

# Reduction of Musical Noise by Weiner Post Processing Method

S.Sulochana, S.Senthil Siva Manikandan.

**Abstract**—A stand-alone noise suppression algorithm is presented for reducing the spectral effects of acoustically added noise in speech. Traditionally de-noising techniques are powerful in terms of noise reduction, have the drawback of generating an annoying musical noise. This paper proposes the problem of enhancing speech in highly noisy environments using non-diagonal audio de-noising algorithm through adaptive time frequency block thresholding. A block thresholding procedure empirically chooses the block size and threshold level at each resolution level by minimizing the Stein's unbiased risk estimate. Numerical experiments show that this adaptive estimator is robust to signal type variations and improves the SNR.

**Index Terms**— Spectral de-noising, Musical Noise, Block thresholding, Stein's risk estimate, spectral subtraction.

## 1. INTRODUCTION

Degradation of the quality of speech caused by the acoustic noise is common in most speech processing applications. In mobile telephony, reducing noise in corrupted speech is a challenging task especially in high noise level. In all speech communication settings the quality and intelligibility of speech is of utmost importance for ease and accuracy of information exchange. The speech processing systems used to communicate or store speech are usually designed for a noise free environment but in a real-world environment, the presence of background interference in the form of additive background and channel noise drastically degrades the performance of these systems, causing inaccurate information exchange and listener fatigue [2], [5]. Restoring the desired speech signal from the mixture of speech and background noise is amongst the oldest, still elusive goals in speech processing.

Speech enhancement algorithms attempt to improve the performance of communication systems when their input or output signals are corrupted by noise [5]. This is important in a variety of contexts, such as in environments with interfering background noise and in speech recognition systems. Over the year, researchers and engineers have developed a number of methods to address this problem. Yet, due to complexities of the speech signal, this area of research still poses a considerable challenge. It is usually difficult to reduce noise without distorting speech and thus, the performance of speech enhancement systems is limited by the tradeoff between speech distortion and noise reduction. In general, the situation where the noise and speech are in the same channel (single channel systems) is the most common scenario and is one of the most difficult situations to deal with. The complexity and ease of implementation of any proposed scheme is another important criterion especially since the majority of the speech enhancement and noise reduction algorithms find applications in real-time portable systems [6].

A large number of speech enhancement techniques have been proposed in the past. They are predominantly based on spectral subtraction [1] and short-time spectral amplitude estimator [3], [4]. But their main drawback is the appearance of an annoying residual noise, often referred to as musical

noise. The noise suppression rules, together with a decision directed recursive estimator of the a priori signal-to-noise ratio(SNR) that efficiently reduces the musical noise [2], [11]. Their suppression rules have been reinvestigated through years and a non-casual a priori SNR estimator has been proposed [3], [14].

The approaches for noise reduction using the wavelet transform have been proposed in [10], [12]. It employs the thresholding in the wavelet domain and has shown to have very broad asymptotic near-optimal properties for a wide class of signals corrupted by additive white Gaussian noise. A semi-soft thresholding is used to remove noise components from the wavelet coefficients of noisy speech [6]. There should be however some considerations in applying the threshold method directly to speech signal. Since the speech signal in the unvoiced region may contain relatively lots of high frequency components that can be eliminated during the thresholding process. Eliminating them in the wavelet domain can cause severe degradation of intelligibility in the reconstructed signal.

Many signal de-noising techniques are based on attenuation in time-frequency signal representation. Diagonal time-frequency audio de-noising algorithms attenuate the noise by processing each window Fourier or wavelet coefficient independently. These algorithms create isolated time-frequency structures that are perceived as a "musical noise" This musical noise is strongly attenuated with nondiagonal time-frequency estimators that regularize the estimation by recursively aggregating time-frequency coefficients. This paper investigates a non-diagonal audio de-noising algorithm with the help of adaptive time-frequency block thresholding [7]. The block sizes and the threshold level are in redundant time-frequency signal representations and the block thresholding eliminates the residual noise artifacts through a temporal regularization and it provides good approximation of the attenuation with oracle.

The rest of the manuscript is organized as follows. In section II, the proposed speech enhancement scheme is briefly outlined and useful backgrounds are given. Section III describes the work of residual noise reduction. In section IV, the results and performance of the time-frequency block thresholding are demonstrated.

## 2 BASELINE SPEECH ENHANCEMENT SYSTEM

The methods of de-noising audio signals have assumed that the signal is stationary over a specified interval of time. The entire time-domain signal to be processed is divided up into a series of these time intervals (often referred to as 'windows'), and the DSP algorithm processes each interval separately. However, it is a fact that audio signals (both speech and music) are generally not stationary; they cannot always be said to be stationary over each of these windows.

### 2.1 Notations

The properties of sounds are revealed by transforms that decompose signals over elementary functions that are well concentrated in time and frequency. Windowed Fourier transforms and wavelet transforms are two important classes of local time-frequency decompositions. Time-frequency audio-de-noising procedures compute a short-time Fourier transform or a wavelet transform or a wavelet packet transform of the noisy signal, and processes the resulting coefficients to attenuate the noise.

The audio signal  $x$  is corrupted by a noise  $d$  that is often modeled as a zero-mean Gaussian process independent of  $x$ :

$$y[n] = x[n] + d[n] \quad (1)$$

where  $x[n]$  is the clean speech signal and  $d[n]$  is the noise signal. The processing is done on a frame-by-frame basis. A Fourier transform is normally performed on each frame to obtain the Short Time Fourier Transform (STFT). The data to be transformed could be broken up into a chunks or frames. Each frame is Fourier transformed, the complex result is added to a matrix, which records magnitude and phase for each point in time and frequency.

$$\text{STFT}\{y[n]\} = Y[l, k] = \sum_{n=-\infty}^{\infty} y[n]w(n-l)e^{-j\omega n} \quad (2)$$

where  $w(n)$  is a time window. In this work  $w(n)$  is the square root of hanning window.

A time-frequency transform decomposes the audio signal  $y$  over a family of time-frequency atoms  $\{s_{l,k}\}_{l,k}$  where  $l$  and  $k$  are the time and frequency localization indices. The resulting coefficients shall be written as

$$Y[l, k] = \langle y, s_{l,k} \rangle = \sum_{n=0}^{N-1} y[n]s_{l,k}[n]^* \quad (3)$$

where  $*$  denotes the conjugate. These transforms define a complete and often redundant signal representation and a tight frame which means that there exists  $A > 0$  such that,

$$\|y\|^2 = \frac{1}{A} \sum_{l,k} |\langle y, s_{l,k} \rangle|^2 \quad (4)$$

The constant  $A$  is a redundancy factor and if  $A=1$ , then a tight frame is an orthogonal basis.

A de-noising algorithm modifies time-frequency coefficients by multiplying each of them by an attenuation factor to attenuate the noise component. The resulting "de-noised" signal estimator is Spectral domain procedure.

$$\hat{x}[n] = \frac{1}{A} \sum_{l,k} \hat{x}[l, k]s_{l,k}[n]$$

$$\hat{x}[n] = \frac{1}{A} \sum_{l,k} a[l, k]Y[l, k]s_{l,k}[n] \quad (5)$$

It is the objective to provide an alternative to the more traditional spectral domain model-based approaches to de-noising by investigating whether wavelet packet or trigonometric packet bases can be used to successfully decompose the signal for processing. Fig.1 shows the basic procedure for spectral domain removal of broad-band continuous noise [15]. The time-domain signal is first broken up into a series of overlapping windows.

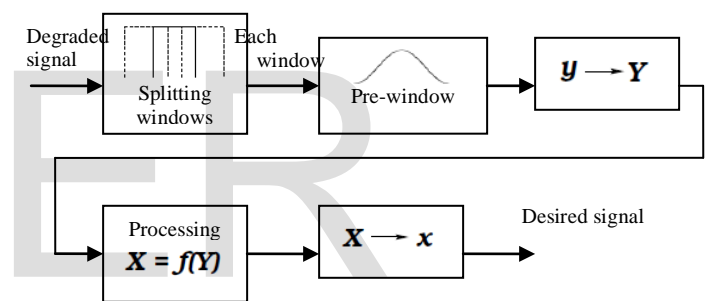


Fig.1 Spectral domain procedure for removing broad-band noise.

Each window is then processed individually in the spectral domain. Before performing the spectral transform, the window is multiplied by a smooth pre-windowing function in order to reduce spectral artifacts caused by the discontinuities at the edges of the window. Once in the spectral domain, the discrete spectral coefficients are adjusted using some function. The modified spectral components are then transformed back into the timed domain, and multiplied by a post-windowing function. This post-windowing function is to once again ensure that no discontinuities are introduced at the edges of the window. All of the overlapping windows are then added back together, and multiplied by a gain compensation function which corrects for the variations in signal amplitude introduced by the pre- and post-windowing functions.

### 2.2 The musical noise phenomenon

In each frame there will be some spectral components due to noise whose power is greater than the estimated noise power. Ideally, these spectral components should be set to zero, but they are instead merely scaled down, leaving a spectral component in the de-noised signal which was not part of the original music or speech. This is often referred to as an 'artifact' or 'residual' of the spectral domain processing [15].

### 3 BLOCKTHRESHOLDING

The block thresholding algorithm of Cai and Silverman [6], [12] regularizes diagonal thresholding estimations by grouping coefficients in blocks and computing a single attenuation factor for all coefficients in each block. This estimator is a general context of orthogonal bases and frames before applying it to spectrograms for audio de-noising. By regularizing the thresholding estimation over blocks of coefficients, the musical noise is almost completely removed and the SNR is improved. Block thresholding has the practical advantage of providing spatial adaptivity to relatively subtle changes in the target function.

#### 3.1 Algorithm for block thresholding

A time-frequency block thresholding regularizes the power subtraction estimation by calculating a single attenuation factor over time-frequency blocks. The time-frequency plane  $\{l,k\}$  is segmented in  $I$  blocks  $B_i$  whose shape may be chosen arbitrarily. The signal estimator  $x$  is calculated from the noisy data  $y$  with a constant attenuation factor  $a_i$  over each block  $B_i$ .

$$\hat{x}[n] = \sum_{i=1}^I \sum_{(l,k) \in B_i} a_i Y[l,k] s_{l,k}[n] \quad (6)$$

To understand how to compute each  $a_i$ , one relates the risk  $r = E\{\|x - \hat{x}\|^2\}$  to the frame energy conservation and obtains

$$r = E\{\|x - \hat{x}\|^2\}$$

To minimize an upper bound of the quadratic estimation risk

$$r \leq \frac{1}{A} \sum_{i=1}^I \sum_{(l,k) \in B_i} E\{|a_i Y[l,k] - X[l,k]|^2\} \quad (7)$$

Since  $Y[l,k] = X[l,k] + d[l,k]$  one can verify that the upper bound of (7) is minimized by choosing

$$a_i = 1 - \frac{1}{\varepsilon_i + 1} \quad (8)$$

where  $\varepsilon_i = \frac{\overline{X_i^2}}{\sigma_i^2}$  is the average a priori SNR in  $B_i$ . It is calculated from

$$\overline{X_i^2} = \frac{1}{B_i^\#} \sum_{(l,k) \in B_i} |X[l,k]|^2$$

and

$$\sigma_i^2 = \frac{1}{B_i^\#} \sum_{(l,k) \in B_i} |\sigma[l,k]|^2$$

which are the average signal energy and noise energy in  $B_i$ , and  $B_i^\#$  is the number of coefficients  $(l,k) \in B_i$ . The oracle block attenuation coefficients  $a_i$  in (8) cannot be calculated because a priori SNR  $\varepsilon_i$  is unknown. Cai and Silverman [14] introduced block thresholding estimators that estimate the SNR over each  $B_i$  by averaging the noisy signal energy

$$\hat{\varepsilon}_i = \frac{\overline{Y_i^2}}{\sigma_i^2}$$

where  $\overline{Y_i^2} = \frac{1}{B_i^\#} \sum_{(l,k) \in B_i} |Y[l,k]|^2$

The resulting attenuation factor  $a_i$  is calculated with a power subtraction estimator by

$$a_i = \left(1 - \frac{\lambda}{\hat{\varepsilon}_i + 1}\right)_+ \quad (9)$$

where  $\hat{\varepsilon}_i$  is an unbiased estimator of  $\varepsilon_i$ . A block thresholding estimator can thus be interpreted as a nondiagonal estimator derived from averaged SNR estimators over blocks. Each attenuation factor is calculated from all coefficients in each block, which regularizes the time-frequency coefficient estimation.

#### 3.2 Block Thresholding in Short-Time Fourier Frames

The time-frequency block thresholding can be applied directly with short-time Fourier frames. Some specifications about choice of parameters are discussed below.

##### 3.2.1 Choice of Block

We group time-frequency contiguous short-time Fourier coefficients in disjoint rectangular blocks. The block size is  $B_i^\# = L_i \times W_i$ , where  $L_i$  and  $W_i$  are respectively the block length in time and the block width in frequency. For simplicity, dyadic lengths  $L_i = 8, 4, 2$  and widths  $W_i = 16, 8, 4, 2, 1$  will be used (the unit being the time-frequency index in spectrogram).

##### 3.2.2 Choice of Thresholding Level $\lambda$

One can tolerate the thresholding level  $\lambda$  by the given values of choice of block size and the residual noise probability level  $\delta$ . For each block width and length,  $\lambda$  is estimated with  $\delta = 0.1\%$ . For a block width  $W > 1$ , blocks that contain same number of coefficients  $B_i^\# = L \times W$  have close  $\lambda$  values. For adaptive block thresholding the value of  $\lambda$  has been chosen as

1.5. Here the size of the macro block is set to be equal to the maximum block size  $8 \times 16$ .

### 3.3 Adaptive Block Thresholding

A block thresholding segments the time-frequency plane in disjoint rectangular blocks of length  $L_i$  in time and width  $W_i$  in frequency. The adaptive block thresholding chooses the sizes by minimizing an estimate of the risk. The risk  $E\{\|x - \hat{x}\|^2\}$  can be estimated with a Stein risk estimate [15].

$$r = E\{\|x - \hat{x}\|^2\}$$

$$r \leq \frac{1}{A} \sum_{i=1}^I \sum_{(l,k) \in B_i} E\{|a_i Y[l, k] - X[l, k]|^2\} \quad (10)$$

To estimate the block thresholding risk Cai [6] uses the Stein estimator of the risk when computing the mean of a random vector, which is given by Stein theorem [15].

### 3.4 Stein Unbiased Risk Estimate (SURE)

Let  $Y = (Y_1, \dots, Y_p)$  be a normal random vector with the identity as covariance matrix and mean  $X = (X_1, \dots, X_p)$ . Let  $Y + h(Y)$  be an estimator of  $X$ , where  $h = (h_1, \dots, h_p)$ , then

$$R = E\|Y + h(Y) - X\|^2 = p + E\{\|h(Y)\|^2 + 2\sum h_i(Y)\} \quad (11)$$

So

$$\hat{R} = p + \|h(Y)\|_2^2 + 2\sum h_i(Y) \quad (12)$$

is an unbiased estimator of the risk  $R$  of  $Y+h(Y)$ , called Stein unbiased risk estimator [15]. The adaptive block thresholding groups coefficients in blocks whose sizes are adjusted to minimize the Stein risk estimate and it attenuates coefficients in those blocks.

To regularize the adaptive segmentation in blocks, the time-frequency plane is first decomposed in macroblocks. Each macroblock is segmented with 15 possible block sizes  $L \times W$  with a combination of block length  $L=8, 4, 2$  and block width  $W=16, 8, 4, 2, 1$ . The size of macroblocks is set to be equal to the maximum block size  $8 \times 16$ . In particular adaptive block thresholding eliminates pre-echo artifacts on signal onsets and results in less signal distortion.

## 4 RESULTS AND DISCUSSION

A number of experiments have been performed on various music signals. The adaptive block attenuation performs well against the conventional thresholding operators. It counters the effects of musical noise that is present in diagonal and non-diagonal estimators.

The results presented below have been performed on various types of audio signals: "Mozart" is a musical signal that contains respectively quick notes played by a solo oboe and by some drums; "TIMIT-M" is a male utterances taken from the TIMIT database. It is sampled at 16 kHz whereas Mozart is sampled at 11 kHz. They were corrupted by white Gaussian noise of different amplitude. Short-time Fourier

transforms with half-overlapping windows is used. These windows are square root of Hanning windows of size 50 ms for "Mozart", and 20 ms for "TIMIT-M".

TABLE 4.1 COMPARISON OF PERFORMANCE OF DE-NOISING

S I. N O	Audio Signal	Without Empirical Wiener		With Empirical Wiener	
		SNR of the noisy signal	SSNR of the de-noised signal	SNR of the noisy signal	SSNR of the de-noised signal
1	Mozart Signal	5.10	14.38	5.10	15.13
2	TIMIT-M Signal	0.49	13.07	0.49	13.98
3	Crumbling paper2	0.91	2.44	0.94	2.84
4	Electric Static Signal	0.29	0.51	0.32	0.62
5	TIMIT-M & F Signal	14.26	14.40	14.28	15.17
6	Piano Signal	2.09	8.07	2.06	8.53

In our simulations, we used a clean speech which was artificially corrupted with white Gaussian noise. Table 4.1 compares the performances of the classical de-noising scheme based on Wiener filtering and the block thresholding approach for different values of SNR. One can observe that the nondiagonal Wiener postprocessing based algorithms achieved systematically a better SNR than the adaptive block thresholding algorithm. For all the signals the SNR value of the noisy signal is same, but the segmented SNR value of the de-noised signal is improved better than the block thresholding method. This fact is predictable since we used spectral and perceptual considerations to enhance speech.

Although the improvement in term of WSS, the de-noising approach didn't reach the original quality (between clean and noisy). However, it is well improved when compared to that of de-noised speech using Wiener. Spectrograms are considered. The noisy speech signal is a speech sequence corrupted by a Gaussian noise whose SNR = 10 dB. It is worth pointing out that the de-noised signal by the classical method is affected by a musical noise (isolated points randomly distributed in time and frequency). The amount of such noise is dramatically reduced by the proposed approach.

The fig. 1 shows the time waveform of original signal, noisy signal and the de-noised signal. These results indicate that considerable noise rejection has been achieved. From these figures, it can be seen that the enhanced signal becomes very close to the original signal.

In order to preserve the high frequency components during the de-noising process, block thresholding is applied. The length of the block is chosen as  $L=8$  and the width of the block is chosen as  $W=16$  with the thresholding level  $\lambda=1.5$ .

The desired signal is obtained by taking the inverse Short Time Fourier Transform.

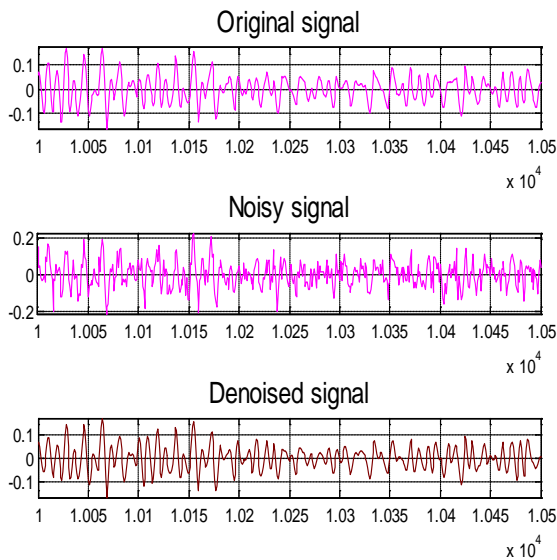


Fig. 1 Time waveform of original signal, noisy signal and enhanced signal

## 5 CONCLUSION


This paper introduces a time-frequency block-thresholding algorithm that adapts all parameters to the time-frequency regularity of the audio signal. The adaptation is performed by minimizing a Stein unbiased risk estimator calculated from the data. It eliminates the residual noise artifacts and preserves the transients of the signal. The resulting algorithm is robust to variations of signal structures such as short transients and long harmonics. The performance was demonstrated using short time spectra with and without noise suppression. Results indicate the overall significant improvements in the quality of audio signal.


## REFERENCES

- [1] Boll S, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transaction on Acoustics, Speech, Signal Processing, vol. ASSP-27, no. 2, pp. 113-120, 1979.
- [2] Cappe, O, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Transactions on Speech, Audio Processing, vol. 2, pp. 345-349, 1994.
- [3] Ephraim, Y, and Malah, D, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," IEEE Transactions on Acoustics, Speech, Signal Processing vol. ASSP-33, no. 2, pp. 443-445, 1984.
- [4] Ephraim, Y, and Malah, D, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," IEEE Transactions on Acoustics, Speech, Signal Processing, vol. 32, no. 6, pp. 1109-1121, 1985.
- [5] Ephraim, Y, Lev-Ari, H, and Roberts, W. J. J, "A brief survey of speech enhancement," in The Electronic Handbook. Boca Raton, FL: CRC Press, 2005.
- [6] Ephraim, Y, and Trees, H.L.V, "A signal subspace approach for speech enhancement," IEEE Transactions on Speech Signal Processing, vol. 3, no. 4, pp. 251-266, Jul. 1995.
- [7] Ghanbari, Y, and Karami, M. R, "A new approach for speech enhancement

- based on the adaptive thresholding of the wavelet packets," Speech Communication, vol. 48, no. 8, pp. 927-940, 2006.
- [8] Ghael, S, Sayeed, A, and Baraniuk, R, "Improved wavelet de-noising via empirical Wiener filtering," Processing, SPIE, Math. Imag., San Diego, 1997.
- [9] Israel Cohen, "On the decision-directed estimation approach of Ephraim and Malah," IEEE Proc. ICASSP, Montreal, Canada, 17-21 May 2004.
- [10] Joachim, T, and Peter, K, "Noise Suppression using a Perceptual Model for Wideband Speech Signals," Proc. Biennial Symposium Communication. (Kingston, ON), pp. 516-519, June 2002.
- [11] Mallat S. "A Wavelet Tour of Signal Processing", 2nd ed. New York: Academic, 1999.
- [12] Matz, G, Hlawatsch, F, and Raidl, A, "Signal-adaptive robust time-varying Wiener filters: Best subspace selection and Statistical analysis," Proc. IEEE ICASSP-01, pp. 3945-3948, May 2001.
- [13] McAulay R. J. and Malpass M. L. (1980), "Speech enhancement using soft decision noise suppression filter," IEEE Transactions on Acoust., Speech, Signal Processing, vol. ASSP-28, no. 2, pp. 137-145.
- [14] Seok, J.W, and Bae, S, "Speech enhancement with reduction of noise components in the wavelet domain," in Proc. IEEE international conference on acoustics, speech and signal Processing Vol. 2, pp. 1323-1326, 1997.
- [15] Silverman, B. W, and Cai, T, "adaptive wavelet estimation: a block thresholding and oracle inequality approach," The annuals of Statistics, vol. 27, pp. 898-924, 1999.
- [16] Sik Park, Y, and JoonHyuk, C, "A novel approach to a robust a priori SNR estimator in speech enhancement," IEICE Transactions communication, vol. E-90B, no. 8, 2007.
- [17] Sorensen, K. V, and Andersen, S. V, "Speech enhancement with natural sounding residual noise based on connected time-frequency Speech presence regions," EURASIP Journal, Applied Signal Processing, vol. 18, no. 18, pp. 2954-2964, 2005.

## AUTHORS PROFILE

	<p><b>Mrs S.SULOCHANA</b> Received the B.E. degree from Francis Xavier Engineering College, Tirunelveli, Tamil nadu, India and the M.E. degree in Communication Systems from PET Engineering College, Tirunelveli, Tamil Nadu, India. She is currently working as Assistant Professor, Department of ECE, Dr.Sivanthi Aditanar College of Engineering, Tiruchendur, Tuticorin District, Tamil Nadu, India. Her research interests include Signal Processing, Wave Propagation Techniques.</p>
--	---

	<p><b>Mr S.SENTHIL SIVA MANIKANDAN</b> Received the B.E. degree from Sree Sowdambika College of Engineering, Aruppukottai, Tamil nadu, India and the M.E. degree in Applied Electronics from SCAD College of Engineering and Technology, Tirunelveli, Tamil Nadu, India. He is currently working as Assistant Professor, Department of ECE, Dr.Sivanthi Aditanar College of Engineering, Tiruchendur, Tuticorin District, Tamil Nadu, India. His research interests include Wave Propagation, communication and VLSI design.</p>
--	--