An improved time-frequency noise reduction method using a psycho-acoustic Mel model

Samir Ouelha, Abdeldjalil Aissa El Bey, Boualem Boashash

An improved time-frequency noise reduction method using a psycho-acoustic Mel model

Samir Ouelha[a,*], Abdeljalil Aïssa-El-Bey[b], Boualem Boashash[a]

[a] *Qatar University, Department of Electrical Engineering, Doha, Qatar*
[b] *IMT Atlantique, UMR CNRS 6285 Lab-STICC, Université Bretagne Loire, F-29238 Brest, France*

Abstract

This paper addresses the problem of noise reduction in non-stationary signals. The paper first describes an improved human physiology based time-frequency (TF) representation (HPTF) using Mel filterbanks. Then, an improved noise reduction algorithm is presented, which does not require any *a priori* information about the signal of interest and the noise. This efficient noise reduction algorithm is implemented using an original wavelet shrinkage method. The overall method results in an original denoised TF representation called denoised HPTF (DHPTF). From this representation one can reconstruct a denoised time-domain signal and therefore define a new noise reduction algorithm, whose performance is evaluated and compared with state-of-the-art methods. The performance assessment uses several criteria: (1) signal-to-noise-ratio (SNR), (2) segmental SNR (SSNR) and (3) mean square error (MSE). The results indicate an improvement of up to 4.72 dB

*Corresponding author
Email addresses: samir_ouelha@hotmail.fr (Samir Ouelha), abdeldjalil.aissaelbey@imt-atlantique.fr (Abdeljalil Aïssa-El-Bey), boualem.boashash@gmail.com (Boualem Boashash)

w.r.t the SNR, 2.79 dB w.r.t the SSNR and 4.72 dB w.r.t the MSE for a speech database signals corrupted with four different noises. In addition, other applications such as EEG signal enhancement show promising results.
Keywords:
Time-frequency analysis, psycho-acoustic model, noise reduction, signal enhancement, wavelet thresholding, Mel filterbank.

1. Introduction

Most real signals are non-stationary, however traditional time-domain or frequency-domain representations are inadequate to analyze such signals because they assume the signal as stationary. Instead, one can use joint time-frequency (t, f) representations as they were found to be better to process such signals. Two family of time-frequency (TF) methods have been widely used in the state-of-the-art: (1) linear TF and (2) quadratic TF representations [1, 2, 3]. Linear methods such as short-time Fourier transform (STFT) are the most used in practice because they are cross-terms free (when components are spaced enough in the TF domain [1, Section 4.1]) and computationally efficient [4]. The main drawback of these types of representations are their poor resolution performance. Quadratic methods have shown improved resolution performance but generally they required the setting of several parameters to obtain a good trade-off between resolution performance and cross-terms suppression [1]. Therefore, it could be difficult for a non-expert to get the best TF representation; in addition optimal parameters are generally signal dependent, therefore such methods are not suitable for an automatic classification system (e.g. automatic speech recognition). To

overcome the last limitation, signal dependent kernel methods have been developed with automatic parameters selection [5, 6], however these methods are not computationally efficient for long duration signals (e.g. speech signals).

Another difficulty for the processing of real signals is that they are generally corrupted by noise. In many applications, such as geophysics [7, 8], EEG abnormalities detection [9] or speech recognition [4, 10], efficient signal enhancement techniques are required [11]. In the literature, there are several methods to suppress noise that depend on the knowledge of characteristics of the useful signal and/or the noise. Some algorithms require *a priori* knowledge about the signal and noise second order statistics [12], while others only require knowledge of the noise spectral density (e.g. Wiener filtering) [13]. Unfortunately, in real applications these information are not available and must be estimated [14]. Other studies made the assumption of Gaussian or sub-Gaussian noise in order to use wavelet based denoising approaches [15, 16]. This is a rough assumption, as in real-life there are various noise sources [17]. Furthermore, in mobile communications, the signal of interest is speech and it often arises from conversations that take place in noisy and non-stationary environments such as inside a car, in the street, or inside airports. In such a case there is no justification to assume Gaussian noise. Therefore, noise reduction methods based on this *ideal* assumption may likely fail in real life applications [18]. Many authors proposed modelling the noise, but these techniques are application dependent and cannot be used in different situations [19, 20, 21].

This paper describes an improved denoised TF representation and blind noise

reduction method that performs well without prior information about the signal and noise. The proposed TF representation is based on a psycho-acoustic TF model and it deals effectively with the non-stationarity of signal and noise. It is based on the finding that the basilar membrane inside the cochlea is usually conceived as a bank of band-pass filters that have logarithmically increasing bandwidth [22]. In this study, a Mel filterbank is used to construct the resulting TF representation as it has shown promising results in modelling the human cochlea [22]. Some of the material presented in this paper has been presented in [23, 24]; the main contribution of this study is to design improved algorithm for noise variance estimation with performance supported by extensive experimental comparisons.

This paper is organized as follows; Section 2 reviews the main principles of the TF representation based on Mel filters called HPTF . Section 3 describes a method to reduce noise in the HPTF . After that, Section 4 discusses reconstructing the signal of interest from the denoised HPTF (DHPTF). Section 5 presents experiments and discusses the results. Finally, section 6 concludes the study and summarizes main findings.

2. HPTF representation

2.1. Principle

Previous studies observed that the human ear acts like filters, which are concentrated only on certain frequencies [25]. Mel filterbank is a psychoacoustic model which represents how humans perceives the sound [22]. These Mel filters are non-uniformly spaced on the frequency axis, with more filters in the low frequency regions and less in high frequency regions. More pre-

cisely, Mel filters are triangular shaped filters with respect to the Mel scale. This scale is given by the following formula for a given frequency f in Hz [22]:

$$\text{mel}\,(f) = 2595\,\log_{10}\left(1 + \frac{f}{700}\right). \tag{1}$$

Thus, the Mel frequency scale is almost linear below 1000 Hz and logarithmic above. If we consider M Mel filters, $H_m(f)$, each of them is centered on a frequency f_m, for $m = 2, 3, \ldots, M-1$, and has a bandwidth $B(m)$ defined as follows:

$$B(m) = f_{m+1} - f_{m-1}, \quad \forall m = 2, 3, \ldots, M-1. \tag{2}$$

The center frequency f_m is calculated from its corresponding center frequency on the Mel scale using the following inverse formula obtained from Eq. (1):

$$f_m = 700\,\left(10^{\frac{\text{mel}\,(f_m)}{2595}} - 1\right), \tag{3}$$

where:

$$\text{mel}\,(f_m) = \frac{m}{M+1}\left(\text{mel}\,(f_{max}) - \text{mel}\,(f_{min})\right), \forall m = 1, 2, \ldots, M, \tag{4}$$

where f_{max} and f_{min} correspond respectively to the highest and the lowest frequencies of the input signal (generally $f_{min} = 0$ and $f_{max} = \frac{F_s}{2}$, where F_s is the sampling frequency).

Therefore, the impulse response $h_m(t)$ that corresponds the Mel filter $H_m(f)$ can be expressed as:

$$\begin{aligned} h_m(t) &= \int_{-\infty}^{\infty} H_m(f)\, e^{j2\pi\, ft}\, df \qquad (5)\\ &= \frac{1}{2\pi^2 t^2}\left(\frac{\cos\left(2\pi t f_{m-1}\right) - \cos\left(2\pi t f_m\right)}{f_{m-1} - f_m} + \frac{\cos\left(2\pi t f_{m+1}\right) - \cos\left(2\pi t f_m\right)}{f_m - f_{m+1}}\right). \end{aligned}$$

Fig. 1 shows an example of Mel filter bank for $M = 10$, $f_{min} = 0$ Hz and $f_{max} = 11025$ Hz, while Fig. 2 presents the impulse responses corresponding to $h_2(t)$ and $h_8(t)$ respectively.

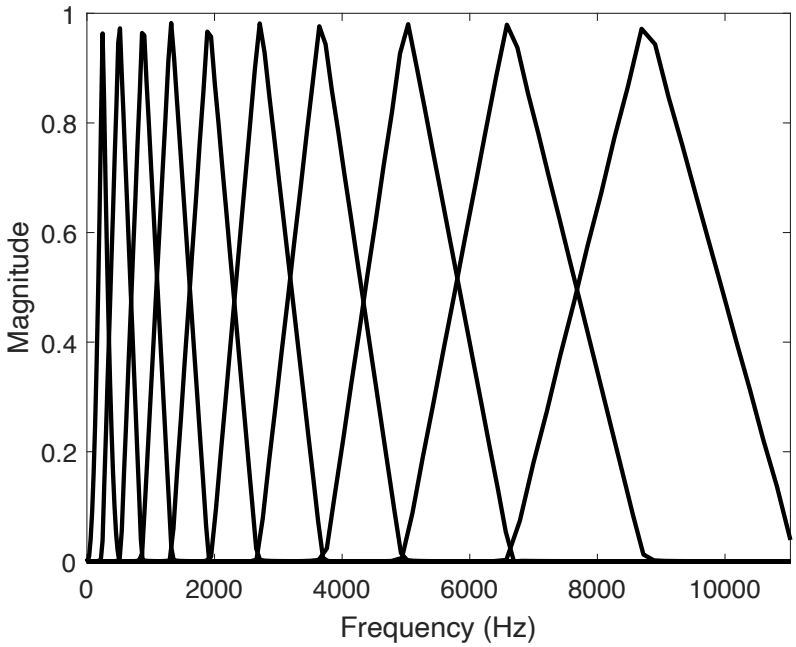
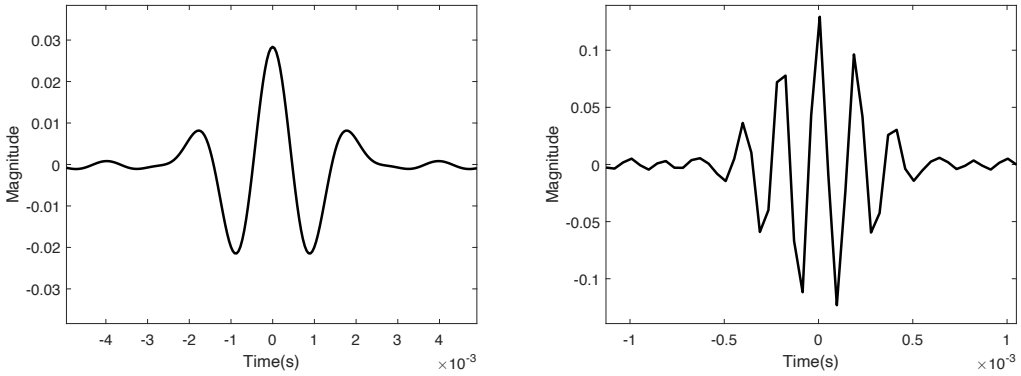


Figure 1: Representation of Mel filterbank $H_m(f) \forall m = 1 \dots 10$ with $M = 10$



(a) (b)

Figure 2: Impulse response corresponding to (a) $h_2(t)$ and (b) $h_8(t)$

2.2. HPTF construction

Let $\boldsymbol{z} \in \mathbb{R}^N$ be a vector of N samples containing data, obtained from an analog signal recorded by sensors and sampled at frequency F_s. This observation is a superposition of signal of interest $\boldsymbol{s} \in \mathbb{R}^N$ and noise $\boldsymbol{\epsilon} \in \mathbb{R}^N$:

$$\boldsymbol{z} = \boldsymbol{s} + \boldsymbol{\epsilon}. \tag{6}$$

The m^{th} row of the HPTF shown in Fig. 3, denoted by $\boldsymbol{z}_m$, is the convolution product between observation $\boldsymbol{z}$ and the sampled impulse response $\boldsymbol{h}_m, \quad \forall\{m = 1 \dots M\}$ such that:

$$\boldsymbol{z}_m = \boldsymbol{z} * \boldsymbol{h}_m. \tag{7}$$

By using the linear property of the convolution, $\boldsymbol{z}_m$ is the sum of the filtered signal of interest and the filtered noise, such that:

$$\boldsymbol{z}_m = \boldsymbol{s} * \boldsymbol{h}_m + \boldsymbol{\epsilon} * \boldsymbol{h}_m = \boldsymbol{s}_m + \boldsymbol{\epsilon}_m. \tag{8}$$

Eq. (7) corresponds to a filtering process in the $H_m(f)$ bandwidth, where $H_m(f)$ is the Mel filter centered on the f_m frequency, according to Mel's scale (see Fig. 1). Therefore, $\boldsymbol{z}_m$ represents the spectral information of the input signal $\boldsymbol{z}$ around the frequency f_m in the time-domain.

One can notice that the number of samples used to describe the impulse response $\boldsymbol{h}_m$ depends on the frequency f_m. Fig. 1 shows that $H_m(f)$ bandwidth is small for low frequencies, and conversely. As a consequence, the impulse response time support is smaller for high frequencies than for low frequencies; this is in accordance with the Heisenberg uncertainty principle [1, Chapter 2]. Hence, to take into account this specificity, if L_m denotes the number of samples needed to describe the impulse response, each filter satisfies the following constraint[1]:

$$p > q \Rightarrow L_p < L_q, \forall (p,q) \in \{1 \dots M\} \times \{1 \dots M\}. \tag{9}$$

Therefore, it is possible to build a TF representation that extends the sonogram method [1, Chapter 2]. Each frequency channel corresponds to the center frequency f_m, $\forall m = 1 \dots M$, by taking the square magnitude of each filtered data. Applying this process for each time lag $n \in [1, L_m]$ provides the instantaneous power distribution of the signal filtered by the Mel filter bank:

$$\rho_z[k,m] = \left| \sum_{n=1}^{L_m} z[k-n] h_m[n] \right|^2 . \tag{10}$$

Eq. (10) corresponds to the square modulus of a convolution product in its discrete form. Fig. 3 presents the HPTF representation obtained by

[1] the number of samples L_m is computed such as 99.9 % of the energy is conserved

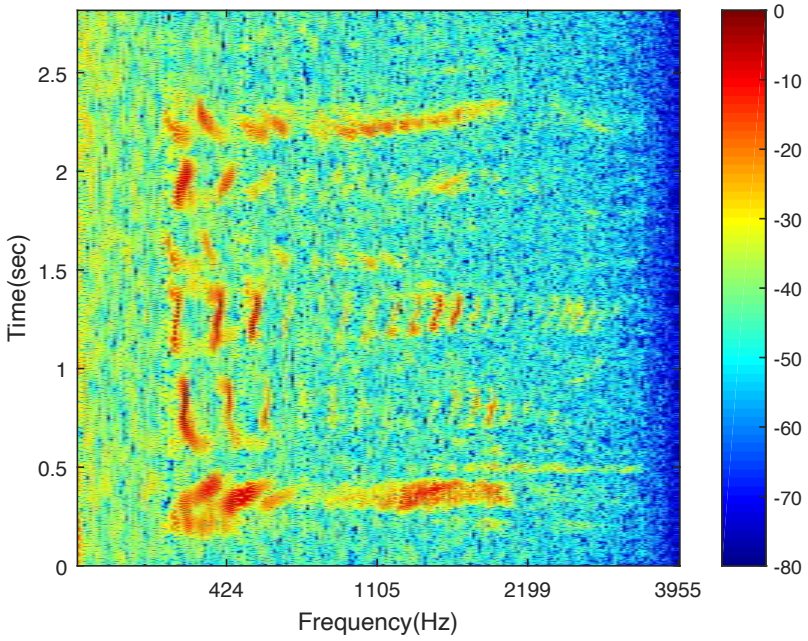


Figure 3: Hearingogram of a speech signal sampled at 8 kHz and corrupted by pink noise (SNR = 10 dB), with $M = 200$ Mel filters. This plane contains $M \times N$ pixels.

applying the proposed process to a speech signal sampled at 8 kHz. For this experiment, the number M of Mel filters considered equals 200.

2.3. Effect of the number of filters

The number of filters M is the only parameter needed for the construction of the HPTF . There are two ways to explain the effect of this parameter: (1) if the number of Mel filters tends towards infinity, the bandwidth $B(m)$ of each Mel filter, described by Eq. (2), tends towards 0. So that the Mel filter $H_m(f)$ becomes similar to a Dirac centered on the frequency f_m. Thus, the impulse response $\boldsymbol{h}_m$ tends towards a sinusoidal function; therefore the HPTF reduces to a spectrogram using a rectangular windowing function.

(2) The higher the number of filters the smaller the width of each filter, which means that the impulse response needs to be described using more samples; this results in losing the temporal resolution. Therefore, to set the parameter M there is a trade-off between time and frequency resolution.

Fig. 4 presents four HPTF plots, with different values of M. For the HPTF , presented in Fig. 4a $M = 10$ (i.e dimension is $10 \times N$), M equals 20 for Fig. 4b, 100 for Fig. 4c and 800 for Fig. 4d.

One can see that even with M equal to 10, significant information can be extracted (see Fig. 4a). Thus it is possible to quickly extract some signal features useful for signal identification purposes. For a higher number of Mel filters, as the frequency resolution is higher, more details are visible; in particular the harmonics are well defined, but at the expense of a poorer time resolution. Fig. 4d highlights the degradation of time resolution caused by the extremely high number of filters ($M = 800$). The next section describes a method to reduce the noise in the HPTF method based on wavelet shrinkage.

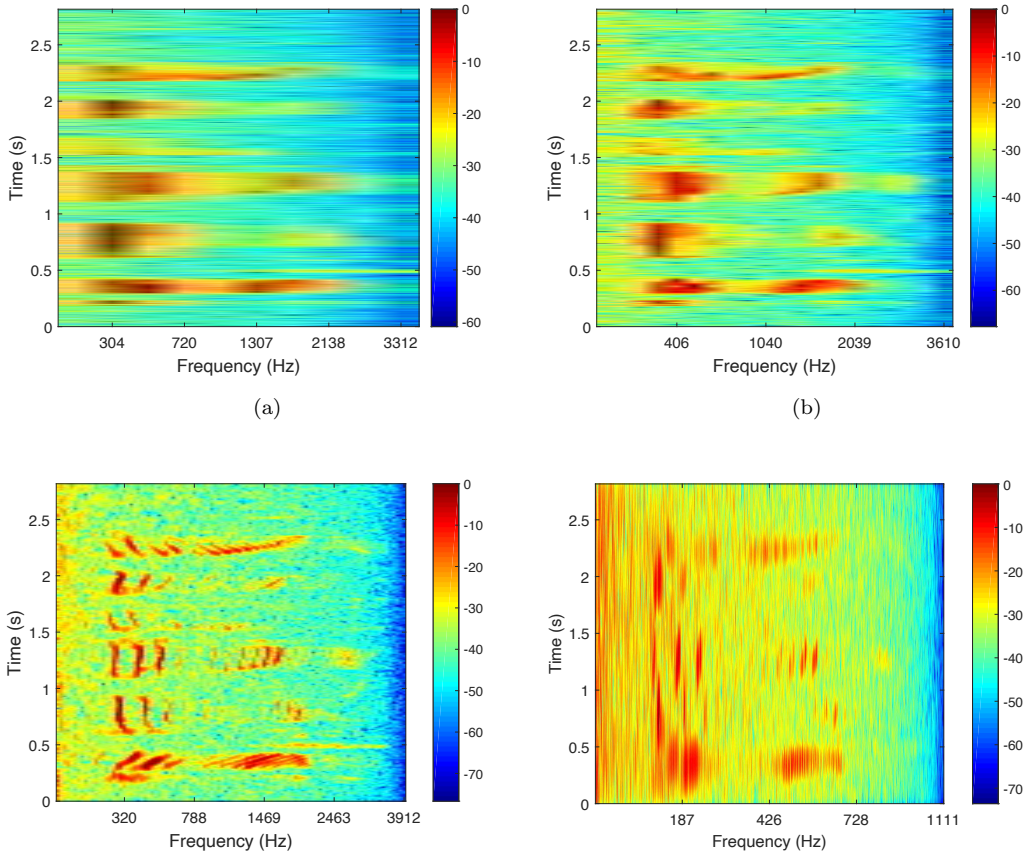


Figure 4: Effect of the number of filters M using (a) 10 Mel filters; (b) 20 Mel filters; (c) 100 Mel filters and (d) 800 Mel filters. (Based on visual inspection one can deduce that $M = 100$ filters is a good trade-off between time and frequency resolution quality for this signal.)

2.4. Characterization of the noise in the HPTF representation

Let us first determine the expected value of the filtered noise $\boldsymbol{\epsilon}_m$ defined by:

$$\boldsymbol{\epsilon}_m = \boldsymbol{\epsilon} * \boldsymbol{h}_m \tag{11}$$

so for a given bandwidth frequency, the noise in the (t, f) plane in its discrete version, is defined by:

$$\epsilon_m[k] = \sum_{n=1}^{L_m} \epsilon[k-n]\, h_m[n]. \tag{12}$$

So the expected value of $\boldsymbol{\epsilon}_m$ is given by :

$$\begin{aligned} \mu_\epsilon^{(m)} &= \mathbb{E}\left\{\epsilon_m[k,m]\right\} \\ &= \mathbb{E}\left\{\sum_{n=1}^{L_m} \epsilon[k-n]\, h_m[n]\right\} \\ &= \sum_{n=1}^{L_m} \mathbb{E}\left\{\epsilon[k-n]\right\}\, h_m[n], \end{aligned} \tag{13}$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation operator. Assuming that the $\mathbb{E}\left\{\epsilon[n]\right\} = \overline{\epsilon}$, $\forall n = 1, \ldots, N$, one can get:

$$\mu_\epsilon^{(m)} = \overline{\epsilon} \sum_{n=1}^{L_m} h_m[n]. \tag{14}$$

Thus from Eq. (5), one can notice that the impulse responses $\boldsymbol{h}_m$ are centered, and therefore the filtered noise $\boldsymbol{\epsilon}_m$ is zero mean. Now, let us characterize the probability density function (PDF) of the filtered noise $\boldsymbol{\epsilon}_m$. Assuming that the smallest impulse response length L_M is still large enough, and given that the convolution product is a sum of random variables, one can invoke the central limit theorem and approximate the PDF of $\boldsymbol{\epsilon}_m$ as Gaussian, such that:

$$\epsilon_m[n] \hookrightarrow \mathcal{N}(0, \sigma^2_{\epsilon_m}) \qquad n = 1, \ldots, N. \tag{15}$$

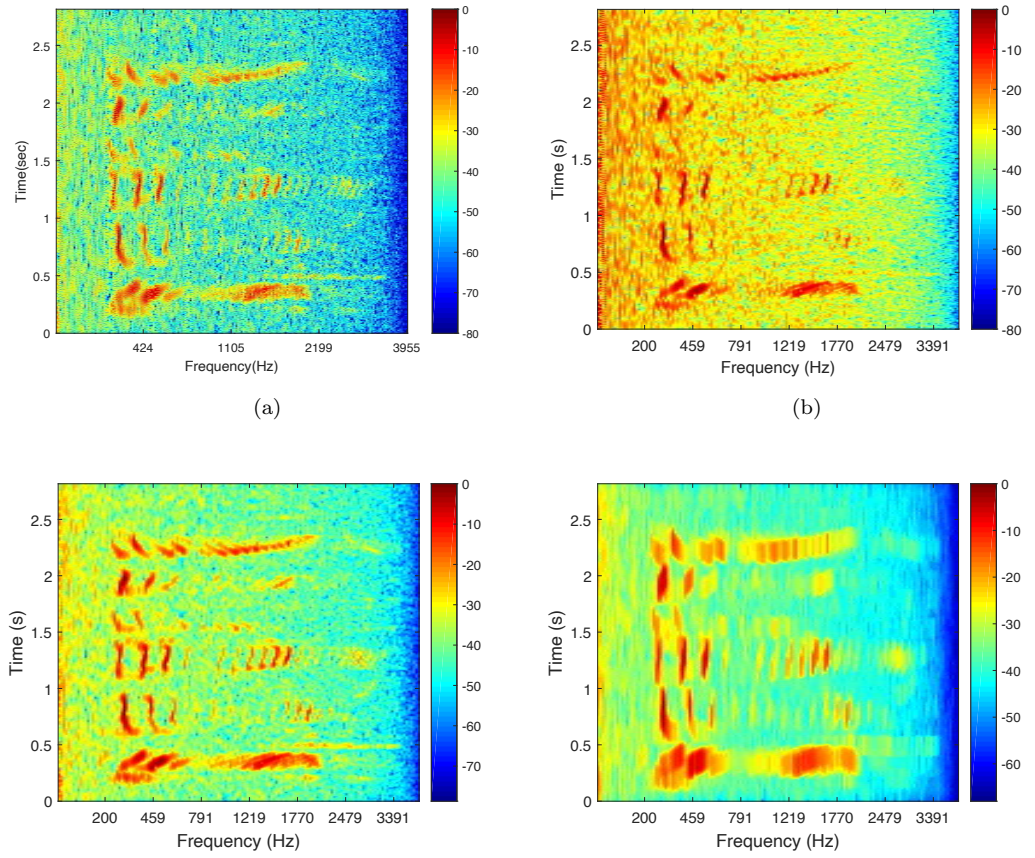
Note that the only assumption made on the noise is that $\mathbb{E}\left\{\epsilon[n]\right\} = \overline{\epsilon}$; the PDF of the filtered noise, given in Eq. (15), is a consequence of the transformation done using the HPTF representation. Section 3.1 describes the method to estimate the variance of $\boldsymbol{\epsilon}_m$, denoted by $\sigma^2_{\epsilon_m}$.

2.5. Noise attenuation based on local integration

To reduce the noise level, one of the most intuitive approaches is to integrate the HPTF along the time axis by locally averaging the instantaneous power distribution such as:

$$\rho_z^{(smooth)}[k_0, m] = \frac{1}{K} \sum_{k=k_0}^{k_0+K-1} \rho_z[k, m], \tag{16}$$

where K corresponds to the number of samples used for the denoising step; the higher this number, the higher the Signal to Noise Ratio (SNR) gain is but with more signal degradation.



(a) (b)

(c) (d)

Figure 5: Illustration of the denoising process based on the smoothing effect as described in Eq. 16. (a) Without smoothing, (b) with a smoothing of length $L_M/64$ (c) of length $L_M/8$ and (d) of length L_M. (One can observe that the best trade-off between the degradation of signal of interest (by smoothing effect) and noise reduction is obtained when the size of the smoothing filter K equals $\frac{L_M}{8}$)

Fig. 5 shows the effect of the proposed smoothing operation using different length K of the filter.

Several experiments led to choose $K = \frac{L_M}{8}$ as a good trade-off between SNR gain and the degradation of the signal of interest. This smoothing operation

reduces the noise power level but also affects the signal of interest; this is not convenient for signal interpretation. For this reason, Section 3 presents an improved approach based on an original wavelet shrinkage method to reduce noise level while retaining the signal of interest.

3. Denoising of the HPTF principle

3.1. Key steps of the denoising method

The denoising algorithm is based on three key steps: (1) Discrete wavelet transform (DWT) of the noisy data (i.e. a multiresolution analysis) to obtain a set of wavelet coefficients [26]; (2) application of a thresholding rule to the wavelet coefficients; (3) estimation of the signal of interest by applying an inverse DWT to the thresholded wavelet coefficients.
This reduction noise process is applied on the filtered data $\boldsymbol{z}_m \, \forall m = 1, \ldots, M$. This method improves the one described in [27] for Gaussian noise reduction; it is based on five key points, as outlined below:

1. Construction of a TF representation using Eqs. (7) (i.e. computation of $\boldsymbol{z}_m$ for $m = 1, \ldots, M$);
2. Estimation of a non-constant threshold for each $\boldsymbol{z}_m$, denoted by $\boldsymbol{\lambda}_m$;
3. Wavelet coefficients thresholding;
4. Multiresolution synthesis from thresholded wavelet coefficients to get a signal approximation $\widehat{\boldsymbol{s}}_m$, within the $H_m(f)$ bandwidth;
5. DHPTF construction taking the square value of $\widehat{\boldsymbol{s}}_m$, for $m = 1, \ldots, M$.

This method finds an estimate of $\boldsymbol{s}_m$ denoted $\widehat{\boldsymbol{s}}_m$. The first stage corresponds to the computation of $\boldsymbol{z}_m$ (see Eq. (7)) for $m = 1, \ldots, M$. The second

stage consists in defining thresholds for wavelet shrinkage. To explain these steps in detail, let us recall that Eq. (15) shows that the PDF of the filtered noise $\boldsymbol{\epsilon}_m$ is centered and Gaussian. Based on this approximation, one can use the universal threshold [27], which is a simple entropy measure solely dependent on the number of samples in $\boldsymbol{z}_m$, denoted N, as a threshold to be applied to the wavelet coefficients.

Then, let us consider a Gaussian white noise samples with unitary variance filtered by Mel filterbank. Each Mel filter is a band-pass filter; therefore a white noise filtered by using one filter of the Mel filterbank becomes a colored noise according to the filter bandwidth $B(m)$. After this step, multiresolution analysis is applied to each filtered white noise to obtain wavelet coefficients $\boldsymbol{\zeta}_m^p$ depending on the m^{th} Mel filter and the p^{th} coarse level. Finally, the standard deviation of $\boldsymbol{\zeta}_m^p$, denoted by $\sigma_\zeta^m[p]$, is estimated using the standard deviation estimator. After this step, one has to estimate the noise power σ_{ϵ_m} in $\boldsymbol{z}_m$. This can be achieved using two different algorithms: (1) the d-dimensional amplitude trimmed estimator (DATE) [28] or (2) the median absolute deviation (MAD). Finally, the threshold is given by:

$$\lambda_m[p] = \sigma_{\epsilon_m}\sigma_\zeta^m[p]\sqrt{2\ln(N)} \tag{17}$$

This is a non-constant threshold that depends on the coarse level p and the bandwidth of the m^{th} filter. Fig. 6a shows an example of threshold $\boldsymbol{\lambda}_m$ (see red plot).

Stage 3 is the application of multiresolution analysis to $\boldsymbol{z}_m$ to get the wavelet coefficients $\boldsymbol{\omega}_m^p$ for the p^{th} coarse level. Previous studies have shown that the useful signal is handled by large wavelet coefficients, whereas the noise is distributed across small coefficients [26]. For this reason, noise in $\boldsymbol{z}_m$ can

be reduced using a thresholding step applied to the wavelet coefficients. The two next subsections present in detail the MAD and the Date algorithm used to estimate σ_{ϵ_m}.

3.1.1. Median Absolute Deviation

The median absolute deviation (MAD) estimator is defined by:

$$\sigma_{\epsilon_m} = C \times \text{median}(|\boldsymbol{z}_m - \text{median}(\boldsymbol{z}_m)|), \tag{18}$$

where C is a constant scale factor which depends on the observation. For Gaussian distribution, C equals 0.6745^{-1} and corresponds to $1/\Phi^{-1}(0.75)$, where Φ^{-1} is the inverse of the cumulative distribution function of the Gaussian distribution [27]. MAD is more robust than the classical moment based standard deviation estimators in the presence of signal (seen in this case as outliers).

3.1.2. DATE algorithm

This technique performs trimming by assuming that the signal norms are above some known lower-bound and that the signal probabilities of occurrence are less than one half [28]. The method is summarized, in Algorithm 1. Note that the two parameters that directly influence the estimate $\hat{\sigma}_{\epsilon_m}$ of σ_{ϵ_m} are the number of observations N and the lower-bound ρ, where ρ can be defined as in [29]. In addition, $\xi(\rho) = \cosh^{-1}\left(e^{\rho^2/2}\right)$ and $\kappa = 0.7979$ as specified in [14].

Algorithm 1 DATE for estimation of noise standard deviation

Inputs:

1. A finite sequence $\boldsymbol{z}_m = [z_m[1], z_m[2], \ldots, z_m[N]]$ of real random variables satisfying the weak-sparseness model
2. A lower-bound ρ
3. A probability value $Q \leqslant 1 - \frac{N}{4(N/2-1)^2}$

Constants: $n_{\min} = N/2 - \sqrt{N/4(1-Q)}$, $\xi(\rho)$, κ

Output: estimate $\hat{\sigma}_{\epsilon_m}$ of σ_{ϵ_m}

Computation of $\hat{\sigma}_{\epsilon_m}$:

Sort $z_m[1], z_m[2], \ldots, z_m[N]$ by increasing norm so that $|z_m^{(1)}| \leqslant |z_m^{(2)}| \leqslant \ldots \leqslant |z_m^{(N)}|$

$n_{\min} = N/2 - \sqrt{N/4(1-Q)}$

if there exists a smallest integer n in $\{n_{\min}, \ldots, N\}$ such that: $|z_m^{(n)}| \leqslant (\mu(n)/\kappa)\,\xi(\rho) < |z_m^{(n+1)}|$

$n^* = n$

where $\mu(n)$ is defined by:

$$\mu(n) = \begin{cases} \frac{1}{n}\sum_{k=1}^{n} |z_m^{(k)}| & \text{if } n \neq 0 \\ 0 & \text{if } n = 0. \end{cases}$$

else

$n^* = n_{\min}$

end if

$\hat{\sigma}_{\epsilon_m} = \mu(n^*)/\kappa$

3.1.3. Thresholding methods

Many approaches exist for wavelet thresholding including the following most popular methods:

- The hard-thresholding rule [27] which consists in zeroing coefficients smaller than the threshold while keeping the other ones unchanged:

$$\nu_m^p[k] = \begin{cases} 0, \text{ if } \quad |\omega_m^p[k]| < \alpha\lambda_m[p] \\ \omega_m^p[k], \text{ otherwise,} \end{cases} \tag{19}$$

  where k is running from 1 to $\frac{N}{2^p}$ (see Figs. 6b and 7c).

- The soft-thresholding rule [15], which scales the remaining coefficients in order to form a zero mean continuous distribution (see Fig. 7d):

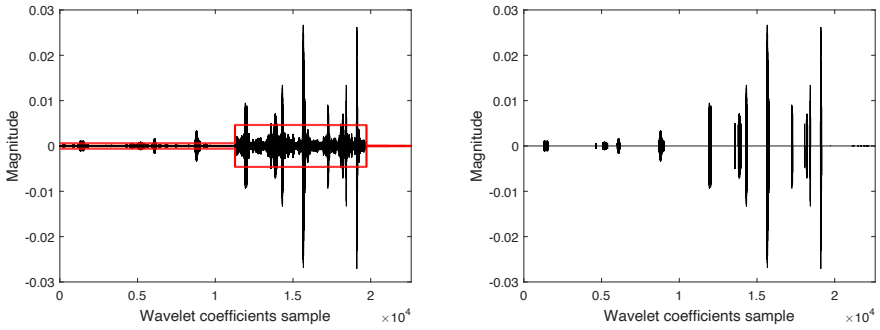
$$\nu_m^p[k] = \begin{cases} 0, \text{ if } \quad |\omega_m^p[k]| < \alpha\lambda_m[p] \\ \omega_m^p[k] - \alpha\lambda_m[p]\,\text{sign}(\omega_m^p[k]), \text{ otherwise,} \end{cases} \tag{20}$$

where k is taking its values in the same range as in the case of hard-thresholding rule.

The constant α is a parameter used to further adjust the threshold; the smaller the SNR gain and the higher α, the stronger the signal of interest degradation. Therefore, selected coefficients preserve useful signal information while noise is strongly attenuated.

Let us now illustrate the output of step 4 in Section 3.1. Let us consider a noisy speech signal $\boldsymbol{z}$ corrupted by pink noise. Fig. 7a displays the filtered noisy signal $\boldsymbol{z}_{65}$; Fig. 7b shows the clean signal $\boldsymbol{s}_{65}$; Fig. 7c shows the estimate of the clean signal $\widehat{\boldsymbol{s}}_{65}$ obtained by using and hard-thresholding, and

finally Fig. 7d shows $\widehat{\boldsymbol{s}}_{65}$ obtained by using soft-thresholding. In this illustration, α equals 1, and Daubechies wavelet of order 4 is used. One can see the efficiency of the proposed denoising process using the hard-thresolding rule compared to the result obtained when using the soft-thresholding rule as the later degrades too much signal of interest.



(a) (b)

Figure 6: (a) Wavelet coefficients $\boldsymbol{\omega}_{65}^{p}$ and in red the threshold; (b) Thresholded wavelet coefficients $\boldsymbol{\nu}_{65}^{p}$

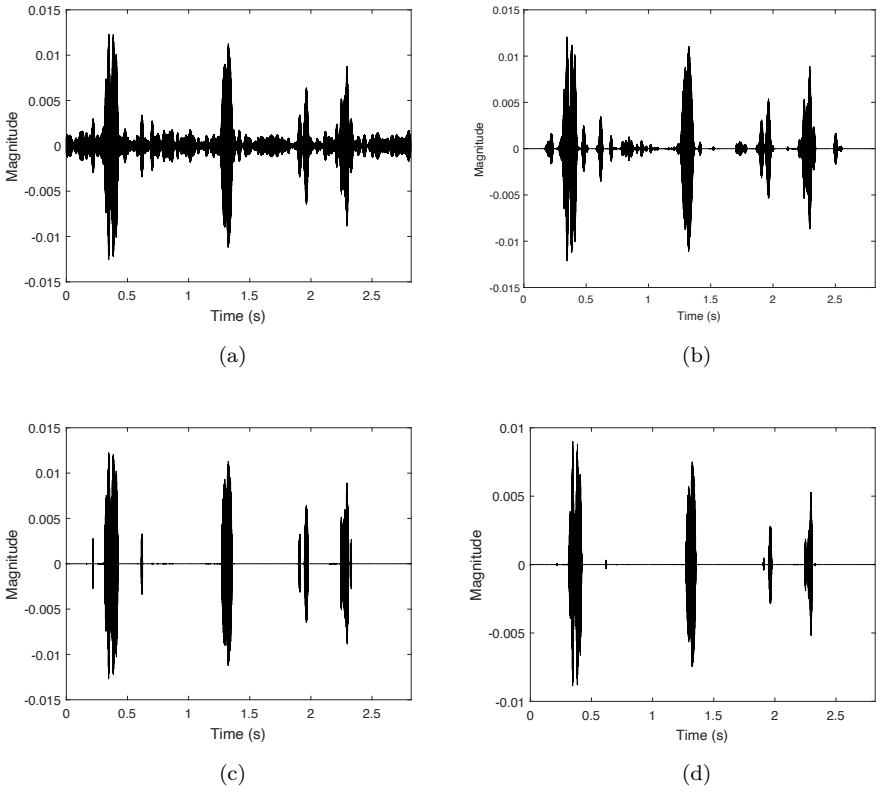


Figure 7: (a) Noisy data $\boldsymbol{z}_{65}$; (b) Clean data $\boldsymbol{s}_{65}$; (c) Denoised data $\widehat{\boldsymbol{s}}_{65}$ using hard-thresholding and (d) Denoised data $\widehat{\boldsymbol{s}}_{65}$ using soft-thresholding

In the part experiments, objective criteria are used to confirm the visual inspection given in this section.

3.2. DWT algorithms

Two main algorithms can be used for multiresolution analysis: the pyramidal decomposition algorithm [26] and the *a trous* ("with holes") algorithm [30]. The *a trous* algorithm requires large amount of data to be computed and stored, which could involve memory problems. Indeed if P denotes the

number of multiresolution planes, the *a trous* algorithm applied to $\boldsymbol{z}_m$ requires the computation of $P \times N$ wavelet coefficients. In addition, unlike the pyramidal decomposition, the *a trous* algorithm does not comply with the translation invariance property due to its principle based on zero insertion. For these reasons, pyramidal decomposition is used in this study.

3.3. Multiresolution synthesis to get the DHPTF representation

Stage 4 of the DHPTF construction is the multiresolution synthesis of the thresholded coefficients $\boldsymbol{\nu}_m^p$ to provide $\widehat{\boldsymbol{s}}_m$. Finally, in step 5 we square $\widehat{\boldsymbol{s}}_m$ for $m = 1, \ldots, M$ to get the DHPTF .

An example of DHPTF of the noisy speech signal is depicted in Fig. 8; the HPTF of the same signal is given in Fig. 3. One can observe that only the signal of interest is retained while the noise is strongly attenuated; therefore this representation is very useful to extract signal relevant features.

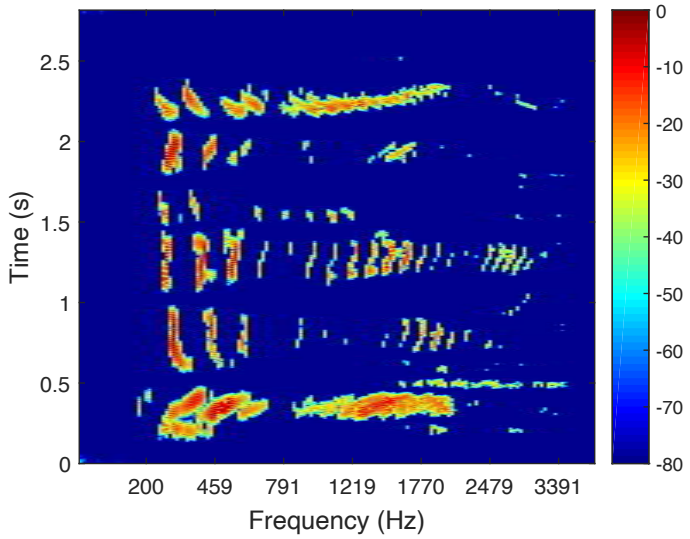


Figure 8: DHPTF of the speech signal presented in Fig. 3

4. Reconstruction of the denoised signal of interest synthesis

In this section, the process of reconstructing a time-domain signal from the presented Mel (t, f) domain is presented. If we consider the filter $H(f)$ associated with the whole Mel filterbank, it corresponds to a band-pass:

$$H(f) = \sum_{m=1}^{M} H_m(f) = 1, \forall f \in [f_1; f_M], \tag{21}$$

where $[\text{mel}\,(f_{min}); f_1[$ and $]f_M; \text{mel}\,(f_{max})]$ are the transition widths. In order to ensure energy conservation, it is necessary to add to $H(f)$, two filters, $H_0(f)$ and $H_{M+1}(f)$, so that:

$$G(f) = H_0(f) + H_{M+1}(f) + H(f) = 1, \forall f \in \left[0; \frac{F_s}{2}\right]. \tag{22}$$

The impulse responses, $h_0(t)$ and $h_{M+1}(t)$, associated with theses filters are defined by:

$$h_0(t) = f_1 \,\mathrm{sinc}^{\,2}(f_1 t), \tag{23}$$

and

$$h_{M+1}(t) = \frac{1}{\pi^2 t^2}\left(\pi t \sin(\pi t F_s) + \frac{\cos(\pi t F_s) - \cos(2\pi t f_M)}{F_s - 2f_M}\right). \tag{24}$$

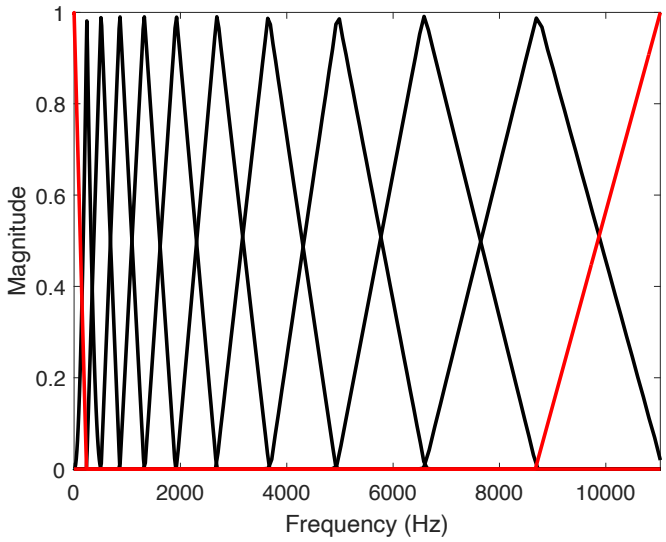


Figure 9: Filter bank (black lines: Mel filters, red lines: additional filters, $H_0(f)$ and $H_{M+1}(f)$, to ensure energy conservation)

Fig. 9 presents the frequency responses associated with $G(f)$, with the Mel filterbank $H(f)$ in black lines and $H_0(f)$ and $H_{M+1}(f)$ in red lines. Thus, the new filterbank $G(f)$ is an all-pass filter, with an impulse response $g(t)$ associated with this filter bank approximated by:

$$g(t) = TF^{-1}[H(f)] \approx \delta(t), \tag{25}$$

where δ denotes the Dirac function.

It follows that one can access to data $\boldsymbol{z}$ from the knowledge of its associated filtered signal $\boldsymbol{z}_m$ by way of a simple summation. Following the same idea, an approximation of the useful signal $\widehat{\boldsymbol{s}}$ can be obtained from $\widehat{\boldsymbol{s}}_m$ as:

$$\widehat{\boldsymbol{s}} = \sum_{m=0}^{M+1} \widehat{\boldsymbol{s}}_m. \tag{26}$$

This denoising process is depicted in Fig. 10 can be summarized as follows:

1. Initialize $\widehat{\boldsymbol{s}} \in \mathbb{R}^N$ as a null vector;
2. For m equal to 0 to $M+1$ do:
   - Computation of impulse response $\boldsymbol{h}_m$;
   - Computation of $\boldsymbol{z}_m$ provided by the convolution product between $\boldsymbol{z}$ and $\boldsymbol{h}_m$;
   - Computation of $\boldsymbol{\omega}_m^p$, for $p = 1, \ldots, P$ obtained from the DWT of $\boldsymbol{z}_m$, where P is the number of resolution levels;
   - Computation of $\boldsymbol{\nu}_m^p$ obtained after the thresholding operation applied to $\boldsymbol{\omega}_m^p$;
   - Construction of $\widehat{\boldsymbol{s}}_m$ by applying an inverse DWT to the coefficients $\boldsymbol{\nu}_m^p$;
   - Iterative estimation of $\widehat{\boldsymbol{s}}$ such that: $\widehat{\boldsymbol{s}} \leftarrow \widehat{\boldsymbol{s}} + \widehat{\boldsymbol{s}}_m$
3. end For loop

Fig. 10 shows the flowgraph of this algorithm.

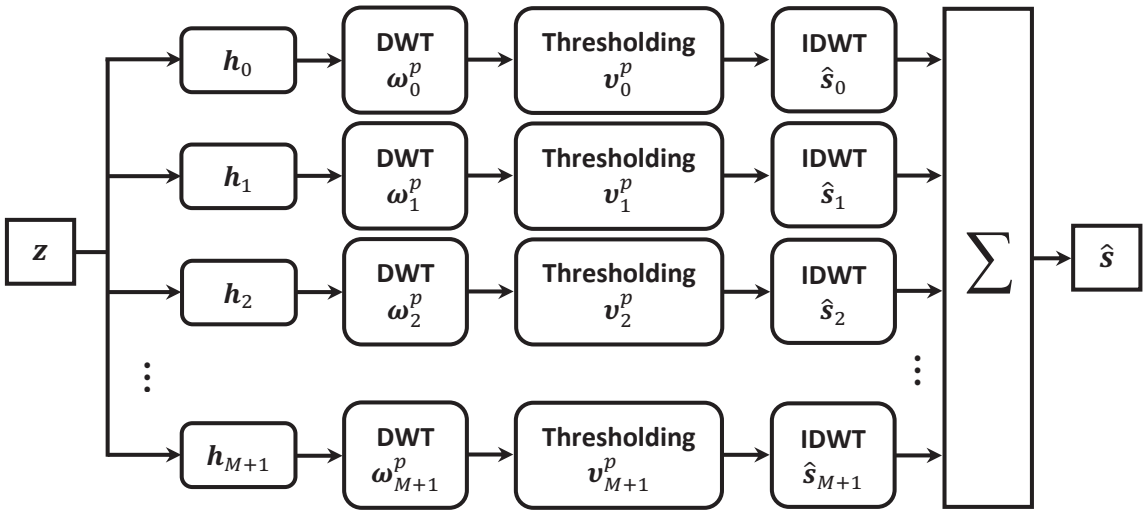


Figure 10: Flowgraph of the denoising process

5. Experiments

5.1. Objective criteria

In order to establish an objective comparison, let us define three measures to evaluate the denoising process efficiency:

- The SNR defined by:

$$\mathrm{SNR} = 10 \log_{10} \frac{\sum_{n=1}^{N} s^2[n]}{\sum_{n=1}^{N} \left(s[n] - \widehat{s}[n]\right)^2}.$$

- The mean square error (MSE) in dB defined by:

$$\mathrm{MSE} = 10 \log_{10} \frac{1}{N} \sum_{n=1}^{N} (s[n] - \widehat{s}[n])^2.$$

  where N represents the number of samples of $\boldsymbol{s}$.

- The Segmental SNR (SSNR), which is defined as the average of the SNR_l values computed over segments with useful signal activity defined by:

$$\mathrm{SSNR} = \frac{1}{L}\sum_{l=1}^{L} 10\log_{10}\frac{\sum_{n=1}^{N} s^2\left[\frac{n+l(N_w-N_o)}{2}\right]}{\sum_{n=1}^{N_w}\left(s\left[\frac{n+l(N_w-N_o)}{2}\right]-\widehat{s}\left[\frac{n+l(N_w-N_o)}{2}\right]\right)^2},$$

where L represents the number of frames in the signal, N_w the number of samples per frame and N_o the number of overlapping samples between two successive windows.

5.2. Parameters setting

Let us recall the set of parameters of the proposed algorithm: (1) number of filters M, (2) constant α to adjust the threshold, (3) thresholding rule and (4) noise standard deviation estimation method (MAD or DATE). In order to set the parameters of the described DHPTF , we use the clean speech signals available in the NOIZEUS database [31] (sampled at 8 kHz). We have then corrupted the speech signals by four different types of noise at SNR level of 5 dB.

First we have fixed $\alpha = 1$ and the number of filters have been set to $M = 100$ because it is a good trade-off between time and frequency resolution as shown in Fig. 7 which shows the effect of the number of filters. The comparison criteria are: (i) SNR after denoising, (ii) SSNR after denoising and (iii) MSE between the clean and estimated clean signal. Then, we have compared the four possible configurations: (1) MAD and hard-thresholding (MADH), (2) MAD and soft-thresholding (MADS), (3) Date and hard-thresholding (DateH) and (4) Date and soft-thresholding (DateS). The results, reported

Table 1: Comparison of the different combinations obtained between the Date, the MAD, the hard- and soft- thresholding rule. (*The bold entries represent the best results obtained for each criterion*).

	Noisy	MADH	MADS	DateH	DateS
SNR (dB)	5	7.17	3.66	**7.66**	4.05
SSNR (dB)	-1.85	1.13	0.23	**1.37**	0.37
MSE (dB)	-32.38	-34.55	-31.04	**-35.03**	-31.43

in Table 1, show that the best combination is obtained when using the Date algorithm combined with hard-thresholding rule.

5.3. Experiment 1: Speech enhancement

5.3.1. Setup

Using the same database as the previous section, this experiment first corrupted the speech signals with four different types of noise, at six SNR levels which are 0, 3, 5, 8, 10 and 15 dB. We have compared the proposed algorithm based on DHPTF with three state-of-the-art denoising methods. The first one is the combination of the MMSE-LSA attenuation rule [32] with decision-directed (DD) *a priori* SNR estimator described in [33]. The second one combines the log-MMSE and E-Date noise estimation algorithm as described in [14]. Finally, the third one is a simple thresholding done by using a constant threshold fixed empirically and where the noise PSD is estimated using the median. The algorithms are tuned as follows: $\alpha = 0.98$ for the DD estimator (as advised in [33]), while the third algorithm threshold is set at $th = 6 \times \sigma_n$, where σ_n the estimated noise PSD. For the computation of the segmental SNR, the window length is 30 ms ($N_w = 240$) and the

Table 2: Comparison of the proposed method with three state-of-the-art denoising methods by using the SNR, SSNR and the MSE metrics. (*The bold entries represent the best results*).

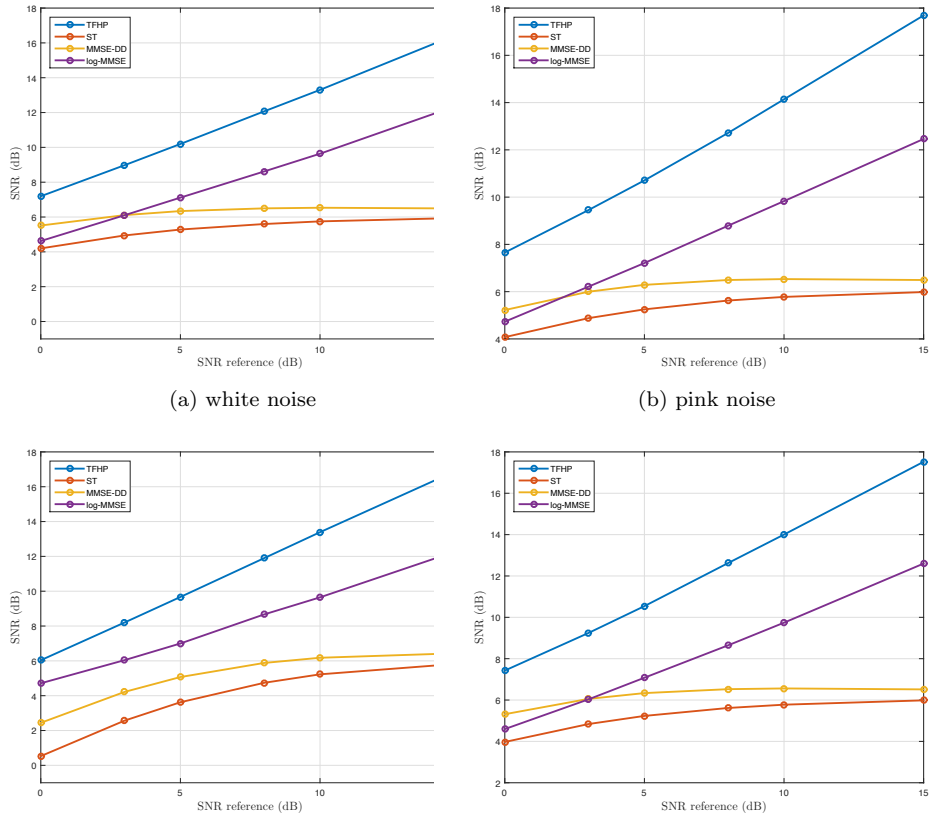
	Noisy	TFHP	Simple thresholding	MMSE˙DD	logmmse
SNR (dB)	0	**7.08**	3.19	4.62	4.66
SSNR (dB)	-4.60	**0.43**	-2.29	-1.72	-1.27
MSE (dB)	-27.38	**-34.46**	-30.57	-32.01	-32.04
SNR (dB)	5	**10.28**	4.85	6.01	7.10
SSNR (dB)	-1.79	**2.48**	-0.67	-0.27	0.44
MSE (dB)	-32.38	**-37.66**	-32.23	-33.39	-34.48
SNR (dB)	10	**13.71**	5.63	6.45	9.71
SSNR (dB)	1.37	**4.88**	0.69	0.94	2.40
MSE (dB)	-37.38	**-41.08**	-33.01	-33.83	-37.09
SNR (dB)	15	**17.19**	5.94	6.49	12.47
SSNR (dB)	4.86	**7.45**	1.84	1.94	4.66
MSE (dB)	-42.38	**-44.57**	-33.32	-33.86	-39.85

overlap between two consecutive windows is $N_o = \frac{N_w}{4}$.

5.3.2. Results

Table 2 presents the results of the experiment; one can see that the HPTF method outperforms all the state-of-the-art methods for all the criteria and all original SNR (before denoising). E.g for an original SNR of 0 dB the SNR of the denoised signal using HPTF is 7.08 dB while it is 4.66 dB when using the log-mmse based method, the SSNR is 0.43 dB while it is

-1.27 dB for the log-mmse and the MSE is -34.46 dB while it is -32.04 dB for the log-msse. Figs. 11, 12 and 13 show respectively the SNR, the SSNR and the MSE after denoising for different SNR and different noises. These figures confirm the results observed in Table 2.



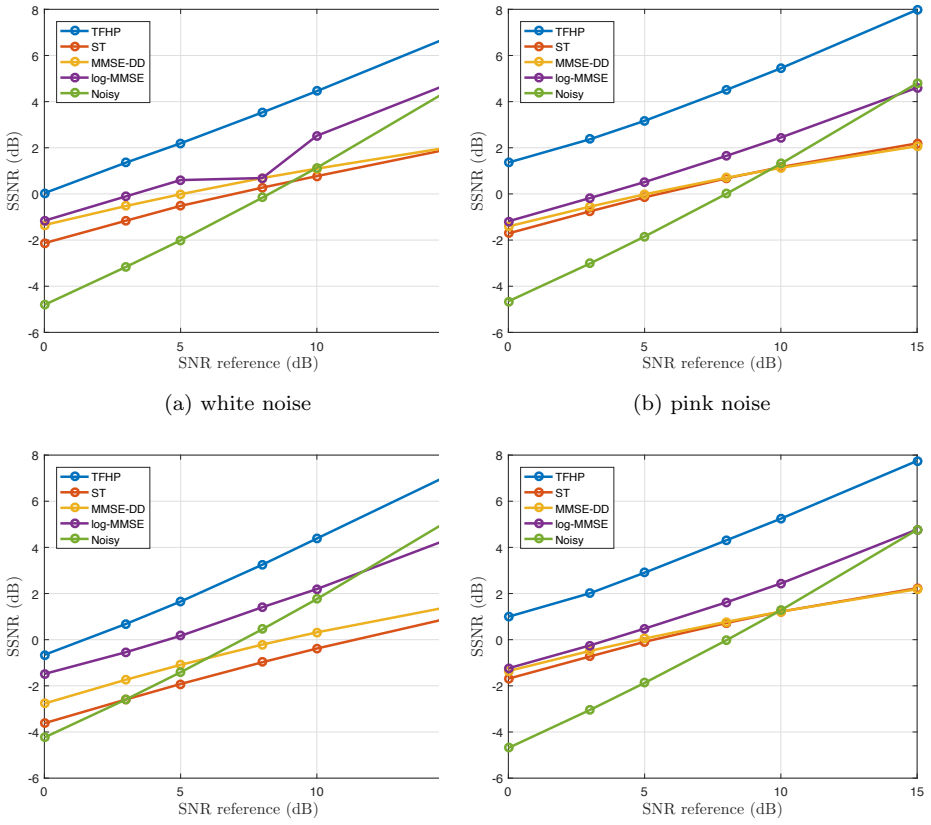
(a) white noise

(b) pink noise

(c) babble noise

(d) f16 noise

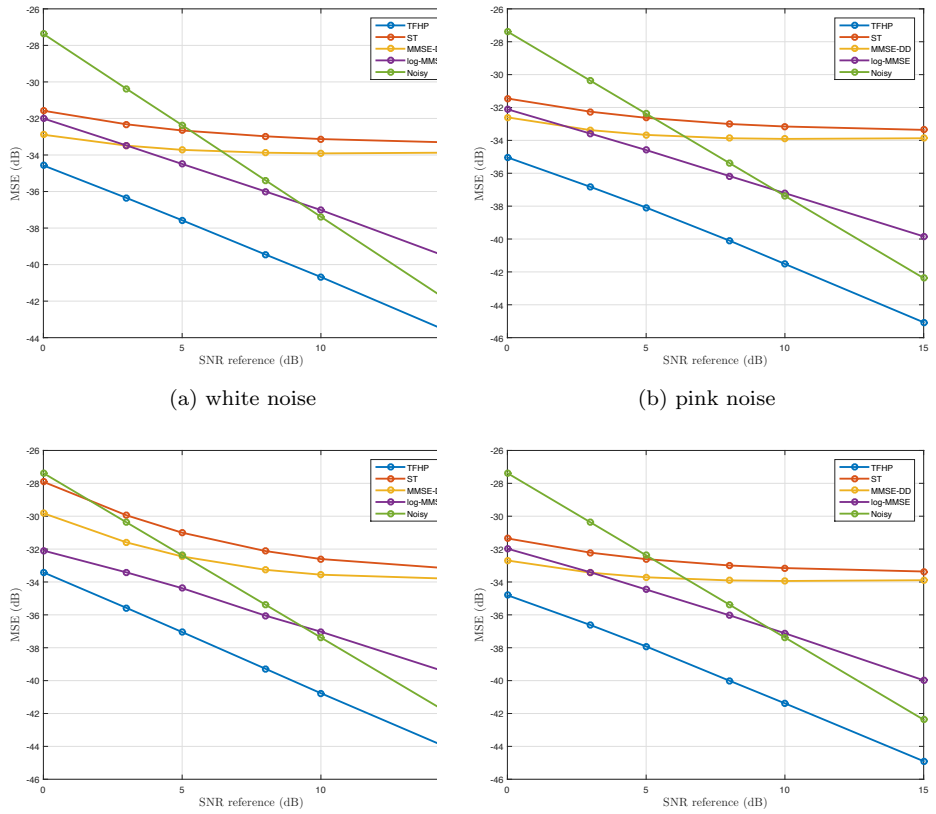
Figure 11: Comparison of the SNRs after denoising process with different methods for signals corrupted by four types of noise: (a) white noise, (b) pink noise, (c) babble noise and (d) f16 noise.



(a) white noise (b) pink noise

(c) babble noise (d) 16 noise

Figure 12: Comparison of the segmental SNRs after denoising process with different methods for signals corrupted by four types of noise: : (a) white noise, (b) pink noise, (c) babble noise and (d) f16 noise.



(a) white noise

(b) pink noise

(c) babble noise

(d) f16 noise

Figure 13: Comparison of the MSE after denoising process with different methods for signals corrupted by four types of noise: : (a) white noise, (b) pink noise, (c) babble noise and (d) f16 noise.

5.4. Experiment 2: Simulated EEG signal

Some particular abnormalities of EEG signals (such as seizures) can be modeled as a sum of multiple chirps of varying amplitudes and IF rates. Such model was used for EEG seizure detection by TF matched filtering in [34] and for defining a new high-resolution TFD named multi-directional TFD (MDD) in [5]. In Experiment 2 we have considered a synthetic multi-

component signal with significant variation in the instantaneous frequency (IF) laws of its components. The signal has 256 s duration and is sampled at 32 Hz. In addition, EEG noise has often been modeled using power spectral density (PSD) that are power law functions of the form $\frac{1}{f^\eta}$ for $0 \leq \eta \leq 2$ [19]. In this study, we have set $\eta = 0.6$ and corrupted the simulated EEG signal such that the SNR is 5 dB. In this experiment, in order to assess only the denoising process method we propose to use it as a "black box" and represent the (t, f) domain of the signals using the spectrogram.

Fig. 14 shows the result of the denoising process where each row contains the time-domain signal and its corresponding spectrogram. The figures in the first row represent the clean EEG signal, while the figures on the second row show the noisy EEG signal and finally the figures on the last row represent the denoised EEG signal using the HPTF process described in this paper. One can see on Fig. 14 the efficiency of the denoising; even in the case of low SNR and colored noise, the signal of interest is well preserved while the noise is strongly reduced.

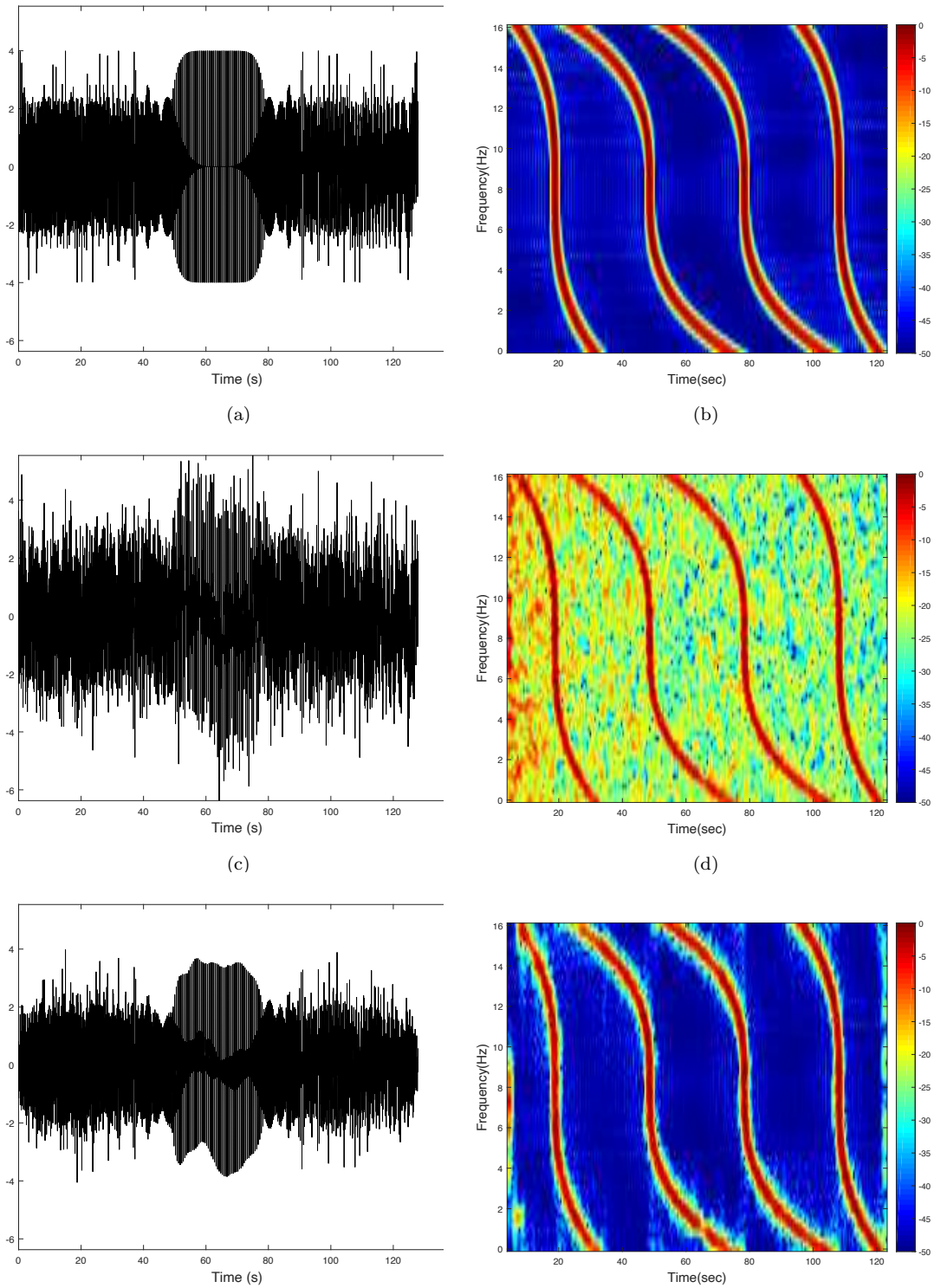


Figure 14: HPTF denoising applied to a simulated EEG signal. (a)-(b) show the time-domain and the spectrogram of the clean EEG signal; (c)-(d) show the time-domain and the spectrogram of the noisy signal (5 dB) and (e)-(f) show the time-domain and the spectrogram of the denoised signal using the proposed HPTF denoising process.

6. Conclusion

This paper describes an innovative approach to reduce the noise level in non-stationary signals. It relies on a time-frequency (TF) representation based on a psycho-acoustic model which describes human perception (HPTF) and its denoised version (DHPTF). The efficiency of this method is demonstrated on several types of signals, including speech and simulated biomedical EEG signals. These experiments show that the DHPTF yields a good information recovery, thus inducing a better signal interpretation. One can see on each experiment given that the signal of interest is preserved while the noise is clearly reduced. With this approach, it becomes possible to extract patterns of interest more precisely for the purpose of signal identification and classification, resulting in a feature extraction stage that provides useful features based on human perception.

The second part of this study is to reconstruct the signal from the DHPTF ; it is based a simple and useful process.

The findings of this study indicate that the proposed denoising technique outperforms three current state-of-the-art algorithms in terms of signal-to-noise-ratio (SNR), segmental SNR (SSNR) and mean squared error (MSE) and can be applied to a large class of signals; e.g. the improvment is up to 4.72 dB w.r.t the SNR and the MSE, 2.79 dB w.r.t the SSNR. Therefore, this algorithm could be of great interest to improve the performance of identification systems dealing with non-stationary noisy signals. In future works, it would be interesting to quantify the improvement in terms of classification compared to other TF representations and measure the quality of extracted features from the DHPTF representation [35].

Acknowledgment

This work was supported by QNRF grants NPRP 6-885-2-364.

References

[1] B. Boashash, Time-frequency signal analysis and processing: a comprehensive reference, Academic Press: ED.2, Elsevier, ISBN: 9780123984999, 2015.

[2] J. O. Toole, B. Boashash, Fast and memory-efficient algorithms for computing quadratic time–frequency distributions, Applied and Computational Harmonic Analysis 35 (2) (2013) 350–358.

[3] M. G. Amin, B. Jokanovic, Y. D. Zhang, F. Ahmad, A sparsity-perspective to quadratic time–frequency distributions, Digital Signal Processing 46 (2015) 175–190.

[4] S. Ouelha, S. Touati, B. Boashash, An efficient inverse short-time fourier transform algorithm for improved signal reconstruction by time-frequency synthesis: Optimality and computational issues, Digital Signal Processing 65 (2017) 81–93.

[5] B. Boashash, S. Ouelha, An improved design of high-resolution quadratic time-frequency distributions for the analysis of non-stationary multicomponent signals using directional compact kernels., IEEE Transactions on Signal Processing.

[6] R. G. Baraniuk, D. L. Jones, A signal-dependent time-frequency representation: Optimal kernel design, IEEE Transactions on Signal Processing 41 (4) (1993) 1589–1602.

[7] K. Baba, L. Bahi, L. Ouadif, Enhancing geophysical signals through the use of savitzky-golay filtering method, Geofísica internacional 53 (4) (2014) 399–409.

[8] M. Bekara, M. Van der Baan, Local singular value decomposition for signal enhancement of seismic data, Geophysics 72 (2) (2007) V59–V65.

[9] H. Maki, T. Toda, S. Sakti, G. Neubig, S. Nakamura, EEG signal enhancement using multi-channel wiener filter with a spatial correlation prior, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 2639–2643.

[10] R. Miyahara, A. Sugiyama, Gain relaxation: A useful technique for signal enhancement with an unaware local noise source targeted at speech recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 2234–2238.

[11] H. Hassanpour, A time–frequency approach for noise reduction, Digital Signal Processing 18 (5) (2008) 728–738.

[12] B. Xerri, B. Borloz, An iterative method using conditional second-order statistics applied to the blind source separation problem, IEEE Transactions on Signal Processing 52 (2) (2004) 313–328.

[13] J. Chen, J. Benesty, Y. Huang, S. Doclo, New insights into the noise

reduction wiener filter, Audio, Speech, and Language Processing, IEEE Transactions on 14 (4) (2006) 1218–1234.

[14] V. K. Mai, D. Pastor, A. Aissa-El-Bey, R. Le-Bidan, Robust estimation of non-stationary noise power spectrum for speech enhancement, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (4) (2015) 670–682. doi:10.1109/TASLP.2015.2401426.

[15] D. L. Donoho, De-noising by soft-thresholding, IEEE transactions on information theory 41 (3) (1995) 613–627.

[16] S. E. Ferrando, R. Pyke, Ideal denoising for signals in sub-gaussian noise, Applied and Computational Harmonic Analysis 24 (1) (2008) 1 – 13. doi:http://dx.doi.org/10.1016/j.acha.2007.03.004.
URL http://www.sciencedirect.com/science/article/pii/S1063520307000449

[17] R. J. Webster, Ambient noise statistics, IEEE Transactions on Signal Processing 41 (6) (1993) 2249–2253.

[18] H. Ou, J. S. Allen, V. L. Syrmos, Frame-based time-scale filters for underwater acoustic noise reduction, IEEE Journal of Oceanic Engineering 36 (2) (2011) 285–297.

[19] A. Paris, G. Atia, A. Vosoughi, S. A. Berman, A new statistical model of electroencephalogram noise spectra for real-time brain-computer interfaces, IEEE Transactions on Biomedical Engineering.

[20] R. Sameni, M. B. Shamsollahi, C. Jutten, G. D. Clifford, A nonlinear bayesian filtering framework for ecg denoising, IEEE Transactions on Biomedical Engineering 54 (12) (2007) 2172–2185.

[21] M. S. Crouse, R. D. Nowak, R. G. Baraniuk, Wavelet-based statistical signal processing using hidden markov models, IEEE Transactions on Signal Processing 46 (4) (1998) 886–902.

[22] F. Zheng, G. Zhang, Z. Song, Comparison of different implementations of mfcc, Journal of Computer Science and Technology 16 (6) (2001) 582–589.

[23] S. Ouelha, Représentation et reconnaissance des signaux acoustiques sous-marins (in french), Ph.D. thesis, Université de Toulon (2014).

[24] P. Courmontagne, S. Ouelha, U. Moreaud, F. Chaillan, A blind denoising process with applications to underwater acoustic signals, in: Oceans-San Diego, 2013, IEEE, 2013, pp. 1–7.

[25] R. Mill, G. Brown, Auditory-based time-frequency representations and feature extraction techniques for sonar processing, Speech and Hearing Research Group, Sheffield, England.

[26] S. G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (7) (1989) 674–693.

[27] D. L. Donoho, J. M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika 81 (3) (1994) 425–455.

[28] D. Pastor, F.-X. Socheleau, Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurrences, IEEE Transactions on Signal Processing 60 (4) (2012) 1545–1555.

[29] S. M. Aziz-Sbaï, A. Aïssa-El-Bey, D. Pastor, Contribution of statistical tests to sparseness-based blind source separation, EURASIP Journal on Advances in Signal Processing 2012 (1) (2012) 169.

[30] M. Holschneider, R. Kronland-Martinet, J. Morlet, P. Tchamitchian, A real-time algorithm for signal analysis with the help of the wavelet transform, in: Wavelets, Springer, 1990, pp. 286–297.

[31] P. C. Loizou, Speech enhancement: theory and practice, CRC press, New-York, 2013.

[32] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, IEEE Transactions on Acoustics, Speech and Signal Processing 33 (2) (1985) 443–445.

[33] Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, IEEE Transactions on Acoustics, Speech and Signal Processing 32 (6) (1984) 1109–1121.

[34] B. Boashash, G. Azemi, A review of time–frequency matched filter design with application to seizure detection in multichannel newborn EEG, Digital Signal Processing 28 (2014) 28–38.

[35] B. Boashash, S. Ouelha, Automatic signal abnormality detection using time-frequency features and machine learning: A newborn EEG seizure case study, Knowledge-Based Systems 106 (2016) 38–50.