Semi-Parametric Joint Detection and Estimation for Speech Enhancement based on Minimum Mean Square Error

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

Semi-Parametric Joint Detection and Estimation for Speech Enhancement based on Minimum Mean Square Error

Van-Khanh MAI, Dominique PASTOR, Abdeldjalil AISSA-EL-BEY, Raphaël LE BIDAN

IMT Atlantique, UMR CNRS 6285 Lab-STICC, UBL, 29238 Brest, France

Abstract

We propose a novel estimator for estimating the amplitude of speech coefficients in the time-frequency domain. In order to avoid a phase spectrum estimator of complex coefficients when using the Fourier transform, we consider the discrete cosine transform (DCT). This estimator aims at minimizing the mean square error of the absolute values of the speech DCT coefficients. In order to take advantage of both parametric and non-parametric approaches, the proposed method combines block shrinkage and Bayesian statistical estimation. First, the absolute value of the clean coefficient is estimated by block smoothed sigmoid-based shrinkage (Block-SSBS). The block size required by the block-SSBS is obtained by statistical optimization. This step enables us to reduce the negative impact on speech intelligibility of classical denoising methods similarly to smoothed binary masking. Second, for refining the estimation, an optimal statistical estimator is added to handle musical noise. For evaluating the performance of the proposed method, objective criteria are used. The experiments enhance the relevance of the approach, in terms of speech quality and intelligibility.

Keywords: Speech enhancement, noise reduction, Bayesian estimation, non-parametric estimation, Smoothed sigmoid-based shrinkage.

1. Introduction

1.1. Motivation

Machine learning approaches can provide good performance in speech enhancement (see [1, 2, 3] among others). However, unsupervised techniques are still needed when available databases hardly cover all types of noise and speech signals met in practice [4, 5, 6, 7] as, for example, in assisted listening for hearing aids, cochlear implants or voice communication applications. Indeed, in such cases, unsupervised approaches can achieve a good trade-off between intelligibility and quality with low complexity.

It then turns out that many results in non-parametric and robust statistical estimation established in the last two decades [8, 9, 10, 11, 12, 13] and based on sparse thresholding and shrinkage, are general enough to suggest their use in unsupervised speech denoising. Generally speaking and as recalled below, the interest of non-parametric speech denoising is twofold. First, it can be applied without any knowledge or assumption on the signal distribution. Second, it achieves gain in intelligibility [14]. Since Bayesian approaches are known to improve speech quality [15], the idea is to combine the two approaches. Nonetheless, this combination requires some care. Indeed, most non-parametric estimators force to 0 small magnitude coefficients obtained after transformation into a certain domain. Although much background noise is canceled by doing so, removing small noisy coefficients pertaining to the signal of interest generates musical noise and reduces speech quality [16]. This problem is well known in image processing where zero-forcing of small coefficients induces artifacts [10].

Therefore, if we want to improve quality by eliminating residual musical noise, the non-parametric denoising should be a smooth shrinkage merely aimed at attenuating small coefficients. A Bayesian estimator can then be used right after the non-parametric one to retrieve speech information in small coefficients and thus improving the overall quality. Note that if the Bayesian estimator were used before the non-parametric one, the latter would tend to shrink small coefficients estimated by the former, which is not desirable because even small coefficients after Bayesian estimation may pertain to relevant speech contents for overall quality.

With respect to the foregoing, the problem addressed in this paper is the design and combination of non-parametric and Bayesian estimations for speech denoising. We restrict our attention to the single-channel case. Indeed, a technique designed in the single-channel case can always be used after beamforming on a microphone array. Since we focus on statistical approaches, no psycho-acoustic knowledge on speech signals in noise is considered below.

1.2. Contributions

In this paper, similar to the other methods mentioned above, we estimate the amplitudes of the clean signal coefficients in the time-frequency domain. The estimation is based on the MMSE criterion. However, instead of the DFT, we focus on the discrete cosine transform (DCT), which avoids estimating

the phase spectrum as in [5, 17] and may reduce complexity [18, 19]. To this end, we will consider the following strategy.

We begin by reducing the negative impact on speech intelligibility of the denoising methods by a non-parametric approach based on smoothed sigmoid-based shrinkage (SSBS) [11], originally introduced for image denoising. Two main features of the approach are: 1) it attenuates DCT coefficients that are very likely to pertain to noise only or to speech with small amplitude in noise; 2) it tends to keep unaltered large-magnitude DCT coefficients. However, such a non-parametric approach can be regarded as an approximated Wiener filtering and, as such, introduces musical noise. We then modify the original SSBS approach and propose the SSBS block estimator, hereafter named Block-SSBS. Block-SSBS is relevant to eliminate isolated points in the time-frequency domain that may induce musical noise. Basically, Block-SSBS applies the same SSBS gain function to time-frequency blocks. The sizes of these blocks are determined by adaptive Stein's Unbiased Risk Estimate (SURE) [20] so as to minimize the unbiased estimate of the mean square error over regularly distributed time-frequency regions. In addition, other parameters of Block-SSBS can be optimized by resorting to recent results in non-parametric statistical signal processing [21]. A nice feature of the proposed parameter optimization procedure is the level of control offered on the denoising performance, which yields a good compromise between speech quality and intelligibility. This is made possible by discriminating speech components with significant contents from speech components with lesser interest.

For reasons detailed below, the outcome of Block-SSBS is assumed to satisfy the same hypotheses as those generally used for Bayesian estimation. Therefore, in a second step, to further reduce musical noise and, above all, improve speech quality, a Bayesian statistical estimator is devised for application to the smoothed short-time spectral amplitudes (STSA) provided by Block-SSBS. This Bayesian estimator is hereafter called STSA-MMSE.

In a nutshell, the main contributions of this paper are the following ones. To begin with, the whole method is carried out in the DCT domain, so as to get rid of the phase estimation problem. It introduces Block-SSBS in the DCT domain for speech denoising in presence of stationary or non-stationary noise. Block-SSBS is then optimized via automatic and adaptive statistical methods tailored to speech enhancement. The derivation of STSA-MMSE in the DCT domain is another contribution. The paper also propounds and studies the combination of Block-SSBS and STSA-MMSE and shows that this combination is very promising for speech denoising in presence of various types of noise. It must also be pointed out that these tests include situations where the noise spectrum is known, as well as cases where this spectrum is estimated via an up-to-date estimator.

1.3. Paper organization

Section 2 introduces the signal model, the notation and makes some general recalls on the DCT. In Section 3, we present semi-parametric speech enhancement by Block-SSBS, derive the Bayesian STSA-MMSE in the DCT domain and then combine the two. Experimental results are reported and analyzed in Section 4. Finally, Section 5 concludes this paper with prospects opened by this work.

2. Signal Model and notation in the DCT domain

As announced above, the DCT will hereafter be used for denoising. Therefore, this section recalls the principle of DCT and the reasons why DCT can be applied to speech enhancement.

DCT is analyzed from a general point of view in [22]. Originally developed for pattern recognition and Wiener filtering in image processing, its application to speech enhancement is more specifically studied in [18, 19]. Basically, given a sequence $\{y[n]\}$ with $0 \leq n \leq K-1$, the DCT coefficients are calculated as:

$$Y[k] = \alpha_k \sum_{n=0}^{K-1} y[n] \cos \frac{(2n+1)k\pi}{2K}, \tag{1}$$

with $\alpha_0 = \sqrt{1/K}$ and $\alpha_k = \sqrt{2/K}$ for $1 \leq k \leq K-1$ [23]. The inverse DCT is then given by:

$$y[n] = \sum_{k=0}^{K-1} \alpha_k \, Y[k] \cos \frac{(2n+1)k\pi}{2K}. \tag{2}$$

The DCT defined by (1) and (2) can be advantageously used in speech enhancement or noise reduction for the subsequent reasons. As discussed in [18, 23, 22], DCT has higher energy compaction than DFT. The signal of interest can thus have a sparse representation in the DCT domain. That is why DCT is widely used in image compression [22] and dictionary learning [24]. Second, the DCT coefficients are real, whereas the DFT coefficients are complex. The DCT coefficients have binary phase, whereas phases of the DFT coefficients are often assumed to follow the uniform distribution in the range $[-\pi, \pi]$. Therefore, the DCT phase [18] does not need to be estimated because error in the DCT phase has no important impact for estimating the signal of interest. Third, DCT is known to be better than DFT for approximating the Karhumen-Loève transform (KLT), which is optimal in terms of variance distribution, rate distortion function and mean-square estimation error. Moreover, DCT and inverse DCT (IDCT) can be also calculated by fast computation algorithms.

For estimating clean speech from its noisy observation, the latter is often segmented, windowed and transformed by computational harmonic analysis. In the present framework, for the reasons evoked above, this harmonic analysis will be performed by DCT. Formally, let us denote the noisy signal in the DCT domain by:

$$Y[m,k] = S[m,k] + X[m,k], \tag{3}$$

where m and $k \in \{0, 1, \ldots, K-1\}$ are the time and frequency-bin indices, respectively. As an extension of (1) and similarly to the expressions of the DFT coefficients, the DCT coefficients are obtained as [16]:

$$Y[m,k] = \sum_{n=0}^{K-1} \alpha_n \, w[n] \; y[mK^* + n] \; \cos \frac{(2n+1)k\pi}{2K}, \tag{4}$$

where K is the frame length, K^* is the number of the shifted samples and $w[n]$ is a window function such as the Hamming or Hanning windows with length K. For the sake of simplicity, the indices m and k will be omitted unless for clarification. Wide hat symbols are henceforth used to denote estimates. Moreover, lower case letters denote realizations of random variable. The absolute value (resp. sign) of the DCT coefficients of the noisy signal, signal of interest and noise are denoted by A_Y, A_S, A_X (resp. Φ_Y, Φ_S, Φ_X), correspondingly.

The signal of interest and the noise are assumed to be independent and zero mean, so that $\mathbf{E}[Y^2] = \mathbf{E}[S^2] + \mathbf{E}[X^2] = \sigma_S^2 + \sigma_X^2$, where the spectra of the clean signal and noise are denoted by $\mathbf{E}[S^2] = \sigma_S^2$, $\mathbf{E}[X^2] = \sigma_X^2$, respectively, and where $\mathbf{E}(.)$ is the expectation. We also define the *a priori* signal-to-noise ratio (SNR) ξ and the *a posteriori* SNR γ as $\xi = \sigma_S^2/\sigma_X^2$, $\gamma = A_Y^2/\sigma_X^2$. As usual [25], the DCT coefficients $Y[m,k]$ with $k \in \{0, 1, \ldots, K-1\}$ are assumed to be uncorrelated. The notation introduced above is used throughout with always the same meaning.

3. Block speech estimation of Discrete Cosine Coefficients

Our purpose is to design a method that achieves a good trade-off between intelligibility and quality. To this end, we combine an SSBS-based method with a Bayesian statistical estimator. The rationale for this combination is the following. Bayesian statistical estimators of STSA in the DCT domain can be expected to provide good performance in speech enhancement, especially to improve quality without introducing musical noise. Therefore, a Bayesian estimator placed right after an SSBS-based approach which cancels most of the background noise and attenuates small coefficients pertaining to speech, should then contribute to retrieving information on clean speech and thus enhancing speech quality.

In this respect, the next subsection reviews basics on non-parametric thresholding methods originally developed for image denoising, with a particular emphasis on SSBS. Then, Subsection 3.2 introduces the Block-SSBS approach. Based on the SSBS estimator, it is designed for audio denoising. Section 3.3 then presents STSA-MMSE, a Bayesian estimation of STSA in the DCT domain. The combination of Block-SSBS and STSA-MMSE is described in Section 3.4.

3.1. Sparse thresholding and shrinkage for detection and estimation

Denoising by shrinkage involves estimating the signal of interest by thresholding the coefficients obtained by projection of the noisy observation onto an orthogonal basis. Given an observation coefficient Y in the wavelet, DCT or DFT domain, the estimate $\widehat{S}$ is obtained by $\widehat{S} = GY$, where G is a gain or shrinkage function. In the sequel, G will be expressed as a function of γ or an estimate of γ. For instance, the hard thresholding gain function [8, 9] with threshold λ is:

$$G_\lambda(\gamma) = \begin{cases} 1 & \text{if} \quad \gamma \geq \lambda^2, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Smooth shrinkage performs estimation of the clean transformed coefficient by the soft thresholding gain function [9, 26]:

$$G_\lambda(\gamma) = \begin{cases} 1 - \frac{\lambda}{\sqrt{\gamma}} & \text{if} \quad \gamma \geq \lambda^2, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Soft thresholding combines detection and estimation. Indeed, by comparing the *a posteriori* SNR γ to λ^2 and setting $\widehat{S}$ to zero if the *a posteriori* SNR γ falls below this threshold, a kind of speech detection is realized. In addition, soft thresholding provides a transformed coefficient estimate of the desired signal by subtracting the threshold from the noisy coefficients.

The SSBS approach [11, 10] performs another type of smoothed shrinkage. The original SSBS gain function [11] reads:

$$G_{\tau,\lambda}(\gamma) = \frac{1}{1 + e^{-\tau(\sqrt{\gamma} - \lambda)}}, \tag{7}$$

where parameter λ influences the detection performance and τ controls the shrinkage. Such a function achieves smoothness, penalized shrinkage and vanishing attenuation at infinity. It is a trade-off between hard and soft thresholding. In particular, SSBS functions attenuate in a continuous manner values of $\sqrt{\gamma}$ that are below λ, instead of setting them to zero as in hard and soft thresholding.

The attenuation factors or gain functions $G_\lambda(\gamma)$ and $G_{\tau,\lambda}(\gamma)$ are independently evaluated for each (time, frequency) pair. In the sequel, we will extend these functions so as to incorporate neighboring time-frequency atoms, as in [27] or [28].

3.2. Non-parametric estimation by Block-SSBS

The original SSBS estimation is a pointwise method which may yield isolated spectral amplitudes and, thus, musical noise in speech enhancement. We can eliminate these isolated points by performing SSBS by blocks of time-frequency neighboring atoms. Such an approach is very similar to that proposed in [27] for denoising signals in the wavelet domain. However, the method we propose has some specific features.

First, it is carried out in the DCT domain for reasons evoked before. Second, speech is not stationary but can be considered stationary on relatively small time-frequency zones. The same may hold for non-stationary noise as well. It follows that we must choose time-frequency zones in which speech and noise can reasonably be expected to be stationary. Such zones are unknown and highly dependent on the signal and noise of interest. The design of algorithms dedicated to the detection of such zones is postponed to future work. In this work, we restrict our attention to a regular splitting of the time-frequency domain in rectangular time-frequency boxes with same size $(\Delta T, \Delta F)$, where ΔT is the number of time frames and ΔF is the number of frequency bins in each box. Values for ΔT and ΔF will hereafter be chosen so that speech and noise can acceptably be regarded as stationary in the resulting time-frequency boxes. If the speech distribution in a given box is assumed to be unknown, the general methodology exposed in [27] can be adapted as follows for noisy speech estimation in the DCT domain.

Since the signal distribution in a given box is unknown, the idea is to divide the box into non-overlapping rectangular blocks so that the signal can reasonably be considered to be deterministic and unknown in each block. To reduce computational cost, we look for blocks with same size inside a given box. The issue is then to find the optimal blocks size such that the overall estimation error in the box containing these blocks is minimal. On the one hand, when the box is filled with noise only, it makes sense to divide this box into small blocks. In this case, the optimal block size should be the minimum block size. On the other hand, when the box contains speech, it is expected that the local stationarity of the speech signal will lead to a relatively big optimal block size. In order to comply with the aforementioned considerations, the following estimation algorithm arises.

3.2.1. Block-SSBS gain function

Consider a given box $\mathfrak{B}$ and a block B within this box. As mentioned above, speech is assumed to be deterministic but unknown in B. Noise is assumed to be zero-mean and Gaussian distributed in the box under consideration, so that the noise variance is supposed to be the same in all blocks within this box. Let $\sigma_X^2(\mathfrak{B})$ stand for the noise power spectrum in $\mathfrak{B}$. Under these assumptions, in block B, the estimated *a posteriori* SNR $\widehat{\gamma}$ can be calculated by averaging the instantaneous noisy signal energies $Y^2[m,k]$ divided by the noise variance, so that:

$$\widehat{\gamma}_B = \frac{1}{\sigma_X^2(\mathfrak{B}) \times |B|} \sum_{(m,k)\in B} Y^2[m,k] \tag{8}$$

where $|B|$ is the number of time-frequency points (m,k) within B. Since we want to remove isolated time-frequency points, we proceed similarly to [27] and [28] by choosing the SSBS gain function in block B equal to $G_{\tau,\lambda}(\widehat{\gamma}_B)$ where $G_{\tau,\lambda}$ is given by (7). To implement this gain function, we must choose the sizes of the boxes and blocks as well as parameters τ and λ.

3.2.2. Size of the time-frequency boxes

With the notation introduced above, the larger ΔT, the greater the time delay. Therefore, for real time processing applications, the length ΔT should be small enough. We have chosen ΔT = 8 frames (*i.e* 128 ms in our implementation) as a good trade-off between performance and time-delay. Furthermore, taking into consideration that non-stationary noise impacts differently distinct frequency bands, we follow [29], which recommends to choose more than 6 bands, linearly spaced within the bandwidth $[0, 8]$ kHz, to get good speech quality. Accordingly, and as a good trade-off between performance and computational load, we set $\Delta F = 16$, which corresponds to 8 bands linearly spaced.

3.2.3. Time-frequency splitting by SURE

We now address the computation of the optimal block size within a given box $\mathfrak{B}$. The common size of the blocks is a pair henceforth denoted by (L, W). The number of DCT coefficients pertaining to any block is thus $N = LW$. The computation of the optimal size (L^*, W^*) for the blocks within a given box $\mathfrak{B}$ can be performed as in [27, 28], by resorting to the SURE approach derived from Stein's Theorem [20]. However and in contrast to [27, 28], the SURE approach is hereafter limited to the estimation of the optimal block size (L^*, W^*) and will not be used to estimate λ or τ. Indeed, these two parameters can be evaluated more appropriately via other means, as explained later.

For a given τ and λ, consider a box $\mathfrak{B}$. Split this box in J non-overlapping rectangular blocks $B_1, \ldots, B_J$. The overall estimation risk for $\mathfrak{B}$ and its partition into J boxes is thus:

$$R = \sum_{j=1}^{J} R_j, \tag{9}$$

where

$$R_j = \sum_{(m,k)\in B_j} \mathbf{E}\left[|\widehat{S}[m,k] - S[m,k]|^2\right]$$

and

$$\widehat{S}[m,k] = G_{\tau,\lambda}(\widehat{\gamma}_{B_j}) Y[m,k]$$

for $(m,k) \in B_j$. Since the SSBS gain function is constant in each block and the blocks are constrained to have same size, the overall risk depends on the block size (L, W). The SURE Theorem now provides us with an unbiased estimate of R_j. Therefore, we can calculate an unbiased estimate of the overall risk R. It is then possible to look for the block size (L^*, W^*) that minimizes this unbiased estimate of R.

Specifically, we proceed as follows. Let $Y[m,k]$ with $(m,k) \in B_j$ be the N available DCT values in block B_j. We can rearrange these DCT values so as to form an N-dimensional random vector $\mathbf{Y}$. Since speech is supposed to be deterministic unknown and noise to be Gaussian in B_j with variance σ_X^2, we assume that

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{S}, \sigma_X^2(\mathfrak{B})\mathbf{I}_N) \tag{10}$$

where $\mathbf{S}$ models the unknown speech signal in B_j and $\mathbf{I}_N$ is the $N \times N$ identity matrix.

Now, define $\widehat{\mathbf{S}} : \mathbb{R}^N \to \mathbb{R}^N$ for any $\mathbf{y} \in \mathbb{R}^N$ by

$$\widehat{\mathbf{S}}(\mathbf{y}) = G(\mathbf{y})\mathbf{y}$$

and use Eq. (8) so that:

$$G(\mathbf{y}) = G_{\tau,\lambda}\left(\frac{\|\mathbf{y}\|_2^2}{N\sigma_X^2(\mathfrak{B})}\right)$$

where $\|\cdot\|_2$ stands for the usual Euclidean norm in $\mathbb{R}^N$. Readily, $\widehat{\mathbf{S}}$ is differentiable. Therefore, [30, Section 2] applies and the Stein's unbiased risk estimate of R_j is given by:

$$\widehat{R}_j(\mathbf{y}) = -N\sigma_X^2(\mathfrak{B}) + \|\mathbf{y} - \widehat{\mathbf{S}}(\mathbf{y})\|_2^2 + 2\sigma_X^2(\mathfrak{B}) \sum_{n=1}^{N} \frac{\partial \widehat{S}_n}{y_n}(\mathbf{y}) \tag{11}$$

with $\mathbf{S} = (S_1, \ldots, S_N)$. Straightforward algebra leads to:

$$\begin{aligned}\widehat{R}_j(\mathbf{y}) = {} & N\sigma_X^2\left(2G_{\tau,\lambda}(\widehat{\gamma}_{B_j}) - 1\right) + \left(1 - G_{\tau,\lambda}(\widehat{\gamma}_{B_j})\right) \\ & \times \left(1 + \tau\, G_{\tau,\lambda}(\widehat{\gamma}_{B_j}) / \left(N\sqrt{\widehat{\gamma}_{B_j}}\right) - G_{\tau,\lambda}(\widehat{\gamma}_{B_j})\right) \|\mathbf{y}\|_2^2,\end{aligned} \tag{12}$$

We can then estimate R by:

$$\widehat{R} = \sum_{j=1}^{J} \widehat{R}_j \tag{13}$$

As enhanced at the beginning of this section, the overall risk R depends on the size (L, W) of the blocks composing the box. Therefore, we carry out an exhaustive search among all possible pairs (L, W) so as to find the pair (L^*, W^*) that minimizes $\widehat{R}$. Note that the value of $\widehat{R}_j$ does not only depend on N but also on L and W through $\mathbf{y}$. With respect to the values ΔT and ΔF chosen above for the boxes $\mathfrak{B}$, it turns out that the set of all possible sizes (L, W) contains 20 values only, which is easily tractable in practice. In addition, the noise variance $\sigma_X^2(\mathfrak{B})$ within a given box $\mathfrak{B}$ is estimated according to:

$$\sigma_X^2(\mathfrak{B}) = \frac{1}{|\mathfrak{B}|} \sum_{(m,k)\in\mathfrak{B}} \sigma_X^2[m,k], \tag{14}$$

where $|\mathfrak{B}|$ is the number of the time-frequency bin $[m, k]$ in $\mathfrak{B}$ and the values $\sigma^2[m, k]$ are chosen equal to the actual noise power spectrum if it is known, or estimated otherwise.

Fig. 1 shows an example of box and block tiling obtained by minimization of the overall risk (13) on some noisy speech. In this figure, boxes have size 8×16 and the color of each box corresponds to the size determined by the SURE approach for the blocks within this box. For example, the rectangular box that spans from frames 17 to 24 and from frequency bins 17 to 32 is divided into blocks of size 16. Note that, as expected, the SURE approach yields a block size equal to the box size in time-frequency zones that are mainly occupied by speech. This is normal since, within such boxes, speech is homogeneous thanks to the chosen size for boxes. In contrast, in boxes where mainly noise is present, the SURE approach returns smaller block sizes because variations of speech inside these boxes require a finer analysis. This was expected as well.

3.2.4. Random distortion threshold (RDT) based selection of Block-SSBS parameters τ and λ

For speech enhancement applications, the two parameters τ and λ in (7) are also key elements for controling the performance of the proposed method and reaching the desired trade-off between signal distortion and noise reduction. As mentioned above, it is possible to estimate τ and λ via the SURE approach. Such a possibility has not been tested in this work for reasons detailed in the next two paragraphs.

Choice of λ. This parameter plays the role of a threshold that can be used to make a decision on speech presence or absence. This threshold may therefore vary significantly in the time-frequency domain with respect to the type of speech signal under observation. Thence the idea to estimate this threshold in each block, once (L^*, W^*) has been calculated. Additionally, it is desirable to keep some control on the estimation performance, which is not actually feasible via the SURE approach. Whence the interest of the following non-parametric approach, since it ensures that the proposed choice for λ is optimal while upper-bounding the false alarm probability of erroneously deciding that significant speech is present.

The method we propose is based on the following rationale. Parameter λ influences the performance of shrinkage by SSBS gain function because it affects the level of noise reduction applied to the noisy DCT coefficients. Although the SSBS gain function is smoother than the hard thresholding gain function, parameter λ must however be carefully chosen to enhance speech quality. Indeed, suppressing too many speech components for reducing noise will necessary induce loss in speech quality. Otherwise said, when one aims at improving not only speech quality but also speech intelligibility, missing some important speech-carrying time-frequency channels may be more detrimental to speech enhancement than conserving more noise-only channels than strictly required. This favors the choice of small values for λ. On the other hand, the smaller λ, the smaller the signal distortion and musical noise, but the larger the residual background noise. Therefore, we cannot choose too small a value for λ. A means to achieve such a trade-off is to control the denoising by taking the outcome of some speech detector into account [31, 32].

We follow a similar strategy by choosing λ such that DCT coefficients with amplitudes above λ pertain to relevant speech signal components with high probability, whereas DCT coefficients below λ are more certainly components of noise only or noisy speech coefficients that can be safely discarded. Since we accept that observations with amplitudes below λ may contain information merely attenuated by the SSBS function, the choice of λ is not derived hereafter from a detection problem as in [10, 33] for denoising images by wavelet shrinkage. Instead, we resort to the random distortion testing (RDT) approach [21].

Basically, with the notation and hypotheses of (10), the RDT approach amounts to testing whether $\|\mathbf{S}\|_2 \leqslant \delta$ or not when we observe $\mathbf{Y}$, where δ is a tolerance that is specified by the application. For a better understanding of the sequel, it must be noticed that this binary hypothesis test is invariant by orthogonal transform, in the sense that it remains identical under any any orthogonal transform of $\mathbb{R}^N$ applied to $\mathbf{Y}$. This basically derives from the properties of the Gaussian distribution.

Let us decide that $\|\mathbf{S}\|_2 \leqslant \delta$ if $\|\mathbf{Y}\|_2 \leqslant \sigma_X \eta_\alpha(\delta/\sigma_X)$ and that $\|\mathbf{S}\|_2 > \delta$ otherwise, where $\eta_\alpha(\delta/\sigma_X)$ is the unique solution in x to the equation $Q_{N/2}(\delta/\sigma_X, x) = \alpha$, where $Q_{N/2}(\cdot,\cdot)$ stands for the Generalized Marcum function [21] . According to [21, Proposition 2], this thresholding test satisfies several optimality properties for testing whether $\|\mathbf{S}\|_2 \leqslant \delta$ or not when the observation is $\mathbf{Y}$ given by (10). In particular, it is Uniformly Most Powerful Invariant (UMPI) with size α among all of the tests with level α that are invariant by orthogonal transforms. The reader is invited to refer to [21] for further details.

According to these properties, the threshold $\eta_\alpha(\delta/\sigma_X)$ makes it possible to control the false alarm probability via α and guarantees optimal power or correct decision probability without prior knowledge on the signal of interest, an appealing feature for speech enhancement. For homogeneity of the physical quantities in Eq. (8), we choose

$$\lambda = \eta_\alpha(\delta/\sigma_X)/\sqrt{N}. \tag{15}$$

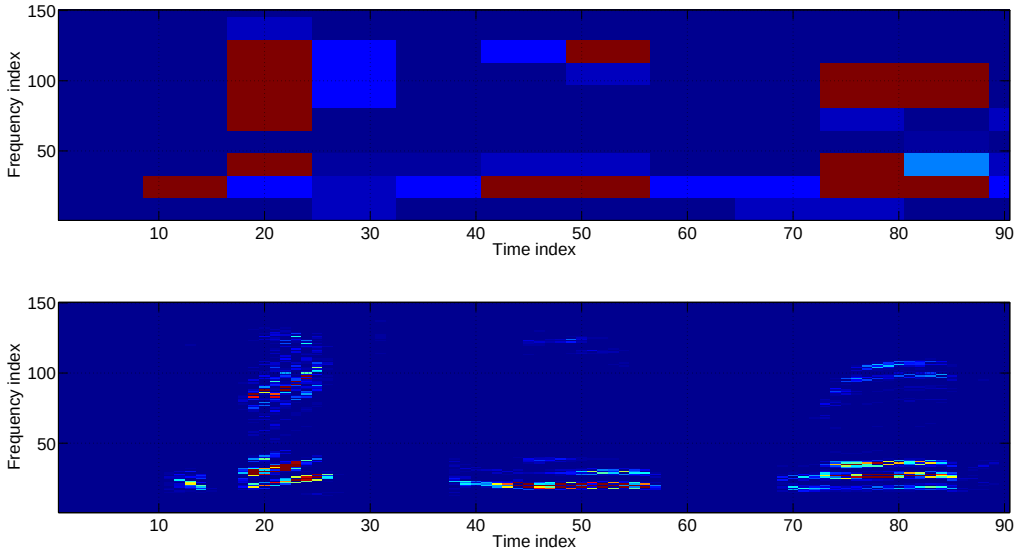


Figure 1: A typical division of the time-frequency domain into boxes and blocks inside boxes shown in the upper figure. This division is obtained by risk minimization for noisy white speech at SNR = 5dB. The time-frequency domain is first divided into non-overlapping rectangular boxes of size $2^3 \times 2^4$. Then, each box is split into blocks whose size is determined by minimizing the overall risk (13) via the SURE approach. We can see that this division matches rather well to the DCT spectrogram displayed in the lower figure.

Choice of τ. The SURE approach is particularly relevant to estimate local parameters. However, the authors' experience with speech and images [33] suggest that τ can be adjusted as a global parameter. Indeed, although τ is a slope that may vary from one signal to another, a global or average value for this parameter is not really detrimental. Some informal tests then led to choose

$$\tau \approx 4/\lambda, \tag{16}$$

as recommended in [11] for images.

To clarify the use of RDT theory in speech denoising and the choice for τ, Fig. 2 shows spectrograms when denoising is performed by SSBS on blocks and two different levels α are tested. The smaller α, the smaller the background noise. However, with $\alpha = 0.05$, some important frequency-time atoms are ignored (for instance, see the rectangle in Fig.2 (c).

3.3. STSA-MMSE in the DCT domain

Similarly to standard Bayesian MMSE-based methods in the DFT domain [34], we compute the MMSE Bayesian estimator of the absolute value of the DCT clean signal coefficients. To this end, we need a model for the clean speech distribution. Motivated by the central limit theorem when the frame length is large enough, we assume that the DCT coefficients of the clean signal have Gaussian prior density. Based on this assumption, the probability of each event $\Phi_S = 1$ or $\Phi_S = -1$ is equal to $1/2$. Thus, the probability density function of the amplitude of a given clean speech DCT coefficient A_S has half-normal distribution:

$$f_{A_S}(a) = \frac{\sqrt{2}}{\sigma_S \sqrt{\pi}} \exp\left(-\frac{a^2}{2\sigma_S^2}\right) \mathbf{1}_{[0,\infty)}(a), \tag{17}$$

where $\mathbf{1}_{[0,\infty)}$ is the indicator function $\mathbf{1}_{[0,\infty)}(x) = 1$ if $x \geq 0$ and $\mathbf{1}_{[0,\infty)}(x) = 0$ otherwise. Moreover, noise is assumed to be Gaussian. Thus, we can write

$$\begin{aligned} f_{Y|A_S}(y|a) =& \mathbb{P}(\Phi_S = 1) f_{Y|A_S,\Phi_S}(y|a, 1) \\ &+ \mathbb{P}(\Phi_S = -1) f_{Y|A_S,\Phi_S}(y|a, -1) \end{aligned} \tag{18}$$

where $f_{Y|A_S}(y|a)$ (resp. $f_{Y|A_S,\Phi_S}(y|a, \phi_S)$) is the conditional probability density function of Y at y given $A_S = a$ (resp. $A_S = a$ and $\Phi_S = \phi_S$). It follows that $f_{Y|A_S}$ can be rewritten as:

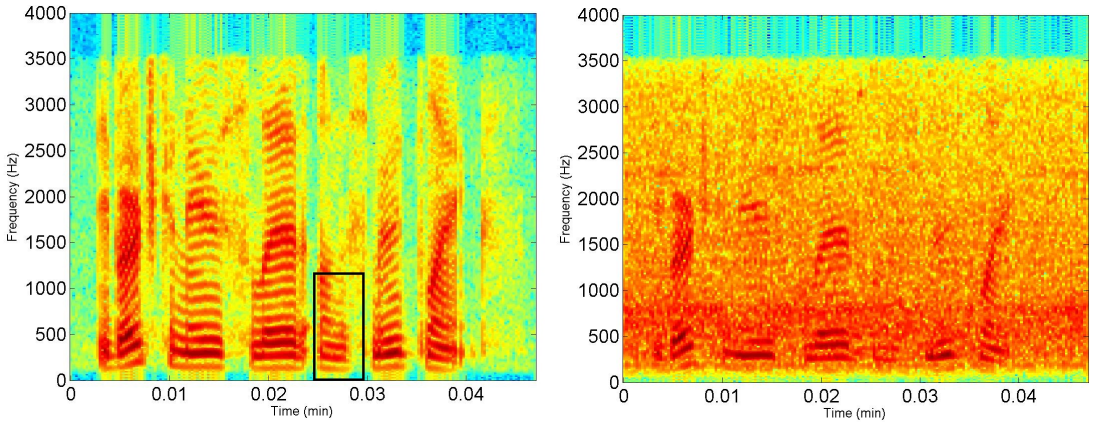
$$\begin{aligned} f_{Y|A_S}(y|a) =& \frac{1}{2\sigma_X \sqrt{2\pi}} \times \\ & \left(\exp\left(-\frac{(y-a)^2}{2\sigma_X^2}\right) + \exp\left(-\frac{(y+a)^2}{2\sigma_X^2}\right) \right). \end{aligned} \tag{19}$$

The Bayesian estimator of the speech short-time spectral amplitude (STSA) is a map ψ of $\mathbb{R}$ into $[0, \infty)$ aimed at minimizing the mean-square error between the estimated and the true amplitude. According to [35] (among others), it is known to be the conditional mean $\mathbf{E}[A_S|Y = y]$ and is given for every $y \in \mathbb{R}$ by :

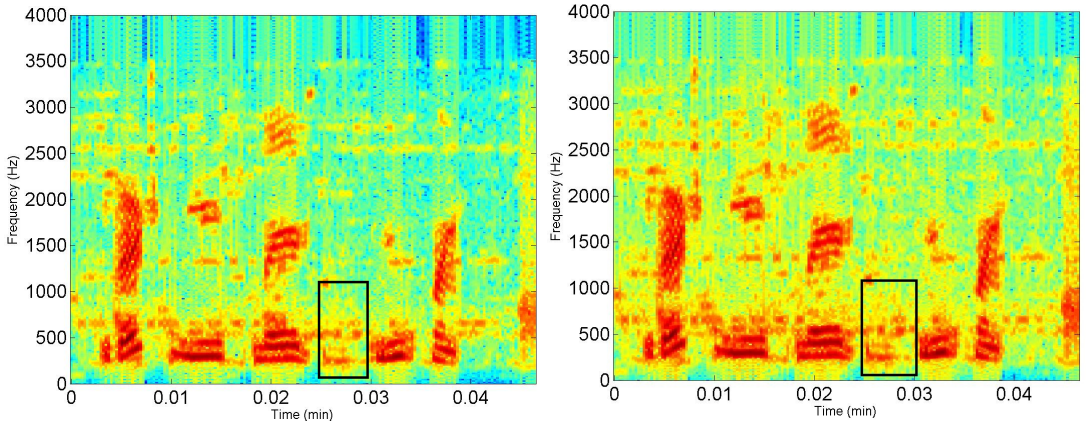
$$\psi(y) = \frac{\displaystyle\int_0^\infty a f_{Y|A_s}(y|a) f_{A_s}(a) da}{\displaystyle\int_0^\infty f_{Y|A_s}(y|a) f_{A_s}(a) da}. \tag{20}$$

Given the DCT coefficient Y, the estimate $\widehat{A_S}$ of A_S is then:

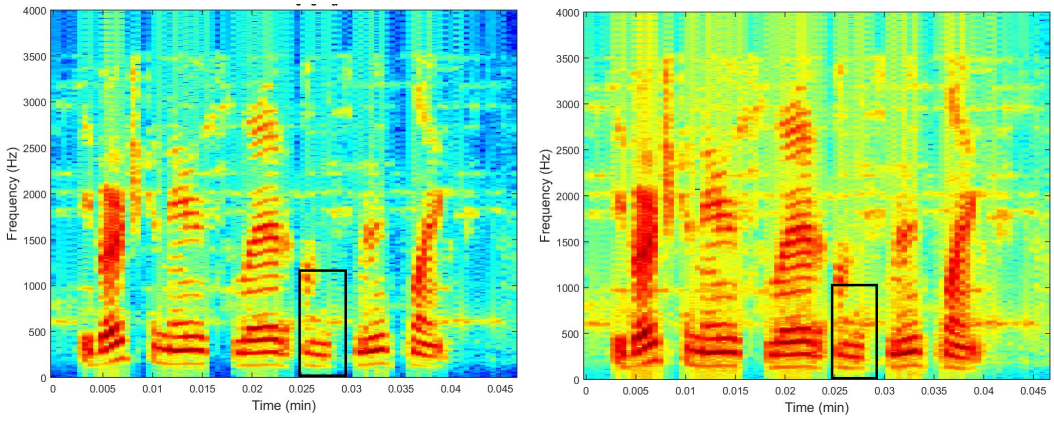
$$\widehat{A_S} = \psi(Y), \tag{21}$$



(a) Clean signal (b) Noisy signal at 5dB



(c) $\alpha = 0.05$ (d) $\alpha = 0.15$



(e) $\alpha = 0.05$ (f) $\alpha = 0.15$

Figure 2: Spectrogram of clean speech (a), corresponding noisy car speech (b), denoised speech by Block-SSBS with two different levels: level = 0.05 (c) and level = 0.15 (d) and denoised speech by BSSBS-MMSE with two different levels: level = 0.05 (e) and level = 0.15 (f).

Injecting (17) and (19) into (20) yields:

$$\psi(y)=\frac{\int_0^\infty a\left[\exp\left(\frac{ay}{\sigma_X^2}-\frac{a^2}{2\sigma^2}\right)+\exp\left(-\frac{ya}{\sigma_X^2}-\frac{a^2}{2\sigma^2}\right)\right]da}{\int_0^\infty\left[\exp\left(\frac{ay}{\sigma_X^2}-\frac{a^2}{2\sigma^2}\right)+\exp\left(-\frac{ya}{\sigma_X^2}-\frac{a^2}{2\sigma^2}\right)\right]da}, \quad (22)$$

where $\sigma = \sigma_S\sigma_X/\sqrt{\sigma_X^2+\sigma_S^2}$. As in [34], we can obtain a closed-form expression of the gain function. By direct computation from (22) or by using [36, Eqs. 3.462.1, 9.254.1, 9.254.2] successively, the map function $\psi(y)$ is given by:

$$\psi(y) = G(\xi,\gamma)|y|, \quad (23)$$

where $G(\xi,\gamma)$ is the gain function of the STSA-MMSE in the DCT domain defined as

$$G(\xi,\gamma) = \frac{\sqrt{\nu}}{\gamma}\frac{\sqrt{2}+\sqrt{\pi\nu}\,\mathrm{erf}(\sqrt{\nu/2})\exp(\nu/2)}{\sqrt{\pi}\exp(\nu/2)}, \quad (24)$$

where

$$\nu = \frac{\xi}{1+\xi}\gamma. \quad (25)$$

and $\mathrm{erf}(\cdot)$ is the error function. This gain function depends on the *a priori* SNR ξ and the *a posteriori* SNR γ. The *a posteriori* SNR is directly given by the observed amplitude A_Y. In contrast, the *a priori* SNR is unknown. This variable ξ can be estimated via the decision directed approach [34]:

$$\xi[m,k] = \beta\frac{\widehat{A}_S^2[m-1,k]}{\sigma_X^2[m-1,k]} + (1-\beta)(\gamma[m,k]-1)_+, \quad (26)$$

where $0<\beta<1$ is the smoothing parameter, $\widehat{A}_S[m-1,k]$ is the estimated STSA at the previous frame and $(x)_+$ is x if $x\geqslant 0$ and 0 otherwise.

For comparative purpose, Fig. 3 (a) displays both the STSA-MMSE in the DCT domain (24) and the STSA-MMSE in the DFT domain [34] as functions of the *a posteriori* SNR γ for fixed values of $\xi = 5, -5, -10$ dB. Alternatively, in Fig 3 (b), these same gain functions are plotted as functions of ξ for fixed values of $\gamma = 5, -5, -10$ dB. In the two cases, the gain function of the STSA estimator in the DCT domain is shifted down by 2 dB with respect to the gain function of the STSA estimator in the DFT domain. This suggests that denoising in the DCT domain tends to reduce more the background noise.

3.4. Combination method

After Block-SSBS, the transformed signal and noise are assumed to be Gaussian distributed. We then apply the Bayesian STSA-MMSE in the DCT domain established in the preceding section. By doing so, prior knowledge on speech is incorporated to improve speech quality beyond speech intelligibility improvement achieved by Block-SSBS. Whence the following combination of these parametric and non-parametric methods, which is summarized by Fig 4.

(i) **Signal decomposition**: The observed signal is segmented and transformed using DCT.

(ii) **Noise reduction**: The transformed coefficients are shrunk by the block SSBS gain function $G^B_{\tau,\lambda}(\widehat{\gamma})$ in each block B. Given a DCT coefficient Y in this block, the estimate $\widehat{A}_S^*$ of the amplitude A_S of the clean signal is calculated by:

$$\widehat{A}_S^* = G^B_{\tau,\lambda}(\widehat{\gamma})A_Y \quad (27)$$

(iii) **Refined Estimation**: The Bayesian MMSE statistical estimator is applied to the coefficients shrunk by Block-SSBS so that the final estimate of the clean signal amplitude is:

$$\widehat{A}_S = G(\xi,|\widehat{A}_S^*|^2/\sigma_X^2)\widehat{A}_S^*, \quad (28)$$

where G is the gain function of the STSA-MMSE Bayes estimator given by (24) and ξ is calculated by the decision-directed approach (26).

(iv) **Signal reconstruction**: The enhanced signal is finally obtained from the estimated STSA $\widehat{A}_S$ and the noisy phase ϕ_Y by the overlap-add method [16].

Figs. 2 (e) & (f) illustrate the gain brought by the combination. More precisely, the reader will notice that components erased by Block-SSBS in the rectangles enhanced in Figs. 2 (c) & (d) have been recovered in Figs. 2 (e) & (f).

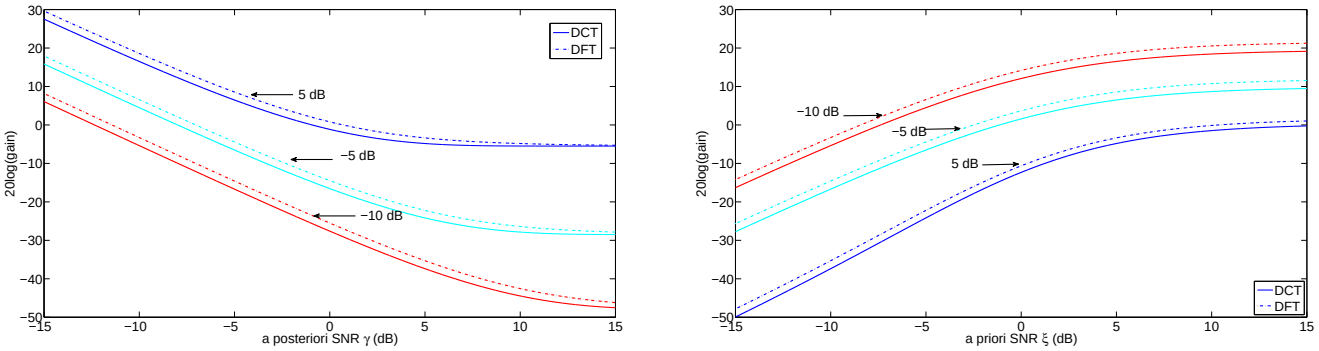
4. Experimental Results

4.1. Experimental setting and parameter adjustment

Experiments have been conducted on the NOIZEUS database to evaluate the performance of the proposed methods for speech enhancement. The NOIZEUS database contains speech sentences degraded by noise environments from the AURORA database at various levels, namely 0, 5, 10 and 15 dB. The speech signals are sampled at 8 kHz. In our experiments, the noisy signals were Hamming-windowed into 32-ms frames with 50% overlap, and then transformed by DCT or DFT. The methods under test are:

- STSA-MMSE(DCT): STSA-MMSE in the DCT domain,
- STSA-MMSE(DFT): STSA-MMSE in the DFT domain,
- Block-SSBS
- BSSBS-MMSE: the combination of Block-SSBS and STSA-MMSE(DCT)

With respect to the theorical framework, the purpose of the experiments is threefold. first, we want to assess the non-parametric method Block-SSBS in the DCT domain to its Bayesian — and thus parametric — counterpart STSA-MMSE(DCT) calculated in the Section 3.3. Second, we aim to evaluate the gain brought by the combination BSSBS-MMSE in the DCT domain, in comparison to the Bayesian reference STSA-MMSE(DCT). Third, we want to assess the interest of working in the DCT domain and therfore, benchmark Block-SSBS and BSSBS-MMSE to the standard baseline STSA-MMSE(DFT) [34].



(a) Fixed ξ (b) Fixed γ

Figure 3: Gain functions of the STSA-MMSE estimators in the DCT and DFT domains as functions of ξ and γ. In Fig. 3 (a) the gain functions vary with γ at fixed values of ξ whereas, in Fig. 3 (b), the gain functions vary with ξ at fixed values of γ.

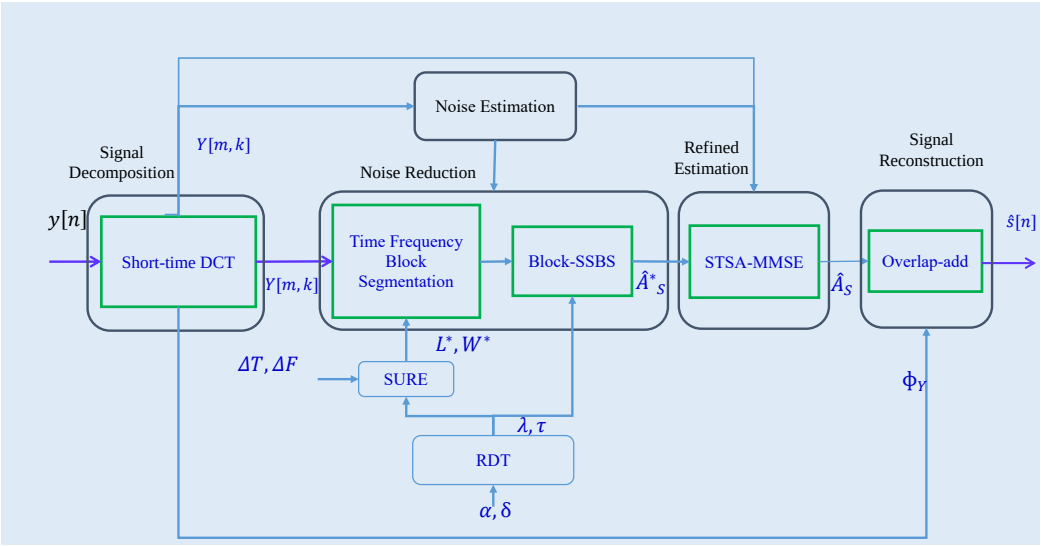


Figure 4: Block overview of the combination method, where $y[n]$ is the input and ΔT, ΔF, δ and α are the parameters of the proposed combination method.

For Block-SSBS, the tolerance δ and the level α were chosen by maximizing the segmental SNR (SSNR) on a small set of sentences randomly chosen and corrupted by car noise at SNR level of 5 dB. These preliminary tests led to choose $\alpha = 0.05$ and $\delta = 4$ dB for Block-SSBS.

The performance of all methods were evaluated in two scenarios. In the first one, denoising is performed by using the reference noise power spectrum. This one is simply the theoretical power spectrum if noise is stationary. Otherwise, the reference noise power spectrum in a given bin m is estimated as in [37] by:

$$\sigma_X^2[m,k] = \mu\sigma_X^2[m-1,k] + (1-\mu)A_X^2[m,k], \qquad (29)$$

where $\mu = 0.9$ and $\sigma_X^2[0,k] = A_X^2[0,k]$. In the second scenario, the noise power spectrum was estimated for all methods using the B-E-DATE algorithm recently introduced in [38].

4.2. Speech objective Test

Speech quality and intelligibility were evaluated via objective quality and intelligibility criteria. Speech quality was assessed using the standard segmental SNR (SSNR) and the overall quality of speech criteria. SSNR values were trimmed so as to remain within the range [−10,35 dB] and avoid the use of a silence/speech detector [16]. The overall speech quality was measured by the multivariate adaptive regression spline (MARS_ovl) criterion. This metric combines the Itakura-Saito distance (IS) and the perceptual evaluation of speech quality (PESQ) [39]. It has been shown to strongly correlate with subjective assessments [39].

Speech intelligibility was first estimated by the short-time objective intelligibility (STOI) criterion. Basically, the STOI criterion computes the mean correlation between clean and estimated speech [40]. It is known to be highly correlated with intelligibility scores obtained by listening tests. We applied the

logistic function [40, Eq. (8)] to map the STOI measure to a meaningful intelligibility score.

The results are displayed in Figures 5 to7. Each figure has the same legend where STSA-MMSE(DFT), STSA-MMSE(DCT), Block-SSBS, and BSSBS-MMSE are designated by the red, green, black and blue lines with the circle, x-mark, plus and star makers, respectively, as specified once for all in Fig. 5a. Moreover, all measures obtained with the reference noise power spectrum and with B-E-DATE are drawn by dashed and solid lines, correspondingly. All algorithms have been benchmarked at four SNR levels and against various noise models, namely white Gaussian noise (White), 2nd-order auto-regressive (AR) noise, 4 usual types of quasi-stationary noise (car, train, station and street) and 4 kinds of non-stationary noise (airport, exhibition, restaurant and babble). AR noise was obtained by filtering white Gaussian noise by the discrete filter with transfer function $1/(1 + az^{-1})$ and $a = 0.5$.

Fig. 5 shows the segmental SNR improvement obtained with the different denoising methods employing the reference noise power spectrum (dashed lines) as well as the noise power spectrum estimated by B-E-DATE (solid lines). We first consider the scenario where the reference noise power spectrum is used. The results for white and AR noise are given in Fig. 5a and 5b, respectively. The proposed BSSBS-MMSE method yields the highest segmental SNR improvement at 0, 5 and 10 dB, whereas the non-parametric Block-SSBS method achieves the best SSNR at 15dB. For non-stationary environment with slowly-varying noise spectrum like car, train, station and street noises, similar results are obtained: BSSBS-MMSE provides the largest SSNR improvement at low and medium SNRs, but Block-SSBS performs better than BSSBS-MMSE at 15 dB. Yet the difference is small, as shown by Figures 5c to 5f. Figures 5g to 5j present SSNR improvement for non-stationary noises. In this case, BSSBS-MMSE yields the best score at low and medium SNRs. At high SNR level, Block-SSBS and BSSBS-MMSE both lead to the same best measure. Remarkably, in comparison to STSA-MMSE in the DFT domain, the BSSBS-MMSE method has a gain of around $2.5 - 3$dB in this case.

The SSNR improvement, obtained in the more realistic case where the noise power spectrum is estimated by B-E-DATE, is also shown in Fig. 5 by solid lines. In this case, the BSSBS-MMSE method still yields the best score for all noise types from 0 dB to 10 dB, whereas Block-SSBS achieves the highest score at 15 dB. The gain now is about $0.5 - 1$ dB. Such results basically relate to the sensitivity of STSA-MMSE and Block-SSBS in the DCT domain to noise estimation errors. In comparison with Fig. 5, STSA-MMSE(DCT), Block-SSBS and BSSBS-MMSE undergo performance loss by using B-E-DATE for noise power spectrum estimation. This loss is negligible for white and AR Gaussian noise and around 3 dB for other types of noise. Generally, although BSSBS-MMSE is sensitive to noise estimation errors, it keeps on yielding the best SSNR improvement.

In term of speech quality estimated by the MARS overall criterion, Fig. 6 (dashed lines) shows the score improvement when the reference noise power spectrum is used. With small *a priori* information about speech, the Block-SSBS method yields

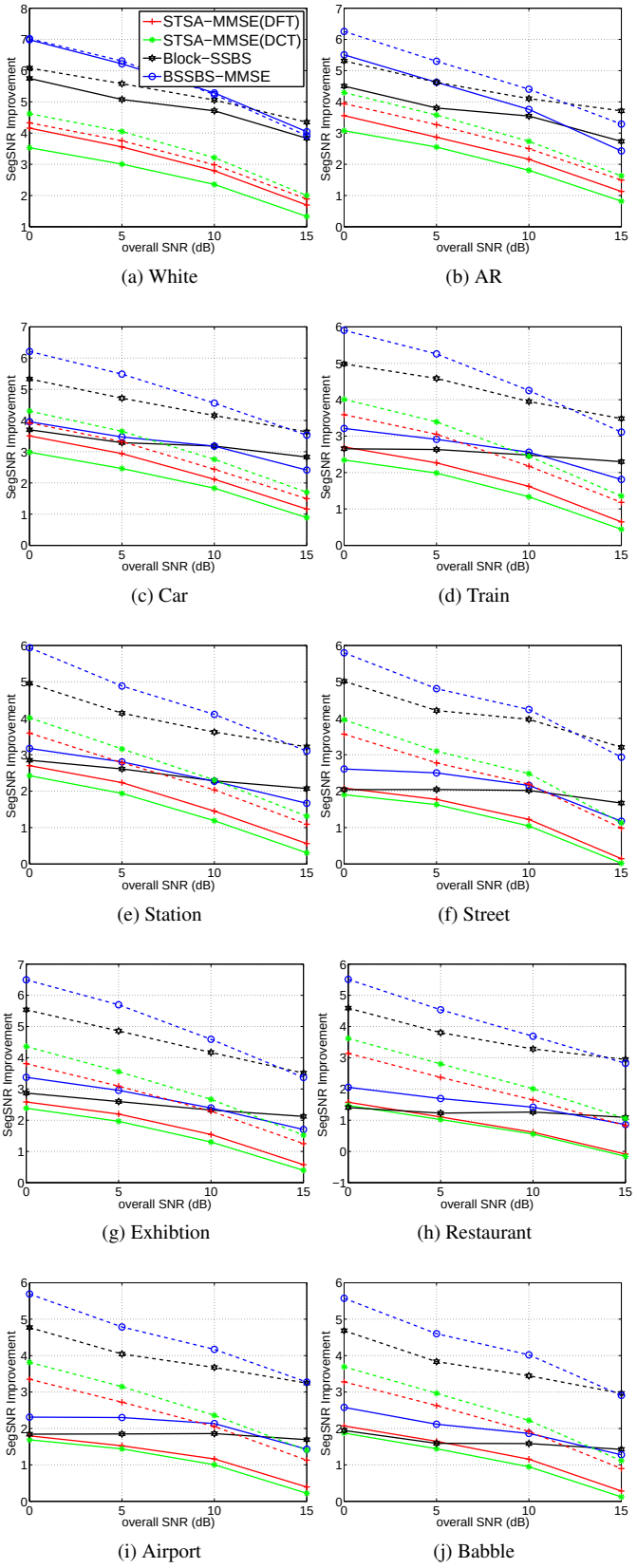
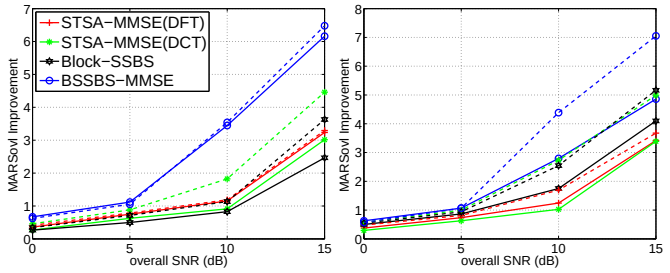
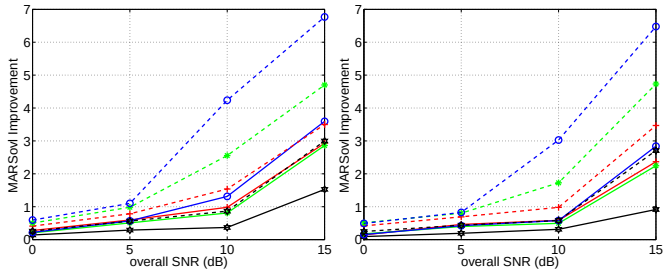


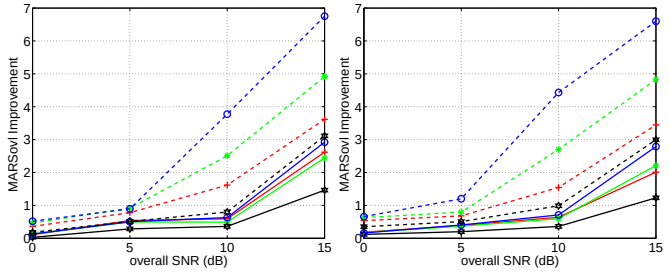
Figure 5: Speech quality evaluation after speech denoising: improvement of segmental SNR criterion. The result is displayed first for synthetic noise (White,AR) then quasi-stationary noise (train, car, station and street) and finally non-stationary noise (restaurant, exhibition, babble and airport). The legend is the same for all sub-figure like Fig. 5a.



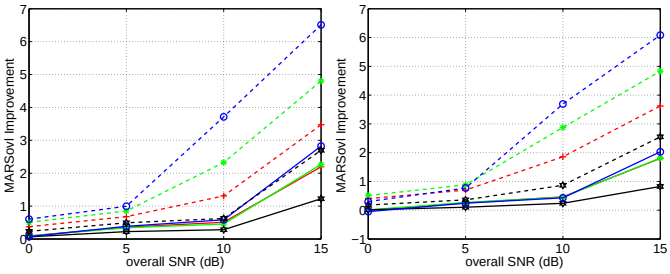
(a) White (b) AR



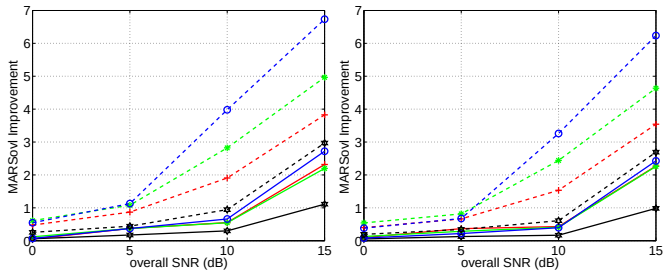
(c) Car (d) Train



(e) Station (f) Street

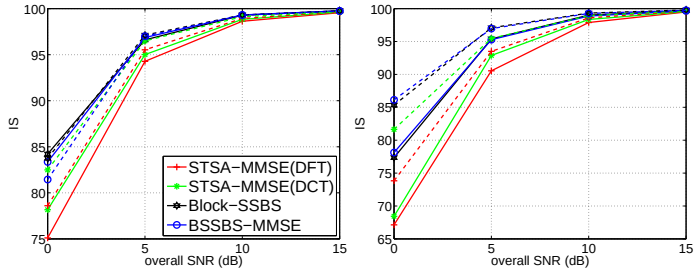


(g) Exhibition (h) Restaurant

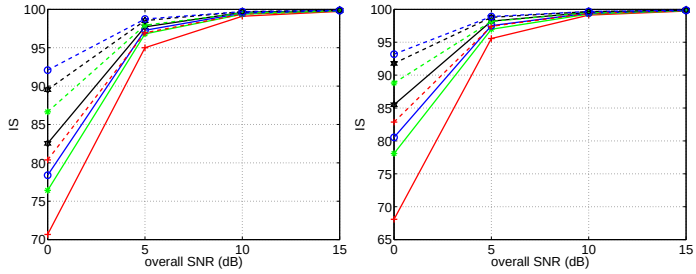


(i) Airport (j) Babble

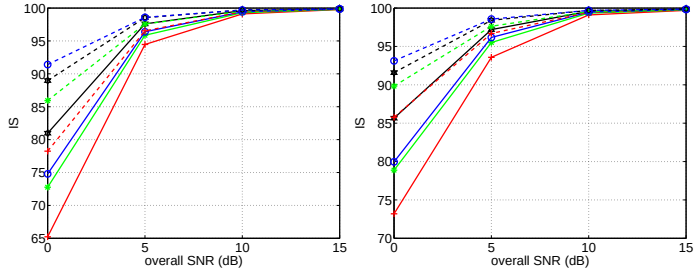
Figure 6: Speech quality evaluation after speech denoising: improvement of MARS OVL composite criterion. The legend is also the same for all sub-figure like Fig. 6a.



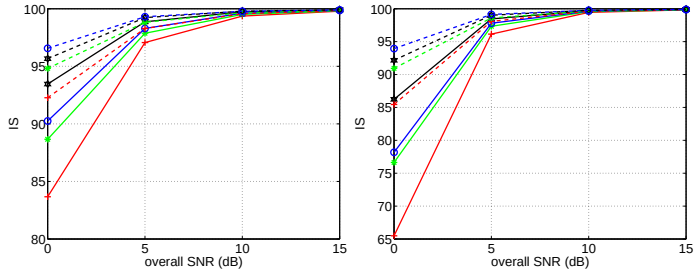
(a) White (b) AR



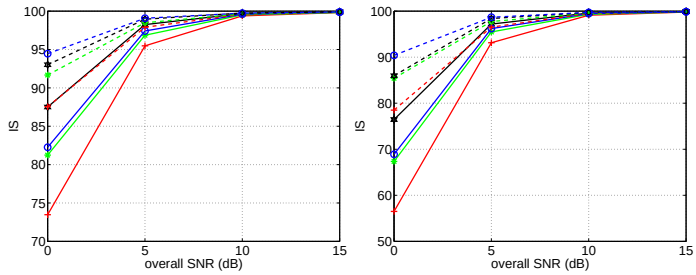
(c) Car (d) Train



(e) Station (f) Street



(g) Exhibition (h) Restaurant



(i) Airport (j) Babble

Figure 7: Speech intelligibility evaluation after speech denoising: Intelligibility score obtained by mapping STOI criterion.

the lowest score in all situations except for AR noise. This remains true when B-E-DATE is employed to estimate the noise power spectrum (see the solid lines in Fig. 6). By taking into account the speech distribution at the refined estimation step, the good performance of the BSSBS-MMSE method is confirmed by the MARS improvement measured in the case of white Gaussian noise and AR noise (see Figs. 6a and 6b). For all types and levels of noises, BSSBS-MMSE provides the best MARS scores, except for babble and restaurant noises at low noise levels. However, in these cases, the MARS scores of the BSSBS-MMSE method is not significantly different from the best ones obtained by STSA-MMSE(DCT) (see Fig. 6c-6j).

When combining denoising with B-E-DATE noise estimation, the MARS overall improvement is presented in Fig. 6 by solid lines. It turns out that the speech quality obtained by STSA-MMSE(DFT) is not really affected by errors in the noise spectrum estimation (compare the dashed lines to the solid ones in Fig. 6). In contrast, for non-stationary noise, methods performing in the DCT domain are more sensitive to noise estimation errors, as mentioned earlier. Thereby, BSSBS-MMSE, STSA-MMSE(DCT) and STSA-MMSE(DFT) yield very similar results for this type of noise, especially at low and medium SNR levels (see Figs. 6d-6j). For stationary noise, Figs. 6a and 6b show that the BSSBS-MMSE method remains better than the other methods, without real performance loss due to noise spectrum estimation by B-E-DATE.

In terms of speech intelligibility, the intelligibility score (IS) obtained by mapping the STOI measure is shown in Figs. 7. At high SNR, the scores obtained by all the methods are not significantly different. At low SNR and in presence of AR and white Gaussian noise, Block-SSBS and BSSSBS-MMSE behave similarly in the two scenarios (with and without reference noise power spectrum). For non-stationary noises and when using the reference noise spectrum, BSSBS-MMSE yields the highest scores. In comparison with the worst results, the gain is around $10-15\%$. When noise spectrum is estimated by B-E-DATE, the best performance is attained by Block-SSBS. In comparison to the STSA-MMSE(DFT) method, BSSBS-MMSE provides often better score in the case of non-stationary noise with a gain between 5 and 15%, whereas Block-SSBS method leads to an improvement of $10-20\%$ (see Figs. 7h, 7i and 7j).

In a nutshell, the proposed BSSBS-MMSE combination achieves a better overall trade-off between speech quality and intelligibility than the other methods under consideration.

5. Conclusion

In this paper, we have introduced several speech denoising methods. All have in common to operate in the DCT domain, which makes it possible in particular to get rid of the phase estimation problem. These methods are Block-SSBS, STSA-MMSE(DCT) and BSSBS-MMSE. Block-SSBS is non-parametric and can be seen as a smooth shrinkage of DCT coefficients. Its parameters are optimized by the SURE and RDT approaches, which are also non-parametric methods for statistical inference. STSA-MMSE(DCT) is a Bayesian estimator. BSSBS-MMSE combines Block-SSBS and STSA-MMSE(DCT) so as to benefit from the advantages of each of these methods. Namely, Block-SSBS achieves good performance in terms of speech intelligibility by background noise reduction; STSA-MMSE improves speech quality by enhancing speech contained in small coefficients after shrinkage.

The performance evaluation was conducted on the NOIZEUS database, with and without noise power spectrum reference. Various types of stationary and non-stationary noises were considered. When the noise spectrum is unknown, it is estimated by an up-to-date method. In addition, objective and subjective tests were used to assess the speech estimators, in comparison to the reference approach STSA-MMSE [34]. The experimental results show that BSSBS-MMSE performs better than the other methods in most situations. These experiments also confirm the relevance of working in the DCT domain. In this respect, transposition to the DCT domain of Bayesian frameworks such as that considered in [31, 32] could be investigated.

As specified in the introduction, the approach proposed in this paper concerns applications where sufficiently large databases are not available for designing deep and recurrent neural networks. However, in the case where large databases are available for more or less specific applications, it would be relevant to benchmark deep and recurrent neural networks to Block-SSBS and BSSBS-MMSE in terms of development costs, complexity, robustness and denoising performance in various contexts.

The approach proposed in this paper could apply to transforms other than DCT. For instance, although the phase spectrum estimation remains an issue for STFT-based transforms, combining Block-SSBS and BSSBS-MMSE with Mel or Bark transforms could be addressed. For instance, Bayesian speech estimation in the Mel domain as proposed in [41] yields good performance for Automatic Speech Recognition (ASR). It can thus be wondered whether Block-SSBS and BSSBS-MMSE can be used in such domains for speech enhancement in audio applications. In the same way, denoising by Block-SSBS and BSSBS-MMSE after cochlear transforms [42] could also be considered.

The STSA-MMSE used in BSSBS-MMSE is devised under the Gaussian assumption for the Block-SSBS outcome. Asymptotic statistics could perhaps help justify this assumption. However, the task seems rather difficult and, in any case, the experimental results provide evidence that such a Gaussian assumption leads to an STSA-MMSE estimator good enough to retrieve relevant speech contents, even in small DCT coefficients as emphasized by Figs. 2 (e) & (f). However, extensions to super-Gaussian [43] or Gamma distributions [44] in the DCT domain could be envisaged. B-E-DATE could still be used since it performs noise estimation regardless of the signal and its distribution, even in presence of non-stationary and non-Gaussian noise [38].

All these results have been obtained for speech signals. However, the underlying theoretical framework is based on very general assumptions. Therefore, it can be wondered whether the proposed methods could not be used to denoise other types of signals as well. On-going work could involve a study ded-

icated to audio denoising solely via the methods introduced in this paper. Moreover, in comparison with the performance measurements obtained when the noise spectrum reference is given, all the methods tested in the DCT domain undergo a loss in speech quality when the noise power spectrum is estimated by B-E-DATE. Thus, the design of efficient noise power spectrum estimation in the DCT domain must be further investigated.

References

References

[1] J. Xu, Y.and Du, L. R. Dai, C. H. Lee, An experimental study on speech enhancement based on deep neural networks, IEEE Signal Process. Lett. 21 (1) (2014) 65–68.

[2] T. T. Vu, B. Bigot, E. S. Chng, Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition, in: Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process., IEEE, 2016, pp. 499–503.

[3] S. Mavaddaty, S. M. Ahadi, S. Seyedin, Modified coherence-based dictionary learning method for speech enhancement, IET Signal Process. 9 (7) (2015) 537–545.

[4] I. Andrianakis, P. R. White, Speech spectral amplitude estimators using optimally shaped gamma and chi priors, Speech Communication 51 (1) (2009) 1–14.

[5] M. Krawczyk-Becker, T. Gerkmann, On mmse-based estimation of amplitude and complex speech spectral coefficients under phase-uncertainty, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (12) (2016) 2251–2262. doi:10.1109/TASLP.2016.2602549.

[6] B. Chen, P. C. Loizou, A laplacian-based MMSE estimator for speech enhancement, Speech commun. 49 (2) (2007) 134–143.

[7] P. Mowlaee, J. Stahl, J. Kulmer, Iterative joint map single-channel speech enhancement given non-uniform phase prior, Speech Communication 86 (2017) 85–96.

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

[9] D. L. Donoho, De-noising by soft-thresholding, IEEE Trans. Inf. Theory 41 (3) (1995) 613–627.

[10] A. M. Atto, D. Pastor, G. Mercier, Detection threshold for non-parametric estimation, Signal, Image and Video processing 2 (3) (2008) 207–223.

[11] A. M. Atto, D. Pastor, G. Mercier, Smooth sigmoid wavelet shrinkage for non-parametric estimation., in: Proc. IEEE Int. Conf. Acoust. Speech, Signal Process., 2008, pp. 3265–3268.

[12] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, M. Muma, Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts, IEEE Signal Processing Magazine 29 (4) (2012) 61–80.

[13] B. P. Rao, Nonparametric functional estimation, Academic press, 2014.

[14] P. C. Loizou, G. Kim, Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions, IEEE Trans. Audio, Speech, Lang. Process. 19 (1) (2011) 47–56.

[15] Y. Hu, P. C. Loizou, Evaluation of objective measures for speech enhancement., in: Proc. Interspeech, 2006, pp. 1447–1450.

[16] P. C. Loizou, Speech enhancement: theory and practice, CRC press, 2013.

[17] K. P., K. W., B. S., The importance of phase in speech enhancement, speech communication 53 (4) (2011) 465–494.

[18] Y. Soon, S. N. Koh, C. K. Yeo, Noisy speech enhancement using discrete cosine transform, Speech communication 24 (3) (1998) 249–257.

[19] S. Gazor, W. Zhang, Speech enhancement employing laplacian-gaussian mixture, IEEE Trans. Speech, Audio, Process. 13 (5) (2005) 896–904.

[20] C. M. Stein, Estimation of the mean of a multivariate normal distribution, The annals of Statistics (1981) 1135–1151.

[21] D. Pastor, Q. T. Nguyen, Random distortion testing and optimality of thresholding tests, IEEE Trans. Signal Process. 61 (16) (2013) 4161–4171.

[22] K. R. Rao, P. Yip, Discrete cosine transform: algorithms, advantages, applications, Academic press, 2014.

[23] N. Ahmed, T. Natarajan, K. R. Rao, Discrete cosine transfom, IEEE trans., Comput. (1) (1974) 90–93.

[24] M. Elad, Sparse and redundant representations, Springer, 2010.

[25] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, IEEE Trans. Acoust., Speech, Signal, Process. ASSP-33 (2) (Apr. 1985) 443–445.

[26] R. Tavares, R. Coelho, Speech enhancement with nonstationary acoustic noise detection in time domain, IEEE Signal Process. Lett. 23 (1) (2016) 6–10.

[27] T. T. Cai, Adaptive wavelet estimation: a block thresholding and oracle inequality approach, Annals of statistics (1999) 898–924.

[28] G. Yu, S. Mallat, E. Bacry, Audio denoising by time-frequency block thresholding, IEEE Trans. Signal Process. 56 (5) (2008) 1830–1839.

[29] S. Kamath, P. C. Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, in: Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP), Vol. 4, IEEE, 2002, pp. IV–4164.

[30] R. Tibshirani, L. Wasserman, Stein's unbiased risk estimate, Course notes from "Statistical Machine Learning, Spring 2015" (2015) 1–12.

[31] A. Abramson, I. Cohen, Simultaneous detection and estimation approach for speech enhancement, IEEE Trans. Audio, Speech, Lang. Process. 15 (8) (2007) 2348–2359.

[32] H. Momeni, H. R. Abutalebi, A. Tadaion, Joint detection and estimation of speech spectral amplitude using noncontinuous gain functions, IEEE/ACM Trans. Audio, Speech, Lang. Process. 23 (8) (2015) 1249–1258.

[33] D. Pastor, A. M. Atto, Wavelet shrinkage: from sparsity and robust testing to smooth adaptation; In Fractals and Related Fields, Eds: J. Barral & S. Seuret, Birkhaüser, 2010.

[34] Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, IEEE Trans. Acoust. Speech, Signal Process. 32 (6) (1984) 1109–1121.

[35] S. M. Kay, Fundamentals of statistical signal processing, volume i: estimation theory.

[36] A. Jeffrey, D. Zwillinger, Table of integrals, series, and products, Academic Press, 2007.

[37] R. C. Hendriks, J. Jensen, R. Heusdens, Noise tracking using DFT domain subspace decompositions, IEEE Trans. Audio, Speech, Lang. Process. 16 (3) (Mar. 2008) 541–553.

[38] V. K. Mai, D. Pastor, A. Aïssa-El-Bey, R. Le-Bidan, Robust estimation of non-stationary noise power spectrum for speech enhancement, IEEE/ACM Trans. Audio, Speech, Lang. Process. 23 (4) (2015) 670–682.

[39] Y. Hu, P. C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Trans. Audio, Speech, Lang. Process. 16 (1) (2008) 229–238.

[40] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time–frequency weighted noisy speech, IEEE Trans. Audio, Speech, Lang. Process. 19 (7) (2011) 2125–2136.

[41] C. W. Han, S. J. Kang, N. S. Kim, Reverberation and noise robust feature compensation based on imm, IEEE Transactions on Audio, Speech, and Language Processing 21 (8) (2013) 1598–1611.

[42] C.-T. Do, D. Pastor, A. Goalic, A novel framework for noise robust asr using cochlear implant-like spectrally reduced speech, Speech Communication 54 (1) (2012) 119 – 133.

[43] T. Lotter, P. Vary, Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model, EURASIP J. applied, signal, process. 2005 (2005) 1110–1126.

[44] R. Martin, Speech enhancement based on minimum mean-square error estimation and supergaussian priors, IEEE Trans. Speech, Audio, Process. 13 (5) (2005) 845–856.