Joint Soft Threshold and Statistical Estimation for Speech Enhancement

Van Khanh Mai, Dominique Pastor, Abdeldjalil Aissa El Bey

Joint Soft Threshold and Statistical Estimation for Speech Enhancement

Van Khanh MAI, Dominique PASTOR and Abdeldjalil AISSA-EL-BEY
IMT Atlantique, UMR CNRS 6285 Lab-STICC, UBL, F-29238 Brest, France
Email: firstname.lastname@imt-atlantique.fr

***Abstract*—This paper presents a novel method for speech enhancement based on the combination of sigmoid shrinkage and bayesian estimator. The main idea is to apply a joint detection and estimation to noisy speech before using a standard minimum-mean-squared-error (MMSE) estimator. Hence, the proposed method can take advantage of two basic approaches for improving the quality of noisy speech. Experiments performed on stationary and non-stationary noisy speech signals show that the proposed approach is promising when compared to classical methods, in terms of objective and pseudo-subjective measurements.**

***Index Terms*—Sigmoid shrinkage, minimum mean squared error (MMSE), speech enhancement, noise reduction.**

I. INTRODUCTION

Single channel speech enhancement algorithms based on time-frequency transforms are often used to reduce background noise of noisy speech signals. These techniques aim at improving not only the quality but also the intelligibility of speech. By assuming a statistical model and using various cost functions, many estimators have been proposed in [1]. Among these traditional algorithms with continuous gain functions, minimum mean square error (MMSE) of short-time spectral amplitude (STSA) [2] or log-spectral amplitude (LSA) [3], together with modifications, produce significantly high performance in terms of speech quality [4]. In usual approaches, speech presence is assumed in every time-frequency bin, which may entail some performance loss. Therefore, some studies try to join detection and estimation for improving speech quality [5]–[8]. However most algorithms do not improve speech intelligibility [9]. A different approach based on the so-called binary masking makes it possible to overcome this drawback [10], [11]. In the binary masking approach, some speech spectrum bins are retained while other ones are discarded. The binary mask can be refined by combining its computation with a speech signal estimator [9], [12]. Nevertheless, such a technique based on hard binary masking may generate musical noise, which degrades speech quality.

In this paper, we propose a new approach that can take advantage of binary masking and usual estimation. Our strategy joints the smoothed binary mask provided by smoothed sigmoid based shrinkage (SSBS), initially proposed for image processing [13], and bayesian estimation. According to objective criteria, this combination reduces musical noise and improves speech intelligibility.

The remainder of this paper is organized as follows. Section II presents a basic overview of smoothed sigmoid shrinkage and traditional MMSE estimator. Section III details the proposed algorithm. The experimental results are presented in Section IV. Finally, Section V concludes this paper and discusses prospects.

II. BACKGROUND

One of the most important tasks in signal processing is the removal of additive noise from an observed signal $y = s + x$, where s, x are the clean signal and noise, respectively. In speech enhancement applications, the noisy speech signal is segmented into a set of short frames and is transformed via short-time Fourier transform (STFT). The contaminated signal in the time-frequency domain becomes

$$Y[m,k] = S[m,k] + X[m,k], \tag{1}$$

where m and k denote the frame and frequency indices, respectively. To ease notation, the indices m and k will be omitted unless required for clarification. In the same way, the estimated signals are denoted by using the wide hat symbol: *e.g* $\widehat{\xi}$ is an estimate of ξ. The noisy signal can also be written in polar form as follows:

$$Re^{i\phi_Y} = Ae^{i\phi_S} + Ne^{i\phi_X} \tag{2}$$

where $\{R, A, N\}$ and $\{\phi_Y, \phi_S, \phi_X\}$ are the short-time spectral amplitudes (STSA) and the associated phases of the STFT coefficients of the observed signal, clean signal and noise, respectively. Due to the importance of STSA, many researches aim to estimate it. In order to retrieve the clean signal, a gain function $\mathbb{G}$ is often determined. The enhanced STSA signal is then calculated as

$$\widehat{A} = \mathbb{G}R. \tag{3}$$

The estimated speech signal is then obtained by keeping the phase of the observation so that $\widehat{S} = \widehat{A}e^{i\phi_Y}$.

A. Standard bayesian STSA

For the sake of self-completeness, this subsection presents the standard Bayesian STSA [2] and its modification, that is, the log-spectral amplitude (LSA) [3]. Assuming uncorrelation

978-1-5386-8173-2/18/$31.00 ©2018 IEEE

between STFT coefficients, the STSA is obtained by minimizing the expectation of the error defined by a cost function $C(A, \widehat{A})$ [14]. The Bayes risk $\mathbf{R}$ is then defined by

$$\mathbf{R}(\widehat{A}) = \mathbf{E}[C(A, \widehat{A})], \tag{4}$$

where $\mathbf{E}$ denotes mathematical expectation. By defining various cost functions and minimizing the Bayes risk $\mathbf{R}$ with respect to $\widehat{A}$, many estimators can be proposed.

The most usual STSA estimator is derived by minimizing the Bayes risk of the cost function [2]

$$C(A, \widehat{A}) = (A - \widehat{A})^2. \tag{5}$$

Assuming that the STFT coefficients have complex Gaussian distribution with zero-mean, the Bayesian estimator is given by

$$\widehat{A} = \mathbf{E}[A|Y] = \mathbb{G}_{\text{STSA}}(\xi, \gamma)R. \tag{6}$$

This gain function $\mathbb{G}_{\text{STSA}}(\xi, \gamma)$ is given by [2]:

$$\mathbb{G}_{\text{STSA}}(\xi,\gamma)=\frac{\sqrt{\pi}\nu}{2\gamma}\exp\left(-\frac{\nu}{2}\right)\left[(1+\nu)I_0\left(\frac{\nu}{2}\right)+\nu I_1\left(\frac{\nu}{2}\right)\right], \tag{7}$$

where $I_0(\cdot)$, $I_1(\cdot)$ denote the modified Bessel function of zero and first order and

$$\nu = \frac{\xi\gamma}{1+\xi},$$

with *a priori* signal to noise ratio (SNR) $\xi = \mathbf{E}[A^2]/\mathbf{E}[N^2] = \sigma_S^2/\sigma_X^2$ and the *a posteriori* SNR $\gamma = R^2/\sigma_X^2$, where σ_S^2 and σ_X^2 denote the spectral speech and noise power, respectively. In the same way, the LSA estimator is obtained by using the cost function [3]

$$C(A, \widehat{A}) = \left(\log(A) - \log(\widehat{A})\right)^2. \tag{8}$$

Using the moment-generating of the STSA A, it can be proved that the Bayesian estimator is now $\widehat{A} = e^{\mathbf{E}[\log A|Y]} = \mathbb{G}_{\text{LSA}}(\xi, \gamma)R$ where the gain function $\mathbb{G}_{\text{LSA}}(\xi, \gamma)$ is still a function of the two variables ξ and γ and given by [3]:

$$\mathbb{G}_{\text{LSA}}(\xi, \gamma) = \frac{\xi}{1+\xi}\exp\left(\frac{1}{2}\int_{\nu}^{\infty}\frac{e^{-t}}{t}dt\right). \tag{9}$$

Note that, the unknown SNR ξ in both methods is evaluated by the decision-directed approach [2].

STSA estimators are also generalized in [15], where the cost function is defined as the square error of the β power amplitude $C(A, \widehat{A}) = \left(A^\beta - \widehat{A}^\beta\right)^2$. The cost functions are also proposed by following and incorporating perceptual measures in [14], [16]. In other approaches as in [1], the STSA is not assumed to be Rayleigh distributed but to follow the super-Gaussian or generalized Gamma distributions.

B. Thresholding estimation: Sigmoid shrinkage

Shrinkage functions are frequently used in image processing for estimating signal coefficients provided by the projection of the noisy signal on an orthogonal basis. The main difference with respect to Bayesian estimators is that shrinkage does not require prior information about the probability distribution of the signal of interest. The original idea is presented in [17] and developed in [18]. With the same notation as above, denoised STSA coefficients can be obtained via hard thresholding as

$$\widehat{A} = \begin{cases} R & \text{if} \quad R \geq \lambda, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where λ is a suitable threshold to choose. This can be rewritten in the form of (3) by defining the gain function as

$$\mathbb{G}_\lambda(R) = \begin{cases} 1 & \text{if} \quad R \geq \lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

The gain function in (11) can also be interpreted or used as a binary mask or a channel selection function [4, Section 13.2, pp. 618].

Shrinkage can be smoothed by using soft thresholding instead of hard thresholding. The soft thresholding function [18] is defined by

$$\mathbb{G}_\lambda(R) = \begin{cases} 1 - \frac{\lambda}{R} & \text{if} \quad R \geq \lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Smoothed shrinkage can also be performed by Smoothed Sigmoid-Based Shrinkage (SSBS) functions propounded and analyzed in [13]. Among the two types of SSBS functions considered in [13], we hereafter consider the family of functions defined by

$$\mathbb{G}_{\tau,\lambda}(R) = \frac{1}{1 + e^{-\tau(R-\lambda)}}. \tag{13}$$

It is worth noticing that the hard thresholding function is a limit case of SSBS function. In other words, SSBS functions of type (13) make it possible to attenuate amplitudes below λ instead of forcing them to 0 as the hard-thresholding function, without introducing a bias as the soft-thresholding function. An SSBS function achieves a trade-off between the hard and soft thresholding functions. In the above references, parameter λ is taken equal to the universal threshold, the minimax threshold or the detection threshold [19], whereas τ controls the attenuation achieved by the SSBS function [13].

III. Proposed joint detection and estimation method

As mentioned above, binary mask improves speech intelligibility but degrades speech quality. On the contrary, statistical estimation is a favorable strategy to enhance speech quality but does not ameliorate speech intelligibility. Therefore, in this section, we propose a combination technique that makes it possible to take advantage of the two approaches. The proposed method can be regarded as a joint detection and estimation algorithm.

A. Proposed denoising algorithms

Figure 1 shows an overview of the proposed method with the several possible combinations it employs. This type of speech enhancement involves the following steps:

- The noisy signal is segmented into short-frames and is decomposed by either an FFT, or another transform like the discrete cosine, filter-bank or wavelet transforms.

- Shrinkage by SSBS performs a binary mask. On the one hand, the SSBS function tends to keep unaltered STSAs that are large enough above λ because such STSAs probably pertain to noisy speech. On the other hand, the SSBS function attenuates small STSAs because they are probably due to noise alone or noisy signals with small amplitudes.
- A noise reduction algorithm, *e.g.* MMSE-STSA, spectral subtraction, etc... is applied to STSAs after binary masking by SSBS.
- The enhanced signal is synthesized from these STSAs by inverse transform.

In this combination, shrinkage by SSBS functions can also be considered as a channel selection that yields performance improvement, at least in terms of speech intelligibility. Indeed, SSBS automatically selects frequency bins where STSAs exceeding λ probably means speech presence. The main difference with binary masking is that the SSBS gain function is smoothed around the threshold value. In addition, SSBS slightly reduces noise by shrinkage and returns the smooth STSA estimate

$$\widehat{A}_1 = \mathbb{G}_{\tau,\lambda}(R)R, \tag{14}$$

where $\mathbb{G}_{\tau,\lambda}(R)$ is given by (13). The choice of λ is addressed in Section III-B. At this stage, it suffices to mention that, in a given frequency bin, this value is proportional to the noise power in this bin and that the noise power is estimated by a noise estimation algorithm [20].

Because SSBS does not involve any smoothing in time or frequency, SSBS may introduce musical noise resulting from isolated time-frequency bins insufficiently attenuated. Thence, the introduction of an MMSE Bayesian estimator based on the decision-directed approach to smooth in time the estimates returned by SSBS. The smoothing in frequency is postponed to some further work.

Note that the gains of the STSA and the LSA estimator are the function of two variables ξ and γ (see (7) and (9)). The *a priori* SNR ξ is thus determined by decision-directed approach whose input is the noisy STSA R. As shown in Figure 1 the *a posteriori* SNR γ_1 is calculated from the rough enhanced STSA $\widehat{A}_1$ provided by SSBS. Therefore, we obtain the gain function

$$\widehat{A} = \mathbb{G}(\xi, \gamma_1)\widehat{A}_1, \tag{15}$$

where $\mathbb{G}(\xi, \gamma_1)$ is the gain function of the Bayesian estimator — either $\mathbb{G}_{\text{LSA}}$ or $\mathbb{G}_{\text{STSA}}$ — with

$$\gamma_1 = \frac{\widehat{A}_1^2}{\sigma_X^2} = \mathbb{G}_{\tau,\lambda}^2(R)\frac{R^2}{\sigma_X^2} = \mathbb{G}_{\tau,\lambda}^2(R)\gamma. \tag{16}$$

It follows from the foregoing equalities that SSBS enables to select and modify the *a posteriori* SNR γ input of the Bayesian gain function.

B. Discussion on the impact of parameters

The two parameters τ and λ play an important role in influencing the performance of the proposed method. Let us first discuss the impact of each parameter. Basically, λ affects the selection of the channels where speech is present and the attenuation degree imposed to the noisy signal.

On the one hand, with high value of λ, speech components can be ignored, which degrades speech quality. On the other hand, with low value of λ, we cannot eliminate noise-only components, which limits also the algorithm performance. However, SSBS provides a smoothed gain function that attenuates such adverse effects. In comparison to binary masking or hard and soft threshold functions, the range where λ must be chosen can be expected to be less crucial. In this sense, the SSBS functions are robust.

The attenuation degree applied to the noisy signal is regulated by τ. It is known that using a continuous gain function cannot improve the output SNR in each frequency and thus, no speech intelligibility improvement should be expected [4, pp.613] if the SSBS function is not similar to a discontinuous shrinkage. Fortunately, for fixed λ, the SSBS function tends to the hard-thresholding function (10), which is similar to a binary mask. Since the hard thresholding function can be regard as an approximation of the optimal — in the mean square sense — diagonal linear estimator, an SSBS function with sufficiently big attenuation can thus be expected to improve speech intelligibility for the same reasons as above [4, pp.525]. However, a too high value for τ can damage speech quality because some important components of speech could be suppressed.

The choice of λ and τ thus relates to the traditional trade-off between residual noise and signal distortion. The greater λ, the smaller the background noise but the larger the signal distortion and musical noise. In the same way, the greater τ, the better the speech intelligibility and the lesser the speech quality.

According to [13], [17], [18], [19] and [21], several values for λ can be thought up. The adaptation to speech enhancement of the several theoretical aspects and results developed in these papers requires a dedicated study. For now, we can however make the following remarks that will guide our choice for this parameter to get the experimental results of the next section.

The objective of speech enhancement is to remove background noise without distorting too much speech, so as to maintain speech quality and improve speech intelligibility. Therefore, even if the SSBS function may compensate the effect of a too large threshold, our final recommendation is to prefer a relatively small value for λ and not value too large for τ. Indeed, missing important speech components can be more detrimental to speech quality than keeping noise-only components that can anyway be filtered by the Bayesian estimator.

IV. Experimental results

We assessed our proposed method on the NOIZEUS database [4] to evaluate its performance. This database contains IEEE sentences corrupted by noise coming from the AURORA database, at four levels, namely 0, 5, 10 and 15 dB. Two combinations between SSBS and MMSE estimators were tested. The first combination, the bayesian estimator

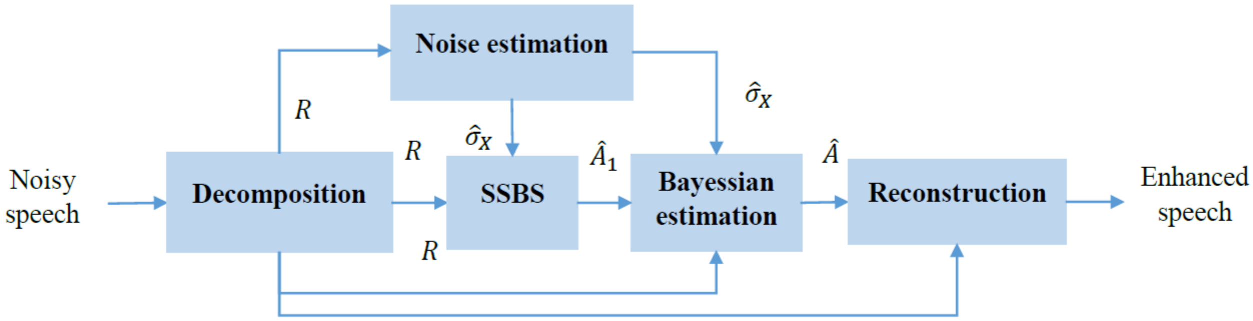

Fig. 1. Processing chain of the proposed algorithm using a smoothed rough binary masking

TABLE I
PERFORMANCE EVALUATION WITH THREE CRITERIA: MARS_OVL, STOI, SSNR

Noise	Method	SSNR				MARS_ovl				STOI (%)			
		0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
AR	STSA	-0.83	1.44	3.85	6.42	2.26	2.94	4.15	6.92	70.98	93.52	98.65	99.62
	SBSA	**-0.64**	**1.69**	**4.15**	**6.74**	**2.27**	**3.00**	**4.63**	**7.69**	**73.15**	**94.17**	**98.78**	**99.65**
	LSA	-0.16	2.14	4.55	7.06	**2.42**	**3.22**	5.49	8.40	75.83	94.72	98.81	99.64
	SBLA	**-0.10**	**2.23**	**4.66**	**7.17**	2.41	3.20	**5.67**	**8.72**	**76.65**	**94.96**	**98.87**	**99.65**
Train	STSA	-0.83	1.62	3.86	6.42	2.24	2.75	3.52	6.67	82.87	97.42	99.39	99.79
	SBSA	**-0.63**	**1.87**	**4.15**	**6.73**	**2.25**	**2.78**	**3.79**	**7.29**	**84.28**	**97.60**	**99.42**	**99.80**
	LSA	-0.07	2.37	4.61	7.10	**2.34**	**2.88**	4.67	8.19	85.40	97.69	99.44	99.80
	SBLA	**-0.02**	**2.45**	**4.71**	**7.21**	2.33	**2.88**	**4.77**	**8.42**	**85.94**	**97.77**	**99.45**	**99.81**
Airport	STSA	-0.85	1.27	3.79	6.25	2.43	3.20	4.82	7.86	87.54	97.89	99.59	**99.86**
	SBSA	**-0.65**	**1.51**	**4.07**	**6.56**	**2.46**	**3.28**	**5.23**	**8.40**	**88.43**	**98.03**	**99.61**	**99.86**
	LSA	-0.07	2.08	4.55	7.02	**2.56**	3.40	5.95	9.16	88.80	98.00	99.58	**99.86**
	SBLA	**-0.02**	**2.15**	**4.64**	**7.10**	**2.56**	**3.43**	**6.14**	**9.47**	**89.17**	**98.06**	**99.59**	99.85
Babble	STSA	-1.27	0.99	3.57	6.06	2.34	3.03	4.37	7.62	78.51	96.53	99.49	**99.84**
	SBSA	**-1.09**	**1.21**	**3.85**	**6.35**	**2.38**	3.07	**4.75**	**8.13**	**80.12**	**96.81**	**99.50**	**99.84**
	LSA	-0.56	1.66	4.30	6.77	2.44	**3.16**	5.35	8.75	80.97	96.94	**99.50**	**99.83**
	SBLA	**-0.51**	**1.73**	**4.39**	**6.86**	**2.47**	**3.16**	**5.50**	**8.97**	**81.74**	**97.05**	**99.50**	**99.83**

is MMSE-STSA [2]. In the second combination, we use MMSE-LSA [3]. The first combination is hereafter called SBSA for Smoothed Binary Spectral Amplitude and the second SBLA for Smoothed Binary Log-Amplitude. MMSE-STSA and MMSE-LSA are also considered as the references methods. In our experiments, speech signals with sampling rate at 8 kHz were segmented into sets of 256 sample frames, transformed using STFT with 50% overlapped Hamming windows. The parameters τ and λ of the SSBS gain function were chosen after preliminary tests on 20 randomly chosen sentences corrupted by car noise with SNR equal to 5 dB. According to these tests, τ was set to 61 and, for each given frequency bin, λ was fixed to $1.1\sqrt{\sigma_X^2}$ where σ_X^2 is the noise power spectral. This power spectral is estimated by the up-to-date B-E-DATE method [20].

For assessing speech quality and intelligibility yielded by the denoising algorithms, some objective and pseudo-subjective quality and intelligibility criteria were used. Speech quality is firstly measured by the segmental SNR (SSNR) objective criterion and the overall quality pseudo-subjective criterion based on multivariate adaptive regression splines (MARS_ovl). The SSNR values were clipped to $[-10, 35 \text{ dB}]$ to bypass the use of a silence/speech detector [4, pp. 480]. Criterion MARS_ovl enables to predict the rating of overall speech quality [22]. It is the function of some widely used measures Itakura-Saito distance (IS) and perceptual evaluation of speech quality (PESQ) [22].

Secondly, speech intelligibility was initially evaluated by the short-time objective intelligibility measure (STOI), which has high correlation with intelligibility measured by listening tests. In general, STOI measures the mean correlation of clean and enhanced speech coefficients calculated by regrouping DFT-bin coefficients in the time-frequency domain [23]. A logistic function is applied to STOI measures to map intelligibility scores :

$$f(\text{STOI}) = \frac{100}{1 + \exp(a \times \text{STOI} + b)}, \tag{17}$$

where, for fitting with IEEE sentences, $a = -17.4906$ and $b = 9.6921$ [23].

The average results for different noise types and SNR values are shown in table IV. For each SNR, each type of noise, each given criterion and each pair of results provided by STSA and SBSA or by LSA and SBLA, the value in boldface points out the best result. For each SNR, each type of noise and each given criterion, the value in red emphasizes the best result.

The proposed SBSA and SBLA techniques outperform the reference methods in almost every case. Especially, the SSNRs obtained by SBSA and SBLA are always higher than those yielded by STSA and LSA. The influence of the proposed methods in terms of STOI is more emphasized at low SNRs.

V. Conclusion

In this paper, we have proposed a novel method to enhance speech contaminated by non-stationary noise. This method takes advantage of two familiar approaches include sigmoid shrinkage and MMSE. Initial tests conducted on the NOIZEUS database confirm that this approach is promising. Although the performance gain compared to reference methods is not very large, the approach however tends to provide some gain in almost every situation with parameters roughly fixed via some preliminary experiments. This indicates that the combination is a good strategy. Since the algorithms involved in our approach are based on a theoretical background, which is not dedicated to speech signals, it can be wondered whether the parameters of the SSBS cannot be chosen theoretically or even adaptively. Hence, future work will focus on the computation of these parameters with regard to the unavoidable trade-off between false alarm and miss detection probabilities. In addition, for more improvement, some estimators other than STSA, could be considered. The idea would be that combining the SSBS smooth binary masking with most non-parametric estimators bring better results and robustness.

References

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, “Dft-domain based single-microphone noise reduction for speech enhancement: a survey of the state of the art,” *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–80, 2013.

[2] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] ——, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans., Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

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

[5] J. Jensen and R. C. Hendriks, “Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 92–102, 2012.

[6] A. Abramson and I. Cohen, “Simultaneous detection and estimation approach for speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2348–2359, 2007.

[7] H. Momeni, H. R. Abutalebi, and A. Tadaion, “Joint detection and estimation of speech spectral amplitude using noncontinuous gain functions,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 8, pp. 1249–1258, 2015.

[8] Y. Lu and P. C. Loizou, “Estimators of the magnitude-squared spectrum and methods for incorporating snr uncertainty,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1123–1137, 2011.

[9] P. C. Loizou and G. Kim, “Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 47–56, 2011.

[10] D. Wang, M. S. Kjems, U.and Pedersen, J. B. Boldt, and T. Lunner, “Speech intelligibility in background noise with ideal binary time-frequency masking,” *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2336–2347, 2009.

[11] G. Kim and P. C. Loizou, “Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints,” *IEEE Signal Process. Lett.*, vol. 17, no. 12, pp. 1010–1013, 2010.

[12] R. Tavares and R. Coelho, “Speech enhancement with nonstationary acoustic noise detection in time domain,” *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 6–10, 2016.

[13] A. M. Atto, D. Pastor, and G. Mercier, “Smooth sigmoid wavelet shrinkage for non-parametric estimation.” in *Proc. IEE Int. Conf. Acoust. Speech, Signal Process.*, 2008, pp. 3265–3268.

[14] E. Plourde and B. Champagne, “Generalized bayesian estimators of the spectral amplitude for speech enhancement,” *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 485–488, 2009.

[15] C. H. You, S. N. Koh, and S. Rahardja, “β-order mmse spectral amplitude estimation for speech enhancement,” *IEEE Trans. Speech, Audio, Process.*, vol. 13, no. 4, pp. 475–486, 2005.

[16] P. C. Loizou, “Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum,” *IEEE Trans. Acoust. Speech, Process.*, vol. 13, no. 5, pp. 857–869, 2005.

[17] D. L. Donoho and J. M. Johnstone, “Ideal spatial adaptation by wavelet shrinkage,” *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[18] D. L. Donoho, “De-noising by soft-thresholding,” *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[19] A. M. Atto, D. Pastor, and G. Mercier, “Detection threshold for non-parametric estimation,” *Signal, Image and Video processing*, vol. 2, no. 3, pp. 207–223, 2008.

[20] V. K. Mai, D. Pastor, A. Aïssa-El-Bey, and R. Le-Bidan, “Robust estimation of non-stationary noise power spectrum for speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 670–682, 2015.

[21] D. Pastor and Q. T. Nguyen, “Random distortion testing and optimality of thresholding tests,” *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 4161–4171, 2013.

[22] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.