A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets

Yasser Ghanbari, Mohammad Reza Karami-Mollaei *

University of Mazandaran, Faculty of Electrical Engineering, Department of Electrical Engineering, 484 Babol, Iran

Received 15 December 2004; received in revised form 20 December 2005; accepted 20 December 2005

Abstract

In this paper, we propose a new speech enhancement system using the wavelet thresholding algorithm. The basic wavelet thresholding algorithm has some defects including the assumption of white Gaussian noise (WGN), malfunction in unvoiced segments, bad auditory quality, etc. In the proposed system, we introduce a new algorithm which does not require any voiced/unvoiced detection system. Also, in this proposed method adaptive wavelet thresholding and modified thresholding functions are introduced to improve the speech enhancement performance as well as the automatic speech recognition (ASR) accuracy. A new voice activity detector (VAD) was designed to update noise statistics in the proposed speech enhancement system when facing to the colored and non-stationary noises. The proposed method was evaluated on several speakers and under various noise conditions including white Gaussian noise, pink noise, and multi-talker babble noise. The SNR and ASR results show that the new method highly improves the performance of speech enhancement algorithm based on the wavelet thresholding.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Speech processing; Speech enhancement; Wavelet thresholding; Noisy speech recognition

1. Introduction

In many speech processing applications such as mobile communications, speech recognition and hearing aids, speech has to be processed in the presence of a background noise. The performance of such applications is highly dependent on how much the noise is removed. During the last decades, various approaches have been proposed to solve this problem, such as spectral subtraction (Deller et al., 2000; Boll, 1979; Berouti et al., 1979; Kamath and Loizou, 2002; Ghanbari et al., 2004; Ghanbari and Karami, 2004), hidden Markov modeling (Sameti et al., 1998), signal subspace methods (Klein and Kabal, 2002), and wavelet-based methods (Donoho, 1995; Sheikhzadeh and Abutalebi, 2001; Seok and Bae, 1997).

Generally, the approaches can be classified into two major categories of single-channel and multi-channel methods. There are many practical situations that one is limited to use a single microphone. In these applications, single-channel based methods are used for noise reduction among which the spectral subtraction is one of the most popular methods. This method is capable of removing the background...
noise but the main disadvantage is the musical noise generated through this method.

Donoho (1995) introduced a new algorithm based on wavelet thresholding for denoising the signals corrupted by WGN. After that, employing this new method in speech enhancement was widely studied (Soon et al., 1997; Chang et al., 2002; Sheikhzadeh and Abutalebi, 2001; Seok and Bae, 1997). In the most techniques which use the wavelet thresholding for speech enhancement, they may suffer from a main problem that is the detection of the voiced/unvoiced segments of the speech signals (Sheikhzadeh and Abutalebi, 2001; Seok and Bae, 1997). For the incorrectly classified segments, the enhancement performance drastically decreases. The other controversial subjects affecting the enhancement performance, are the thresholding function and the threshold value. The main objective of this work is to propose a new algorithm which does not need to detect the unvoiced segments. In this algorithm, a new thresholding function with an adaptive threshold value to be used in the wavelet packet domain. Also, for enhancement of speech corrupted by colored and non-stationary noises, we propose a new voice activity detector (VAD) system which works in the wavelet packet domain. It should be noted that although VAD and voiced/unvoiced detection systems approximately have the same complexity, but both of them are usually required for a wavelet-based enhancement system and we try to eliminate the need to use the voiced/unvoiced detection.

In this paper, we first describe the wavelet transform, the wavelet packet transform, and their applications in denoising of speech signals. Then, we propose a technique to improve the method, and finally our VAD system is described. The implementation results of the proposed algorithm to some noisy speech signals are also reported. Finally, the proposed algorithm is compared with some wavelet-based speech enhancement algorithms as well as based on spectral subtraction. We also compare the proposed algorithm with some enhancement algorithms, when they are used as a preprocessing block in automatic speech recognition (ASR) system in the noisy environment.

2. Wavelet thresholding

2.1. Wavelet transform

Since processing signals in the frequency domain is often easier to implement, most of single-channel speech enhancement algorithms are implemented in the frequency domain using short time Fourier transform (STFT). In comparison to STFT, wavelet transform has the advantage of using a variable length window for different frequency components. This allows the use of long-time intervals to obtain more precise low-frequency information and shorter intervals for high-frequency information (Soon et al., 1997). There are numerous types of wavelets to choose which are scaled versions of the “mother wavelet”. The continuous wavelet transform is defined as (Seok and Bae, 1997):

\[
\text{CWT}(p, \tau) = \int f(t) \psi_{p,\tau}(t) dt, \tag{1}
\]

\[
\psi_{p,\tau}(t) = \frac{1}{\sqrt{p}} \psi\left(\frac{t - \tau}{p}\right), \tag{2}
\]

where \( p \) and \( \tau \) are real and \( \psi(t) \) is the mother wavelet. The wavelets are contracted (\( p < 1 \)) or dilated (\( p > 1 \)) and moved with shift time \( \tau \) over the signal to be analyzed. Contraction and dilation scale the frequency response to allow the set of wavelets to span the desired frequency ranges (Seok and Bae, 1997).

In the wavelet analysis, we often speak of approximations and details. The approximations are the high-scale, low-frequency components of the signal. The details are the low-scale, high-frequency components. For the discrete wavelet transform, the decomposition process is shown in Fig. 1. In this figure, the symbol 2 \( \downarrow \) represents downsampling by 2.

The decomposition process can be iterated with successive approximations being decomposed in turn. Fig. 2 shows the resulting spectral characteristics of the filter bank for 3-level decomposition.

After the processing of the wavelet coefficients, the processed components can be assembled back

![Fig. 1. The decomposition process of the fast wavelet transform.](image-url)
into the enhanced signal. This process is called reconstruction. The reconstruction process is shown in Fig. 3. In this figure, the symbol $2 \uparrow$ represents upsampling by 2.

2.2. Wavelet packet transform

The wavelet packet transform is a generalization of the decomposition process that offers a richer range of possibilities in signal analysis. In the wavelet analysis, a signal is split into an approximation and a detail. The approximation is then itself split into a second-level approximation and detail and the process is repeated. In the wavelet packet analysis, the details as well as the approximations can be split. Fig. 4 shows a 3-level wavelet packet decomposition tree and Fig. 5 shows the spectral characteristics for a 3-level wavelet packet decomposition.

2.3. Denoising by thresholding

Removing the noise components by thresholding the wavelet coefficients is based on the observation that in many signals (such as the speech signals), the energy is mostly concentrated in a small number of wavelet dimensions (Sheikhzadeh and Abutalebi, 2001). The coefficients of these dimensions are relatively large in comparison to the other dimensions or to any other signals (especially noise) that have their energy spread over a large number of coefficients. So by thresholding the coefficients to set the smaller ones to zero, we can eliminate the noise components from the noisy signal components. The basic denoising algorithm can be summarized as follows.

Let $s$ be clean speech with finite length and $y$ be the corrupted speech, by white Gaussian noise with variance $\sigma^2$, so

$$ y = s + \text{noise}. \quad (3) $$

If $W$ denotes the wavelet transform matrix, Eq. (3) can be written in the wavelet domain as

$$ Y = S + \text{NOISE}, \quad (4) $$

where

$$ Y = W \cdot y, \quad S = W \cdot s, \quad \text{NOISE} = W \cdot \text{noise}. \quad (5) $$

We can obtain the wavelet coefficients of the estimated speech signal $\hat{S}$ from the wavelet coefficients of the noisy speech signal by

$$ \hat{S} = \text{THR}(Y, T), \quad (6) $$

where $\text{THR}(\cdot, \cdot)$ denotes the thresholding function and $T$ is a scalar value as the threshold.

The standard thresholding functions used in the wavelet-based methods are hard and soft thresholding functions defined as the following equations, respectively.

$$ \text{THR}_H(Y, T) = \begin{cases} Y, & |Y| \geq T, \\ 0, & |Y| < T \end{cases} \quad (7) $$

and

$$ \text{THR}_S(Y, T) = \begin{cases} \text{sign}(Y)(|Y| - T), & |Y| \geq T, \\ 0, & |Y| < T \end{cases} \quad (8) $$
Also some other thresholding functions were proposed in Chang et al. (2002) and Sheikhzadeh and Abutalebi (2001).

The proper value for the threshold can be determined in many ways. A universal threshold for the fast wavelet transform (FWT) has been introduced by Donoho (1995) as

\[
T = \hat{\sigma}\sqrt{2\ln(N)}
\]

and for the wavelet packet transform (WPT), the threshold value is determined as

\[
T = \hat{\sigma}\sqrt{2\ln(N\log_2(N))},
\]

where \(N\) is the length of the noisy signal (\(Y\)) and \(\hat{\sigma}\) is the standard deviation of zero-mean additive WGN estimated by Donoho and Johnston (1994):

\[
\hat{\sigma} = \frac{\text{MAD}}{0.6745} = \frac{\text{Median}(|c|)}{0.6745},
\]

where \(c\) is the coefficient sequence of the noise wavelet transform.

For the correlated noise situation, a level dependent threshold was proposed by Johnston and Silverman (1997):

\[
T_j = \hat{\sigma}_j\sqrt{2\ln(N_j)},
\]

where \(N_j\) is the number of the samples in the scale \(j\) and \(\hat{\sigma} = \text{MAD}_j/0.6745\) in which \(\text{MAD}_j\) represents the median absolute deviation estimated on the scale \(j\).

### 3. Proposed modifications

There are some defects with the basic wavelet thresholding method when it is faced to the noisy speech corrupted by real-life noises. First, the basic method assumes that the noise spectrum is white. However, not only the white noise does not exist in the real environment but also we are faced with colored noises in most practical systems. Therefore, the basic wavelet shrinkage does not result in good speech quality and cannot remove the non-stationary noises.

The next problem is the shrinkage of the unvoiced speech segments which contain many noise-like speech components. This leads to degradation of the quality of the enhanced speech. Also use of a single threshold for all wavelet packet bands is not reasonable and use of the classic thresholding functions like the Hard and Soft thresholding functions often brings about time-frequency discontinuities. Therefore, we propose some modifications to solve these problems. Block diagram of our proposed system is illustrated in Fig. 6.

As can be seen in Fig. 6, wavelet packet transform is applied to each input frame and the coefficients are processed for denoising. At the end of the processing, the inverse wavelet packet transform (IWPT) is applied to provide the enhanced speech. Basic blocks of the processing part of this diagram will be described here.

#### 3.1. Calculating subbands SNRs

As mentioned before, we apply the wavelet packet transform (WPT) on each input frame to produce \(2^J\) subband wavelet packets, where \(J\) is the number of levels for WPT. Then a posteriori segmental signal-to-noise ratio (SSNR) is computed for each subband:

![Block diagram of the proposed system.](image-url)
\[ \text{SSNR}_{j,k} = 10 \log_{10} \left( \frac{\sum_{l=0}^{N_j-1} \text{YWC}_{j,k}^2(l)}{\sum_{l=0}^{N_j-1} \text{NWC}_{j,k}^2(l)} \right), \]  

where \( \text{YWC}_{j,k} \) and \( \text{NWC}_{j,k} \) denotes the wavelet packet coefficients (in the level \( J \)) of the \( k \)th subband of the processing frame (noisy frame) and the estimated noise respectively. \( \text{NWC}_{j,k} \) is updated during the silence segments detected by the VAD block. The SSNR of each subband will help us sense how much the current subband is corrupted by noise. Therefore, we will use this information for denoising. In fact, the more noise a subband is contained, the more WPT coefficients of that subband will be shrunk.

### 3.2. Noise update

During the silence segments, we update the estimated subband noise energy as

\[ \text{NE}_{k,n} = \alpha \text{NE}_{k,n-1} + (1 - \alpha) \sum_{l=0}^{N_j-1} \text{NWC}_{k}^2(l), \]  

where \( \text{NE}_{k,n} \) denotes the \( k \)th subband energy of the \( n \)th silence segment and \( \text{NWC}_{k} \) denotes the wavelet packet coefficients (in the level \( J \)) of the \( k \)th subband of the silence segment and \( 0 \ll \alpha < 1 \). Update in the noise energy can help us follow the noise alterations along the time. The reason behind which the \( \alpha \) coefficient is much larger than zero, is to conquer the sudden changes of the noise in the silence segments and to smooth the noise level. The greater the \( \alpha \) coefficient, the less the effect of the new estimated subband noise energy on the overall estimated subband noise energy. The effect of the \( \alpha \) coefficient on enhancement performance does not have considerable variations when its value is around 0.9.

### 3.3. Node-dependent threshold update

During a non-speech segment, one can obtain an estimation of the noise spectrum profile. In our proposed system, we use the node-dependent threshold (Chang et al., 2002) which is defined as

\[ A_{j,k} = \hat{\alpha}_{j,k} \sqrt{2 \ln(N_j)}, \]  

where \( \hat{\alpha}_{j,k} = \text{MAD}_{j,k}/0.6745 \) and \( \text{MAD}_{j,k} \) represents the median absolute estimated on the scale \( j \) and the subband \( k \). Use of this threshold is because of the different noise level of each subband for colored and non-stationary noises. In real environment, the noise level almost differs from one frequency-band to another. So, because of the threshold dependency to the noise level (\( \hat{\alpha} \)), it is reasonable to use different threshold values for different subbands and consequently, we can more efficiently control the denoising of noisy speech signal in the presence of colored and non-stationary noises.

When the input frame is detected as a silence frame, we will update the node-dependent threshold as

\[ A_{j,k,n} = \alpha A_{j,k,n-1} + (1 - \alpha) \left( \frac{\text{MAD}_{j,k,n}}{0.6745} \right) \sqrt{2 \ln(N_j)}, \]  

where \( n \) is the index of the silence segments, \( A_{j,k,n} \) represents the estimated threshold value of the scale \( j \) and the subband \( k \) updated at the \( n \)th detected silence segment.

### 3.4. Adaptive node-dependent threshold update

In this section, we propose an adaptive node-dependent threshold. As has been described, using a single threshold value for all the wavelet packet bands is not reasonable. So we propose to apply different threshold values for different values of SSNR\(_{j,k}\) for each subband. The function which we propose to determine the adaptive node-dependent threshold value (\( T_{j,k} \)) is defined as

\[ T_{j,k} = \begin{cases} A_{j,k} + (B_{j,k} - A_{j,k})e^{\text{SSNR}_{j,k}}, & \text{SSNR}_{j,k} > 0, \\ B_{j,k}, & \text{SSNR}_{j,k} < 0, \end{cases} \]  

where \( \tau \) determines at what SSNR\(_{j,k} (=5\tau)\), the threshold should be approximately equal to \( A_{j,k,n} \), so \( 2 < \tau \) seems to be adequate so that for \( \text{SSNR} > 10 \text{ dB} \) the threshold values equal to \( A_{j,k,n} \). Also, \( A_{j,k,n} \) and \( B_{j,k,n} \) updated at the last detected silence segments are determined as

\[ A_{j,k,n} = \left( \frac{\text{MAD}_{j,k,n}}{0.6745} \right) \sqrt{2 \ln(N_j)}, \]  

\[ B_{j,k,n} = 2A_{j,k,n}. \]

Fig. 7 shows the threshold value as a function of the posteriori segmental SNR (SSNR\(_{j,k}\)).

As can be seen from Fig. 7, when the SSNR of each subband decreases (which means that the noise energy increases in comparison to the signal energy), we adapt the threshold value to shrink more WPT coefficients. In fact, we apply an over-thresholding method to WPT coefficients to remove
more noise coefficients of each subband. Thus, the more the noise energy in comparison to the signal energy in a subband, the more threshold value is applied to the WPT coefficients of that subband and so, the more noise components will be removed.

For each subband, for computation of \( T_{j,k} \) by Eq. (17), we need the posteriori segmental SNR (SSNR) of that subband which is determined by Eq. (13) and updated using Eq. (14). We also need \( A_{j,k,n} \) (the node-dependent threshold) which is determined by Eq. (15) and updated using Eq. (16). Therefore, \( T_{j,k} \) is also updated and adapted through its calculation.

3.5. Thresholding function

For the thresholding function, we propose to use a modified version of the hard thresholding function instead of the standard form which sets the wavelet coefficients to zero when lower than the threshold value (that causes time–frequency discontinuities in the enhanced speech spectrum). So we apply a nonlinear function to the threshold value. The proposed thresholding function is determined as

\[
\text{THR}(Y, T_{j,k}) = \begin{cases} 
Y, & |Y| \geq T_{j,k}, \\
\text{sign}(Y) \cdot \frac{|Y|}{T_{j,k}}, & |Y| < T_{j,k}.
\end{cases}
\]

(20)

Fig. 8 shows the input–output characteristics of the modified thresholding function. As can be seen, the \( \gamma \) parameter can be determined by optimization. We have seen that this modified function can improve the enhancement performance alone.

3.6. Voice activity detector (VAD)

Voice activity detector permits distinction between the speech and the non-speech segments. VAD should be robust to the noisy context. Many different algorithms of voice activity detection have been proposed for several applications including mobile communication services (Freeman et al., 1989), real-time speech transmission on the internet (Sangwan et al., 2002) and noise reduction for digital hearing aids (Itoh and Mizushima, 1997). The detection principles are based fundamentally on the signal subband energy (Marzinzik and Kollmeier, 2002), its spectrum (Sohn et al., 1999; Cho et al., 2001), zero crossing rates (ZCR) (ITU-T Recommendation G, 1996), etc.

Our proposed VAD algorithm is based on the wavelet packets energy and uses the proposed denoising algorithm described here. For convenience, we assume that the first input frame is silence. If this assumption does not hold, it will converge slower and misclassify a few frames. The VAD algorithm can be described by the following steps (as shown in Fig. 9):

1. Update the node-dependent threshold value and the subband noise energy using Eqs. (16) and (14), respectively.
2. Calculate the posteriori segmental SNR for each subband using Eq. (13).
3. Update the adaptive node-dependent threshold value using Eq. (17).
4. Apply the thresholding function in Eq. (20) to the wavelet packet coefficients of each subband.
5. For each of denoised wavelet packets, calculate its energy and name it VADSE$_k$.

6. If for all bands, VADSE$_k$ is less than the last updated noise subband energy, the current frame will be classified as a silence frame.

After implementing the VAD, the enhancement process should be continued according to the block diagrams of Figs. 6 and 9. This means that the enhancement algorithm is used for two times, once for VAD and another for the main procedure. It should be noted that one can use any other VAD from other literature. The proposed VAD does not make or fit anything in this algorithm.

4. Experimental results

The proposed algorithm has been tested on the spoken English sentence chosen from the TIMIT database. The sentence is about 2.7 s with the sampling rate of 16 kHz and spoken by a male speaker. The used wavelet was “db8”, but for most of the wavelets, the results did not have any considerable changes. Also some parameters like $\gamma$ and $J$ were optimized through the experiments. As mentioned in Section 3.2, the $\alpha$ coefficient should be around 0.9 and its variation around this value does not have considerable effects on the enhancement performance. Thus, this coefficient was set to 0.9 and also each frame contained 512 samples. The experiments showed that $\tau$ factor also did not affect considerably on the enhancement performance when it is between 2 and 4. For optimization, 60 speech signals were chosen from TIMIT database as the development set as well as 60 signals as the test set. Then we varied $\gamma$, $J$ and the input SNRs so that the average SNR results can be considered to survey the feasibility of optimization. In order to optimize the factor “$\gamma$”, it was varied from 2 to 6 for corrupted speech signals (by WGN, pink, and babble noises) with input SNRs of $-5$ dB to $15$ dB (with $5$ dB steps). The average SNRs of enhanced speech signals for both of the development and the test set did not have considerable differences. So, the average SNRs of 360 (120 signals $\times$ 3 noise types) enhanced speech signals are shown in Table 1. As can be seen, the best value for $\gamma$ is $\gamma = 3$, when the speech signals are corrupted with different kinds of noises with different SNRs. Also the number of wavelet packet decomposition levels ($J$) was varied from 1 to 5 and the average SNRs of the 360 enhanced speech signals (corrupted by WGN, pink, and babble noises) are displayed in Table 2. As can be seen, the best value is $J = 3$. The global signal-to-noise ratio (SNR) values are determined by the following equation as the objective evaluation criterion:
\[
\text{SNR} = 10 \times \log_{10}\left(\frac{\sum_{n=1}^{N}s^2(n)}{\sum_{n=1}^{N}|s(n) - \hat{s}(n)|^2}\right),
\]

where \(N\) is the number of the samples in the clean and enhanced signals.

In order to show the VAD function, it was applied to the speech corrupted by WGN and pink noise with SNR of 10 dB. Fig. 10(a)–(c) shows the clean speech, noisy speech by WGN, and VAD output for it. Fig. 10(d) and (e) shows the noisy speech by pink noise and VAD output for it.

Then the proposed algorithm was applied to the speech corrupted by WGN with 10 dB SNR. To have more sense about the threshold values, we show the graph of threshold values vs. SSNR for all subbands in all frames. Fig. 11 shows this graph. As can be seen from this figure, the pattern of the threshold values follows that of Fig. 7.

Fig. 12(a)–(c) shows the results in the time domain and Fig. 13(a)–(c) shows the spectrograms of clean, noisy and enhanced speech by the proposed algorithm.

For the colored noise, the algorithm was tested on the speech corrupted by 10 dB pink noise.

---

Table 1
Average SNRs of 360 enhanced speech signals (corrupted by WGN, pink, and babble noises) as a function of \(\gamma\) and the input SNRs (for \(''J=3''\))

<table>
<thead>
<tr>
<th>Input SNRs (dB)</th>
<th>Output SNRs (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\gamma)</td>
<td>2</td>
</tr>
<tr>
<td>−5</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>5</td>
<td>8.2</td>
</tr>
<tr>
<td>10</td>
<td>12.5</td>
</tr>
<tr>
<td>15</td>
<td>17.0</td>
</tr>
</tbody>
</table>

Table 2
Average SNRs of 360 enhanced speech signals (corrupted by WGN, pink, and babble noises) as a function of \(J\) and the input SNRs (for \(''\gamma=3''\))

<table>
<thead>
<tr>
<th>Input SNRs (dB)</th>
<th>Output SNRs (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(J)</td>
<td>1</td>
</tr>
<tr>
<td>−5</td>
<td>1.6</td>
</tr>
<tr>
<td>0</td>
<td>3.9</td>
</tr>
<tr>
<td>5</td>
<td>7.1</td>
</tr>
<tr>
<td>10</td>
<td>11.3</td>
</tr>
<tr>
<td>15</td>
<td>15.9</td>
</tr>
</tbody>
</table>

---

Fig. 10. Voice Activity Detector (VAD) output for speech corrupted by 10 dB SNR: (a) clean speech; (b) noisy speech by WGN; (c) VAD output for noisy speech of (b); (d) noisy speech by pink noise; (e) VAD output for noisy speech of (d).
Fig. 14 (a)–(c) shows the results in the time domain and Fig. 15 (a)–(c) shows the spectrograms of the clean, noisy, and enhanced speech by the proposed algorithm.

The proposed wavelet packet thresholding algorithm has been implemented to enhance sixty speech signals corrupted by WGN, pink, and multi-talker babble noises with several SNRs for
several speakers. In this section, first, the influence of the adaptive thresholding function and the noise update method are analyzed to clarify which of them mainly affects the speech quality improvement.

In order to survey this matter, the average enhancement results of the proposed algorithm on sixty corrupted signals are shown in Table 3, when it includes the adaptive thresholding function and does not include the noise update method. Also, Table 4 displays the average enhancement results of the proposed algorithm when it includes the noise update method and does not include the adaptive thresholding function. As can be seen, the adaptive thresholding function is more effective on the SNR improvement than that of the noise update method.

Also, the performances of the proposed algorithm have been compared with six algorithms including the basic spectral subtraction algorithm proposed by Boll (1979) (named “BSS”), the proposed algorithm by Kamath and Loizou (2002) (named “MBSS”), our previous proposed algorithm (Ghanbari and Karami, 2004) (named “SSWD”), the basic wavelet thresholding algorithm proposed by Donoho (1995) (named “BWT”), the proposed algorithm by Sheikhzadeh and Abutalebi (2001) (named “IWBS”), and the proposed algorithm by Seok and Bae (1997) (named “SERNWD”). The mentioned algorithms have been implemented and tested on the same 60 utterances used before. The average global SNR results of the noisy signal by WGN are depicted in Fig. 16. The average performance for noisy signals by pink and babble noises are reported in Table 5. As can be seen, the proposed algorithm has considerable performance improvements. Another noteworthy point which should be noticed is that the six algorithms which have been compared with the proposed algorithm have relatively bad spectrograms in comparison to the new algorithm. So, the new algorithm was shown to be much better in comparison to the other algorithms.

The speech recognition was also considered to evaluate the proposed algorithm in comparison to the others. The enhancement system was evaluated on the TIMIT database which contains 6300 sentences spoken by 630 speakers. TIMIT database has 2432 distinct sentences among which two sentences are shared among all of the speakers. This
Fig. 14. The time domain results for (a) clean speech; (b) noisy speech corrupted by pink noise (SNR = 10 dB); (c) enhanced speech by the proposed method (SNR = 12.2 dB).

Fig. 15. The spectrograms for (a) clean speech; (b) noisy speech corrupted by pink noise (SNR = 10 dB); (c) enhanced speech by the proposed method (SNR = 12.2 dB).
The database is divided into the training and the test sets containing 462 and 168 speakers, respectively. In the current work, both of the shared sentences among all of the speakers were selected (for the 8 dialect regions). Therefore, 21 words were achieved which all of the speakers had spoken them. The employed training and test set also contained 462 and 168 speakers.

The recognition system was based on the hidden Markov model (HMM) (Deller et al., 2000; Rabiner, 1989) with 8 state and 10 mixtures. The number of states and mixtures were selected according to some experiments which showed that this model can adequately perform the recognition of clean signals in the test set. The models of the 21 words were trained under the clean signals in the training set. The employed features were the Mel frequency cepstral coefficients (MFCC) and their first and second derivative coefficients. In order to evaluate and compare some of the enhancement algorithms with the proposed one, the test signals were corrupted by WGN, pink, and babble noises with 0 dB, 5 dB, 10 dB, and 15 dB SNRs. Then, the noisy test signals were enhanced by the BSS, MBSS, BWT, SERNCWD, and the proposed algorithms and the enhanced signals were evaluated by the recognition system. The evaluation was based on the performance defined as the following equation:

\[
\text{Performance} = \frac{N - D - S - I}{N} \times 100
\]

where \(N\) is the total number of labels in the reference transcriptions, \(S\) is the number of substitution errors, \(D\) is the number of deletion errors, and \(I\) is the number of insertion errors.

The recognition test on the clean test dataset led to 85% performance. Table 6 displays the ASR performances for the enhanced test signals averaged over three kinds of noises.

As can be seen, the proposed algorithm has better performances in comparison with the other mentioned algorithms. It should be noticed that the proposed VAD is used just for the proposed speech enhancement algorithm and it does not eliminate the recognition of silence segments. So, the increased WRR in noisy conditions is mainly due
to the behaviour of both adaptive thresholding and voice activity detector (VAD) in the proposed algorithm.

5. Conclusion

The adaptive wavelet packet thresholding algorithm has shown that it can considerably enhance the noisy speech corrupted by white and colored noises. This approach uses a wavelet signal processing strategy and controls the threshold values based on a posteriori subband SNRs to remove more noise components from the subbands which are further corrupted by noise level. The SNR and ASR evaluations demonstrated the ability of this system to improve the performance of the speech enhancement system for both SNR and ASR improvement targets. Also, the other advantage of this algorithm is that unlike most of the other wavelet-based algorithms in which the detection of unvoiced segments highly affects their performances, it does not require any voiced/unvoiced detection method. Although the SNR performance of some spectral subtraction based methods is near to that of the proposed method, but the ASR tests prove its strength in reduction of the noise components as well as increasing the recognition rate in noisy environment.

References


Table 5
The average performance of seven algorithms for sixty noisy signals by pink and babble noises tested on 60 TIMIT speech signals

<table>
<thead>
<tr>
<th>Noisy input</th>
<th>Enhanced output</th>
<th>BSS</th>
<th>MBSS</th>
<th>SSWD</th>
<th>BWT</th>
<th>IWBSE</th>
<th>SERNCWD</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>By pink noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>–5</td>
<td></td>
<td>–2.5</td>
<td>–0.6</td>
<td>0.9</td>
<td>–4.7</td>
<td>–4.1</td>
<td>–3.9</td>
<td>1.6</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>2.3</td>
<td>3.7</td>
<td>4.5</td>
<td>0.2</td>
<td>0.8</td>
<td>0.8</td>
<td>5.1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>7.0</td>
<td>8.0</td>
<td>8.3</td>
<td>5.1</td>
<td>5.7</td>
<td>5.5</td>
<td>8.4</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>11.7</td>
<td>12.4</td>
<td>12.5</td>
<td>9.9</td>
<td>10.6</td>
<td>10.2</td>
<td>12.5</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>16.3</td>
<td>16.8</td>
<td>16.8</td>
<td>14.7</td>
<td>15.5</td>
<td>14.9</td>
<td>16.8</td>
</tr>
<tr>
<td>By babble noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>–5</td>
<td></td>
<td>–4.1</td>
<td>–3.5</td>
<td>–1.5</td>
<td>–5.0</td>
<td>–4.6</td>
<td>–4.9</td>
<td>–0.9</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>0.8</td>
<td>1.3</td>
<td>2.8</td>
<td>–0.0</td>
<td>0.4</td>
<td>0.1</td>
<td>3.0</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>5.7</td>
<td>6.0</td>
<td>7.2</td>
<td>5.0</td>
<td>5.3</td>
<td>5.1</td>
<td>7.5</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>10.5</td>
<td>10.8</td>
<td>11.6</td>
<td>10.0</td>
<td>10.2</td>
<td>10.0</td>
<td>11.8</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>15.4</td>
<td>15.6</td>
<td>16.2</td>
<td>15.0</td>
<td>15.2</td>
<td>15.0</td>
<td>16.5</td>
</tr>
</tbody>
</table>

Table 6
ASR results for five algorithms averaged over three kinds of noises

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>ASR performances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Input SNR→</td>
</tr>
<tr>
<td>Noisy</td>
<td></td>
</tr>
<tr>
<td>BSS</td>
<td></td>
</tr>
<tr>
<td>MBSS</td>
<td></td>
</tr>
<tr>
<td>BWT</td>
<td></td>
</tr>
<tr>
<td>SERNCWD</td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
</tr>
</tbody>
</table>


