNR on average by about 0.2 dB and improves the audio quality of denoised signals.

Retaining a low-amplitude noise is sometimes desirable to mask artifices generated by an estimation procedure [2], [51]. Following [2], one can retain a masking noise by setting a floor value to the attenuation factor:

$$\tilde{a}_M[l, k] = \max(\tilde{a}[l, k], a_0) \quad (29)$$

where $0 < a_0 \ll 1$ is the minimum attenuation factor of the noise.

IV. EXPERIMENTS AND RESULTS

The experiments presented below have been performed on various types of audio signals: “Piano” is a simple example that contains a single clear clavier stroke; “Mozart” is a musical excerpt that contains relatively quick notes played by a solo oboe; “TIMIT-M” and “TIMIT-F” are respectively male and female utterances taken from the TIMIT database [27]. “TIMIT-M” and “TIMIT-F” are sampled at 16 kHz whereas all the other signals are sampled at 11 kHz. They were corrupted by Gaussian white noise of different amplitude. Short-time Fourier transform with half-overlapping windows were used in the experiments. These windows are square root of Hanning windows of size 50 ms for “Piano” and “Mozart” and 20 ms for “TIMIT-M” and “TIMIT-F”.

For each sound, denoising with “partial noise removal” and “maximum noise removal” were applied: the former retains some low-amplitude residual noise; the latter removes almost all the original noise.

Block thresholding was configured as described in Sections III-C and III-D. For partial noise removal and maximum noise removal, we respectively set $a_0 \approx 0.05$ (the residual noise was calibrated to have similar energy for all methods under comparison) and $a_0 = 0$ in (29).

MMSE-LSA attenuation rule [22] of Ephraim and Malah was also used in our evaluation. Combined with the decision-directed *a priori* SNR estimator (11) with $\alpha = 0.98$ as proposed in [21], [22], this algorithm (referred to as LSA-DD) led to satisfactory results for partial noise removal. However, it resulted in too much signal distortion for maximum noise removal as a larger α was configured. Consequently, for this case, we substituted the decision-directed SNR estimator by the noncausal SNR estimator recommended in [10] which has been shown more effective in noise reduction. The so-obtained algorithm is referred to as LSA-NC.

Power subtraction (9) was configured with $\lambda = 5$, $\beta_1 = \beta_2 = 1$ as recommended in [2]. The floor value a_0 in (29) has the same values as the ones chosen for block thresholding ($a_0 \approx 0.05$ for partial noise removal and $a_0 = 0$ for maximum noise removal).

Both objective and subjective evaluations have been performed. The objective measures are respectively the SNR and the segmental SNR [49] defined as

$$SNR = 10 \log_{10} \frac{\sum_{n=0}^{N-1} f^2[n]}{\sum_{n=0}^{N-1} (f[n] - \hat{f}[n])^2} \quad (30)$$

$$SegSNR = \frac{1}{H} \sum_{l=0}^{H-1} \mathcal{T} \left(10 \log_{10} \frac{\sum_{n=0}^{S-1} f^2[n + lS/2]}{\sum_{n=0}^{S-1} (f[n + lS/2] - \hat{f}[n + lS/2])^2} \right) \quad (31)$$

where H represents the number of frames in the signal, S is the number of samples per frame that corresponds to 32 ms, and $\mathcal{T}(x) = \min[\max(x, -10), 35]$ confines the SNR in each frame to a perceptually meaningful range between 35 dB and -10 dB. Segmental SNR has been shown to have a higher correlation with perceived quality than SNR does [49].

Table II compares the SNR and the segmental SNR of the three denoising algorithms : block thresholding (BT), MMSE-LSA based algorithms (LSA-DD or LSA-NC) and power subtraction (PS). One can observe that the MMSE-LSA based algorithms achieved systematically a better SNR than the power subtraction method, the average gain being 0.3 dB for partial noise removal and 1.3 dB for maximum noise removal. Yet another systematic SNR improvement was achieved by block thresholding over MMSE-LSA, with an average gain of 0.9 dB for partial noise removal and 0.8 dB for maximum noise removal. With respect to segmental SNR, though the average gains are smaller, these results are confirmed: block

thresholding outperformed MMSE-LSA based algorithms which performed better than power subtraction.

The subjective evaluation was performed by a large group of 200 adult listeners. All subjects claimed to have normal hearing, 151 claimed to listen to music regularly, 58 claimed to have some general knowledge on signal processing and 26 claimed to have had experience using audio processing softwares. The authors were obviously excluded from this test.

Each subject participated in an evaluation of successively the 7 sounds mentioned above. The evaluation of each sound consisted in 3 consecutive steps: partial noise removal, maximum noise removal and a comparison between these two noise removals. For the first two steps, each subject had to rank the 3 denoising results (block thresholding, MMSE-LSA and power subtraction) according to their global appreciation of the sounds. Let us note that they had the possibility to give a same rank to several methods each time. In the third step, each subject had to select between the 2 previously top ranked denoising results (i.e., the top ranked partial denoising result and the top ranked maximum denoising result) the one they appreciated the most. In all cases, the subjects could listen to the denoising results as well as to the noisy sounds as many times as they wished. The order of the sounds and of the denoising results were randomized in order to minimize any bias. The overall test for a single subject lasted for about 15 minutes.

The subjective evaluation showed clearly that the power subtraction algorithm is by far the least favored as it obtained less than 4% top ranking votes for each of the sounds. The major complaint the subjects had about it was the strong musical noise artifact.

Table III concentrates on the comparison between block thresholding and the MMSE-LSA algorithms. Confirming the previous (segmented) SNR results, in the case of musical sounds, the subjects showed a clear preference for block thresholding over MMSE-LSA for both partial noise removal and maximum noise removal. Again, for the male speech sound TIMIT-M, block thresholding is very clearly preferred over the MMSE-LSA algorithm in the case of partial noise removal. Besides, a slight preference for block thresholding is shown for the female sound TIMIT-F in the case of maximum noise removal. On the other speech sounds (TIMIT-M with maximum noise removal and TIMIT-M with partial noise removal), the results do not show any significant difference. Table III also displays the 95% confidence intervals of the overall votes on music and speech signals. For example, the statistics show that one is 95% confident that between 48.2% and 54.5% of subjects favor block thresholding for music signals in the case of partial noise removal. These small confidence intervals, nonoverlapping in most cases, demonstrate the high reliability of this subjective evaluation and confirm the preference for block thresholding.

For musical sounds, one can explain the improvement of block thresholding over MMSE-LSA based

algorithms as follow. For partial noise removal, the residual noise is more uniform, closer to a white noise and less “metallic” than the one obtained by LSA-DD. For maximum noise removal, block thresholding produces less musical noise than LSA-NC, and it results in less distortion on signal transients. With the Piano sound for instance, which corresponds to one of the highest vote in favor of block thresholding, the clavier stroke is much less muffled by block thresholding than by LSA-NC, due to its adaptive block size adjustment as explained in Section III-C. These improvements are not significant enough for speech sounds (except for the partial noise removal of the male voice TIMIT-M for which the vote is clearly in favor of block thresholding) to lead to a clear distinction between the two algorithms.

Finally, the third step of the evaluation showed that maximum noise removal was most of the time preferred to partial noise removal. A little musical noise does not seem to be as annoying as a small residual noise. However, such preference is much stronger for musical sounds (99.2% v.s. 9.8%) than for speech sounds (71.7% v.s. 29.3%) for which intelligibility and a clear articulation (i.e., clear transients) appear to be one of the main criteria.

The block size distribution presented in Table IV shows the adaptivity of the block thresholding algorithm. The largest block size $L \times W = 8 \times 16$ is most frequently selected because it is optimal for large time-frequency regions where the signal energy is uniformly dominated by the noise energy. The blocks of size 8×1 having a narrow frequency width occur relatively often for musical signals such as Mozart recording because it matches their narrow frequency harmonics. On the contrary, the speech signal TIMIT-M privileges 2×16 blocks having a narrow time width because speech signals contain many short transients. As expected, the adaptive window size adjustment follows the signal time-frequency energy distribution properties.

V. CONCLUSION

Non-diagonal time-frequency estimators are more effective than diagonal estimators to remove noise from audio signals because they introduce less musical noise. These non-diagonal estimators are derived from a time-frequency SNR estimation performed with parameterized filters applied to time-frequency coefficients. This paper introduces an adaptive audio block-thresholding algorithm that adapts all parameters to the time-frequency regularity of the audio signal. The adaptation is performed by minimizing a Stein unbiased risk estimator calculated from the data. The resulting algorithm is robust to variations of signal structures such as short transients and long harmonics. Numerical experiments demonstrate improvements with respect to state of the art time-frequency audio denoising procedures through objective and subjective evaluations.

REFERENCES

- [1] M. Bahoura, J. Rouat, "Wavelet speech enhancement based on time-scale adaptation," *Speech Communication*, Vol 48, Issue 12, Dec. 2006, .
- [2] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP* Vol. 4, pp. 208- 211, 1979.
- [3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.* ASSP-27, pp. 113-120, 1979.
- [4] T. Cai, "Adaptive wavelet estimation: a block thresholding and oracle inequality approach," *Ann. Statist.*, 27, 898-924, 1999.
- [5] T. Cai and B.W. Silverman, "Incorporation information on neighboring coefficients into wavelet estimation", *Sankhya*, 63, 127-148, 2001.
- [6] T. Cai and H. Zhou, "A data-driven block thresholding approach to wavelet estimation", *Technical Report*, Statistics Department, University of Pennsylvania, 2005.
- [7] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah Noise Suppressor", *IEEE Trans. Speech and Audio Processing*, vol. 2, p.p. 345-349, Apr.1994.
- [8] S. Chang, Y. Kwon; S. Yang, I.Kim, "Speech enhancement for non-stationary noise environment by adaptive wavelet packet," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. 561-564 vol.1.
- [9] S.H. Chen. J.F. Wang, "Speech Enhancement Using Perceptual Wavelet Packet Decomposition and Teager Energy Operator," *The Journal of VLSI Signal Processing*, Volume 36, Numbers 2-3, February 2004, pp. 125-139(15).
- [10] I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator", *Signal Processing Letters, IEEE*, vol. 11, Issue 9, pp. 725-728, Sept. 2004.
- [11] I. Cohen, "Enhancement of Speech Using Bark-Scaled Wavelet Packet Decomposition", *Eurospeech*, 2001, Scandinavia.
- [12] I. Cohen, "Speech Enhancement Using Supergaussian Speech Models and Noncausal A Priori SNR Estimation," *Speech Communication*, Vol. 47, No. 3, Nov. 2005, pp. 336-350.
- [13] I. Cohen, "Relaxed Statistical Model for Speech Enhancement and A Priori SNR Estimation," *IEEE Trans. Speech and Audio Processing*, Vol. 13, No. 5, Sep. 2005, pp. 870-881.
- [14] I. Cohen, "Speech Spectral Modeling and Enhancement Based on Autoregressive Conditional Heteroscedasticity Models," *Signal Processing*, Vol. 86, No. 4, Apr. 2006, pp. 698-709.
- [15] I. Cohen, "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator," *IEEE Signal Processing Letters*, Vol. 9, No. 4, Apr. 2002, pp. 113-116.
- [16] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Trans. Speech and Audio Processing*, Vol. 11, No. 5, Sep. 2003, pp. 466-475.
- [17] I. Cohen and B. Berdugo, "Speech Enhancement for Non-Stationary Noise Environments," *Signal Processing*, Vol. 81, No. 11, Nov. 2001, pp. 2403-2418.
- [18] R.R. Coifman, D.L. Donoho, "Translation-Invariant De-Noising", 1995.
- [19] I. Daubechies, A. Grossmann, Y Meyer, "Painless nonorthogonal expansions", *J. Math. Phys.*, Vol. 27, No. 5, pp. 1271-1283, 1986.
- [20] D. Donoho and I. Johnstone, "Idea Spatial Adaptation via Wavelet Shrinkage", *Biometrika*, vol. 81, pp. 425-455, 1994.
- [21] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator", *IEEE. Trans. Acoust. Speech Signal Process*, 32 (6), 1109–1121, Dec. 1984.

- [22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator", *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.
- [23] Y. Ephraim. and H.L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Signal Process.* vol. 3, issue 4, pp. 251-266.
- [24] Y. Ephraim and I. Cohen, "Recent Advancements in Speech Enhancement," in *Richard C. Dorf (Ed.), The Electrical Engineering Handbook*, Edition, CRC Press, 2005, Ch. 15, Sec. 2, pp. 15-12 C 15-26.
- [25] Y. Ephraim, H. Lev-Ari and W. J. J. Roberts, "A brief survey of Speech Enhancement," *The Electronic Handbook*, CRC Press, April 2005.
- [26] O. Farooq, S. Datta, "Wavelet-based denoising for robust feature extraction for speech recognition," *Electronics Letters*, Vol. 39, Issue 1, pp 163- 165, Jan. 2003.
- [27] J. S. Garofolo, "Getting Started with the DARPA TIMIT CD-ROM: an Acoustic Phonetic Continuous Speech Database," National Institute of Standards and Technology (NIST), Gaithersburgh, MD.
- [28] S. Ghael, A. Sayeed and R. Baraniuk, "Improved wavelet denoising via empirical wiener filtering", *Proceedings for SPIE, Mathematical Imaging*, San Diego, July 1997.
- [29] Y. Ghanbari and M. R. Karami-Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets," *Speech Communication*, vol. 48, no. 8, 2006, pp.927-940.
- [30] Y. Hu and P.C. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Processing Letters*, Vol. 11, Issue 2, Part 2, pp. 270- 273, Feb. 2004.
- [31] P. Hall, G. Kerkycharian and D. Picard, "A note on the wavelet oracle", *Statistics and Probability Letters*, 43, 415-420, 1999.
- [32] P. Hall, G. Kerkycharian and D. Picard, "Block threshold rules for curve estimation using kernel and wavelet methods", *Ann. Statist*, 26, 922-942, 1998.
- [33] P. Hall, G. Kerkycharian and D. Picard, "On the minimax optimality of block thresholded wavelet estimators", *Statistica Sinica*, 9, 33-50, 1999.
- [34] M. Johnson, X. Yuan, Y. Ren, "Speech signal enhancement through adaptive wavelet thresholding," *Speech Communication*, Vol. 49 , Issue 2, Feb. 2007.
- [35] N.S.Kim, J.H. Chang, "Spectral enhancement based on global soft decision," *Signal Processing Letters, IEEE*, Vol. 7, Issue 5, pp: 108-110, May 2000.
- [36] I.J.Kim, S.I. Yang, Y. Kwon, "Speech enhancement using adaptive wavelet shrinkage," *Proceedings of IEEE International Symposium on Industrial Electronics, 2001*, vol. 1, pp. 501-504, 2001.
- [37] M.Li, H.G. McAllister, N. D. Black, De Perez T.A., "Perceptual time-frequency subtraction algorithm for noise reduction in hearing aids," *Biomedical Engineering, IEEE Transactions on*, Vol. 48, Issue 9, pp: 979-988, Sept. 2001.
- [38] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proc. of the IEEE*, vol.67, Dec. 1979.
- [39] S.F. Lei, Y.K. Tung, "Speech enhancement for nonstationary noises by wavelet packet transform and adaptive noise estimation," *Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems, 2005*, pp. 41-44, Dec. 2005.
- [40] C.T. Lu, H. C. Wang, "Enhancement of single channel speech based on masking property and wavelet transform," *Speech Communication*, Vol. 41, No. 2, Oct. 2003, pp. 409-427(19).

- [41] C.T. Lu and H.C. Wang, "Speech enhancement using perceptually-constrained gain factors in critical-band-wavelet-packet transform," *Electronics Letters*, March 2004, volume 40, Issue 6, pp. 394-396.
- [42] D. Malah, R.V. Cox and A.J. Accardi, "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments," *Proc. IEEE Int'l Conf. Acoust., Speech, Signal Processing - ICASSP99, Phoenix, Arizona, March 1999*.
- [43] S. Mallat, "A Wavelet Tour of Signal Processing, 2nd edition", New York Academic, 1999.
- [44] R. Martin, "Speech enhancement using MMSE short-time spectral estimation with gamma speech prior," *Proc. ICASSP 02, Orlando, FL, 2002*, pp. I-253-I-256.
- [45] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process*, 9 (5) (July 2001) 504-512.
- [46] G. Matz and F. Hlawatsch, "Minimax robust nonstationary signal estimation based on a p-point uncertainty model," *J. Franklin Inst., special issue on Time-Frequency Signal Analysis and Applications*, vol. 337, no. 4, July 2000, pp. 403-419.
- [47] G. Matz, F. Hlawatsch, and A. Raidl, "Signal-adaptive robust time-varying Wiener filters: Best subspace selection and statistical analysis," *Proc. IEEE ICASSP-01, Salt Lake City (UT), May 2001*, pp. 3945-3948.
- [48] R.J. McAulay, and M.L. Malpass, "Speech enhancement using soft decision noise suppression filter", *IEEE Trans. Acoust., Speech, Signal Process*, ASSP-28, pp.137-145, 1980.
- [49] S. R. Quackenbush, T. P. Barnwell and M. A. Clements, "Objective Measures of Speech Quality", Prentice-Hall, New York, NY, 1988.
- [50] J.W. Seok, K. S. Bae, "Speech enhancement with reduction of noise components in the wavelet domain," *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, Vol.2, pp. 1323-1326.
- [51] Y. Shao and C. H. Chang, "A generalized time-frequency subtraction method for robust speech enhancement based on Wavelet filter bank modeling of human auditory system," *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, Vol.37, No.4, pp. 877- 889, 2007.
- [52] H. Sheikhzadeh and H. R. Abutalebi, "An improved wavelet-based speech enhancement system", *EUROSPEECH*, 2001, 1855-1858.
- [53] K. V. Sørensen and S. V. Andersen "Speech Enhancement with Natural Sounding Residual Noise Based on Connected Time-Frequency Speech Presence Regions," *EURASIP Journal on Applied Signal Processing*, Volume 18, Issue 18, Pages 2954-2964, 2005.
- [54] C. Stein and W. James, "Estimation with quadratic loss", *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1 (Berkeley, University of California Press)*, 361-379, 1961.
- [55] C. Stein, "Estimation of the mean of a multivariate normal distribution", *Ann. Statist.* 9 1135-1151, 1980.
- [56] H. Tolba, "A time-space adapted wavelet de-noising algorithm for robust automatic speech recognition in low-SNR environments," *Proceedings of the 46th IEEE International Midwest Symposium on Circuits and Systems, 2003*, volume 1, pp. 311-314.
- [57] J. S. Walker, "Denoising Gabor Transforms", *submitted*.
- [58] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement", *IEEE Workshop on Statistical Signal Processing*, pp. 496-499, Aug. 2001.
- [59] J. Yang, "Frequency domain noise suppression approaches in mobile telephone systems," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993*, vol. 2, pp. 363-366.

- [60] G. Yu, E. Bacry and S. Mallat, “Audio signal denoising with complex wavelets and adaptive block attenuation”, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hawaii, 2007.*

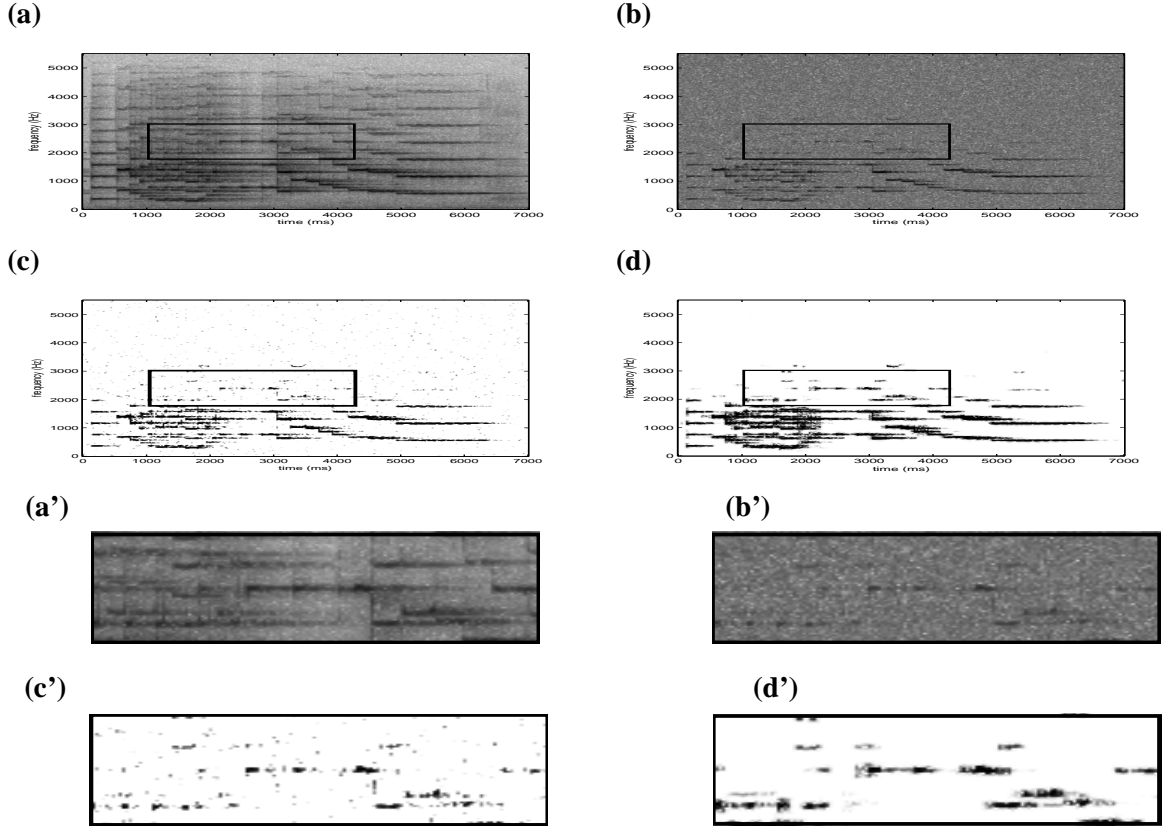


Fig. 1. (a),(b): Log-spectrograms of the original and noisy “Mozart” signals. (c),(d): attenuation coefficients calculated with a power subtraction and a block thresholding. Black pixels correspond to 1 and white to 0. (a’)(b’)(c’)(d’): zooms over rectangular regions indicated in (a)(b)(c)(d).

$B_i^\#$	4	8	16	32	64	128
λ	4.7	3.5	2.5	2.0	1.8	1.5

TABLE I

THRESHOLDING LEVEL λ CALCULATED FOR DIFFERENT BLOCK SIZE $B^\#$ WITH $\delta = 0.1\%$.

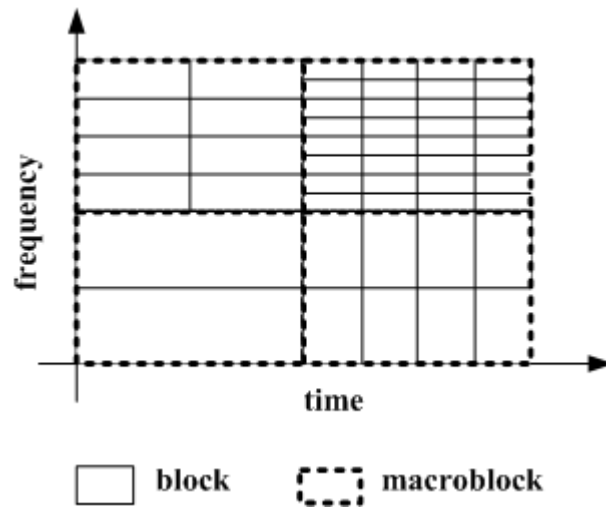


Fig. 2. Partition of macroblocks into blocks of different sizes.

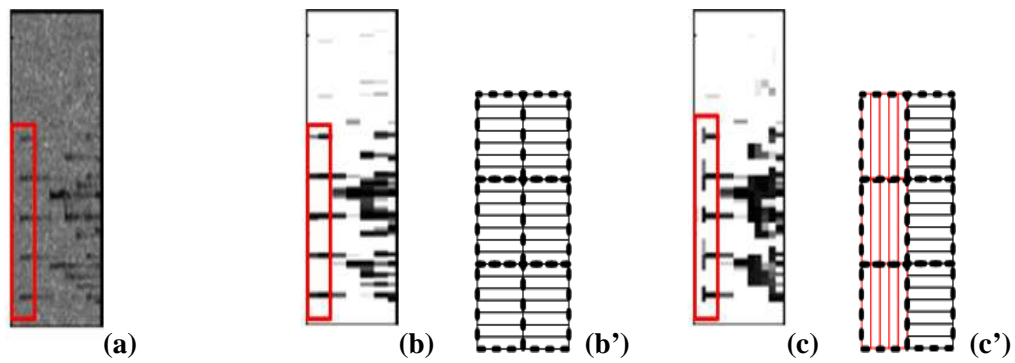


Fig. 3. Zoom on the onset of “Mozart”. (a): Log-spectrogram. (b): Attenuation coefficients of a fixed block thresholding. (b’): Block sizes in the time-frequency rectangle at the signal onset. (c): Attenuation coefficients of an adaptive block thresholding. (c’): Adapted block sizes at the signal onset.

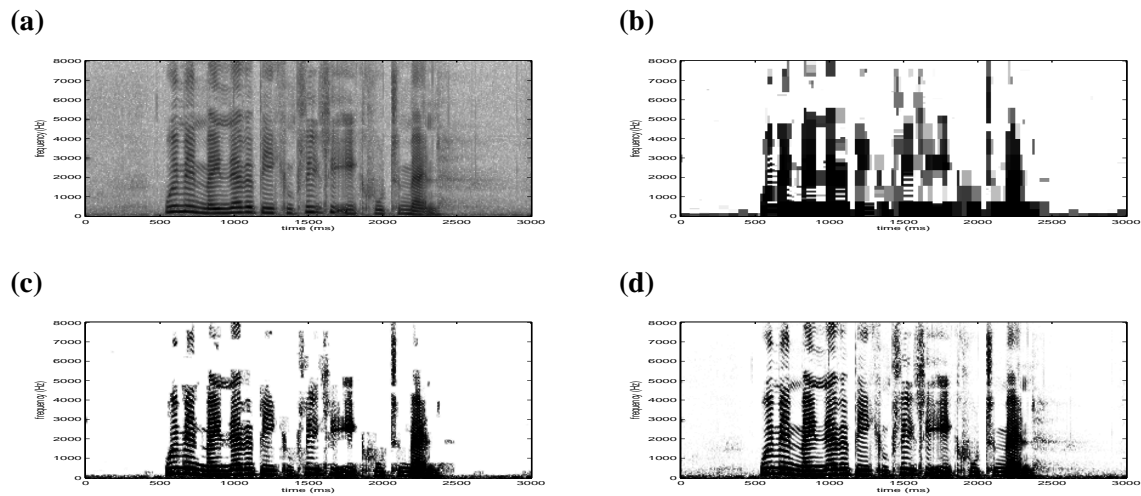


Fig. 4. (a): log-spectrogram of “TIMIT-F”. (b),(c),(d): attenuation coefficients respectively of a block thresholding, of a non-diagonal Wiener estimator, and of an oracle estimator.

Signal & SNR	Partial Noise Removal			Maximum Noise Removal		
	PS	LSA-DD	BT	PS	LSA-NC	BT
Mozart -2.73 dB	8.68	8.91	11.12	8.75	10.15	11.90
Mozart 3.46 dB	13.01	13.21	14.46	12.92	14.01	14.45
Mozart 9.23 dB	17.17	17.93	18.40	16.98	18.10	18.45
Mozart 14.73 dB	21.11	21.12	22.49	20.87	21.99	22.43
Piano 4.75 dB	17.70	18.24	19.95	18.30	19.45	20.47
TIMIT-M 10.76 dB	18.65	18.84	19.46	18.55	19.16	19.70
TIMIT-F 20.63 dB	25.15	25.21	26.46	24.95	25.88	26.38

Signal & SSNR	Partial Noise Removal			Maximum Noise Removal		
	PS	LSA-DD	BT	PS	LSA-NC	BT
Mozart -5 dB	6.32	7.17	8.53	6.80	8.23	9.77
Mozart 0 dB	10.56	11.61	12.12	10.76	12.14	12.24
Mozart 5 dB	14.79	15.87	15.92	14.79	16.01	16.14
Mozart 10 dB	18.68	19.31	19.96	18.52	19.78	19.90
Piano -5 dB	5.74	6.70	7.53	6.77	8.42	8.94
TIMIT-M 0 dB	9.16	9.97	9.98	9.61	10.85	11.02
TIMIT-F 10 dB	15.04	15.70	16.51	14.88	15.67	16.45

TABLE II

COMPARISON OF POWER SUBTRACTION (PS), EPHRAIM AND MALAH (LSA-DD OR LSA-NC) AND BLOCK THRESHOLDING (BT) ALGORITHMS, ON 4 TYPES OF NOISY SIGNALS WITH DIFFERENT NOISE LEVELS. THE TOP TABLE GIVES THE SNR VALUES FOR PARTIAL NOISE REMOVAL AND MAXIMUM NOISE REMOVAL, AND THE BOTTOM TABLE GIVES THE SEGMENTAL SNR VALUES.

Signal & SSNR	Partial Noise Removal			Maximum Noise Removal		
	BT	LSA-DD	<i>EQU.</i>	BT	LSA-NC	<i>EQU.</i>
Mozart -5 dB	47.0	26.0	27.0	80.1	10.5	9.4
Mozart 0 dB	47.3	21.6	31.1	44.1	37.5	18.4
Mozart 5 dB	53.2	22.8	24.0	40.4	38.7	20.9
Mozart 10 dB	54.7	12.0	33.3	41.3	24.7	34.0
Piano -5 dB	54.7	29.3	16.0	70.0	12.1	17.9
TIMIT-M 0 dB	61.9	10.7	27.4	39.4	38.5	22.1
TIMIT-F 10 dB	34.5	30.9	34.5	37.0	26.0	37.0
Music	51.4	22.3	26.3	55.2	24.7	20.1
95% CI	(48.2, 54.5)	(19.8, 25.0)	(23.6, 29.1)	(52.1, 58.3)	(22.1, 27.5)	(17.7, 22.7)
Speech	48.2	20.8	31.0	38.2	32.3	29.5
95% CI	(43.2, 53.2)	(16.9, 25.1)	(26.5, 35.8)	(33.4, 43.1)	(27.7, 37.1)	(25.1, 34.2)

TABLE III

SUBJECTIVE COMPARISON BETWEEN BLOCK THRESHOLDING (BT) AND EPHRAIM AND MALAH (LSA-DD AND LSA-NC), FOR PARTIAL NOISE REMOVAL AND MAXIMUM NOISE REMOVAL. THE COLUMNS BT AND LSA GIVE THE PERCENTAGE OF LISTENERS THAT PREFERRED THE CORRESPONDING ALGORITHM OVER THE OTHER ONE, FOR EACH NOISY SIGNAL. THE COLUMN EQU. GIVES THE PERCENTAGE OF LISTENERS FOR WHOM THE QUALITY OF BOTH ALGORITHMS IS EQUAL. THE LAST TWO TABLE ROWS AGGREGATE THE RESULTS FOR ALL MUSIC SIGNALS (MOZART AND PIANO) AND ALL SPEECH SIGNALS (TIMIT-M AND TIMIT-F), AND THEY GIVE THE 95% CONFIDENCE INTERVAL (CI) DERIVED FROM THE NUMBER OF LISTENERS.

Mozart	$W = 16$	$W = 8$	$W = 4$	$W = 2$	$W = 1$
$L = 8$	25.3	10.4	5.2	4.0	11.5
$L = 4$	10.7	4.2	3.0	1.9	3.6
$L = 2$	5.1	2.5	2.2	3.0	7.3
TIMIT-M	$W = 16$	$W = 8$	$W = 4$	$W = 2$	$W = 1$
$L = 8$	26.4	9.1	6.3	1.7	3.9
$L = 4$	12.3	7.8	1.5	1.3	2.4
$L = 2$	11.9	6.7	3.0	1.7	3.9

TABLE IV

PERCENTAGE OF THE DIFFERENT BLOCK SIZE SELECTED BY THE BLOCK THRESHOLDING ALGORITHM FOR MOZART (TOP) AND TIMIT-M (BOTTOM).