Multimed Tools Appl (2017) 76:23661–23678
DOI 10.1007/s11042-016-4145-0

Noise estimation for speech enhancement algorithms with post-smoothness processor incorporating global posterior SNR

Anis Ben Aicha[1]

Received: 12 May 2016 / Revised: 24 September 2016 / Accepted: 9 November 2016 /
Published online: 22 November 2016

Abstract Performances of speech enhancement algorithms depend greatly on the accuracy of the estimated noise. In this paper, we explain in details the relationship between noise estimation and denoised speech quality. We particularly show the importance of noise smoothing over frames on denoising quality. This study leads to the development of a new technique to smooth the estimated noise power spectrum over frequency bins of the same frame. Compared to inter-frame smoothing, experimental results show that the proposed intra-frame smoothing has a good impact on the denoised speech. Quality is evaluated over three dimensions: speech distortion, residual background noise and overall quality.

Keywords Noise estimation · Noise smoothing · Global Posterior SNR · Perceptual quality

1 Introduction

Generally, denoising techniques seek to resolve two challenges. The first one is how to estimate the noise power spectrum and the second one is how to compute the denoising filters [4, 15]. In real cases, only disturbed speech is observed, hence the estimated noise must be extracted from it. Accurate estimation of the noise is crucial, as an underestimate of the noise leaves an unnecessary amount of residual noise in the enhanced signal, while an over-estimate leads to speech distortions and a potential loss of speech quality [5, 17, 20].

Two main approaches are considered to estimate the noise power spectral density (PSD) from the noisy speech. The simplest one is to estimate and update the PSD during silent

✉ Anis Ben Aicha
anis_ben_aicha@yahoo.fr

[1] University of Carthage, Higher School of Communications of Tunis (SUP'COM), COSIM Research Laboratory, Ariana, Tunisia

segments of the signal using a voice activity detector (VAD) [22, 24]. Assuming that the background noise is locally stationary, its PSD is estimated in the absence of the speech. As a result, the update rate is restricted to the duration of the speech bursts.

However, it was shown that such methods does not work well for realistic environment where the spectral characteristics of the noise might be changing constantly [5, 6, 17, 20]. Hence, new approaches are proposed in order to update the estimated noise continuously over time in every frame. General tendencies exploit the minima of the noisy speech PSD in order to estimate the noise PSD [5, 6, 17, 20]. The principal idea behind such techniques relies on the fact that even during speech activity the power of the noisy signal in a narrow frequency band frequently decays to the level of the noise power. Such behavior of the noisy speech is due to the frequent disappearance of speech components in frequency bands. Hence, as a result, the local minima of the noisy speech PSD correspond to the level of the noise PSD. Since, the minimum is typically smaller than the mean, the noise PSD estimation can be seen as unbiased estimation. Hence, a bias factor is introduced based on the statistics of the minimum estimates.

The common point of all proposed estimators using or not a VAD is the use of a recursive form to estimate the noise spectrum depending on the present frame of the noisy speech and the estimated noise spectrum from the previous frames. In fact, notwithstanding the noise estimators methodologies, the final step involves computing the noise PSD recursively. The recursive form leads to smooth the estimated power spectrum of the noise over frames. We call such smoothness “inter-frame smoothness” since from one frame to the next one the estimated noise PSD becomes more and more smoothed. This point of view will be more detailed in Section 4.

In this paper, we propose to investigate the relationship between “inter-frame smoothness” and the final quality of the denoised speech. Such direct relationship can be used to understand what could be done with the estimated noise to improve the quality of the denoised signal. Hence, we propose in this paper a simple post-processing of the estimated noise which can be implemented downstream any traditional noise estimator. Such post-processing can be seen as an “intra-frame smoothness” of the estimated noise PSD.

This paper is organized as follows. Sections 2 and 3 are reserved respectively to give a background on the denoising techniques and the noise estimation. In Section 4, we study and explain in details the relationship between noise smoothing and denoising quality. In Section 5, a new technique based on to “intra-frame” noise smoothing is developed. In Section 6, we show the importance of global posterior SNR to design smoothing filter. In Section 7, we present the used assessment criteria. In Section 8, we argue experimental results. Section 9 is reserved to conclusion and discussions.

2 Background on speech denoising

In real world case, the noisy speech can be formulated as the sum of the clean speech with uncorrelated additive noise. The sampled observed speech can be written as following.

$$y(n) = s(n) + d(n), \tag{1}$$

where $y(n)$ is the observed noisy speech, $s(n)$ is the clean speech and $d(n)$ is the additive background noise. n denotes the time index.

The processing is done on a frame-by-frame basis in the frequency domain. Under assumption of non correlation between speech and noise and by applying the short-time

Fourier transform (STFT) to the observed signal, we get in the time-frequency domain [8, 15]:

$$Y(m,k) = S(m,k) + D(m,k), \tag{2}$$

where $Y(m,k)$, $S(m,k)$ and $D(m,k)$ represent the STFT coefficients obtained at frequency index k in signal frame m from the noisy speech, clean speech, and noise respectively.

The principal goal of all denoising techniques is to estimate the clean speech amplitude as accurate as possible. The phase of the noisy speech is not modified. Indeed, the estimated amplitude of the clean speech $|\hat{S}(m,k)|$ is combined with the noisy phase $\theta_Y(m,k)$ obtained from $Y(m,k)$ to compute the STFT of the enhanced speech $\hat{S}(m,k)$. The enhanced speech signal $\hat{s}(n)$ in time domain is obtained by inverse STFT and concatenation of overlapped frames.

In order to generalize the denoising process, a common method to estimate the clean speech amplitude can be formulated as a multiplication of the noisy speech amplitudes by a spectral gain function. We notice that any power β of the speech amplitude can be estimated using a suitable gain function [15]:

$$|\hat{S}(m,k)|^{\beta} = G\left[\beta, \xi(m,k), \zeta(m,k)\right] \cdot |Y(m,k)|^{\beta}, \tag{3}$$

where $G\left[\beta, \xi(m,k), \zeta(m,k)\right]$ depends on the criterion that is optimized for accurate estimation of the clean speech amplitude. The computed gain is driven by two principal components; the *prior* signal to noise ratio $\xi(m,k)$ and the *posterior* signal to noise ratio $\zeta(m,k)$:

$$\xi(m,k) = \frac{|S(m,k)|^2}{|D(m,k)|^2} \text{ and } \zeta(m,k) = \frac{|Y(m,k)|^2}{|D(m,k)|^2} - 1. \tag{4}$$

Since the noise PSD $|D(m,k)|^2$ is not directly accessible, it is estimated using a noise estimation technique. Hence, an estimated version of $\zeta(m,k)$ can be computed. $\xi(m,k)$ is also not directly measured. $\xi(m,k)$ is computed using an estimated version of the noise PSD $|\hat{D}(m,k)|^2$ and $\zeta(m,k)$ as it will be explained later.

As it is mentioned in the introduction, the proposed noise estimation method should works as a post-processing of any classic noise estimator. Without loss of generality and in order to validate our method, we have chosen to test it with the well known Wiener filtering technique [4, 15]. The Wiener gain is given by:

$$G_W(m,k) = \frac{\xi(m,k)}{1+\xi(m,k)}, \tag{5}$$

where $\xi(m,k)$ is estimated as following:

$$\xi(m,k) = (1-\tau)\, z\left[\zeta(m,k)\right] + \tau \frac{\left[G_W(m-1,k) \cdot |Y(m-1,k)|\right]^2}{|\hat{D}(m,k)|^2}, \tag{6}$$

τ is a real constant and $z(x) = \dfrac{1}{2}(x + |x|)$.

3 Background on inter-frame noise estimation

3.1 First solutions: noise estimation based on VAD

The simplest way to estimate the noise spectrum is the use of a voice activity detector in order to track the only noise frames of noisy speech [22, 24]. The estimated power spectrum of the noise $|\hat{D}(m,k)|^2$ is updated during pause speech frames according to the general recursive relationship:

$$|\hat{D}(m,k)|^2 = \alpha(m,k)|\hat{D}(m-1,k)|^2 + [1-\alpha(m,k)]\,|Y(m,k)|^2, \tag{7}$$

We call this noise estimation process an “inter-frame” estimation since the update of the noise is done from one frame to another and uses previous estimation and current information.

3.2 Improvements: noise estimation based on speech presence probability

It is shown that noise estimators based on VAD might work satisfactorily in stationary noise but it fails in more realistic environments where the spectral characteristics of the noise might be changing constantly [5, 17, 20]. Hence, there is a need to update the noise spectrum continuously over time. All proposed techniques to estimate noise use (7) to update noise power spectrum by varying the parameter $\alpha(m,k)$ according to the frame index m and the frequency bin index k. The main difference between all proposed technique is how to compute $\alpha(m,k)$.

As recent techniques examples, we briefly recapitulate the minimum statistics (MS) method of Martin [17], the improved minima controlled recursive averaging (IMCRA) method of Cohen [5] and a variant of minimum controlled recursive average algorithm (MCRA2) proposed by Rangachari and Loizou [20].

3.2.1 Martin method (MS) [17]

Speech signal is a random process characterized by it energy which is frequently zero during speech pause and between words and syllables. Moreover, when the clean speech is corrupted by an additive noise, in some cases and in certain frequency bins, the speech power may be much smaller than the noise power. The principle of MS method is a trivial one, it consist firstly on the computing of the smoothed noise speech periodogram in a finite window that is large enough to bridge high-power speech segments. Secondly, the tracking of the minima of the smoothed noisy speech periodogram can be used to estimate the noise floor. We notice that the minimum values have an expected value that is smaller than the mean power level. Hence, a bias correction procedure is used. The major drawback of MS technique is it failure to track a very nonstationary noise sources since the update of the noise for frequency bin is done with a lag equal to the duration of the finite used window for noise estimation which in practical cases can reach 1 second.

3.2.2 Cohen method (IMCRA) [5]

In the IMCRA method, the noise PSD estimation is obtained by recursive form according to (7). The parameter $\alpha(m,k)$ depends on the estimation of the speech presence probability $\widehat{Pr}_s(m,k)$.

$$\alpha(m,k) = \lambda + (1-\lambda)\widehat{Pr}_s(m,k), \tag{8}$$

where λ lies between 0 and 1.

Equation (8) means that $\alpha(m,k)$ always lies between λ and 1. When $\widehat{Pr}_s(m,k)$ is near 1, so is $\alpha(m,k)$ and then the noise estimate is kept close to its previous value, preventing speech power to leak into the noise variance estimate. This means that when speech is absence, the update of the noise is done quickly. The estimation of $\widehat{Pr}_s(m,k)$ in IMCRA is controlled by the minima values of the smoothed PSD of the noisy signal. As for the MS method, the IMCRA method reacts slowly to an increase in the noise level.

3.2.3 Rangachari and Loizou method (MCRA2) [20]

Rangachari and Loizou algorithm updates the noise estimate in each frame using a time-frequency dependent smoothing factor computed based on the speech-presence probability. Inspired from Cohen algorithm, Rangachari and Loizou developed their method according to the following steps.

- Computing the smoothed power spectrum of the noisy speech using recursive form.
- Tracking the minimum of noisy speech by continuously averaging past spectral values using non linear rules.
- Computing the speech presence probability $\widehat{Pr}_s(m,k)$: The ratio of noisy speech power spectrum and its local minimum is computed and compared to a certain threshold $\delta(k)$. If the ratio is found to be greater than the threshold $\delta(k)$, it is taken as a speech present frequency bin else it is taken as a speech-absent.
- Computing frequency-dependent smoothing factor $\alpha(m,k)$ using (8).
- Update of noise spectrum estimate using (7).

The principal improvement introduced by Rangachari and Loizou regarding Cohen is the estimation of speech-presence probability exploiting the correlation of power spectral components in neighbouring frames.

4 Relationship between noise smoothing and speech denoising quality

Equation (7), which can be easily recognized as an infinite impulse response low pass filtering (IIR), provides a smoothed version of estimated noise PSD. To illustrate the effect of (7) on the estimated noise PSD, we have done the following experiment. A clean speech sequence, extracted from TIMIT database [7], was corrupted by a white noise selected from Noisex92 database [23]. The overall signal to noise ratio (SNR) of the noisy speech is equal to 0 dB. The corrupted speech was denoised using Wiener filtering as denoising method and MCRA2 as noise estimator. Figure 1 shows the PSD of the estimated noise at three different instants, namely, 0 (a), 1.6 (b) and 3.2 (c) seconds. It shows effectively that the PSD is smoother when time grow up.

We recall that the denoising process is done frame by frame. The enhanced speech frame is computed according to (3) that depends essentially from the estimated noise PSD. If we look closely to the gain function $G\left[\beta, \xi(m,k), \zeta(m,k)\right]$ and the estimated noise PSD $|\hat{D}(m,k)|^2$, we can easily conclude that any fluctuations observed in the estimated noise PSD introduce unavoidable fluctuations in the gain function too. This fact might generate an artificial background noise well known in the literature as musical noise [3]. On the contrary, less observed fluctuations on the estimated noise PSD, as it is shown in Fig. 1c, might lead to fewer fluctuations in the final enhanced speech. This fact could reduce the amount of the artificial residual background noise and hence improving the entire quality of the enhanced speech.

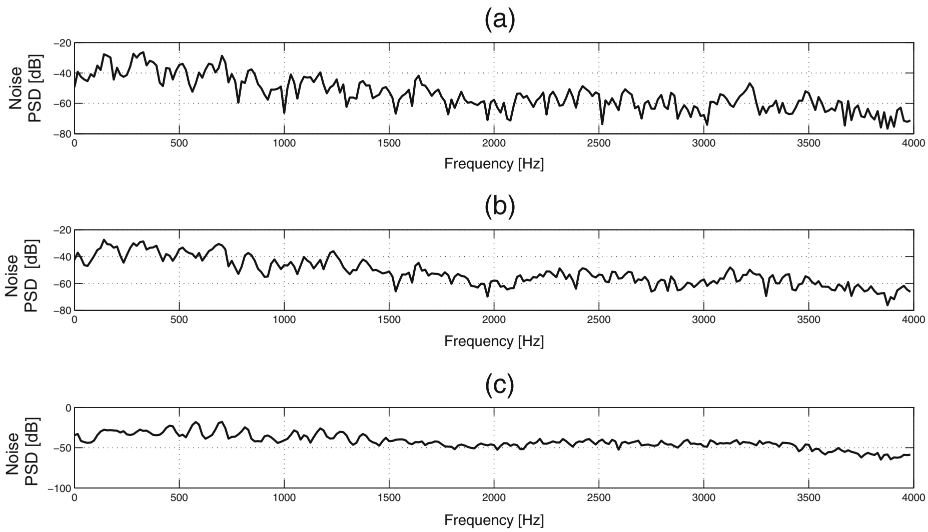


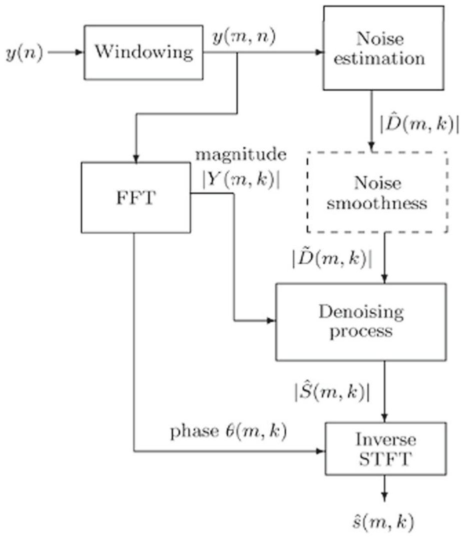
Fig. 1 Illustration of noise smoothing over time

5 Proposed idea of intra-frame noise smoothing

5.1 The idea of noise smoothness

In order to smooth the estimated noise PSD quickly, we propose to do it twice. First, one of the previous and classic way is used (called inter-frame smoothing). Next, the resulting spectrum is smoothed differently by considering only this spectral content and arranging differently the frequency bins. It is called intra-frame smoothing. The idea of the proposed method is depicted in Fig. 2. The solid boxes represent the conventional steps of denoising

Fig. 2 The proposed idea scheme



process and the dashed one represents the proposed smoothing process that is introduced after noise estimation. Since the speech is a non-stationary random signal, the denoising process is done frame by frame. The sampled noisy speech signal $y(n)$ is windowed to isolate the frame $y(m,n)$ that is used to estimate the noise PSD and to calculate the STFT $Y(m,k)$. Under assumption of non-sensitivity to the phase components, only the amplitude is processed. The phase is kept unchanged. Wiener algorithm is then applied using the estimated noise PSD $|\hat{D}(m,k)|^2$, (3) and (6). As initial condition of (6), τ is set to 0.98 and $G_W(m,k)$ is a zero vector. The enhanced frame signal $\hat{s}(m,n)$ is than obtained by inverse STFT of the enhanced amplitude $|\hat{S}(m,k)|$ combined with the noisy phase $\theta(m,k)$.

5.2 Smoothness constraint

The major challenge is how to smooth the estimated noise PSD. Without lost of generality, we can formulate the smoothness process as a filtering of the estimated noise PSD denoted $|\hat{D}(m,k)|^2$ by a filter $H_S(m)$. The filtering is done in frequency domain, so it is formulated as a convolution between the noise PSD and $H_S(m)$:

$$|\tilde{D}(m,k)|^2 = |\hat{D}(m,k)|^2 * H_S(m). \tag{9}$$

Several ideas can arise such as using low pass filter, median filter, etc. The obvious filter to smooth the noise PSD is the average filter.

$$H_S(m) = \left[\frac{1}{P}, \frac{1}{P}, \ldots, \frac{1}{P}\right], \tag{10}$$

where P is the number of the filter coefficients. $P \in \mathbb{N}^*$.

However, it is important to recall that we seek to improve the perceptual quality without introducing neither additional speech distortion nor residual noise in the denoised speech signal $\hat{s}(n)$. Hence, some precautions must be made. In fact, smoother filter $H_S(m)$ must obey the following constraints:

i Preserve noise dynamics. In fact if we smooth so much, we risk to lose noise dynamic. This may have dramatically consequences especially in the case of coloured noise. Hence, the order of $H_S(m)$ must not be very large.

ii Take into account the noise power and nature. It is obvious, that the order of the smoother filter must change from a frame to another and not chosen as fixed value.

iii Smoothing the estimated noise PSD to get the “best smoothed” version in perceptual sense. It means the smoothed version which leads to the best perceptual quality of denoised speech.

6 Filter design incorporating global posterior SNR

6.1 Importance of global posterior SNR

In this section, we propose to design our smoother filter in order to satisfy mentioned constraints. At this stage, we recall that the proposed smoother filter operate frame by frame. However, the background noise may change from one frame to another. So, we have to make the order of the filter variable $P(m)$ depending on the frame index m. To do this, we adopt the following strategy.

- If the level of the background noise is low, the denoising processor can enhance efficiently the noisy signal [9, 10]. In fact, even if the classical inter-frame estimators do not estimate exactly the background noise, there is no significant effects on the perceptual quality of the denoised signal [1, 9, 10]. Hence, we think that there is no need to "much" smooth the estimated noise obtained by a classical inter-frame noise estimator.
- In the opposite case, when the level of the background noise is so high, denoising algorithms do not suppress the noise efficiently [9, 10]. Moreover, it is shown that the speech signal is distorted as well as the residual background noise [1]. Generally, an annoying residual noise appear after denoising process well known as "musical noise" due essentially to the bad estimation of the background noise. More precisely, to the arbitrary fluctuations in the PSD of the estimated noise. Hence, there is a great need to smooth as "much" as possible the estimated noise without loss of noise dynamic.

To materialize these ideas, we propose to compute the Global Posterior Signal to Noise Ratio ($SNR_{GP}(m)$) of the frame m :

$$SNR_{GP}(m) = E\left\{\frac{\Gamma_Y(m)}{\Gamma_D(m)}\right\} \simeq \frac{\Gamma_Y(m)}{\Gamma_{\hat{D}}(m)}, \tag{11}$$

where $E\{\cdot\}$ denote the expectation operator. $\Gamma_Y(m)$ and $\Gamma_{\hat{D}}(m)$ are computed as following:

$$\begin{cases} \Gamma_Y(m) = \frac{1}{N}\sum_{k=1}^{N} |Y(m,k)|^2, \\ \Gamma_{\hat{D}}(m) = \frac{1}{N}\sum_{k=1}^{N} |\hat{D}(m,k)|^2. \end{cases} \tag{12}$$

where N denotes the frequency bins number.

By using $SNR_{GP}(m)$, we can derive the following remarks:

- if $SNR_{GP}(m)$ is high, which means low level of background noise, $P(m)$ must tend toward zero ;
- if $SNR_{GP}(m)$ is low, which means high level of background noise, $P(m)$ must tend toward great values ;
- between the two limit cases, $P(m)$ must be variable and take into account $SNR_{GP}(m)$.

6.2 Filtering principle

The principle of the idea of incorporating $SNR_{GP}(m)$ is depicted in Fig. 3. First, $|\hat{D}(m,k)|^2$ obtained by a classic inter-frame noise estimator and the noisy speech PSD $|Y(m,k)|^2$ are used to compute $SNR_{GP}(m)$. Then, depending on the value of $SNR_{GP}(m)$, we compute $P(m)$ and thus the filter $H_S(m)$. Next, convolution of $|\hat{D}(m,k)|^2$ and $H_S(m)$ gives us the

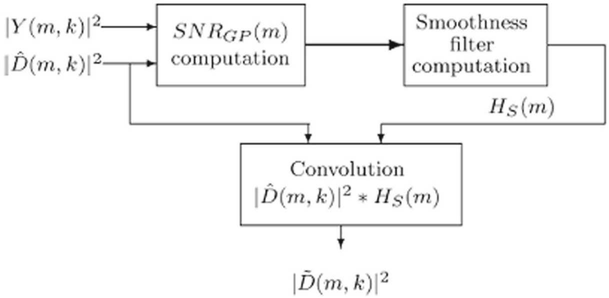

$|Y(m,k)|^2$
$|\hat{D}(m,k)|^2$
$SNR_{GP}(m)$ computation
Smoothness filter computation
$H_S(m)$
Convolution $|\hat{D}(m,k)|^2 * H_S(m)$
$|\tilde{D}(m,k)|^2$

Fig. 3 Smoothness processor incorporating Global Posterior SNR

smoothed version of noise estimation which will be considered in the denoising algorithms instead of the original estimated version $|\hat{D}(m,k)|^2$ as following:

$$|\tilde{D}(m,k)|^2 = \sum_{i=0}^{P(m)-1} H_S(m,i)|\hat{D}(m,k-i)|^2. \tag{13}$$

Equation (13), shows that the present frequency bin (index k) of the smoothed noise $|\tilde{D}(m,k)|^2$ depends from the present frequency bin of $|\hat{D}(m,k)|^2$ and from the past frequency bins $|\hat{D}(m,k-1)|^2$, $|\hat{D}(m,k-2)|^2$,..., $|\hat{D}(m,k-P(m)+1)|^2$. This fact, means that a simple average filtering combined the present frequency bin index k and the frequency bins situated at the left side. It is obvious that is not the reasonable way to smooth the estimated noise and keeping it's dynamics and nature. Hence we think to take into account both sides (left and right) to compute the smoothed version of the noise using the following expression:

$$|\tilde{D}(m,k)|^2 = \sum_{i=-Q(m)}^{Q(m)} H_S(m,i)|\hat{D}(m,k-i)|^2. \tag{14}$$

where $Q(m)$ designs the number of neighbourhood bins of the both sides (left and right).

6.3 Filter order computation

In this section, we propose to find the relationship between the degree of the smoothness, in other word $Q(m)$, and the $SNR_{GP}(m)$. Indeed, as it is mentioned in Section 6.2, we want to govern $Q(m)$ using $SNR_{GP}(m)$. Mathematically, the problem can be formulated as following:

$$\begin{cases} Q(m) = f\left\{SNR_{GP}(m)\right\} \\ Q(m) = \underset{Q}{\operatorname{argmax}}\left\{S_{quality}(m)\right\}, \end{cases} \tag{15}$$

where $S_{quality}(m)$ denotes the subjective quality of the denoised speech for the frame index m.

The first part of (15) can be identified as a mapping function that we want to establish. However, this is could be done only if we resolve the second part. It is clearly that the second part of (15) is very hard to resolve it analytically for two reasons. Firstly, evaluating subjectively each frame index m for different values of $Q(m)$ is very complicated to implement and maybe unrealistic experiment. Secondly, it is very difficult and we don't know yet how to express analytically the relationship between $Q(m)$ and $S_{quality}(m)$. To overcome these problems we propose to use objective evaluation ($O_{quality}(m)$) instead of subjective one since it can be easily computed using an objective criterion. Secondly, we propose to proceed experimentally to resolve (15) as it will be detailed in the following.

Experimental conditions are summarized in Table 1. We have chosen babble noise because of its speech nature which make it one of the difficult noise to remove from noisy speech. In addition, if we smooth so "much" the estimated noise we risk to lost noise dynamics. So, we propose to compute $Q(m)$ which improve the denoised quality and note damage it. Then, we think that we can generalized the results for the remainder kinds of noise. As objective criterion, we have chosen SNR_{seg} for its good correlation with subjective tests. The experimental protocol was tested with a set of 18 speech sequences extracted from TIMIT database [7].

Different steps of the experiment are as following.

Table 1 Experimental conditions

18 clean speeches	Extracted from TIMIT database [7]
Babble noise	Extracted from Noisex database [23]
Frequency sample	8000 Hz
Frame duration	30 ms
Hanning Window	50 % overlap
FFT length	512
Denoising technique	Wiener [15]

- Step 1 : we corrupt artificially the clean speech with babble noise at input SNR = -20 dB (first iteration).
- Step 2 : we set $Q(m) = 0$ (first iteration) and we filter the estimated noise suing (14) to get the smoothed version of the noise.
- Step 3 : the smoothed version of the noise is used with Wiener algorithm to denoise the noisy speech.
- Step 4 : we chose an active speech frame. From the same frame index of noisy speech, clean speech and denoised speech we compute $SNR_{GP}(m)$ and we assess the quality using SNR_{seg} criterion.
- Step 5 : we repeat the process from step 2 to determine $Q(m)$ which leads to the best quality in terms of SNR_{seg}. Hence, we obtain for a given $SNR_{GP}(m)$ the best $Q(m)$.
- Step 6 : we redone all the process from step 1 to get the best $Q(m)$ for different ranges of $SNR_{GP}(m)$.

Results of the proposed experiment are summarized in Fig. 4. We can remark that $Q(m)$ varies in steps. Such behavior can be explained as following. Since $Q(m)$ is an integer, its variation versus $SNR_{GP}(m)$ cannot be a continues one. As summary, we present in Table 2 the possible values of $SNR_{GP}(m)$ that can be computed for a noisy speech frame index m and related $Q(m)$ values. Once $Q(m)$ is predetermined for the noisy frame index m, the smoother filter $H_S(m)$ can be applied according to the (14).

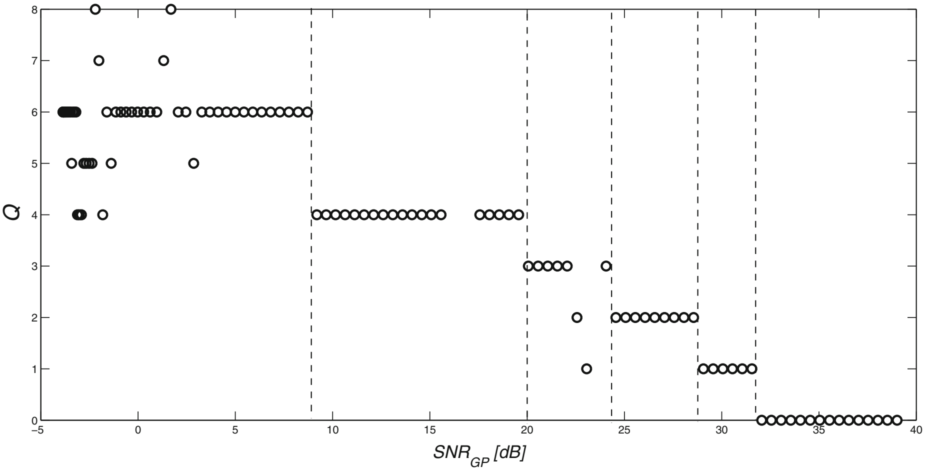


Fig. 4 Relationship between “Q(m)” and $SNR_{GP}(m)$

Table 2 Filter order depending on $SNR_{GP}(m)$ values in dB

$SNR_{GP} < 8.7$	$Q = 6$
$8.7 \leq SNR_{GP}(m) < 19.5$	$Q = 4$
$19.5 \leq SNR_{GP}(m) < 24$	$Q = 3$
$24 \leq SNR_{GP}(m) < 28.5$	$Q = 2$
$28.5 \leq SNR_{GP}(m) < 31.5$	$Q = 1$
$SNR_{GP} > 31.5$	$Q = 0$

In Fig. 5, we represent the power spectrum of the estimated noise using MCRA2 technique and its smoothed version using proposed technique. Experimental conditions are the same listed in Table 1. We can notice that the smoothed version of the estimated noise PSD keeps the global dynamics of the noise PSD when reducing the random fluctuations.

7 Assessment criteria

7.1 Subjective evaluation using P.835 ITU recommendation

The reliable way to assess the denoised speech is the subjective one, when human listeners are asked to judge subjectively the denoised speech over three dimensionalities: overall quality, speech distortion and residual background noise. The P.835 recommendation methodology was designed to reduce the listener's uncertainty in a subjective listening test to the nature of components degradation [13]. This method instructs the listeners to successively attend to rate the enhanced speech signal on:

- the speech signal alone using a five-point scale of signal distortion (SIG);
- the background alone using a five-point scale of background intrusiveness (BAK);
- the overall quality using the scale of the mean opinion score (OVL) as it is mentioned in

The SIG, BAK and OVL scales are described in Table 3.

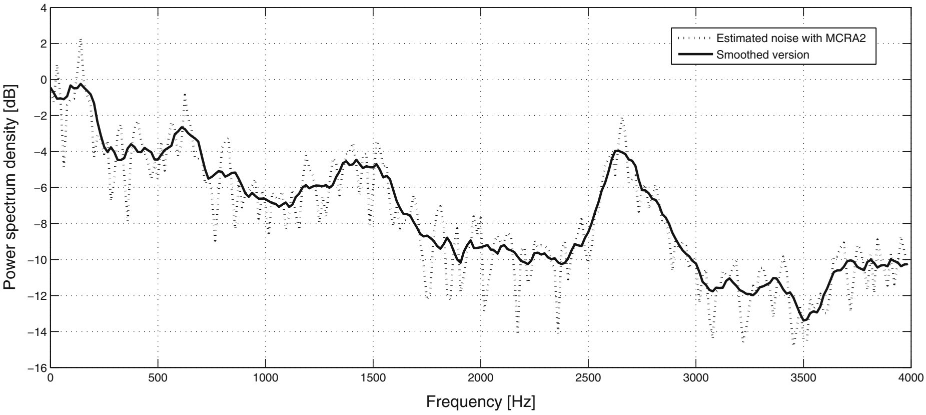


Fig. 5 The smoothness of noise power spectrum. $Q(m) = 4$ and $SNR_{GP}(m) = 12.3$

Table 3 Description of the SIG, BAK and OVL scales used in the subjective listening tests P.835

Rating	Description
SIG scale	
5	Not distorted
4	Slightly distorted
3	Somewhat distorted
2	Fairly distorted
1	Very distorted
BAK scale	
5	Not noticeable
4	Somewhat noticeable
3	Noticeable but not intrusive
2	Somewhat intrusive
1	Very intrusive
OVL scale	
5	Excellent
4	Good
3	Fair
2	Poor
1	Unsatisfactory

7.2 Objective evaluation

Subjective tests are the most accurate way to evaluate denoised speech. However, subjective assessment is an expensive process with specific equipment and large time consuming. Recently, a few attempts are proposed in the literature to develop objective criteria, which can correlate well with subjective tests. We retain, in this framework, two of recent measures. They are composite criteria [9, 11] and perceptual signal to audible noise and distortion ratio [2]. It is found that these criteria are well correlated with subjective test SIG, BAK and OVL when compared with existing objective criteria [2, 11]. The chosen criteria are outlined as follows.

7.2.1 Composite criteria

The basic idea of composite measure is the combination of some known conventional criteria to get new ones more correlated with subjective assessment over three dimensionalities: SIG, BAK and OVL. They are derived by utilizing multiple linear regression analysis.

- $C_{\mathbf{SIG}}$: composite criteria for SIG estimation

$$\begin{aligned} C_{\text{SIG}} = \; & 2.164 - 0.02 \cdot \text{IS} + 0.832 \cdot \text{PESQ} \\ & -0.494 \cdot \text{CEP} + 0.352 \cdot \text{LLR}. \end{aligned} \tag{16}$$

- $C_{\mathbf{BAK}}$: composite criteria for BAK estimation

$$\begin{aligned} C_{\text{BAK}} = \; & 0.985 + 0.848 \cdot \text{PESQ} - 0.319 \cdot \text{CEP} \\ & +0.295 \cdot \text{LLR} - 0.008 \cdot \text{WSS}. \end{aligned} \tag{17}$$

- $C_{\mathbf{OVL}}$: composite criterion for overall quality assessment

$$\begin{aligned} C_{\mathrm{OVL}} = \; & 0.279 - 0.011 \cdot \mathrm{IS} + 1.137 \cdot \mathrm{PESQ} \\ & +0.041 \cdot \mathrm{LLR} - 0.008 \cdot \mathrm{WSS}. \end{aligned} \tag{18}$$

where PESQ is the perceptual evaluation of speech quality [12], IS is the Itakura Saito distance measure [19], CEP is the cepstrum distance measures [19], LLR is the log-likelihood ratio [19], WSS is the weighted-slope spectral distance [14].

7.2.2 Perceptual signal to audible noise and distortion ratio

Measures developed in [2] use human auditory properties to characterize, perceptually, the degradations affecting denoised speech. Accordingly, only audible degradations are considered to evaluate denoised speech quality. The perceptual signal to audible noise and distortion ratio is a set of three measures.

- PSANR: perceptual signal to audible noise ratio. Developed to assess audible background noise.
- PSADR: perceptual signal to audible distortion ratio. Identify audible distortion of the clean speech.
- PSANDR: perceptual signal to audible noise and distortion ratio. Permit to evaluate the overall quality of the denoised speech.

8 Experimental results

8.1 Experimental conditions

In this section, we propose to evaluate and compare the proposed smoothing technique of the estimated noise (NSP: noise smoothing proccessor) with the conventional noise estimation method MCRA2. We limit our simulations with the MCRA2 technique because of its best performances comparable to the other noise estimation techniques [20]. We notice that the same results, remarks and conclusions are obtained with tested noise estimation techniques (VAD, MS, IMCRA, MCRA2).

The proposed method was tested with sentences taken from the TIMIT speech database [7]. The sentences were downsampled to 8 kHz before adding the noises which were extracted from NOISEX database [23]. We have chosen to test the proposed method for 3 realistic types of noise, namely white Gaussian noise, multitalker babble noise and car noise. The noise was added to the speech utterances in such a way that the input SNR range from -5 dB to 10 dB with 1-dB steps. The denoising process was applied to 32 ms (256 samples) frames of noisy speech with a 50 % overlap between adjacent frames. The denoised speech was obtained by the overlap-and-add method.

8.2 Subjective evaluation

8.2.1 Listening tests

The listening tests have been realized with 12 listeners. Each of the listeners assesses the enhanced speech over three dimensions SIG, BAK and OVL by giving a score between one and five according to the P.835 ITU recommendation [13]. Results of the subjective listening tests are summarized in Table 4. Subjective listening tests confirm that the smoothing

Table 4 Subjective listening tests according to P.835 recommendation

Input SNR	0dB	5dB	10dB
	SIG		
Noisy speech	**1.6**	**2.2**	**3.2**
Wiener(MCRA2)	1.2	1.6	3
Wiener(MCRA2)+NSP	1.4	2	3
	BAK		
Noisy speech	1	2	2.8
Wiener(MCRA2)	1.4	2.6	3.2
Wiener(MCRA2)+NSP	**2.2**	**2.8**	**3.4**
	OVL		
Noisy speech	1.2	2	2.8
Wiener(MCRA2)	1.6	2.6	3.2
Wiener(MCRA2)+NSP	**2.2**	**3**	**3.6**

Case of white noise. Bold values indicate better performances

processing of the estimated noise leads to the best results for the human listeners compared to the classic noise estimation techniques MCRA2. Especially for the case of background noise (BAK) and overall quality (OVL). However, the denoising algorithms do not perform the speech distortion (SIG). In fact, denoising speech methods introduce an avoidable distortion to the clean speech signal [16]. Hence, the less distortions of the clean speech are noticed for the noisy speech. Despite this fact, the proposed method seems to perform better than the traditional Wiener denoising technique.

8.2.2 Spectrogram

One way to see the effects of the proposed method on the quality of the denoised signal is the spectrogram. The same experimental conditions as the previous section are carried here. In Fig. 6, we depict the signals spectrograms. The upper subplots present the clean speech spectrogram and the spectrogram of the noisy speech (SNR=10dB). Fig. 6c shows clearly that the quality of the denoised speech is improved, especially in terms of background noise. However, some of speech components are also removed, especially those corresponding to high frequencies. Moreover, the well-known musical noise appears in the spectrogram of the denoised signal as isolated regions. In some cases this kind of artificial noise becomes more annoying than the original background noise [18, 21]. When looking to Fig. 6d, we can see clearly that such kind of noise is completely eliminated without additional speech distortion. This is an expected result. In fact, according to (3), the denoised speech spectrum can be obtained from the spectrum of noisy speech by simply multiplying it by a function gain which depends from the estimated noise spectrum. If the fluctuations of the estimated noise spectrum are reduced to their minimum monitored by the global posterior SNR_{GP}, the risk that musical tones arise is reduced also to its minimum.

8.3 Objective evaluation

– Overall quality evaluation

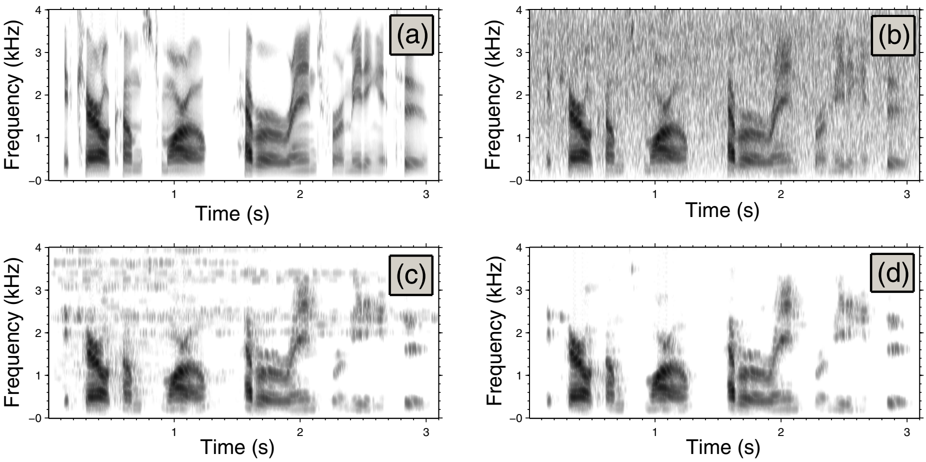


Fig. 6 Spectrograms: (**a**) clean speech, (**b**) noisy speech, (**c**) denoised speech with Wiener(MCRA2), (**d**) denoised speech with Wiener(MCRA2) + NSP

First of all, we have shown in Fig. 7 the assessment of the proposed method using the well-known ITU-recommendation P.862. PESQ measure shows clearly and for the different types of noise the improvement of the global quality of the enhanced speech when the NSP is used. Better performances are noticed for lower input SNR. The criteria C_{OVL} (Fig. 8) and $PSANDR$ (Fig. 9) confirm the interpretations concluded from Fig. 7. These results catch up with listening tests.

– Speech distortion noise evaluation

For the remainder of the paper, we present the case of white noise, but we notice that the same conclusions and interpretations are conducted for the three types of noise. We

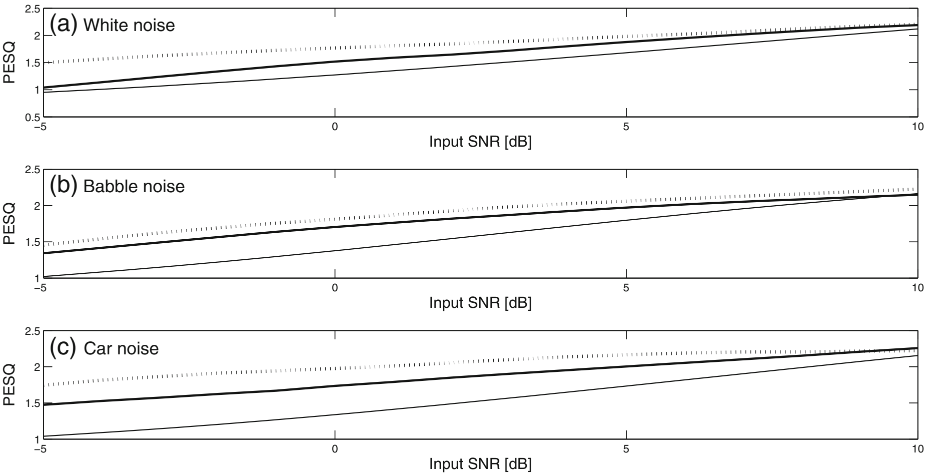


Fig. 7 Objective evaluation using PESQ criterion. *Solid line*: noisy speech, *Bold solid line*: Wiener(MCRA2), *Bold dashed line*: Wiener(MCRA2) + NSP

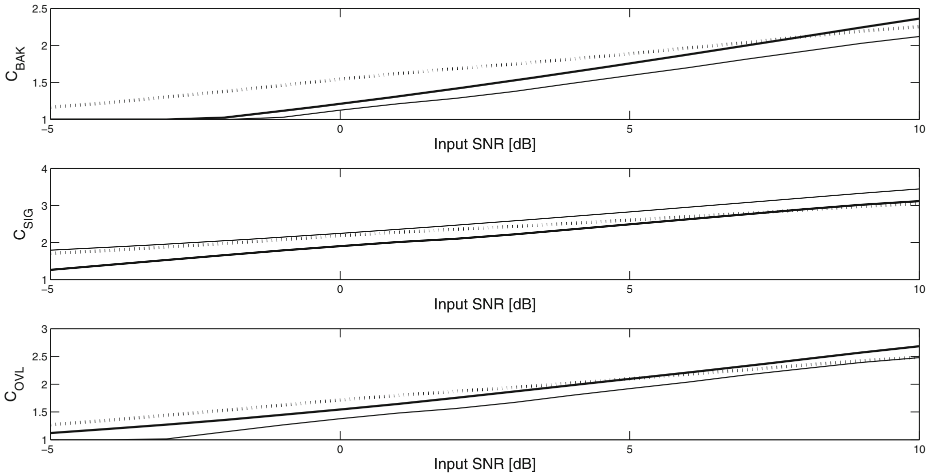


Fig. 8 Objective evaluation using composite criteria (C_{BAK}, C_{SIG} and C_{OVL}). Case of white noise. *Solid line*: noisy speech, *Bold solid line*: Wiener(MCRA2), *Bold dashed line*: Wiener(MCRA2) + NSP

present in Figs. 8 and 9 the objective assessment of the speech distortion using respectively C_{SIG} and $PSADR$. As it is expected denoising techniques introduce a speech distortion which can be seen as a degradation of both objective criteria. However, the proposed technique provides less distortion than the conventional Wiener filtering without NSP processing.

– Background noise evaluation

Figures 8 and 9 depict respectively the assessment of the residual background noise using C_{BAK} and $PSANR$. The two criteria show the improvement brought by the NSP processing. Especially for the low input SNR. For the higher SNR the performances of both classical Wiener and proposed method become close.

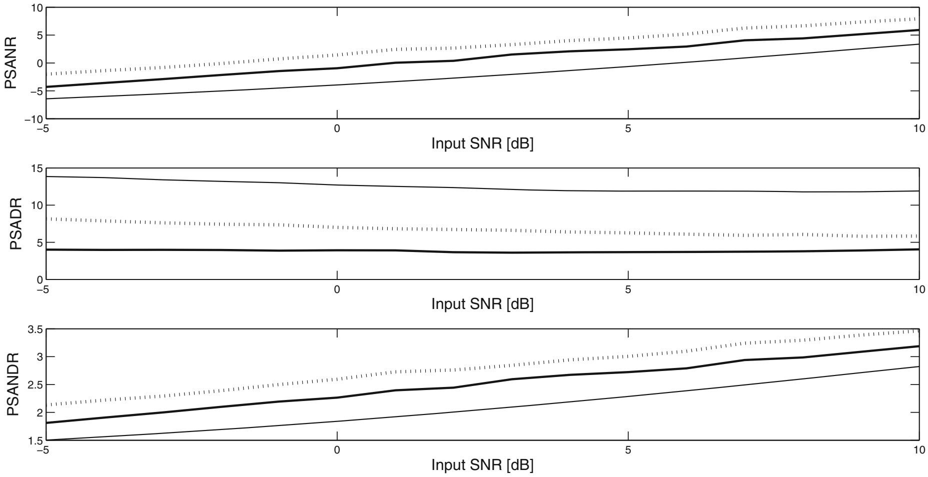


Fig. 9 Objective evaluation using perceptual criteria (PSANR, PSADR and PSANDR). case of white noise. *Solid line*: noisy speech, *Bold solid line*: Wiener(MCRA2), *Bold dashed line*: Wiener(MCRA2) + NSP

9 Conclusion

The quality of the enhanced speech depends from two points; the denoising filter design and the background noise estimation. In this framework, an experimental analysis of the relationship between noise estimation and denoised speech quality is given. It is shown that the smoothness of the estimated noise can improve the denoised quality. Contrary to conventional noise estimators which introduce an inter-frame smoothness, the proposed method acts as an intra-frame smoothing processor. The influence of the noise PSD smoothing on the denoised quality is studied. The proposed post-processing NSP operates as an average filter with variable number of coefficients and governed by the Global Posterior SNR. In such a way, we preserve the dynamics of the noise PSD which leads to track different types of noises. It is shown that the smoothing processing of NSP prevents the occurrence of the residual background musical noise.

The proposed method was evaluated according to three dimensions; background noise, speech distortion and overall quality using recently developed criteria for this purpose. The proposed noise estimation algorithm shows a better performance when compared to traditional noise estimation algorithms. Using the proposed noise estimation algorithm in a noise reduction system, higher noise suppression is achieved together with lower speech distortion.

Compliance with Ethical Standards All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent was obtained from all individual participants included in the study.

Conflict of interests Author Anis Ben Aicha declares that he has no conflict of interest.

References

1. Ben Aicha A, Ben Jebara S (2010) Some reflexion and results about adequate speech denoising technique according to desired qulity. In: International conference on information science, signal processing and their applications
2. Ben Aicha A, Ben Jebara S (2012) Perceptual speech quality measures separating speech distortion and additive noise degradations. Speech Commun 54(4):517–528
3. Ben Aicha A, Ben Jebara S (2012) Reduction of musical residual noise using perceptual tools with classic speech denoising techniques. Signal Image Video Process 6(1):85–97
4. Benesty J, Chen J, Huang Y, Cohen I (2009) Noise reduction in speech processing. Springer
5. Cohen I (2003) Noise spectrum estimation in adverse environnements: improved minima controlled recursive averaging. IEEE Trans Speech Audio Process 11(5):466–475
6. Derakhshan N, Akbari A, Ayatollahi A (2009) Noise power spectrum estimation using constrained variance spectral smoothing and minima tracking. Speech Commun 51:1098–1113
7. Garofolo JS (1988) Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database. National Institute of Standards and Technology
8. Hansen JH, Radhakrishnan V, Arehart KH (2006) Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system. IEEE Trans Audio Speech Lang Process 14(6):2049–2063
9. Hu Y, Loizou FC (2006) Subjective comparaison of speech enhancement algorithms. In: IEEE international conference on acoustics, speech and signal processing
10. Hu Y, Loizou FC (2007) A comparative intelligibility study of speech enhancement algorithms. In: IEEE international conference on acoustics, speech and signal processing
11. Hu Y, Loizou P (2008) Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Lang Process 16(1):229–238

12. ITU-T Recommendation (2000) Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P. 862
13. ITU-T Recommendation (2003) Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. ITU-T recommendation P. 835
14. Klatt D (1982) Prediction of perceived phonetic distance from critical band spectra. In: IEEE International conference on acoustics, speech and signal process
15. Loizou P (2007) Speech enhancement: theory and practice. CRC Press
16. Loizou PC, Kim G (2011) Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans Audio Speech Lang Process 19(1):47–56
17. Martin R (2001) Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans Speech Audio Process 9(5):504–512
18. Miyazaki R, Saruwatari H, Nakamura S, Shikano K, Kondo K, Blanchette J, Bouchard M (2014) Musical-noise-free blind speech extraction integrating microphone array and iterative spectral subtraction. Signal Process 102:226–239
19. Quanckenbush S, Barnwell T, Clements M (1988) Objective measures of speech quality. Prentice Hall Englewood Cliffs
20. Rangachari S, Loizou P (2006) A noise-estimation algorithm for highly non-stationary environments. Speech Commun 48:220–231
21. Saxena P, Gupta VK, Chandra M (2016) Musical noise reduction capability of various speech enhancement algorithms. Information systems design and intelligent applications. Springer, India
22. Sohn J, Kim NS, Sung W (1999) Statistical model-based voice activity detection. IEEE Signal Process Lett 6(1):1–3
23. Varga A, Steeneken HJM (1993) Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun 12:247–251
24. Vondrášek M, Pollák P (2005) Methods for speech SNR estimation: evaluation tool and analysis of VAD dependency. Radio Eng J 14(1):6–11

Anis Ben Aicha received MSc and PhD (2004 and 2010, respectively, both with honors) in telecommunications from the Higher School of Communications of Tunis (SUP'COM), University of Carthage, Tunisia. Assistant Professor at the Faculty of Sciences of Bizerte (FSB). His research and teaching activities focus on signal processing, speech processing and image processing including speech analysis and synthesis, speech enhancement, speech assessment, image segmentation with elements of data mining and information theory. Author or co-author of over 20 conference and journal papers, supervisor of over 30 bachelor and master theses.