2011 International Conference on Computer Communication and Management
Proc .of CSIT vol.5 (2011) © (2011) IACSIT Press, Singapore

A New Method for Threshold Selection in Speech Enhancement by Wavelet Thresholding

Saeed Ayat [+]

[*]Assistant Professor
Department of Information Technology, Payame Noor University
Najafabad, I. R. of Iran

Abstract— Speech is one of the most important signals in multimedia system. Speech enhancement is improving the quality of speech in real noisy environments for these multimedia systems. In this paper, we propose a new threshold selection method for using in speech enhancement by wavelet thresholding. This method updates the threshold value in each frame of noisy speech. The selection of the wavelet threshold value depends on the estimates of the clean speech signal energy that is estimated by using energy of the noise from the noise-only frames. The evaluation results show that the new method achieves higher performance comparison to other famous threshold selection methods for wavelet-based speech enhancement.

Keywords: speech enhancement, wavelet transform, threshold selection.

1. Introduction

The principle under which the wavelet thresholding operates is to improve wavelet coefficients depends on a specific value, called threshold.

The denoising algorithm is summarized as follow [1]:

i) Compute the discrete wavelet transform for noisy signal.

ii) Based on an algorithm, called thresholding algorithm and a threshold value, shrink some detail wavelet coefficients.

iii) Compute the inverse discrete wavelet transform.

In this paper, we develop a new threshold selection technique using the energy estimation of the clean speech signal in each frame.

This paper is organized as follows: In section 2, we review the most famous threshold selection methods. In sections 3 we propose the new energy estimation method. Finally, in section 4, the simulation results of the proposed methods to different noisy speech signals are reported and compared with basic wavelet-based speech enhancement algorithms.

2. Threshold selection

There are many formulas for obtaining threshold value. In this section we review some of the most popular of them. In all these formulas λ is the threshold value.

2.1. Universal method

Donoho and Johnstone derived a general optimal universal threshold for the white Gaussian noise under a mean square error criterion and its side condition that with high probability, the enhanced signal $\hat{f}$ is at least as smooth as the clean signal f [3, 4]. In this method threshold is selected as:

$$\lambda = \hat{\sigma}\sqrt{2\log_e(n)} \qquad (1)$$

In this formula *n* is number of samples in the noisy signal and σ is the standard deviation of noise that is estimated by the relation [2]:

[+] Corresponding author. Tel.: + (983312727121); fax: +(983312727124).
E-mail address: (ayat@ce.sharif.edu).

$$\hat{\sigma} = \left[\frac{median(|Y_{ij}|)}{0.6745} \right] \tag{2}$$

In which $|Y_{ij}|$ is the first level detail coefficients of wavelet transform of noisy speech.

This method tries to optimize the performance as measured by the risk function as:

$$R(\hat{f}, f) = n^{-1} E \left\| \hat{f} - f \right\|_{2,n}^{2} \tag{3}$$

where $\|v\|_{2,n}^{2} = \sum_{i=1}^{n} v_i^2$ denotes the usual squared l_n^2 norm[4].

In practice, the threshold value obtained by this method is not ideal for speech signals due the poor correlation between MSE and subjective quality and the more realistic presence of correlated noise [7].

2.2. Minimax method

In this method that is also proposed by Donoho and Johnstone, it supposed that $Y = N(\mu, 1)$ is the observation, then λ is selected such that minimizes the following relation:

$$\Lambda_n^* = \inf_{\lambda} \sup_{\mu} \left\{ \frac{E(\sigma_{\lambda}(Y) - \mu)^2}{n^{-1} + \min(\mu^2, 1)} \right\} \tag{4}$$

Where $\delta_{\lambda}(Y)$ is the shrink function or thresholding algorithm and *n* is number of signal samples.

A complete explanation of this method can be found in [3] and [6].

2.3. SURE method

SURE or Stein Unbiased Risk is also introduced by Donoho and Johnstone for wavelet de-noising, this method denoises wavelet coefficients so that the mean squared error is minimized, where MSE is estimated by Stein's unbiased risk estimator based on the variance of the coefficients.

The details of this method can be found in [8] and [5].

3. Proposed Energy-Estimation method

Suppose $\{f_i\}$ and $\{z_i\}$ are the clean speech and noise processes respectively, those are related with noisy signal samples as:

$$y_i = f_i + z_i \tag{5}$$

In which $\{z_i\}$ is a zero-mean, white Gaussian noise with standard derivation σ and independent of $\{f_i\}$. As noise and clean speech are independent and as their standard derivations represent their power, we have:

$$P_y = P_f + P_z \tag{6}$$

Where *P* represents the power of the signal. Therefore for a specific frame we have:

$$E_y = E_f + E_z \tag{7}$$

Where *E* is the signal energy in the current frame.

The relation between the wavelet coefficients for our signal processes is:

$$W_y = W_f + W_z \tag{8}$$

Where *W* is the wavelet transform. Therefore by using the wavelet transform properties and using the (7) we have:

$$E_{W_y} = E_{W_f} + E_{W_z} \tag{9}$$

that shows the relation between the signals' energy in wavelet domain.

Therefore, from the speech and noise signals properties and wavelet transform characteristics we have:

$$E_{W_y} = E_f + E_z \tag{10}$$

That means the energy of the noisy speech frame in the wavelet domain is equal to sum of clean signal energy and noise energy in the time domain.

For estimation of the noise energy in each frame, we use the average of the noise energy for noise-only frames. Unlike in the spectral subtraction method where the accuracy of noise estimation has a great effect on the performance, because it estimates the spectra, here we just estimate the average energy of the noise for each frame. Note that for stationary noise, the variance can be reduced by estimating the PSD as an average of the periodograms from multiple frames[9], and as the variance is the power of the under study signals, therefore we may have better energy estimation per frame by adopting this method.

Using $\bar{E}_z$ that is the noise average, in (10) we have:

$$E_{W_f} = E_y - \bar{E}_z \qquad (11)$$

In the same time the wavelet based methods achieve noise reduction through thresholding, which relies on the fact that only a few wavelet coefficients are significant for the reconstruction of the signal and these coefficients are usually greater than noise coefficients due to sparsity property of wavelet transform.

By using these observations, we propose a new method for threshold selection in each frame. This method selects those largest coefficients of the noisy signal wavelet coefficients that give the estimation of energy for the clean signal regards to (11).

The method flowchart can be seen at figure 1.

4. Simulation results

The proposed improvements have been tested on spoken English sentences chosen from the Wall Street journal database. The sentences were sampled at 16 KHz in Mono mode with 16 pit per sample. Then we added white Gaussian noise to speech to obtain noisy signals with the specific SNRs. The tested noisy speech signals were with input SNRs equal 0 dB to 15 dB (in 5 dB steps).

The wavelet used was "sym8". The symlet families are nearly symmetrical wavelets proposed by Daubechies as modifications to the db family.

Number of decomposition levels was set at six and that gives the resolution of 250Hz in the approximation.

Classically, the DWT is defined for sequences with length of some power of 2, and for other sizes we need to use signal extension to a size that is in the power of 2. Therefore, for framing we used hanning window with 512 samples, which is a power of 2, to avoid the need for signal extension. We also used 216 samples overlap, equal to half frame length, to avoid changing the magnitude due to use of the hanning window.

For the performance evaluation, we computed the PESQ improvement score, which we define it as:

$$PESQ_{imp} = PESQ_{out} - PESQ_{in} \qquad (12)$$

Where $PESQ_{in}$ is the PESQ score computed between the noisy speech and the clean speech, and $PESQ_{out}$ is the PESQ score between the enhanced speech and the clean speech, therefore the $PESQ_{imp}$ shows how much improvement the method obtains.

We also computed the LSD improvement score, which we define it as:

$$LSD_{imp} = \left| LSD_{out} - LSD_{in} \right| \qquad (13)$$

Where LSD_{in} is the LSD computed between the noisy speech and the clean speech, and LSD_{out} is the LSD between the enhanced speech and the clean speech, therefore the LSD_{imp} shows how much the system can decrease the difference between the signals spectra.

The comparisons between average $PESQ_{imp}$ scores for five different enhanced noisy sentences are shown in table 1. These sentences are enhanced with four different systems. Three of them use universal, minimax and sure and the last one that is the proposed system uses Energy-Estimation method for threshold selection. In all simulations, the threshold is updated in each frame of the signal.

Table 1 shows the average $PESQ_{imp}$ score for the enhanced speeches. This table shows the higher average $PESQ_{imp}$ scores in different SNRs for the proposed system.

Table 2 shows the average LSD_{imp} for the enhanced speeches. This table confirms the better LSD_{imp} is obtained by our system in comparison with other systems.

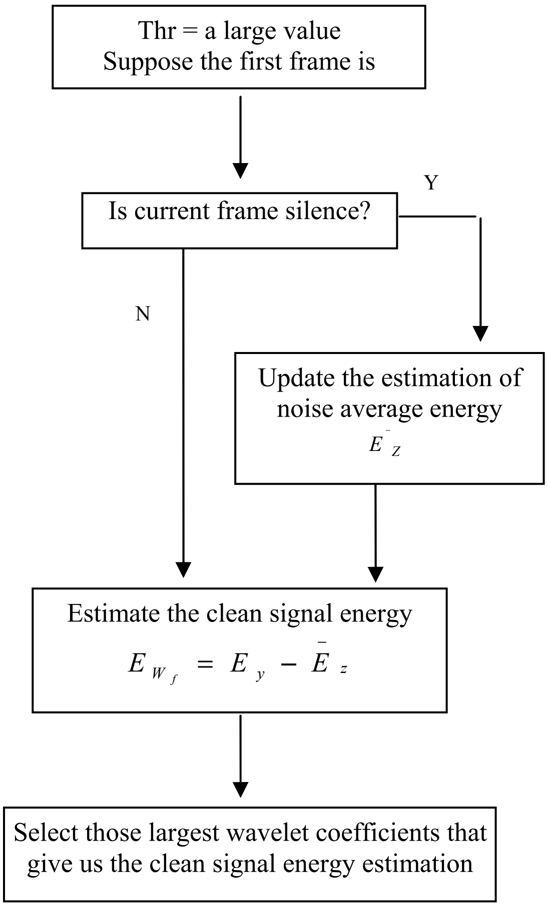
Thr = a large value
Suppose the first frame is
Is current frame silence?
Y
N
Update the estimation of noise average energy
$\bar{E}_Z$
Estimate the clean signal energy
$E_{W_f} = E_y - \bar{E}_z$
Select those largest wavelet coefficients that give us the clean signal energy estimation

Figure 1: Proposed Energy-Estimation threshold selection flowchart

Table 1: Threshold Selection Methods average PESQ improvement in different SNRs

SNRin / Method	0	5	10	15
Universal	0.14	0.16	0.11	0.01
Minimax	0.08	0.11	0.10	0.06
SURE	0.05	0.04	0.05	0.05
Energy-Estimation	0.17	0.23	0.18	0.11

Table 2: Threshold Selection Methods average Log Spectral distance Improvement in different SNRs

SNRin / Method	0	5	10	15
Universal	0.63	0.69	0.65	0.53
Minimax	0.33	0.39	0.40	0.35
SURE	0.11	0.14	0.16	0.15
Energy-Estimation	1.53	1.29	0.90	0.46

5. Conclusions

In this paper, we proposed a new threshold selection method that is used for an adaptive wavelet thresholding speech enhancement system. This method uses the estimation of the clean speech signal energy for each frame to select the threshold for the thresholding algorithm of the current frame.

By applying many tests, we evaluated our system extensively and compared it with the basic wavelet-based denoising methods that use universal, minimax and sure threshold selection. The results confirmed the improvement in performance and achievements of our system.

6. References

[1] S. Ayat. Enhanced Human-Computer Speech Interface Using Wavelet Computing. *IEEE VECIMS*. 2008.

[2] S. G. Chang, B. Yu, M. Vetterli. Adaptive Wavelet thresholding for Image Denoising and Compression. *IEEE Trans On Image processing.* 2000, vol. 9, no. 9.

[3] D. L, Donoho, I.M. Johnstone. Ideal Spatial Adaptation via Wavelet Shrinkage. *Biometrik.* 1995, vol. 81, no.3, pp. 425-455.

[4] D. L. Donoho. Denoising by Soft thresholding. *IEEE Trans on Information Theory*. 1995, vol. 41, no. 3, pp. 613-627.

[5] D. L. Donoho, I.M. Johnstone. Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association*. 1995, vol. 90, pp. 1200–1224.

[6] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, D. Picard. Wavelet Shrinkage: Asymptopia?. *Journal of the Royal Statistics Society*. 1995, Series B, vol. 57, pp.310-369.

[7] Q. Fu, E. A. Wan. Perceptual Wavelet Adaptive Denoising of Speech. *EUROSPEECH2003.*

[8] H.Y. Gao, A.G. Bruce. Waveshrink with Firm Shrinkage.Statistica Sinica. 1997, vol. 7, pp. 855–874.

[9] S. Srinivasan. Knowledge-Based Speech Enhancement. *Phd thesis*. *KTH - Royal Institute of Technology*. Stockholm 2005.