WPD-based Noise Suppression Using Nonlinearly Weighted Threshold Quantile Estimation and Optimal Wavelet Shrinking

Tuan Van Pham and Gernot Kubin.

Signal Processing and Speech Communication Laboratory
University of Technology, Graz, Austria
v.t.pham@tugraz.at; g.kubin@ieee.org; http://spsc.tugraz.at

Abstract

A novel speech enhancement system based on wavelet packet decomposition (WPD) is proposed. Noise level is estimated based on quantile in wavelet threshold domain. To handle colored and non-stationary noises, the universal thresholds are weighted by a time-frequency dependent nonlinear function. Two nonlinear weighting methods using temporal threshold variation and kernel smoothing are proposed. The weighted thresholds are smoothed and employed for wavelet shrinking with an adaptive factor to compress noise while preserving speech quality. The proposed system is evaluated and compared with other algorithms based on spectral subtraction via objective measures and subjective tests to demonstrate its superior performance.

1. Introduction

One of the challenges in speech enhancement is noise estimation which is very difficult for non-stationary noise. With the common spectral subtraction method in [1], the noise spectrum is calculated from non-speech frames which are detected previously by voice activity detection (VAD). The so-called ”musical noise” related to short-time spectral subtraction remains after the processing and has a very unnatural and disturbing quality. The optimal non-linear spectral amplitude estimation in [2] achieves a significant noise reduction while reducing the musical noise and maintaining good speech quality.

The disadvantage of these approaches is low temporal consistency because the VAD is unreliable and the noise estimate cannot be updated during speech periods. A new method called quantile-based noise estimation in [3] which is based on minimum statistics [4] overcomes the problems of VAD. In 1995, Donoho [6] proposed the wavelet shrinking method as a powerful tool to enhance noisy signals. However, the soft-thresholding does not have high-order derivatives and results in some artifacts in the denoised signal. More sophisticated shrinking functions have been studied in [7, 8] to improve the denoised signal quality.

In this paper, the denoising process is considered in a sequence of buffers consisting of the overlapping speech frames. The wavelet coefficients of 128 WPD channels are extracted by performing the WPD on each speech frame at the 7th analysis level. Next, these coefficients are thresholded by optimal wavelet shrinking which is controlled by a nonlinear mechanism. Then the enhanced speech frames are reconstructed by the wavelet packet reconstruction (WPR) of denoised coefficients (see Fig. 1).

For every WPD channel of all frames in the overlapping buffer, the universal thresholds determined in [6] are calculated.

Then the thresholds related to the noise levels are estimated as quantiles in the recursive buffers which store the sorted thresholds along each WPD channel. The wavelet shrinking from [8] is applied to shrink the coefficients below the noise thresholds to zero. The major contributions to the system are two new nonlinear weighting functions and an adaptive factor of the wavelet shrinking to control non-stationary and colored noise in the time-frequency domain. These nonlinear weighting functions are based on the temporal threshold variation and kernel smoothing of frame indices after sorting. Finally, the smoothed weighted thresholds are used for optimal wavelet shrinking.

2. Wavelet packet shrinking

2.1. Wavelet packet decomposition

Any signal \( s[n] \) with length \( N \) in space \( V_0 \) can be decomposed into an approximation and a detail in spaces \( V_j \) and \( W_j \), \( j \in \mathbb{Z} \). In this research, we apply a full WPD which is represented by a full binary tree. Each node \( (j, k) \) at a deeper level is covered by the subspace \( W_{j,k} \) where \( j \) denotes the decomposition level \( j < \log_2 N \), \( k \in \mathbb{Z} \). This subspace is spanned by an orthonormal basis \( \{ \psi_{j,k}(-2^j n + m) \}_{m \in \mathbb{Z}} \). The WPD coefficients of the signal \( s[n] \) is:

\[
D_{j,k}[m] = \langle s[n], \psi_{j,k}(-2^j n + m) \rangle
\]

The proposed denoising system is executed on all WPD channels at \( j = 7 \) and \( k = 1, \ldots, 2^j \). From now on, \( j \) is discarded in the notation. Using the full WPD constitutes a computational load but gains a high accuracy of the decomposition. The Daubechies wavelet db40 is selected as a good choice because of its high performance in an informal listening test.

2.2. Wavelet packet shrinking

The universal threshold rule [6] using a robust estimate of standard deviation is applied to estimate the thresholds of the WPD coefficients as:

\[
T_{k,i} = \frac{1}{\gamma_{MAD}} \text{Median}(|D_{k,i}|) \sqrt{2 \log N_{k,i}}
\]
where $\gamma_{\text{MAD}} = 0.6745$ and $D_{k,i}$ is the WPD coefficient sequence having length $N_{k}$ in the $k^{th}$ channel at the $i^{th}$ frame. The WPD coefficients of noisy speech $X_{k,i}$ can be expressed as the sum of WPD coefficients of clean speech $S_{k,i}$ and noise $E_{k,i}$ as:

$$X_{k,i} = S_{k,i} + E_{k,i}. \quad (3)$$

With the hard thresholding procedure [9], all wavelet coefficients which are lower than the threshold are removed while the others are left untouched:

$$\tilde{S}_{k,i} = \begin{cases} X_{k,i}, & \text{if } |X_{k,i}| > T_{k,i} \\ 0, & \text{if } |X_{k,i}| \leq T_{k,i}. \end{cases} \quad (4)$$

Because of the strong discontinuity of the input-output characteristics, artifacts are introduced in the outputs such as annoying blips. The soft thresholding from [6] can reduce these artifacts but is still sub-optimal because of the strict setting to zero of the coefficients whose absolute values are below the threshold. An enhanced shrinking is proposed as a smoothed hard thresholding from [8]:

$$\tilde{S}_{k,i} = \frac{X_{k,i}, \text{sign}(X_{k,i})}{\mu_{k,i}} A_{k,i} \mu_{k,i}, \text{if } |X_{k,i}| \leq T_{k,i} \quad \text{(5)}$$

where $A_{k,i}$ and the adaptive parameter $\mu_{k,i}$ are defined as:

$$A_{k,i} = (1 + \mu_{k,i}) T_{k,i} - 1, \quad (6)$$

$$\mu_{k,i} = \frac{\max \{ |X_{k,i}(m)| \} \theta}{T_{k,i}}, \quad (7)$$

where $\theta$ is a constant factor and $m$ is the WPD coefficients in the WPD channel. This shrinking preserves more coefficients which are below the thresholds, (see Fig. 2). In the next sections, a new adaptation method for the factor $\theta$ is proposed to maintain the pleasantness of speech.

![Figure 2: Wavelet shrinking as the smoothed hard thresholding.](image)

### 3. Quantile-based threshold estimation

#### 3.1. Recursive buffering

As reported in [3], the energy in each WPD channel is on the noise level over a significant part of the time. Thus, the noise level can be estimated by taking the $q^{th}$ quantile observed over the duration of the utterance in every WPD channel. From this observation, we develop a quantile-based algorithm to estimate the thresholds related to the noise level. The buffers (960ms length and 480ms overlap) which store 47 speech frames (40ms length and 20ms overlap) are used because of the good results from informal listening tests and their suitability to real-time applications.

After sorting the thresholds for every WPD channel, we observe that the thresholds related to noisy frames occupy up to 60% of the quantiles, (see Fig. 3). This lower quantile part should be memorized and transferred into the next buffer to maintain correlation with the estimated noise level of a previous buffer. A recursive buffer is constructed from the overlapping buffer by a recursive buffering scheme as described in Fig. 4.

![Figure 3: Quantiles of sorted threshold values.](image)

#### 3.2. Noise threshold estimation algorithm

- First, the $T_{k,i}$ is calculated for all WPD channels of all frames from the current buffer (e.g. the 2nd buffer in Fig. 4) and stored in the corresponding threshold buffer.
- Second, the new thresholds $T_{k,i}$ of the frames $i = 24, \ldots, 47$ are selected and merged with the sorted thresholds $T_{k,i}$ which are selected from quantile range $q = 0.1, \ldots, 0.6$ of the previous sorted threshold buffer to form a recursive buffer. The selection of this quantile range guarantees the 'fading memory' property for the recursive buffering scheme.
- Then, for each WPD channel, the thresholds of all speech frames in the recursive buffer are sorted in ascending order which lead to $T_{k,i}$, where $i = 1, \ldots, N_f$ is the frame indices after sorting with $N_f = 47$. This sorted recursive buffer will be used in next loop.
- Finally, the threshold related to the noise levels, $\Gamma_{k}$, for all frames in the sorted recursive buffer at the $k^{th}$ channel is determined as the $q^{th}$ quantile:

$$\Gamma_{k} = T_{k,i} |i = \lceil q N_f \rceil \quad (8)$$

We have performed the pre-formal listening tests over the range of possible values $q = 0.0, 0.1, \ldots, 0.6$. The quantile $q = 0.2$ is a good choice which yields the best performance.

![Figure 4: Recursive buffering scheme.](image)
4. Nonlinear threshold weighting function

To handle non-stationary and colored noise, the noise threshold $\Gamma_k$ of each frame $i$ in each channel $k$ is weighted by a time-frequency dependent nonlinear function $W_{\text{opt}}$ as:

$$\tilde{\Gamma}_{k,i} = W_{\text{opt}}(\Gamma_k) = \lambda_{k,i} \eta_k \Gamma_k,$$

where $\lambda_{k,i}$, $\eta_k$ are nonlinear parameters in time and frequency domains. $\tilde{\Gamma}_{k,i}$ is the weighted estimate of the noise threshold.

4.1. Nonlinear weighting in frequency domain

To suppress colored noise which has high variation of noise levels in different WPD channels effectively, the noise threshold $\Gamma_k$ is weighted by the nonlinear parameter $\eta_k$ as:

$$\eta_k = (10\Gamma_k)^{-q} + d_1,$$

where $\alpha_2 = 0.55$, $d_1 = 0.6$ are constants. With this static nonlinear weighting, the WPD channels which have a large $\Gamma_k$ or high noise level are compressed by strong weighting. (see Fig. 5). Vice versa, a reduced weighting on the WPD channels having small $\Gamma_k$ is necessary to maintain speech quality.

![Image](image.png)

Figure 5: Frequency-nonlinear weighting $\eta_k \Gamma_k$.

4.2. Nonlinear adaptive weighting in time domain

Two time adaptive weighting methods are proposed as follows:

* In the first method, the time-variant threshold dependent curve (TDC) $\lambda_{k,i}$ is built to track where the speech or noise appears along the buffer. The frames with smaller thresholds $T_{k,i}$ might correspond to noise and will be impacted by stronger thresholding. The frames with large $T_{k,i}$ always contain more speech information and are treated in the reverse way to preserve speech quality. (see Fig. 6). This is expressed as:

$$\lambda_{k,i} = T_{k,i}^{a_2} - a_2 + d_2,$$

where $\alpha_2 = 0.14$, $d_2 = 0.2$ are constants, $T_{k,i}$ is the universal threshold in the current threshold buffer (step (*) in Fig. 4).

![Image](image.png)

Figure 6: Time adaptive weighting based on TDC.

* The second method applies kernel smoothing (KS) on frame indices. After sorting the thresholds of all frames in ascending order, we observe that most of speech frames lie in the right half of the recursive buffer for every channel (step (**)) in Fig. 4). This is expected because the speech frames always have higher thresholds $T_{k,i}$ than the noise frames. To track the positions of speech frames in the buffer, we construct the discrete histogram $f(i)$ over all channels for the frames of which the thresholds $T_{k,i}$ fall into the upper 50% quantile range of the buffer as shown in Fig. 7a, (the area drawn by dotted line).

Then the envelope of the discrete histogram is computed using kernel smoothing with a gaussian window $f_w(i)$:

$$\tilde{f}(i) = f_w(i) * f(i).$$

The time adaptive parameter $\lambda_{k,i}$ is constructed as:

$$\lambda_{k,i} = e^{a_3(f(i) - d_3)},$$

where $a_3 = 2$, $d_3 = 0.6$ are constants. All of the above constants are selected from the informal listening tests.

![Image](image.png)

(a) at one channel (b) all channels in the buffer

Figure 7: Time adaptive weighting based on KS.

After the nonlinear weighting by the frequency parameter and one of the two time parameters, the weighted thresholds $\tilde{\Gamma}_k$ are smoothed by locally weighted linear regression to reduce the fluctuations between neighboring frames, see Fig. 8.

![Image](image.png)

Figure 8: Time-frequency nonlinear weighting.

4.3. Optimal wavelet shrinking with adaptive factor $\theta$

From our initial listening tests, we felt that the processed speech sound is not as pleasant as natural speech if we keep the factor $\theta$ constant. A high value of $\theta$ eliminates the noise effectively with a rather quiet residual in the non-speech frames. But it also creates some discontinuities in the speech frames that result in a harsh voice perception. To overcome this phenomenon, the factor $\theta$ is adapted by the normalized smoothed thresholds $\tilde{\Gamma}_k$:

$$\theta_{k,i} = \exp \left( \alpha \frac{\tilde{\Gamma}_{k,i}}{\max_i \tilde{\Gamma}_{k,i}} \right),$$

where $\alpha = 5.8$ is a slope constant. By this adaptive factor $\theta_{k,i}$, at each WPD channel, the curve of part below the threshold in Fig. 2 is flatter for the speech frames to preserve more coefficients and steep for the non-speech frames to compress noise. This improves the perceptual impression for the speech part significantly and keeps the background noise at a very small level for the non-speech parts.
5. Performance evaluation

The two WPD-based algorithms using the threshold dependent curve (WP-TDC) and kernel smoothing (WP-KS) approaches have been implemented on each set of 40 utterances of the AURORA3 and NTIMIT databases. The systems are evaluated and compared with other algorithms such as spectral subtraction (SS) [1], nonlinear spectral subtraction (NSS) [2], and the Ephraim-Malah algorithm with the noise estimator from R. Martin (NSS-RM) [4] which has been implemented in [5]. The samples of denoised sounds are available on our website [11].

5.1. Objective test using SSNR

As shown in Fig. 9, the WPD-based algorithms have higher output segmental signal-to-noise ratios (SSNR) than NSS-RM for noisy speech with input SSNR $\geq$ 5dB and a little bit lower SSNR in case of input SSNR $\leq$ 0dB. The SS and NSS algorithms gain higher output SSNR. However, the SSNR cannot faithfully express the speech quality. Thus, we have also performed three subjective tests which have been adapted from [10]. Each of them is done by 10 native male listeners.

![Figure 9: Input-output SSNR for different algorithms.](image)

5.2. Comparison Diagnostic Test (CDT)

The test material is focused on consonants only because they are more problematic than vowels. The listeners hear the unprocessed and processed sentences, and concentrate on a single phonetic class which is marked. Then they vote by Comparison Category Rating (CCR), [10]. As shown in Fig. 10a, the three latter algorithms maintain consonants better than the two former. The WP-based algorithms are the best ones in preserving plosives as expected from the nonlinear time adaptive weighting.

5.3. Overall quality evaluation using CCR

In this test, 8 car noise and 4 telephone noise utterances are selected randomly. The listeners hear the noisy and enhanced utterances, then judge which of the utterances are better overall and how much it is improved in comparison with the original sound. From Tab. 1, we see that WP-based algorithms achieve a higher performance in cases of car and telephone noises with 10dB SSNR. In case of 5dB SSNR, more than 2/3 of listeners prefer the output of our system over the original noisy speech.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Scales AURORA3</th>
<th>M</th>
<th>S</th>
<th>Mean</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>-2.25</td>
<td>1.10</td>
<td>2.45</td>
<td>0.61</td>
<td>-2.25</td>
</tr>
<tr>
<td>NSS</td>
<td>-1.28</td>
<td>1.26</td>
<td>0.55</td>
<td>1.51</td>
<td>-1.35</td>
</tr>
<tr>
<td>NSS-RM</td>
<td>0.45</td>
<td>1.75</td>
<td>1.00</td>
<td>1.44</td>
<td>0.75</td>
</tr>
<tr>
<td>WP-TDC</td>
<td>0.85</td>
<td>1.35</td>
<td>1.30</td>
<td>1.24</td>
<td>1.30</td>
</tr>
<tr>
<td>WP-KS</td>
<td>0.80</td>
<td>1.55</td>
<td>1.53</td>
<td>1.18</td>
<td>1.50</td>
</tr>
</tbody>
</table>

Table 1: Overall evaluation, (M: mean, S: standard deviation).

5.4. Overall quality test according to ITU-T (P.85)

By using absolute scales per category, the listeners assess the speech quality of noisy speech and denoised speech via 6 categorical ratings, [10]. As resulted in Fig. 10b, the three later algorithms get the similar or lower mean values compared with the ones of the noisy speech. In almost all categorical ratings with various SSNR, the performance of the two WP-based algorithms always exceeds the one of the NSS-RM.

![Figure 10: Subjective test results.](image)

6. Conclusions

A WPD-based speech enhancement system is proposed. A time-frequency dependent nonlinear weighting function for quantile-based threshold estimates has been designed to manage the colored and non-stationary noises. An adaptation of the sophisticated wavelet shrinking is proposed to maintain the naturalness of speech sounds and keep the background noise consistently at low level. The good results of subjective tests confirm the possibility of a high-level of noise suppression while preserving the speech intelligibility and naturalness of the system. The computation time reduction will be studied in future research by selecting the optimal WPD tree while maintaining the utmost speech quality.

7. References