EFFICIENT HARD AND SOFT THRESHOLDING FOR WAVELET SPEECH ENHANCEMENT

M. S. Arefeen Zilany, Md. Kamrul Hasan and M. Rezwan Khan
Dept. of Electrical and Electronic Engg., Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh
Tel: +88 02 8611594; fax: +88 02 8613046
e-mail: khasan@eee.buet.edu

ABSTRACT

This paper presents an efficient thresholding technique for wavelet speech enhancement. The signal-bias compensated noise level is used as the threshold parameter. The noise as well as signal level is estimated from the detail wavelet packet (WP) coefficients in the first scale. Both hard and soft thresholding are applied successively. The regions for hard thresholding are identified by estimating their signal to noise ratio (SNR) in the wavelet domain. Soft thresholding is applied to the rest of the regions. The performance of the proposed scheme is evaluated on speech recorded in real conditions with artificial noise added to it.

1 INTRODUCTION

Speech enhancement plays a key role in designing robust automatic speech and speaker recognition systems. As the presence of noise in a speech deteriorates the performance of the recognition systems, several approaches for speech enhancement in additive noise have been proposed [1]-[5]. The spectral subtraction based approaches have been studied by many researchers for the enhancement of speech degraded by additive uncorrelated white noise [2], [3]. The basic idea is to restore the magnitude spectrum or power spectrum of a signal observed in additive noise through subtraction of an estimate of the average noise spectrum from the noisy signal spectrum. The advantage of the spectral subtraction method is its simplicity. However, the main problem in spectral subtraction is the processing distortions caused by the random variations of the noise spectrum and the use of noisy phase. Methods for speech enhancement have also been developed based on extraction of parameters from noisy speech, and synthesizing speech from these parameters [1], [5].

Wavelet transform has recently been evolved as a powerful tool for removing noise from speech and image signal. Bahoura and Rouat [6] have recently proposed a wavelet speech enhancement technique using the teager energy operator. The main idea was to define a discriminative threshold in various scales as a function of speech components. Setting of a threshold criterion requires an accurate estimate/knowledge of the additive noise level in the noisy speech. The higher frequency region of the wavelet coefficients mostly contains noise from which noise level is usually estimated [7]. The effect of the signal components present in this region may be insignificant at low SNR but is not negligible particularly at high SNR. For this reason, this method shows deteriorating performance at a relatively high SNR. As for example, a signal having SNR of 20 dB deteriorated to an SNR of 16.47 dB as shown in Table I of [6].

In this research, we propose an efficient wavelet speech enhancement method applying both hard and soft thresholding successively. Regions in wavelet domain, where average signal strength is less than that of noise, hard thresholding is applied by forcing the WP coefficients to be zero. For the rest of the regions a modified soft thresholding criterion is introduced to further reduce the noise level. Unlike other conventional techniques, we incorporate the effect of the trace of the signal remaining at the high frequency region of the wavelet packet (WP) coefficients of the degraded speech. A method for estimating this signal level is addressed. Using the estimated signal level in the detail WP coefficients, the bias compensated noise level is obtained from the median absolute deviation (MAD) of the detail coefficients of the degraded speech. The value thus obtained is used as the threshold parameter.

2 THEORY

In general, the measurements of a clean speech signal $s(n)$ are corrupted by noise. Usually, the noise $v(n)$ is modeled as an additive white Gaussian process with zero-mean and variance σ_v^2. The noisy speech signal $x(n)$ is then given by

$$x(n) = s(n) + v(n), \quad n = 1, 2, \cdots, N \tag{1}$$

The objective of this research is to extract the speech signal $s(n)$ from the degraded observed signal $x(n)$ by applying new thresholding techniques in the wavelet domain.

For a given level j, the wavelet packet (WP) transform decomposes the noisy signal $x(n)$ into 2^j subbands

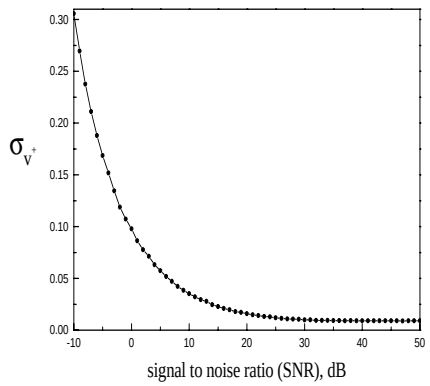

Figure 1: Variation of σ_{v^+} with SNR

corresponding to wavelet coefficients sets $X^j_{k,m}$ as given by [9]

$$X^j_{k,m} = WP\{x(n), j\} \tag{2}$$

In other words, $X^j_{k,m}$ represents the mth coefficient of the kth subband, where $m = 1, 2, \cdots, N/2^j$ and $k = 1, 2, \cdots, 2^j$. For this application, WP decomposes the given signal at level 4 over which the proposed method is applied. But in estimating the noise level, WP transform of the degraded speech signal at the first scale is used.

Donoho and Johnstone [7] proposed the noise level as the median absolute deviation (MAD) of the wavelet coefficients at the finest level divided by 0.6745, i.e.,

$$\sigma_{v^+} = MAD/0.6745 \tag{3}$$

Usually, the coefficients at the finest level are predominantly noise. Because of the presence of a small fraction of signal coefficients we get an estimate of noise that suffers an upward bias [7]. To signify this fact the subscript "v^+" is used instead of just "v" in (3).

Fig. 1 shows the variation of σ_{v^+} with SNR for a given speech degraded by noise of different power. It is interesting to observe that σ_{v^+} shows asymptotically flat behaviour for higher values of SNR. This indicates that the detail coefficient region contains signal whose MAD corresponds to the asymptotic value. Applying σ_{v^+} as a threshold parameter removes some of the signal components which have significant adverse effect at high SNR values.

At a high SNR, all existing methods concerning denoising of the speech signal encounter a strong drawback of SNR reduction of the denoised speech. . As for example, as shown in Table 1 of [6] the SNR of the enhanced speech is found to be 14.47 and 16.47 dB while the original SNR of the noisy speech was 15 and 20 dB, respectively. This is due to the undesired inclusion of signal components in σ_{v^+}. This justifies our observation on the behaviour of σ_{v^+} depicted in Fig. 1 and suggests that introduction of a correction factor in σ_{v^+} is necessary to make the threshold value more effective.

2.1 Calculation of corrected value of noise level

Unlike other approaches [6], [8], the noise level used as the threshold parameter is estimated taking into account the signal remaining at the finest level. For Gaussian distribution of a noise sequence, the noise level is correctly defined as the MAD of the wavelet coefficients at the finest level divided by 0.6745 as defined in (3). It has been observed that when two noise sequences are added together, the MAD of the resultant sequence at the finest level varies as the square root of the sum of their individual (MAD at the finest level) squared values. This is true for random noise sequences only.

The distribution of signal coefficients remaining at the finest level differs from that of the random noise. Hence, the addition of noise to the signal makes the MAD at the finest level to deviate from its behaviour when a random noise sequence is added to another random noise sequence. This deviation can be used as an approximate measure of the signal remaining at the finest level.

Initially $\sigma_{v^+}(0)$ is calculated from the noisy data sequence. To obtain a better estimate of the actual noise level, we generate m sets of known noise sequences and add them in succession to the given noisy speech, where m is a convenient value ($m = 20$ is a good value). Then $\sigma_{v^+}(i)$ is calculated at each step. We calculate another noise parameter, $\sigma_v(i)$, assuming $\sigma_{v^+}(0)$ to be the initial noise level, using the following relation $\sigma^2_v(i) = \sigma^2_{v^+}(0) + \sigma^2_{add}(i)$, where $\sigma^2_{add}(i)$ is the added noise power at i-th step. The deviation which is assumed to be an approximate measure of the signal level ($\hat{\sigma}_s$) at the finest region is then computed as

$$\hat{\sigma}_s = \sqrt{\frac{\sum_{i=1}^{m}[\sigma_{v^+}(i)]^2 - \sum_{i=1}^{m}[\sigma_v(i)]^2}{m}} \tag{4}$$

Once $\hat{\sigma}_s$ is obtained, the correction of noise level given in (3) is made as

$$\sigma_{vc} = \sqrt{\sigma_{v^+}(0)^2 - \hat{\sigma}^2_s} \tag{5}$$

where σ_{vc} denotes the corrected value of $\sigma_{v^+}(0)$.

2.2 Thresholding WP coefficients

In this paper, we simultaneously apply both hard and soft thresholding to devise an improved denoising method. Regions in wavelet domain, where average signal strength is less than that of noise, hard thresholding is applied by forcing the coefficients to be zero. After accomplishing hard thresholding, soft thresholding is applied over the rest of the regions to further reduce the noise level. Details of the thresholding techniques are described in the following.

2.2.1 Application of hard thresholding

The wavelet packet coefficients at a particular level is first divided into a number of blocks consisting of convenient number of consecutive WP coefficients. Then

Table 1: Comparison of actual and estimated values of the signal level ($\hat{\sigma}_s$) in detail WP coefficients at the first level for S2 (refer to the result section). The values of σ_{v^+} and σ_{vc} are also presented.

SNR	Actual signal level	Estimated $\hat{\sigma}_s$	σ_{v^+}	σ_{vc}
-10	0.0090	0.0436	0.2994	0.2962
-5	0.0090	0.0326	0.1708	0.1677
0	0.0090	0.0182	0.0979	0.0962
5	0.0090	0.0143	0.0575	0.0557
10	0.0090	0.0128	0.0354	0.0330
15	0.0090	0.0121	0.0229	0.0194
20	0.0090	0.0118	0.0159	0.0107
25	0.0090	0.0114	0.0123	0.0046
30	0.0090	0.0109	0.0115	0.0036

hard thresholding is applied to a block of WP coefficients where average signal power is less than the average noise power as that block essentially contributes more noise than signal in the denoised speech. To identify the blocks for hard thresholding, a window of convenient length is slided over the whole range. The hard thresholding used in this paper is defined as

$$\bar{X}^j_{k,w} = \begin{cases} X^j_{k,w}, & \text{if } P^w_x \geq 2P^w_v \\ 0, & \text{if } P^w_x < 2P^w_v \end{cases} \tag{6}$$

where $w = m$ to $m+l-1$, l is the length of the window and P^w_x $(= P^w_s + P^w_v)$ represents the total power of the WP coefficients inside a given window and P^w_v denotes the power of the noise component over the same window. An estimated value of P^w_v can be obtained using the relation $P^w_v = l{\sigma_{vc}}^2$ and P^w_x can be estimated simply by taking the sum of the squared value of the WP coefficients for that given window.

2.2.2 Application of soft thresholding

The hard thresholding described in the preceding section eliminates a significant portion of noise from the regions of wavelet coefficients where noise dominates signal. The rest of the regions where signal strength is higher than that of noise, soft thresholding is applied for further enhancement of the noisy signal. As the noise power uniformly penetrates into the actual signal in wavelet domain, subtraction of noise power from the signal power is expected to improve the SNR of the enhanced signal. The coefficients with power less than the average noise power are more susceptible to distortion; their amplitudes are reduced proportionately and the soft thresholding (T1) applied in this paper is defined as

$$\tilde{X}^j_{k,m} = \begin{cases} sign(\bar{X}^j_{k,m})\sqrt{|(\bar{X}^j_{k,m})^2 - \sigma^2_{vc}|}, & \text{if } |\bar{X}^j_{k,m}| \geq \sigma_{vc} \\ \frac{\bar{X}^j_{k,m}|\bar{X}^j_{k,m}|}{\sigma_{vc}}, & \text{if } |\bar{X}^j_{k,m}| < \sigma_{vc} \end{cases} \tag{7}$$

It may be mentioned here that the amplitude subtraction based soft thresholding (T2) technique is defined as [8], [9]

$$\tilde{X}^j_{k,m} = \begin{cases} sign(\bar{X}^j_{k,m})(|\bar{X}^j_{k,m}| - \sigma_{vc}), & \text{if } |\bar{X}^j_{k,m}| \geq \sigma_{vc} \\ 0, & \text{if } |\bar{X}^j_{k,m}| < \sigma_{vc} \end{cases} \tag{8}$$

A comparison of the performance of the two soft thresholding techniques defined by (7) and (8) is presented in Table 2.

2.3 Reconstruction of the original signal

The enhanced signal is synthesized with the inverse transformation WP^{-1} of the modified wavelet packet coefficients ($\tilde{X}^j_{k,m}$) [9], i.e.,

$$\hat{s}(n) = WP^{-1}(\tilde{X}^j_{k,m}, j). \tag{9}$$

3 RESULTS

Two clean speeches, namely, "Should we chase those cowboys?" termed as S1 and "She had your dark suit in greasy wash water all year" termed as S2 from the TIMIT database are used in the simulation for comparing the proposed method with the one described in [6]. The speech signals are sampled at 8 kHz. This generates a total of 21248 samples for S1 and 22221 for S2. Then computer generated white noise sequences are added to the clean speech signals for obtaining different SNRs. Instead of segmenting the given sequence of the speech signal, WP transform (at level 4) is applied over the full range of the data samples.

First, we estimate the SNR of the given signal from the detail coefficients in the first scale. Then the corrected value of σ_{v^+}, i.e., σ_{vc}, is calculated from the detail coefficients of the noisy speech by adding auxiliary white noise sequences such that it results an incremental SNR of -0.5 dB. 20 such points are taken to calculate $\hat{\sigma}_s$. Estimated results of the signal level $\hat{\sigma}_s$ in the detail WP coefficients at the first scale for different SNRs are presented in Table 1 along with σ_{v^+} and σ_{vc}. It can be seen that the estimated value of $\hat{\sigma}_s$ is fairly close to the actual one. Since σ_{vc} determines the threshold level for the noisy WP coefficients, an over-estimation of σ_{vc} would have an adverse effect on the denoised speech. In particular, Table 1 shows that the biased noise level σ_{v^+} is significantly higher than the proposed corrected noise level σ_{vc} at a relatively high SNR.

The clean samples of each of the speech signal are corrupted by additive white noise for various SNRs ranging from -10 to 30 dB. The noisy speech signals are then denoised using the proposed technique. For hard thresholding, a window size of 40 samples is chosen and is slided over the whole WP coefficients at level 4 of the given signal. Application of hard thresholding in the first stage eliminates the predominantly noisy portions in wavelet coefficients domain. The soft thresholding is

Table 2: Results on SNR improvement for real speech (S1 and S2)

SNR (dB)	WP Ref. [6], dB		Proposed WP with T2, dB		Proposed WP with T1, dB	
	S1	S2	S1	S2	S1	S2
−10	−1.28	−1.65	2.15	1.63	1.75	1.37
−5	2.35	1.57	4.54	3.95	4.52	3.85
0	6.01	5.30	7.59	6.64	8.15	6.82
5	9.65	8.91	10.13	9.58	10.78	10.04
10	13.15	12.16	13.71	12.49	14.59	13.34
15	15.99	13.93	17.67	15.92	18.51	17.01
20	18.15	16.05	21.25	20.03	22.37	21.39
25	20.16	17.81	25.88	25.45	26.77	25.63
30	22.06	18.89	28.49	26.89	30.14	29.89

then applied to further enhance the signal quality. The average results of 25 independent runs for each SNR are shown in Table 2. For comparison, the results obtained using a recent method described in [6] are also included in the table. It is evident that the proposed method gives better results than the previous one for all SNRs. However, the results show that the power subtraction based thresholding (T1) performs better than the amplitude subtraction based thresholding (T2). Also notice that the proposed method with T1 prevents the undesired fall of SNR of the denoised speech even when the original signal has a high SNR of 30 dB.

Fig. 2 shows the degraded speech $x(n)$ for SNR = 0 dB and the corresponding enhanced speech resulting from the wavelet packet method described in [6] and the denoising method proposed in this paper. The noise-free speech $s(n)$ is also plotted along with the enhanced speeches for comparison. It is apparent from Fig. 2 that the proposed method eliminates noise in a better way from both the speech-absent and speech-present regions.

4 CONCLUSION

A method for reducing upward bias in the threshold parameter using MAD of WP coefficients for speech enhancement has been investigated. New hard and soft thresholding strategies for WP coefficients have also been proposed using the corrected thresholding parameter defined in this paper. For both very strong and low noise levels, the proposed method shows significant improvement in SNR than the very recent results reported in [6]. In practice, the SNR level of any observed signal is unknown. The noise and signal power estimation scheme proposed here can be used for estimating the SNR value for further processing of the noisy speech signal. The results of this paper may be used as a preprocessor for designing robust speech and speaker recognition systems.

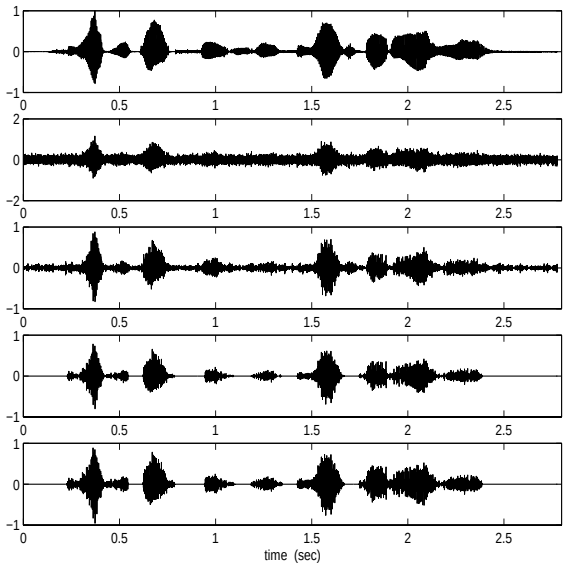

Figure 2: Speech (S2) enhancement results: (a) clean signal; (b) noisy version, (SNR=0 dB); (c) enhanced with WP method described in [7]; d) enhanced with proposed WP using T2; (e) enhanced with proposed WP method using T1.

References

[1] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust. Speech Signal Processing*, Vol. ASSP-26, pp. 197-210, 1978.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. of IEEE*, Vol. 67, No. 12, pp. 1586-1604, 1979.

[3] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. of IEEE*, Vol. 80, No. 10, pp. 1526-1555, 1992.

[4] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy Speech enhancement using discrete cosine transform," *Speech Communication*, Vol. 24, pp. 249-257, 1998.

[5] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, Vol. 28, pp. 25-42, 1999.

[6] M. Bahoura, J. Rouat, "Wavelet speech enhancement based on the teager energy operator," *IEEE Signal Processing Letters*, Vol. 8, No. 1, pp. 10-12, January 2001.

[7] D. L. Donoho, I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrica*, Vol. 81, no. 3, pp. 425-455, 1994.

[8] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, Vol. 41, pp. 613-627, May 1995.

[9] J. C. Goswami, A. K. Chan, *Fundamentals of Wavelets: theory, algorithms and applications,* New York: John Wiley and Sons Inc., 1999.