Block Smoothed Sigmoid-Based Shrinkage in Time-Frequency Domain for Robust Audio Denoising

van Khanh Mai, Dominique Pastor, Abdeldjalil Aissa El Bey

Block Smoothed Sigmoid-Based Shrinkage in Time-Frequency Domain for Robust Audio Denoising

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*—In this paper, we propose a novel robust method for short-time spectral amplitude (STSA) estimation in audio denoising. This method extends the smoothed sigmoid-based shrinkage (SSBS), which does not require any prior information about the probability distribution of the signal of interest. With regard to audio processing, the SSBS method yields better performance in terms of noise reduction but nevertheless introduces significant musical noise. In order to benefit from non-diagonal processing, which removes background noise without introducing musical noise, two non-diagonal SSBS are derived. First, the decision-directed approach is incorporated into the SSBS method. Second, the time-frequency domain is divided into rectangular blocks and then, a SSBS function is applied to estimate the spectral amplitude in each block. The experimental results demonstrate the relevance of the proposed methods both in terms of speech quality and intelligibility via objective criteria.**

***Index Terms*—Block threshold, audio denoising, smoothed sigmoid-based shrinkage.**

I. Introduction

Nowadays, a fundamental task in signal processing especially in audio and speech enhancement, is the elimination of additive noise from the contaminated signal $y[n] = s[n]+x[n]$, where s, x are the signal of interest and noise respectively and $n = 0, 1, \ldots, N-1$. Traditionally, by transforming the signal model into the time-frequency domain, many studies aim to improve not only speech quality but also speech intelligibility.

The first computationally simple method is power spectral subtraction, which can be carried out without having much prior information [1]. Then, assuming a linear relationship between the estimate and observed signals, the optimal Wiener filter is derived [2, Chap.6]. However, the main drawback of these methods is the independent time-frequency generating musical noise or artifact noise as the thresholding methods proposed in [3]–[5]. Alternatively, short-time spectral amplitude (STSA) estimators based on minimum mean square error (MMSE) yield better performance without introducing musical noise by taking into account past estimate of STSA via decision-directed approach [6]–[8]. Nevertheless, supposing that the signal of interest follows a Gaussian probability density function (pdf) is not always respected. Benefiting from the block thresholding introduced in [9], the block threshold denoising for audio signal developed in [10] is able to reduce musical noise for music signal but not for speech signal. In contrast to [10] our approach reduces musical noise for speech denoising.

This paper addresses a novel time-frequency domain non-diagonal audio denoising method that takes into account both advantage of smoothed sigmoid-based shrinkage (SSBS) gain function and block thresholding. Firstly, the modified SSBS is proposed, which fits with short-time spectral amplitude estimator in the time-frequency domain. Then, for eliminating isolated estimated spectral amplitudes, which produce musical noise, two strategies are proposed. The first one, the decision-directed approach is used to estimate the instantaneous a priori signal to noise ratio (SNR), which is incorporated into the SSBS gain function. The second one, the same SSBS gain function is applied to each time-frequency bin block.

The remaining of this paper is organized as follows. Section II presents some notation and background on threshold methods. Section III details the proposed algorithms. Then, experimental results are presented in Section IV. Finally, Section V concludes this paper.

II. Signal Model and Background

In most audio denoising applications, the noisy speech signal is segmented and transformed into the time-frequency domain by short-time Fourier transform (STFT). Thus, the observed signal in the time-frequency domain becomes

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

where, m , k denote respectively the frame and frequency index. The noisy signal can also be formulated in polar form as follows

$$R[m,k]e^{j\phi_Y[m,k]} = A[m,k]e^{j\phi_S[m,k]} + D[m,k]e^{j\phi_X[m,k]},$$

where $R[m,k]$, $A[m,k]$, $D[m,k]$ are the short-time spectral amplitude (STSA) of the observed signal, clean signal, noise STFT coefficients and the associated phases are $\phi_Y[m,k]$, $\phi_S[m,k]$, $\phi_X[m,k]$, respectively. Additionally, the signal of interest and noise are assumed to be independent so that the observed spectral power is $\mathbf{E}\left[R^2[m,k]\right] = \mathbf{E}\left[A^2[m,k]\right] + \mathbf{E}\left[D^2[m,k]\right]$. We set $\sigma^2[m,k] = \mathbf{E}\left[D^2[m,k]\right]$. The *a posteriori* signal to noise ration (SNR) $\gamma[m,k]$ is defined by $\gamma[m,k] = R^2[m,k]/\sigma^2[m,k]$. The *a priori* SNR and the instantaneous *a priori* SNR are also defined as $\xi[m,k] =$

978-1-5386-8173-2/18/$31.00

$\mathbf{E}[A^2[m,k]]/\sigma^2[m,k]$, $\zeta[m,k] = A^2[m,k]/\sigma^2[m,k]$. Note that $\zeta[m,k]$ can be considered as an estimate of $\xi[m,k]$. Due to the importance of the short-time spectral amplitude, many researches aim to estimate it and the associated phase is simply fixed to the noisy phase. In order to retrieve the clean signal, a gain function $G[m,k]$ is often determined so that the enhanced STFT coefficient signal is calculated by

$$\widehat{S}[m,k] = G[m,k]Y[m,k], \tag{2}$$

where for simplifying notation, the estimates are henceforth denoted by a wide hat symbol *e.g* $\widehat{\xi}$ is an estimate of ξ.

A. Thresholding estimation: Sigmoid shrinkage

Shrinkage functions are often applied to image processing to estimate signal coefficients produced by the projection of the noisy signal on an orthogonal basis. The main difference with Bayesian estimators is that shrinkage methods do not require prior information about the pdf of the signal of interest. The first shrinkage function called hard thresholding is presented in [3] and developed in [4]. Denoised STSA coefficients can then be obtained by hard threshold as follows

$$\widehat{A}[m,k] = \begin{cases} R[m,k] & \text{if} \quad R[m,k] \geq \Lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where $\Lambda = \lambda\sigma[m,k]$ and λ is a convenient parameter. Thus, this equation can be written as (2) by introducing the gain function

$$G_\lambda[m,k] = \begin{cases} 1 & \text{if} \quad \gamma[m,k] \geq \lambda^2, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

This gain function (4) is also depicted as binary masking or channel selection function [2, Section 13.2, pp. 618].

In the same way, shrinkage can be smoothed by using soft thresholding instead of hard thresholding. The soft thresholding proposed in [4] is given as follows

$$G_\lambda[m,k] = \begin{cases} 1 - \frac{\lambda}{\sqrt{\gamma[m,k]}} & \text{if} \quad \gamma[m,k] \geq \lambda^2, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

and can be written as:

$$G[m,k] = \left(1 - \frac{\lambda}{\sqrt{\gamma[m,k]}}\right)_+, \tag{6}$$

where $(z)_+ = \max(z,0)$. Smooth shrinkage can be also performed SSBS stated and analyzed in [5]. The gain function is constructed based on three properties of shrinkage functions including smoothness, penalized shrinkage and vanishing attenuation at infinity. This type of shrinkage can be described by the following equation:

$$G_{\lambda,\tau,\lambda}[m,k] = \frac{|1 - \frac{\mu}{\sqrt{\gamma[m,k]}}|}{1 + e^{-\tau\left((\gamma[m,k])^{1/p} - \lambda\right)}}, \tag{7}$$

where μ, τ, p and λ are the parameters of methods. It is worthy noticing that hard and soft thresholding functions are limit cases of SSBS functions. For often, parameter λ can be taken equal to the universal, minimax or detection threshold [11], whereas parameter μ is often set equal to zero and τ controls the attenuation yielded by the SSBS function [5]. Parameter p can control the smooth property of the SSBS function. In [5] $p = 2$.

B. Block Thresholding

From the above methods [3]–[5] the attenuation factors $G[m,k]$ are independently and singly evaluated for each (m,k) atom. Therefore, in order to incorporate the impact of neighboring time-frequency (or scale) atoms, the block thresholding is proposed in [9] for the wavelet transform and is developed in [10] for the STFT transform. Both methods are based on the power subtraction gain function:

$$G[m,k] = \left(1 - \frac{\lambda}{1 + \widehat{\xi}[m,k]}\right)_+. \tag{8}$$

This gain function can be considered as the generalization of the Wiener filter where $\widehat{\xi}[m,k]$ is an estimator of $\xi[m,k]$.

The time-frequency $[m,k]$ plane is divided into non-overlapping time-frequency rectangular blocks. In place of calculating a gain function $G[m,k]$ for each time-frequency $[m,k]$ bin, for each block B_i, the signal of interest is estimated by the same gain function G_i. Thus, the estimated signal in the time-frequency domain is given by

$$\widehat{S}[m,k] = G_i Y[m,k] \qquad (m,k) \in B_i, \tag{9}$$

where the block gain function $G_i = \left(1 - \frac{\lambda}{1+\widehat{\xi}_i}\right)_+$. By considering neighboring time-frequency atoms, the estimated *a priori* SNR $\widehat{\xi}_i$ is calculated as the mean of all $\xi[m,k]$, where (m,k) bin belongs to the given block B_i.

III. Non-diagonal Smoothed Sigmoid-Based Shrinkage

A. Decision-directed SSBS (DSS)

Performance of SSBS method in the image processing is pinpointed in [5]. SSBS function is also called logistic function, which is widely used in the machine learning. In order to devise a robust audio denoising method, the SSBS gain function (7) with $\mu = 0$ is considered. The *a posteriori* SNR is approximated as following:

$$\gamma[m,k] = \zeta[m,k] + 1. \tag{10}$$

Thus, for calculating the gain function $G[m,k]$, we need to estimate the instantaneous *a priori* SNR $\zeta[m,k]$. However, the clean short-time spectral amplitude $A[m,k]$ is not available so that the instantaneous *a priori* SNR can be estimated via the decision-directed approach. First of all, as in [6], for taking into account the previous atoms in the same frequency bin, the estimated a priori SNR $\widehat{\xi}[m,k]$ is given by:

$$\widehat{\xi}[m,k] = \beta\frac{\widehat{A}^2[m,k]}{\sigma^2[m,k]} + (1-\beta)(\gamma[m,k] - 1)_+, \tag{11}$$

where β is tthe smoothing factor. For the same reason to modify the decision-directed approach in [12], we propose to estimate the instantaneous SNR as:

$$\widehat{\zeta}[m,k] = \left(\frac{\widehat{\xi}[m,k]}{1+\widehat{\xi}[m,k]} \right)^2 \gamma[m,k]. \tag{12}$$

This gives more emphasis to the role of current atom $Y[m,k]$. Instead of using the observed absolute value $R[m,k]$ or the square root of the *a posteriori* SNR $\gamma[m,k]$ as in (7), we modify the SSBS gain function as

$$G[m,k] = \frac{1}{1+e^{-\tau\left(\sqrt[4]{\widehat{\zeta}[m,k]+1}-\lambda\right)}}. \tag{13}$$

It mean $p=4$. Because, this method brutally modifies each isolated atoms in the time-frequency domain, the power four can provide more smooth property than using the $\sqrt{\widehat{\zeta}[m,k]}$ or $\widehat{\zeta}[m,k]$ and conserve a large enough reduction of noise.

B. Block SSBS (BSS)

For considering the time-frequency neighboring atoms, the estimated *a priori* SNR $\widehat{\xi}_i$ can be estimated by averaging instantaneous noisy signal energy $R^2[m,k]$ and the averaging noise power spectral $\sigma^2[m,k]$ over block B_i, so that:

$$\widehat{\xi}_i = \left(\frac{\overline{Y_i^2}}{\overline{\sigma_i^2}} - 1 \right)_+, \tag{14}$$

where

$$\overline{Y_i^2} = \frac{1}{N_i} \sum_{(m,k)\in B_i} R^2[m,k], \tag{15}$$

$$\overline{\sigma_i^2} = \frac{1}{N_i} \sum_{(m,k)\in B_i} \sigma^2[m,k], \tag{16}$$

where N_i is the number of the time-frequency bin $[m,k]$ in the given block B_i. The block SSBS gain function now becomes

$$\mathbf{G}_i = \frac{1}{1+e^{-\tau(\sqrt{\widehat{\xi}_i+1}-\lambda)}}. \tag{17}$$

We remain use $p=2$ for have a good attenuation of noise whereas the smooth property is regularized by using block approaches.

The fundamental problem is convenient choice of the block size N_i. For simple implementation, the time-frequency image is tiled in non-overlapping rectangular blocks so that the bin number of rectangular block B_i is $N_i = L_i \times W_i$, where L_i and W_i are the rectangular length and width corresponding to the number of the frames and the number of the frequency bins in the time-frequency domain, respectively. Note that the larger L_i, the greater time delay will appear. For real time processing application, the rectangular length L_i must be small enough. However, the larger L_i, the greater time delay will be presented. For real time processing application, the rectangular length L_i must be small enough. Therefore, in this subsection as in [9], [10], we address the size of the given block with respect to this constraint and aim at minimizing the upper bound on risk r which is given by:

$$r = \mathbf{E}\left[\|\widehat{s}-s\|^2\right]. \tag{18}$$

Based on the property of the frame basis after short-time Fourier transform [13], the risk r is upper bounded as

$$r \le \frac{1}{f} \sum_{i=1}^{I} \sum_{(m,k)\in B_i} \mathbf{E}\left[\|\widehat{S}[m,k]-S[m,k]\|^2\right], \tag{19}$$

where f is the redundant factor and I is the total number of frames. By denoting $\mathbf{Z}_i$ as the upper bound of risk over the given block B_i, we obtain the upper bounded risk as

$$r \le \frac{1}{f} \sum_{i=1}^{I} \mathbf{Z}_i = \mathbf{Z}, \tag{20}$$

where $\mathbf{Z}$ is overall upper bound risk. Applying the SURE theorem [14, Section 2] and after some routine algebra, we are able to get the unbiased estimate $\widehat{\mathbf{Z}}_i$:

$$\begin{aligned} \widehat{\mathbf{Z}}_i = {} & N_i \overline{\sigma_i^2}(2\mathbf{G}_i - 1) + \\ & (1-\mathbf{G}_i)(1+\tau\mathbf{G}_i\sqrt{\widehat{\xi}_1+1}-\mathbf{G}_i)\|\mathbf{Y}\|_2^2., \end{aligned} \tag{21}$$

where $\|\mathbf{Y}\|_2^2 = \sum_{(m,k)\in Z_i} Z^2[m,k]$.

Generally, all parameters τ, λ and N_i can be estimated via minimizing the estimate $\widehat{\mathbf{Z}}$ of $\mathbf{Z}$. For now, a given τ and λ, the block size N_i is obtained by minimizing the estimated risk $\mathbf{Z}$. For reducing the complexity and the time delay, the time-frequency domain is divided into rectangular blocks B_i of size $W_i \times L_i$. Then, we try to split each block into sub-blocks via minimizing the overall risk $\mathbf{Z}_i$ over a given block B_i. Furthermore, for taking into consideration that the impact of non-stationary noise is different from a band to another [15], we need to only find the rectangular length of sub-block whereas the rectangular width W_i is fixed by using the frequency linear, logarithmic or mel spacing.

IV. Experiment and Results

We benchmarked our proposed BSS and DSS methods to the reference Log-Spectral Amplitude (LSA) [7] and Audio Block Thresholding (ABT) [10] methods on the NOIZEUS database [2] to evaluate their performance. This database contains IEEE sentences corrupted by noise coming from the AURORA database, at four levels, namely 0, 5, 10 and 15 dB. In our experiments, speech signals with sampling rate at 8 kHz were segmented into sets of 20-ms duration frames, transformed by STFT with 50% overlapped Hamming windows. The parameters τ and λ of the SSBS gain function were chosen after preliminary tests on 20 randomly selected sentences corrupted by car noise with SNR equal to 5 dB. The result of the test allow us to set $\tau = 5.0725$ and, for each given frequency bin, $\lambda = 0.8$. The noise power spectral σ^2 is estimated by up-to-date method B-E-DATE [16].

The speech quality and intelligibility, yielded by the de-noising algorithms, are evaluated by both objective quality and intelligibility criteria. Speech quality is firstly measured

TABLE I
PERFORMANCE EVALUATION WITH TWO CRITERIA: MARS_OVL AND STOI

Noise	Method	MARS_ovl				STOI(%)			
		0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
White	LSA	2.41	3.08	5.32	8.36	84.97	96.65	99.11	99.67
	ABT	-1.38	-0.67	0.65	3.29	84.76	97.26	99.39	99.80
	DSS	**2.65**	**3.61**	**6.88**	9.67	91.58	98.26	99.51	99.80
	BSS	2.56	3.41	6.86	**10.24**	**91.87**	**98.48**	**99.57**	**99.83**
Train	LSA	**2.34**	**2.88**	4.67	8.19	85.40	97.69	99.44	99.80
	ABT	-1.33	-0.48	0.74	3.27	85.43	98.48	99.62	99.87
	DSS	2.23	2.80	**5.30**	**9.04**	**92.39**	**98.92**	99.69	99.88
	BSS	2.09	2.68	4.25	8.60	88.75	98.87	**99.71**	**99.89**
Airport	LSA	**2.56**	**3.40**	5.95	9.16	88.80	98.00	99.58	99.86
	ABT	-0.49	0.77	2.44	6.35	88.81	98.78	99.79	99.91
	DSS	2.35	3.28	**6.33**	**10.11**	**93.47**	**99.04**	99.77	99.91
	BSS	2.16	3.05	5.44	9.81	91.94	99.03	**99.80**	**99.92**
Babble	LSA	**2.44**	**3.16**	5.35	8.75	80.97	96.94	99.50	99.83
	ABT	-0.84	0.43	1.60	5.10	81.76	98.26	99.73	**99.90**
	DSS	2.19	2.91	**5.48**	**9.41**	**88.51**	98.56	99.74	**99.90**
	BSS	2.01	2.75	4.40	8.95	84.83	**98.57**	**99.76**	**99.90**

by the overall quality pseudo-subjective criterion based on multivariate adaptive regression splines (MARS_ovl) [17]. Secondly, speech intelligibility was initially evaluated by the short-time objective intelligibility measure (STOI), which has the high correlation with intelligibility measured by listening tests. A logistic function is applied to STOI measures to map intelligibility scores [18].

The average results for different noise types and SNR values are shown in Table I. For each SNR, each type of noise and each given criterion, the value in bold face emphasizes the best result. For MARS_ovl criterion, the proposed methods lead the significantly best score at high SNR levels and remain close to the vicinity of the best score at low SNR levels (0 and 5 dB). For speech intelligibility criterion, our proposed methods yield best scores at all SNR levels, especially at low SNR levels.

V. CONCLUSION

In this paper, a novel method has been proposed to enhance the speech corrupted by background noise. By considering advantage of non-diagonal estimator and shrinkage, our method yields promising results conducted on the NOIZEUS database. In future steps, as we discussed above, parameters will be theoretically chosen by minimizing the estimated risk.

REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP).*, vol. 4, 1979, pp. 208–211.

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

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

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

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

[6] 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.

[7] ——, "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.

[8] 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.

[9] T. T. Cai, "Adaptive wavelet estimation: a block thresholding and oracle inequality approach," *Annals of statistics*, pp. 898–924, 1999.

[10] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1830–1839, 2008.

[11] 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.

[12] P. C. Yong, S. Nordholm, and H. H. Dam, "Trade-off evaluation for speech enhancement algorithms with respect to the a priori snr estimation," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP).* IEEE, 2012, pp. 4657–4660.

[13] S. Mallat, *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

[14] R. Tibshirani and L. Wasserman, "Steins unbiased risk estimate," *Course notes from "Statistical Machine Learning, Spring 2015"*, pp. 1–12, 2015.

[15] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 4. IEEE, 2002, pp. IV–4164.

[16] 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.

[17] 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.

[18] 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.