Wavelet-based Perceptual Speech Enhancement Using Adaptive Threshold Estimation

Essa Jafer and Abdulhussain E. Mahdi

Department of Electronic and Computer Engineering
University of Limerick, Limerick, Ireland
essa.jafer@ul.ie hussain.mahdi@ul.ie

Abstract
A new speech enhancement system, which is based on a time-frequency adaptive wavelet soft thresholding, is presented in this paper. The system utilises a Bark-scaled wavelet packet decomposition integrated into a modified Wiener filtering technique using a novel threshold estimation method based on a magnitude decision-directed approach. First, a Bark-Scaled wavelet packet transform is used to decompose the speech signal into critical bands. Threshold estimation is then performed for each wavelet band according to an adaptive noise level-tracking algorithm. Finally, the speech is estimated by incorporating the computed threshold into a Wiener filtering process, using the magnitude decision-directed approach. The proposed speech enhancement technique has been tested with various stationary and non-stationary noise cases. Reported results show that the system is capable of a high-level of noise suppression while preserving the intelligibility and naturalness of the speech.

1. Introduction
With the rapid developments in voice communication systems, speech enhancement using a single channel has become an active and important research area. During the last few decades, various approaches to speech quality by reducing the noise have been proposed. The most widely used methods are those based on spectral subtraction and Wiener filtering and their variants. Although most of these methods have been shown to provide good speech quality particularly in terms of improved signal-to-noise ratio (SNR), they often suffer from an annoying signal distortion caused by a residual effect known as musical noise. In an attempt to reduce this drawback, the use of a human auditory model has recently been proposed in subtractive-type enhancement techniques (e.g. [1], [2] and [3]). This model is based on the fact that the human auditory system can’t detect additive noise as long as it is below some masking threshold. Methods to calculate this threshold are developed according to critical band analysis either in the frequency domain using the short-term Fourier transform (STFT)[1-2], or in the wavelet domain using uniform-band wavelet packet decomposition (UB-WPD) [3]. However, although these methods resulted in good quality speech with reduced level of musical noise, the excessive expansion of high frequency sub-bands, resulting from the use of the STFT or UB-WPD, gives rise to the following drawbacks [4]: (1) increasing computational complexity, (2) degradation of the perceptual quality of the unvoiced sounds, and (3) insufficient spectral resolution for dealing with voiced sounds. In an attempt to address the above problems, this paper proposes a new speech enhancement algorithm based on a time-frequency adaptive wavelet soft thresholding approach [5, 6]. Inspired by recently reported work by Cohen [7], the proposed algorithm is based on threshold estimation method integrated into a modified Wiener filtering technique. The threshold estimation method is uses a magnitude decision-directed technique [4], which is closely related to the decision-directed estimator of Ephraim and Malah [8], and utilises a Bark-scaled wavelet packet decomposition (BS-WPD) [7] and a Quantile-based noise estimation technique [9-10]. Compared to the critical band analysis, the BS-WPD provides higher frequency resolution. It also provides higher time resolution compared to conventional wavelet packet decomposition. The adaptive threshold estimation method in conjunction with the soft-thresholding approach, that is based on a modified Wiener filtering and a magnitude decision-directed technique, achieves a good trade-off between noise reduction and signal distortion. The proposed algorithm has been tested with various types of noise. Results show that the algorithm performs better than conventional wavelet denoising techniques, in terms of improved SNRs, while preserving the naturalness of the speech.

2. The Bark-scaled WPD
Let \( \{\psi_{n}(t) : n \in \mathbb{Z}+\} \) denotes a wavelet packet family, and let \( E \subset \{(l,n): 0 \leq l < L, 0 \leq n < 2^l\} \) represent the terminal nodes of a WPD tree [7]. Then a set of dyadic wavelet packet expansion functions can be expressed as:

\[
\psi_{l,n,k}(t) = 2^{-l/2} \phi_{n}(2^{-l}t - k) \quad (1)
\]

A terminal node \((l,n) \in E\) is associated with a sub-band (subspace) whose centre frequency and bandwidth are roughly given by [7]:

\[
f_{l,n} = 2^{-l}[GC^{-1}(n) + 0.5]F_s / 2
\]

\[
B_{l,n} = 2^{-l}F_s / 2
\]

where \(GC^{-1}\) is the inverse Gray code permutation and \(F_s\) is the sampling frequency in the signal space. To obtain critical bands with the WPD, a decomposition tree has been constructed such that the distance between the centre frequency of one subband to that of the next subband is 1 Bark. The relation between frequency \(f\) in Hz and critical band rate \(c\) in bark is approximately given by [11]:

\[
c = 26.81/(1+1960/f) - 0.53 \quad (3)
\]

Figure 1 shows an approximation of the bark scale by the constructed BS-WPD. The corresponding decomposition tree...
used in this is depicted in Figure 2. As can be seen from this tree, the BS-WPD splits the frequency range 0-8 kHz into 21 sub-bands.

Figure 1: Approximation of the Bark scale (solid line) by

BS-WPD (*).

3. The proposed speech enhancement algorithm

The algorithm can be described as follows. Let \( x(k), k = 0, 1, \ldots, M-1 \), be a finite length observation of a sampled clean speech signal, \( x(k) \), corrupted by uncorrelated additive noise, \( w(k) \), whose noise level is given by \( \sigma_c \), such that:

\[
x(k) = x(k) + w(k)
\]

where \( k \) is the discrete-time index. The proposed algorithm begins by decomposing \( x(k) \) into a number of critical bands using the BS-WPD method explained in Section 3.2, such that:

\[
\{y^{l,n}(k)\} = \text{BS-WPD}(x(k)), \quad \text{for } k = 0, \ldots, M - 1
\]

where \( \{y^{l,n}(k)\} \) represents the set of BS-WPD coefficients in the \((lh, nth)\) band of the decomposition, as per the discussion given in Section 2. The clean speech is estimated based on the outcome of the two following operations: a time-frequency dependent threshold estimation technique using a modified Weiner filtering approach, as will be described in the following Sections. Subsequently, the estimate of the clean speech is transformed back into the time-domain using an inverse BS-WPD, as illustrated in Figure 1.

3.1. Time-frequency dependent threshold estimation

Signals denoising by classical wavelet thresholding can be performed as either ‘hard’ or ‘soft’ thresholding. In hard thresholding, wavelet coefficients that are below a specified threshold value are set to zero. In soft thresholding, the coefficients that are below the threshold value are shrunk or scaled down in an appropriate manner [13]. However, for speech signals, hard/soft thresholding often results in time-frequency discontinuities, which leads to unnatural effects and further degradation of the quality of the enhanced speech. This effect is more pronounced when the speech is corrupted by high noise levels. Considering this drawback, a new adaptive time-frequency dependent threshold estimation method is used here. The method involves first the estimation of the noise level, \( \sigma_c \) for every wavelet band and time frame. This is achieved by utilising a one-pass noise-tracking method which uses the quantile computation algorithm [9, 10], and based on the application of the Ris & Dupont approach [12] in the BS-WPD domain. To do that, the decomposed coefficients, \( y^{l,n}(k) \), are framed into frames of length \( L^{l,n}_{\text{frm}} \). Let \( L^{l,n}_{\text{win}} > L^{l,n}_{\text{frm}} \) be the length of a finite window observation of the coefficients \( y^{l,n}(k) \) spanning a number of frames. Also, let \( \hat{\sigma}_l^{i,n} \) be the noise level of the \( i \)-th frame in the \((lh, nth)\) band estimated using the previous set of coefficients \( \{y^{l,n}(k), k = 0, \ldots, L^{l,n}_{\text{win}} - 1\} \), after being sorted according to the requirement of the quantile-based approach [9, 10], such that:

\[
\hat{\sigma}_l^{i,n} = \beta \sum_{j=0}^{\text{int}(qL^{l,n}_{\text{win}})} y^{l,n}(j) / L^{l,n}_{\text{win}}
\]

where \( q = 0.2 \) and \( \beta \) is an appropriate scaling factor taken here to be equal to 0.4. The corresponding time lengths for \( L^{l,n}_{\text{frm}} \) and \( L^{l,n}_{\text{win}} \) are 64 and 512 ms, respectively, with the frames overlapped by 50 %. Based on this noise estimate, the threshold for the \((lh, nth)\) band at the \( i \)-th frame, is calculated using the optimal universal threshold [5]:

\[
\lambda_l^{i,n} = \sqrt{2\log(L^{l,n}_{\text{seg}}) \log_2(L^{l,n}_{\text{seg}})} \cdot \hat{\sigma}_l^{i,n}
\]

3.2. Soft-thresholding using a modified Weiner filtering technique

The wavelet thresholding technique used here is based on modification of the noise suppression rule of the Ephraim and Malah algorithm [8]. Using the threshold value for the \((lh, nth)\) band at the \( i \)-th frame estimated by the process described in Section (a) above, we first calculate a posteriori signal to threshold ratio, such that:

\[
\left\{y^{l,n}(k)\right\}_{\text{post}} = \left\{y^{l,n}(k)\right\} / \lambda_l^{i,n}
\]

A magnitude decision-directed technique is then applied to estimate the corresponding priori signal to threshold ratio:
\[
\text{Where } L(k)_{\alpha} = \min(0, \alpha \min \{1, (\hat{\eta} + 1)(\hat{\alpha} + 1)\} + 1)
\]

Using the two estimated factors in 8 and 9, a suppression function is computed as:

\[
g^{1,\alpha}(k) = \frac{1}{1 + [\eta(k)]_{\alpha}^{\text{prior}} + 1} \frac{1}{1 + [\hat{\eta}^{\text{prior}}(k)]_{\alpha}^{\text{prior}}}
\]

Finally, to obtain the estimate for the clean speech, the values of the wavelet coefficients of the noisy speech, \( \hat{y}^{1,\alpha}(k) \), are adjusted using the suppression function such that:

\[
y^{1,\alpha}(k) = g^{1,\alpha}(k) \cdot \hat{y}^{1,\alpha}(k)
\]

and transformed back into the time-domain using an inverse BS-WPD to reconstruct the enhanced speech signal.

4. Performance evaluation

The evaluation of the proposed speech enhancement algorithm has been performed using two objective measures: the global SNR of the enhanced speech compared to that of the noisy speech, and the Cepstral distance between the original clean speech signal and both the noisy and enhanced speech signals. The evaluation involved the use of a number of speech records, for male and female speakers, taken from TIMIT database. The speech signals were sampled at 16 kHz, and corrupted by different types of noise taken from the Noisex 92 database. In all the evaluation cases, the BS-WPD is implemented with the discrete Meyer wavelet filters, which provides good separation of bands due to their regularity property. By means of illustration, Figures 3 and 4 show the waveforms resulting from applying the proposed technique to speech signals corrupted by white and pink noise, respectively, both at SNR = 5 dB. For these tests, a value of 0.75 was used for controlling factor \( \alpha \). Figure 5, on the other hand, shows the improvement in global SNR of the enhanced signal obtained from the application of the proposed algorithm to three different noise types. For Cases (1), (2) and (3), full-implementation of the proposed speech enhancement algorithm using White, F16 cockpit and Pink noise respectively as a function of the input SNR. In case (4), however, the magnitude decision-directed soft thresholding method is replaced by a classical soft thresholding process. Using the Cepstral distance measure, Tables 1-4 provide further indication about the performance of the proposed algorithm. In Tables 1 and 2, two different speech signals taken from two different speakers (a male and a female), corrupted by AWGN at different SNRs, were tested. Tables 3 and 4 consider a male speech signal corrupted by two different type of additive noise: pink noise and F16 cockpit noise. For comparison, Tables1-4 also indicate the performance of the speech enhancement using a classical soft thresholding algorithm where the value of \( \alpha \) fixed to be 0.75. Our investigation has shown that for cases of high SNR (above 10dB), setting \( \alpha \) to a value between 0.5 to 0.75 yields the best results. For lower SNR (below 5 dB), however, \( \alpha \geq 0.9 \) has been found to provide the best performance in terms of noise suppression as well as preservation of the naturalness and quality of the original speech signal. This can also be noted in Figure 5, where a little improvement is shown in Case (1), (2) and case (3) for input SNRs above 0 dB. For SNRs below 0 dB, a relatively high improvement in the output SNR is obtained since the suppression factor is 0.9. It is clear from the results given that the proposed speech enhancement algorithm gives a good robust performance over a wide range of noise types and levels.
Table 1: Comparison of Cepstral distances resulting from the application of the Classical Soft Thresholding and proposed algorithm on a Male speech corrupted by White noise.

<table>
<thead>
<tr>
<th>Input SNR (dB)</th>
<th>CEPSTRAL DISTANCE (MALE)</th>
<th>Soft Thresh</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noisy (dB)</td>
<td>BS-WPD</td>
</tr>
<tr>
<td>10</td>
<td>0.85</td>
<td>0.63</td>
</tr>
<tr>
<td>5</td>
<td>1.56</td>
<td>1.10</td>
</tr>
<tr>
<td>0</td>
<td>2.67</td>
<td>1.90</td>
</tr>
<tr>
<td>-5</td>
<td>4.33</td>
<td>3.19</td>
</tr>
<tr>
<td>-10</td>
<td>6.57</td>
<td>5.03</td>
</tr>
</tbody>
</table>

Table 2: Comparison of Cepstral distances resulting form the application of the Classical Soft Thresholding and proposed algorithm on a Female speech corrupted by White noise.

<table>
<thead>
<tr>
<th>Input SNR (dB)</th>
<th>CEPSTRAL DISTANCE (FEMALE)</th>
<th>Soft Thresh</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noisy (dB)</td>
<td>BS-WPD</td>
</tr>
<tr>
<td>10</td>
<td>0.78</td>
<td>0.60</td>
</tr>
<tr>
<td>5</td>
<td>1.38</td>
<td>0.98</td>
</tr>
<tr>
<td>0</td>
<td>2.43</td>
<td>1.69</td>
</tr>
<tr>
<td>-5</td>
<td>4.00</td>
<td>2.83</td>
</tr>
<tr>
<td>-10</td>
<td>6.12</td>
<td>4.59</td>
</tr>
</tbody>
</table>

Table 3: Comparison of Cepstral distances resulting form the application of the Classical Soft Thresholding and proposed algorithm on speech corrupted by pink noise.

<table>
<thead>
<tr>
<th>Input SNR (dB)</th>
<th>CEPSTRAL DISTANCE (PINK)</th>
<th>Soft Thresh</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noisy (dB)</td>
<td>BS-WPD</td>
</tr>
<tr>
<td>10</td>
<td>0.42</td>
<td>0.38</td>
</tr>
<tr>
<td>5</td>
<td>0.67</td>
<td>0.55</td>
</tr>
<tr>
<td>0</td>
<td>1.15</td>
<td>0.85</td>
</tr>
<tr>
<td>-5</td>
<td>3.34</td>
<td>2.30</td>
</tr>
<tr>
<td>-10</td>
<td>4.10</td>
<td>2.85</td>
</tr>
</tbody>
</table>

Table 4: Comparison of Cepstral distances resulting form the application of the Classical Soft Thresholding and proposed algorithm on speech corrupted by F16 cockpit noise.

<table>
<thead>
<tr>
<th>Input SNR (dB)</th>
<th>CEPSTRAL DISTANCE (F16 COCKPIT)</th>
<th>Soft Thresh</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noisy (dB)</td>
<td>BS-WPD</td>
</tr>
<tr>
<td>10</td>
<td>0.31</td>
<td>0.26</td>
</tr>
<tr>
<td>5</td>
<td>0.57</td>
<td>0.47</td>
</tr>
<tr>
<td>0</td>
<td>1.23</td>
<td>0.67</td>
</tr>
<tr>
<td>-5</td>
<td>2.42</td>
<td>1.72</td>
</tr>
<tr>
<td>-10</td>
<td>3.10</td>
<td>1.95</td>
</tr>
</tbody>
</table>

5. Conclusions

A new speech enhancement technique that utilizes the frequency resolution of the critical bands obtained by the BS-WPD has been described and its performance evaluated. The system integrates this efficient auditory representation into a modified Weiner filtering approach using a novel threshold estimation technique and a magnitude decision-directed approach. The performance of the system was evaluated under different noisy conditions using a number of speech signals taken from male and female speakers. Reported results have shown that the proposed algorithm is robust to stationary and non-stationary noise. Evaluation of the proposed system has shown that it offers better performance over the traditional wavelet denoising methods. It has shown that, for most of the considered cases, the proposed system resulted in higher intelligibility and quality of enhanced speech, compared to classical systems which are based on uniform spectral decomposition.

6. References