# AN IMPROVED WAVELET-BASED SPEECH ENHANCEMENT SYSTEM

*Hamid Sheikhzadeh*[1,2] *and Hamid Reza Abutalebi*[1]

[1]Dept. of Electrical Eng., Amirkabir University of Technology (Tehran Polytechnic)
Hafez Ave, Tehran, Iran.
[2]Dspfactory Ltd., 611 Kumpf Drive, Unit 200, Waterloo, Ont., Canada N2V 1K8
Emails: hsheikh@dspfactory.com abutalebi@cic.aku.ac.ir

## Abstract

The problem of speech enhancement using wavelet thresholding algorithm is considered. Major problems in applying the basic algorithm are discussed and modifications are proposed to improve the method. First, we propose the use of different thresholds for different wavelet bands. Next, by employing a pause detection algorithm, noise profile is estimated and the thresholds are adapted. This enables the modified enhancement system to handle colored and non-stationary noises. Finally, a wavelet-based voiced/unvoiced classification is proposed and implemented that can further improve the performance of the enhancement system. To evaluate the system performance, we have used real-life noise types such as multi-talker babble and low-pass noises. Subjective and objective evaluations show that the proposed system improves the performance the wavelet thresholding algorithm.

## 1. Introduction

In many speech processing applications, speech has to be processed in the presence of undesirable background noise, leading to a need to a front-end speech enhancement. During the last decades, various approaches to the problem have been adopted. Generally the approaches can be classified into two major categories of single-microphone and multi-microphone methods. Although multi-microphone algorithms have acceptable performance in some applications, there are still many practical situations that one is limited to use a single microphone. Among various single-microphone algorithms for speech enhancement, spectral subtraction has been mostly employed. Despite its capability of removing background noise, spectral subtraction introduces additional artifacts known as the musical noise, and is faced by difficulties in pause detection. In recent years, several alternative approaches such as extended and iterative Wiener filtering, HMM-based algorithms [1,2] and signal subspace methods [3] have been proposed for enhancing degraded speech.

In 1995, Donoho [4] introduced wavelet thresholding (shrinking) as a powerful tool in denoising signals degraded by additive white noise. Although the application of wavelet shrinking for speech enhancement has been reported in several works (for example [5,6]), there are many problems yet to be resolved for a successful application of the method to speech signals degraded by various noise types. The main objective of this work is to exploit the features of the speech signal to modify the basic wavelet shrinkage method for this specific application. To objectively validate the proposed modifications, the system performance in the presence of different colored and non-stationary noises is evaluated.

In this paper, we first briefly introduce the basic wavelet thresholding method and various problems encountered in its application to speech signals. Next, in Section 3, we propose improvements to the method. The implementation and results of applying the proposed modifications and the system performance are discussed in Section 4. Finally, the conclusion of this research and some suggestions for future work will be mentioned in Section 5.

## 2. Wavelet thresholding

### 2.1. Denoising by thresholding

Wavelet transform has been intensively used in various fields of signal processing. It has the advantage of using variable size time-windows for different frequency bands. This results in a high frequency-resolution (and low time-resolution) in low bands and low frequency-resolution in high bands. Consequently, wavelet transform is a powerful tool for modeling non-stationary signals like speech that exhibit slow temporal variations in low frequency and abrupt temporal changes in high frequency. Moreover, when one is restricted to use only one (noisy) signal (as in single-microphone speech enhancement), generally the use of the subband processing can result in a better performance. Therefore, wavelet transform can provide an appropriate model of speech signal for denoising applications.

Removing noise components by thresholding the wavelet coefficients is based on the observation that in many signals (like speech), energy is mostly concentrated in a small number of wavelet dimensions. The coefficients of these dimensions are relatively large compared to other dimensions or to any other signal (specially noise) that has its energy spread over a large number of coefficients. Hence, by setting smaller coefficients to zero, one can nearly optimally eliminate noise while preserving the important information of the original signal [4]. Let **y** be a finite length observation sequence of the signal **x** that is corrupted by zero-mean, white Gaussian noise **n** with variance $\sigma^2$,

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \qquad (1).$$

The goal is to recover the signal **x** from the noisy observation **y**. If W denotes a discrete wavelet transform (DWT) matrix, equation (1) (which is in time domain) can be written in the wavelet domain as

$$\mathbf{Y} = \mathbf{X} + \mathbf{N} \qquad (2),$$

where

$$\mathbf{Y} = \mathrm{W}\mathbf{y} \;\;,\;\; \mathbf{X} = \mathrm{W}\mathbf{x} \;\;,\;\; \mathbf{N} = \mathrm{W}\mathbf{n} \qquad (3).$$

Let $\mathbf{X_{est}}$ be an estimate of the clean signal $\mathbf{X}$ based on the noisy observation $\mathbf{Y}$ in the wavelet domain. The clean signal $\mathbf{x}$ can be estimated by

$$\mathbf{x} = W^{-1}\,\mathbf{X_{est}} = W^{-1}\,\mathbf{Y_{thr}} \qquad (4),$$

where $\mathbf{Y_{thr}}$ denotes the wavelet coefficients after thresholding.

The proper value of the threshold can be determined in many ways. Donoho [4] has suggested the following formula for this purpose

$$T = \sigma \sqrt{2 \log(N)} \qquad (5),$$

where T is the threshold value and N is the length of the noisy signal ($\mathbf{y}$).

Thresholding can be performed as *Hard* or *Soft* thresholding that are defined as follows, respectively:

$$THR_H\,(Y,T) = \begin{cases} Y & , |Y| > T \\ 0 & , |Y| < T \end{cases} \qquad (6),$$

and

$$THR_S\,(Y,T) = \begin{cases} \mathrm{Sgn}(Y)\,(|Y| - T) & , |Y| > T \\ 0 & , |Y| < T \end{cases} \qquad (7).$$

### 2.2. Problems in applying wavelet thresholding to speech signals

Some major problems arise when the basic wavelet thresholding method is applied to a complex signal such as speech degraded by real-life noises.

First, the basic method assumes that noise spectrum is white. However, in most practical situations we are faced with colored noises. As a result, the basic form of the wavelet shrinkage does not provide a satisfying speech quality for most of the actual types of noise. Moreover, the method encounters problems in removing non-stationary noises like multi-talker (babble) noise since no time-adaptation mechanism is provided in the algorithm.

The next problem is related to the wavelet shrinking of the unvoiced segments of speech. Since the unvoiced parts of speech contain many noise-like high frequency components, eliminating them in the wavelet domain can severely degrade the quality of the enhanced signal. Use of a unique threshold for all wavelet bands is another disadvantage for speech applications. Hard/Soft thresholding (setting some of the wavelet coefficients to zero) often results in time-frequency discontinuities, observed as "blank areas" in the spectrogram of the enhanced speech. This leads to annoying artifacts and further degradation of output speech. Considering these deficiencies, major refinements to the basic wavelet thresholding algorithm are necessary.

## 3. Proposed modifications

In order to solve some of the problems mentioned in Section 2.2, we have proposed a modified version of the wavelet shrinkage method. The block diagram of the proposed system is illustrated in Fig. 1. Basic components of the improved system will be discussed here.

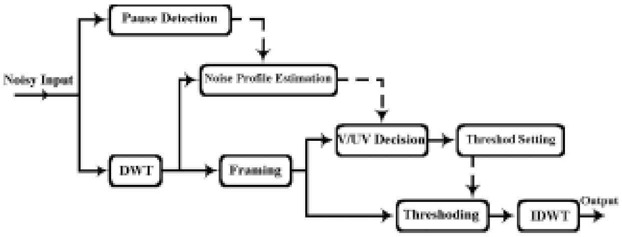

Fig. 1: Block diagram of the proposed system

### 3.1. Pause detection block

Long pauses (lasting a few hundred milliseconds) naturally occur in human speech and can be used to obtain an estimate of the noise profile. Since the noise dynamics is often much (at least ten times) slower than the speech dynamics, we assume that the estimated profile remains constant between two consecutive long pauses. These long segments of pause can be detected in many possible ways. We have already reported a powerful pause detection method that has an acceptable performance in locating long pauses in speech corrupted by diverse environmental noises [7].

### 3.2. Noise profile estimation

During a long pause, one can obtain an estimate of the noise spectrum profile. Specifically, in each wavelet band, we calculate the variance of the noisy coefficients. Using Equation 5, each variance then can then be employed to set the threshold to a value based on the noise energy in the band.

In practical situations, one is often encountered with colored rather than white noises. The proposed noise profile estimation enables the algorithm to cope with colored noises.

Let $\mathbf{c_i}$ be the coefficient sequence of the noise i$^{th}$ wavelet band. Assuming zero-mean Gaussian noise, the coefficients will be Gaussian random variables of zero mean and variance $\sigma_i^2$. The standard deviation $\sigma_i$ is thus estimated by [8]

$$\sigma_i = (1/0.6745)\ \mathrm{Median}(\,|\,\mathbf{c_i}\,|\,) \qquad (8).$$

The set of standard deviation values can now be used as the "noise profile" for threshold setting.

### 3.3. Voiced/Unvoiced decision making

In order to prevent degradation of unvoiced portions in the wavelet shrinking of noisy speech, one has to first classify speech segments into voiced/unvoiced (V/UV) categories. The problem of V/UV decision has been discussed extensively in the literature. Here, we have implemented a simple algorithm based on wavelet transform for V/UV detection. The method uses frequency distribution of the average energy in each time-segment. First, for each wavelet band, the average energy is calculated. By accumulating the average energy of the bands below 2 kHz, we can compute energy of low-bands (EL) of that segment. Similarly, energy of high-bands (EH) of the segment can be calculated by accumulating the average energy of the bands above 2 kHz. The EL to EH ratio (EL/EH) is the fundamental parameter used in our V/UV decision.

When a long pause is detected, the EL to EH ratio for that pause (ELp/EHp) is computed. For the segments between that pause and the consecutive one, ELp/EHp constitutes the decision threshold. If EL/EH is higher than ELp/EHp, the segment is labeled as *Voiced*, otherwise it is considered as

*Unvoiced*. Since this method of V/UV detection considers the noise spectral properties, it is robust to noise variations and works for different noise types.

### 3.4. Threshold adaptation (for UV segments)

The threshold setting based on the noise profile estimation is an important part of the proposed system that enables it to handle non-stationary and colored noises. However, the problem of over-filtering of high-frequency components of the UV segments has to be addressed. Employing the information obtained from the two added blocks (noise profile estimation and V/UV decision), the threshold values can be adapted to alleviate the problem. This adaptation is motivated by the fact that in many speech enhancement systems, the voiced and unvoiced segments are treated differently.

As discussed in Section 3.2, for each band, the threshold value, $T_i$ is calculated by Equation (5) in which we replace $\sigma$ with $\sigma_i$. When a new long pause is detected, the noise profile, and subsequently the threshold value are updated. Generally the voiced segments are more low-pass than the unvoiced segments. Thus, once a voiced segment is detected, we positively bias (increase) the threshold values for high-bands relative to the values for low-bands. On the contrary, for unvoiced segments the threshold values for high-bands are negatively biased. As a result, high-frequency components of the UV speech segments are less probable to be filtered out.

### 3.5. Modification of Hard thresholding algorithm

As another improvement, we have used a refined version of hard thresholding function instead of the standard form of equation (6). More precisely, instead of setting some wavelet coefficients to zero (which causes observable sharp time-frequency discontinuities in the speech spectrogram), we attenuate the coefficients that are smaller than the threshold value in a nonlinear manner to avoid creating abrupt changes.

The standard form of hard thresholding has an input-output characteristic that is drawn by solid line in Fig. 2. In the implemented system, the dashed line curve of Fig. 2 has been used as input-output characteristic. The nonlinear part of this curve can be approximated by an exponential function.

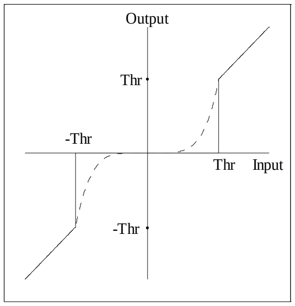

**Fig. 2:** Input-output characteristics for hard thresholding

## 4. Implementation and performance evaluation

The wavelet shrinkage method with proposed modifications was implemented. The V/UV decision procedure was applied to the analysis frames (of 800 samples each, at 16 kHz sampling rate) that overlapped by 50%. The DWT decomposed the input signal into 8 bands (decomposition level equaled to 7). As the pause detector, the one reported in [7] was used. The performance of the proposed V/UV decision block was examined separately. Our extensive observations confirmed an acceptable V/UV separation, even for noisy signals corrupted by colored noises and input signal to noise ratios (SNRs) as low as 5 dB.

The implemented speech enhancement system was evaluated on a set of sentences, spoken by two male and two female speakers. The test data covered major voiced and unvoiced phonemes. The recorded sentences were then corrupted by four types of noise: simulated white, pink and brown noises, and naturally recorded multi-talker babble. These noise types generally occur in real-life applications. Pink and brown noises are two types of low-pass noise and their estimated spectra are shown in Fig. 3.

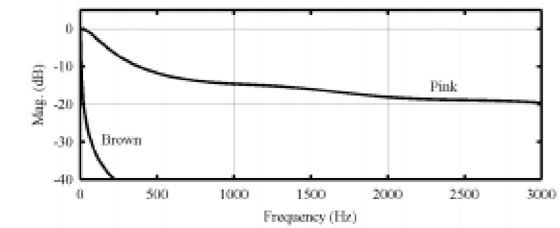

Fig. 3: Estimated spectra for the brown and pink noises

To evaluate the system performance, both objective and subjective tests were employed. In the objective tests, average output SNR of the proposed system was compared to that of the basic wavelet thresholding. Shown in Fig. 4 are the output SNRs for signals degraded by additive white noise at input SNRs of 0, 5, 10, and 15 dB, both for the modified and the basic methods. As demonstrated, due to the applied modifications, the implemented system has a superior performance.

It is well known that the SNR cannot faithfully indicate the speech quality, especially for colored and non-stationary noises. Thus we conducted subjective speech quality tests, employing a preference evaluation similar to one reported in [3]. The tests were performed by a group of 10 listeners, with no previous familiarity with the test material. Each subject participated in two listening sessions. In the first session, for each type of noise, listeners compared the outputs of basic shrinkage algorithm and that of our modified system. In the second session, the comparison was between the output of the modified system and the noisy input signal.

In both sessions, listeners were asked to compare between a pair of speech signals played in random order and vote for one or none of them. Throughout the subjective tests, input SNR was set to 10 dB. Table 1 summarizes the results of the first session comparisons. As shown, for all types of noise, the output of our proposed system has a higher preference percentage compared to the basic algorithm.

The results of the second session tests are shown in Table 2. For all types of noise, more subjects preferred the output of the modified system to the non-processed noisy signal. However, there were cases that some listeners preferred the noisy signal to the enhanced one due to the introduced distortions and artifacts. To summarize, for white and pink noises, the output of the modified system was consistently preferred over the signal enhanced by the basic system and the noisy signal. For the brown noise, the superiority was less consistent, while listeners preferred the enhanced signal to the noisy signal in almost 90% of the trials, the advantage over the basic system was limited to 15%. This is probably due to the fact that the brown noise has most of its energy concentrated at very low frequencies overlapping the frequency components of voiced speech. The multi-talker babble was the toughest cases of all as expected, since it is a speech-like and non-stationary noise.

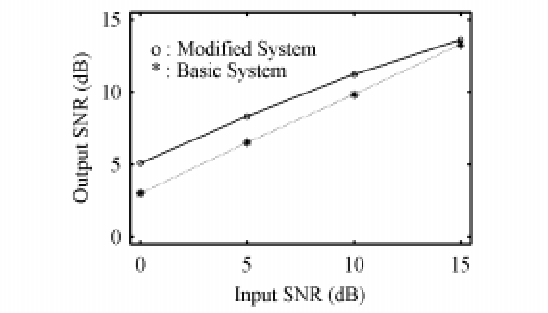

**Fig. 4:** Output SNR for additive white noise

**Table 1:** Preference percentage between outputs of modified and basic methods

| Noise Type | Modified System | Basic System | No Preference |
|---|---|---|---|
| White | 82 % | 10 % | 8 % |
| Pink | 72 % | 18 % | 10 % |
| Brown | 55 % | 5 % | 40 % |
| Multi-talker | 52 % | 30 % | 18 % |

**Table 2:** Preference percentage between output of modified system and noisy signal

| Noise Type | Modified System | Noisy Signal | No Preference |
|---|---|---|---|
| White | 68 % | 20 % | 12 % |
| Pink | 72 % | 18 % | 10 % |
| Brown | 90 % | 5 % | 5 % |
| Multi-talker | 50 % | 28 % | 22 % |

## 5. Conclusion

In this paper the problem of wavelet shrinkage-based speech enhancement was addressed. After examination of basic thresholding method, some modifications were proposed to apply the method for speech enhancement. The basic tool for handling colored noise was the introduction of different thresholds for different bands. The new blocks of pause detection and V/UV decision were added to the system to enable it to cope with the non-stationarities in both speech and noise. Although our simple V/UV classifier has a good performance in different noisy environments, it clearly is not adequate for the phonemes that contain considerable energy in the vicinity of 2-3 kHz. In order to resolve this issue, the V/UV decision rules should be improved. Through pause and V/UV detection, the threshold value could be adapted for the noise type, and also for the speech and noise dynamics. Due to the threshold adaptation, the new system is able to cope better with colored and non-stationary noise types. Both objective and subjective performance evaluations confirm the superiority of the proposed system over the basic shrinking method. We are currently extending our research to achieve to a practically usable system.

## 6. References

[1] Y. Ephraim, “Statistical model based speech enhancement,” *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526-1555, Oct. 1992.

[2] H. Sameti, H. Sheikhzadeh, L. Deng and R.L. Brennan. “HMM-based strategies for enhancement of speech signals embedded in nonstationary noise,” *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 445-455, Sept. 1998.

[3] Y. Ephraim and H.L. Van Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 4, pp. 251-266, July 1995.

[4] D.L. Donoho, “Denoising by soft thresholding,” *IEEE Trans. on Information theory*, vol. 41, no.3, pp. 613-627, May 1995.

[5] J.W. Seok and K.S. Bae, “Speech enhancement with reduction of noise components in the wavelet domain,” in *Proceedings of the ICASSP*, pp. II-1323-1326, 1997.

[6] E. Ambikairajah, G. Tattersall and A. Davis, “Wavelet transform-based speech enhancement,” in *Proceedings of ICSLP*, 1998.

[7] M.H. Ghoreishi and H. Sheikhzadeh, “A hybrid speech enhancement system based on HMM and spectral subtraction,” in *Proceedings of the ICASSP*, pp. III-1855-1858, 2000.

[8] D.L. Donoho and I.M. Johnstone, “Ideal spatial adaptation via wavelet shrinkage,” *Biometrika*, vol. 81, pp. 425-455, Dec. 1994.