A New Approach for Wavelet Speech Enhancement

Mohammed Bahoura, Jean Rouat

ERMETIS, DSA, Université du Québec à Chicoutimi,
555 boul. de l'Université, Chicoutimi, Québec, Canada G7H 2B1.
http://wwwdsa.uqac.uquebec.ca/ermetis

Abstract

We propose a new approach to improve the performance of speech enhancement techniques based on wavelet thresholding. First, *space–adaptation* of the threshold is obtained by extending the principle of the *level–dependent* threshold to the Wavelet Packet Transform (WPT). Next, the *time–adaptation* is introduced using the Teager Energy Operator (TEO) of the wavelets coefficients. Finally, the *time–space* adapted threshold is proposed. Comparisons with the Ephraim and Malah Filter are reported.

1. Introduction

The wavelet theory provides a unified framework for various signal processing applications. They include signal and image denoising, compression, analysis of nonstationary signal, etc.

Recently, a powerful wavelet technique has been proposed for noise reduction [1]. This technique proceeds by thresholding the wavelet coefficients using an estimated discriminatory threshold.

For noisy speech, energies of unvoiced segments are comparable to those of noise. Applying thresholding uniformly to all wavelet coefficients not only suppresses additional noise but also some speech components like unvoiced ones. Consequently, the perceptive quality of the filtered speech is greatly affected.

On the other hand, wavelet transform has been proposed to improve the speech enhancement quality of conventional methods. They comprise the Wiener filtering in the wavelet domain [2], wavelet filter bank for spectral subtraction [3] or coherence function [4, 5].

To prevent the speech quality deterioration during the thresholding process, we propose to adapt the discriminative threshold in space and time. The proposed technique is tested on noisy speech recorded in real environments. Obtained results are closely similar to those from Ephraim and Malah Filter (EMF) [6, 7]. Furthermore, the proposed method does not require *a priori* knowledge of the SNR.

2. Denoising by wavelet thresholding

Wavelet transform (WT) has recently emerged as a powerful tool for noise reduction. The original work of Donoho and Johnstone [1, 8] can be summarized as follows. Let $s(t)$ the noise–free signal and $x(t)$ the signal corrupted with white noise $b(t)$

$$x_i = s_i + b_i \qquad i = 1, \ldots, N \tag{1}$$

This algorithm can be described in three steps

- Wavelet transform of the noisy signal,
- Thresholding the resulting wavelet coefficients,
- Inverse transformation to obtain the denoised signal.

Donoho and Johnstone [8, 9] define the soft thresholding function that has been shown asymptotically near optimal for a wide class of signal corrupted by additive white Gaussian noise [10].

$$T_S(\lambda, w_k) = \begin{cases} sgn(w_k)(|w_k| - \lambda) & \text{if } |w_k| > \lambda \\ 0 & \text{if } |w_k| \leq \lambda \end{cases} \tag{2}$$

where w_k represents the wavelet coefficients.

The authors proposed a universal threshold λ for the WT

$$\lambda = \sigma\sqrt{2\log(N)} \tag{3}$$

where $\sigma = MAD/0.6745$ is the noise level. MAD represents the absolute median estimated on the first scale. In the Wavelet Packets Transform (WPT) case, the threshold becomes:

$$\lambda = \sigma\sqrt{2\log(N\log_2 N)} \tag{4}$$

Johnstone and Silverman [11] studied the correlated noise situation and proposed a *level–dependent* threshold

$$\lambda_j = \sigma_j\sqrt{2\log(N)} \tag{5}$$

with $\sigma_j = MAD_j/0.6745$, and MAD_j represents the absolute median estimated on the scale j. The discriminatory threshold is also defined using other criterion like Minimax and SURE (Stain's Unbiased Risk Estimate) [12, 13].

3. New enhancement method

As pointed out previously, the wavelet thresholding techniques were not successfully applied in speech enhancement. These difficulties are simultaneously associated to the speech signal complexity and to the nature of the noise. To improve their performance, we propose two approaches: 1) extend the concept of the *level–dependent* threshold to the WPT to remove various noise, 2) adapt the thresholds to the speech waveform energy.

The proposed algorithm (Fig. 1) is an extension of the *time* only *adapted* thresholding [14].

3.1. Wavelet packet analysis

The WPT is an extension of the WT. For a given level j, the WPT decomposes the noisy signal $x(n)$ into 2^j subbands corresponding to wavelet coefficient sets $w^j_{k,m}$.

$$w^j_{k,m} = WP\{x(n), j\} \qquad n = 1, \ldots, N \tag{6}$$

In this application, we fix $j = 4$. Thus, $w^4_{k,m}$ defines the m^{th} coefficient of the k^{th} subband. Where $m = 1, ..., N/2^4$ and $k = 1, ..., 2^4$.

This work is partially supported by AUPELF–UREF, FUQAC and CRSNG.

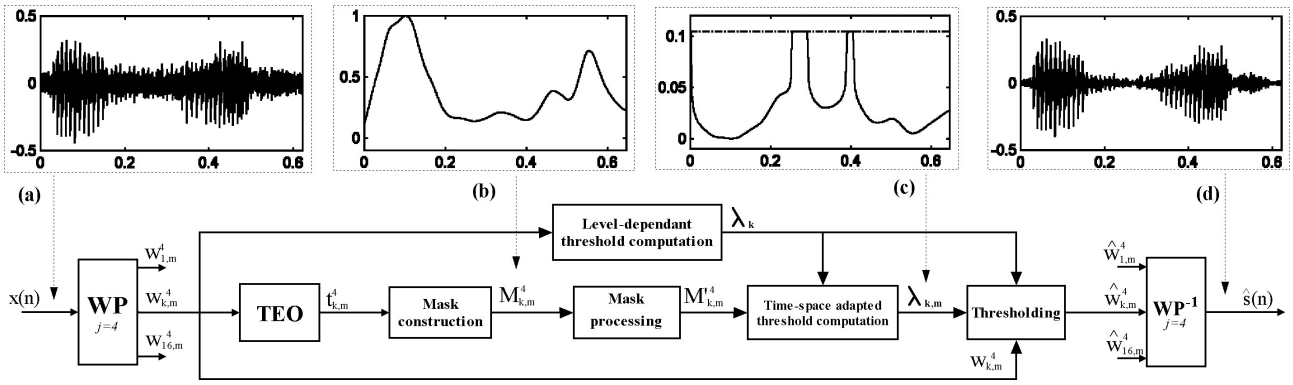

Figure 1: Speech enhancement diagram using *time–space* adapted thresholding in the wavelet packet domain

3.2. Space–adapted threshold

The *space–adapted* threshold is derived from the *level– dependent* threshold (Equation 5). For a given subband k, the corresponding threshold is defined by:

$$\lambda_k = \sigma_k \sqrt{2\log(N)} \qquad k = 1, \ldots, 16 \tag{7}$$

where $\sigma_k = MAD_k/0.6745$ is the noise level and N is the length of the signal. MAD_k represents the absolute median estimated on the subband k.

3.3. Teager energy approximation

The *time–adapting* approach is introduced by using the Teager Energy Operator (TEO) [14]. We applied this operator to the resulting wavelet coefficients $w^4_{k,m}$

$$t^4_{k,m} = [w^4_{k,m}]^2 - w^4_{k,m-1} w^4_{k,m+1} \tag{8}$$

This operation enhances the ability to discriminate speech coefficients from those of noise.

3.4. Masks Construction

We construct an initial mask for each subband k by smoothing the corresponding TEO coefficients (Fig. 1-b)

$$\mathrm{M}^4_{k,m} = t^4_{k,l} * h_k(m) \tag{9}$$

where h_k is an IIR lowpass filter (2^{th} order).

3.5. Time–modulation criterion

For each subband k, the corresponding threshold λ_k should be *time–adapted* only for speech frames and kept unchanged for non speech ones. The speech presence is interpreted by a significant contrast between peaks and valleys of M^4_k, while its absence is observed with a weaker contrast (smoother masks). To distinguish these frames, we define a parameter S^4_k named *offset*, that estimates the valleys level. It is given by the abscissa of the maximum of the amplitude distribution H of the corresponding mask $M^4_{k,m}$, and is estimated over the analyzed frame.

$$S^4_k = abscissa[max(H(M^4_{k,m}))] \tag{10}$$

If S^4_k is below the discriminatory value of $0.35max(M^4_{k,m})$, then we modulate the threshold. Otherwise it remains unchanged.

3.6. Mask processing for the time–adapting threshold

The modulated threshold must be adapted to the speech waveform independently of its energy evolution. In this case, the difference between local maxima must be reduced. We proceed by suppressing the *offset* and normalization, before applying a root power function

$$M'^4_{k,m} = [\frac{M^4_{k,m} - S^4_k}{max(M^4_{k,m} - S^4_k)}]^{\frac{1}{8}} \tag{11}$$

3.7. Time–adapting threshold

For each subband k, the time–space adapted threshold is obtained by adapting the corresponding threshold in the time domain:

$$\lambda_{k,m} = \lambda_k(1 - \alpha M'^4_{k,m}) \tag{12}$$

where λ_k is the *space–dependent* threshold (Equation 7) and α an adjustment parameter $(\alpha = 1)$.

Fig. 1-c represents the *space–dependent* threshold λ_k (dashed line) and the resulting time adapted threshold $\lambda_{k,m}$ (continuous line) for the subband $k = 5$.

3.8. Thresholding process

The soft thresholding (Equation 2) is then applied to the wavelet packet coefficients

$$\widehat{w}^4_{k,m} = T_S(\lambda_a, w^4_{k,m}) \tag{13}$$

where λ_a is the threshold corresponding to the analyzed frame.

$$\lambda_a = \begin{cases} \lambda_{k,m} & \text{if } S^4_k \leq 0.35max(M^4_{k,m}) \\ \lambda_k & \text{if } S^4_k > 0.35max(M^4_{k,m}) \end{cases} \tag{14}$$

3.9. Inverse transformation

The enhanced signal (Fig. 1-d) is synthesized with the back transformation WP^{-1} of the processed wavelet coefficients

$$\widehat{s}_n = WP^{-1}\{\widehat{w}^4_{k,m}, j\} \tag{15}$$

4. Results and discussion

The proposed method is tested and evaluated using speech sound corrupted by white noise and speech recorded in real environments. The speech signals are sampled at 8 kHz.

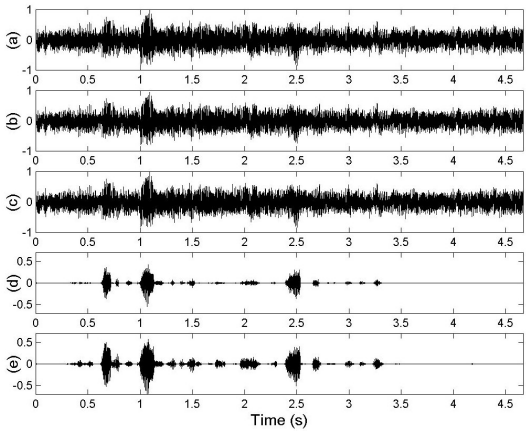

Figure 2: a) noisy speech recorded in an aircraft, enhancement results using b) WPT with a universal threshold, c) *time–adapted* threshold (TAT), d) WPT with *level–dependent*, and e) WPT with *time–space adapted* threshold (TSAT).

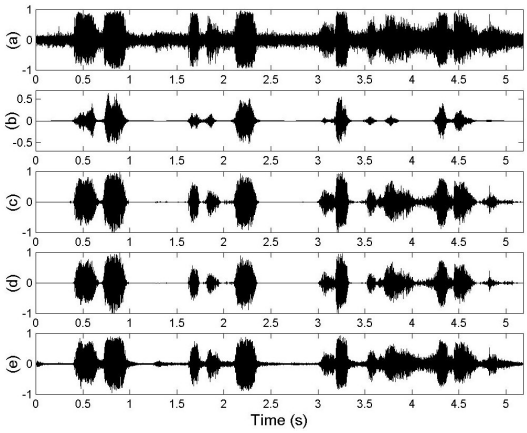

Figure 3: a) noisy speech recorded in a sawmill, enhancement results using b) WPT with a universal threshold (j=7), c) WPT with *time–adapted* threshold (TAT), e) WPT with *time–space adapted* threshold (TSAT), and e) Ephraim and Malah Filter.

4.1. Narrow–band noise

We recall that the wavelet thresholding method has been initially proposed to remove only additional white noise. In this section we use a speech signal corrupted by narrow–band noise (Fig. 2-a), that is far from being white. This example demonstrates , the utility of a *level–dependent* thresholding. In fact, the universal threshold of the WPT (Fig. 2-b) is inefficient to remove this kind of noise. The Time–Adapted Threshold (TAT) is also inefficient (Fig. 2-c). However, the noise is greatly reduced using the *level–dependent* thresholding (Fig. 2-d). The *Time–Space* Adapted Threshold (TSAT) prevents the speech quality deterioration during the thresholding process (Fig 2-e).

4.2. Wide–band noise

In this class, we use a noised speech recorded in a sawmill (Fig. 3-a). The universal threshold method reduces the noise considerably but it is accompanied by speech quality degradation (Fig. 3-b). Our previous solution (TAT) [14] is very efficient to remove this kind of noise (Fig. 3-c). The results obtained by the new approach (TSAT) are also quite efficient (Fig. 3-d), compared to Ephraim and Malah Filter (Fig. 3-e).

Table 1: SNR tests for white noise corrupted speech

Unprocessed (dB)	TAT (dB)	TSAT (dB)	EMF (dB)
-10	2.14	1.99	0.93
-5	4.56	4.35	4.29
0	7.46	7.29	7.06
5	10.28	10.00	9.93
10	12.84	12.54	12.79
15	15.03	14.59	15.58
20	16.94	16.36	18.12

4.3. White noise

The *Time-Adapted* Thresholding (TAT) [14] and the *Time-Space* Adapted Threshold (TSAT) methods are evaluated using a speech signal from the TIMIT database.

Table 1 shows that the proposed methods are well suited to very strong noise with a SNR ranging from -10 to 10 dB and have a better performance than the Ephraim and Malah Filter (EMF).

According to the fact, that universal threshold is asymptotically near optimal to remove this kind of noise, TAT yields a relatively better performance than TSAT (Table 1).

5. Conclusion

The proposed method constitutes a successful application of the wavelet thresholding for speech enhancement. The *space-dependent* threshold allows the removal of various environmental noises (narrow or large frequency–band). Whereas, the *time-adaptation* of the threshold avoids the degradation of speech quality during the thresholding process.

We recall that the EMF algorithm requires an explicit estimation of the noise level or the *a priori* knowledge of the SNR, which is not necessary for our system.

6. References

[1] D.L. Donoho, "Nonlinear wavelet methods for recovering signals, images, and densities from indirect and noisy data," *Proceedings of Symposia in Applied Mathematics*, vol. 47, pp. 173–205, 1993.

[2] D. Mahmoudi, "A microphone array for speech enhancement using multiresolution wavelet transform," in *Proc. Of Eurospeech'97*, Rhodes, Greece, September 1997, pp. 339–342.

[3] T. Gulzow, A. Engelsberg, and U. Heute, "Comparison of a discrete wavelet transformation and nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement," *Signal Processing*, vol. 64, pp. 5–19, 1998.

[4] J. Sika and V. Davidek, "Multi-channel noise reduction using wavelet filter bank," in *EuroSpeech'97*, Rhodes, Greece, Spetember 1997, pp. 2595 – 2598.

[5] D. Mahmoudi and A. Drygajlo, "Combined wiener and coherence filtering in wavelet domain for microphone array speech enhancement," in *ICASSP*, Seattle, USA, 1998, pp. 385 –388.

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

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

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

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

[10] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard., "Wavelet shrinkage: Asymptopia?," *J. Royal Statist. Soc. B.*, vol. 57, no. 2, pp. 301–337, 1995.

[11] I.M. Johnstone and B.W. Silverman, "Wavelet threshold estimators for data with correlated noise," *J. Roy. Statist. Soc. B*, vol. 59, pp. 319–351, 1997.

[12] D.L. Donoho and I.M. Johnstone, "Adapting to unknow smoothness via wavelet shrinkage," *J. Amer. Stat. Assoc.*, pp. 1200–1224, 1995.

[13] X.P. Zhang and M.T. Desai, "Adaptative denoising based on SURE risk," *IEEE Signal Processing Letters*, vol. 5, no. 10, pp. 265–267, October 1998.

[14] M. Bahoura and J. Rouat, "Wavelet speech enhancement using the Teager energy operator," *IEEE Signal Processing Letters*, vol. 8, no. 1, pp. 10–12, January 2001.