SPECTRAL SUBTRACTION IN THE WAVELET DOMAIN FOR SPEECH ENHANCEMENT

Yasser Ghanbari

Mazandaran University, MSc. Student of Engineering, Department of Electrical Eng., 744, Babol, Iran

e-mail: ghanbari@ee.sharif.edu

Mohammad Reza Karami

Mazandaran University, Faculty of Engineering, Department of Electrical Eng., 744, Babol, Iran

e-mail: mkaramifr@yahoo.fr

Key words: Speech Processing, Speech enhancement, Spectral Subtraction, Wavelet denoising

ABSTRACT

In this paper we propose a new approach for speech enhancement. The method used to remove the noise components is a combination of two methods: Wavelet de-noising and spectral subtraction. The idea is to apply the spectral subtraction to wavelet approximations and details coefficients. A new parameter for spectral subtraction in unvoiced speech frames is introduced and the existing power factor in spectral subtraction method is improved. Also, for reduction of musical noise, we propose to use iterative Wiener filtering. Experimental results demonstrate that the proposed speech enhancement algorithm is very promising.

I. INTRODUCTION

In many speech communication systems, background noise causes the quality of speech to degrade. In most of speech processing applications such as mobile communications, speech recognition and hearing aids [6], removing the background noise in a noisy environment is inevitable. So, speech enhancement as a necessity for related applications has been widely studied in recent years.

There are several techniques for speech enhancement such as Wiener filtering[5], hidden Markov modelling[6], adaptive filtering[8], signal subspace methods[9], spectral subtraction[2,3,4] and wavelet-based methods[7,10,11]. The enhancement methods based on wavelets are almost implemented by thresholding on wavelet coefficients[7,10,12]. In such methods, choice of a good threshold and thresholding function for reduction of noise in the wavelet domain is the challenging subject[7,10,11,12]. Recently, a novel approach for noise reduction using the wavelet thresholding has been proposed by Donoho[7]. Spectral subtraction is also the most popular method on which has been widely studied in recent years.

II. ORIGINAL POWER SPECTRAL SUBTRACTION (PSS)

The original spectral subtraction method is based of three assumptions: (1) the noise is additive (2) speech signal and noise are uncorrelated (3) one channel is available. Based on the assumption that noise is additive, the corrupted speech signal is modeled by following equation:

\[ y(n) = s(n) + Gd(n) \]  (1)

Where \( s(n) \) is clean speech signal, \( d(n) \) is noise and \( G \) is the term for SNR control. We assume that the noise signal is uncorrelated:

\[ r_d(\eta) = D_\eta \delta(\eta) \]  (2)

Where \( r_d \) is the autocorrelation function of noise signal and \( D_\eta \) is a constant [1]. Because \( d(n) \) is uncorrelated process, we can show:

\[ \hat{\Gamma}_s(\omega) = \Gamma_s(\omega) + \Gamma_d(\omega) \]  (3)

Where \( \Gamma \) is the power spectral density (PSD). So, if we can estimate \( \hat{\Gamma}_d(\omega) \), we will be able to estimate \( \hat{\Gamma}_s(\omega) \):

\[ \hat{\Gamma}_s(\omega) = \Gamma_s(\omega) - \hat{\Gamma}_d(\omega) \]  (4)

Note that noise is estimated from silence frames. Because of PSD is related to Discrete-Time Fourier Transform (DTFT) [1] as:

\[ \Gamma_s(\omega) = \frac{\hat{Y}(\omega)\hat{Y}^*(\omega)}{N^2} = \frac{|\hat{Y}(\omega)|^2}{N^2} \]  (5)

We can conclude from (4) & (5):

\[ \left| \hat{S}(\omega) \right|^2 = \left| \hat{Y}(\omega) \right|^2 - \left| \hat{D}(\omega) \right|^2 \]  (6)

In order to estimate the speech signal frame, the other necessary factor is \( \hat{\varphi}_s(\omega) \) which is the estimated phase spectrum of speech frame. Boll has shown [3] that in practical applications, it is sufficient to use the noisy phase spectrum as an estimation of clean speech phase spectrum:

\[ \hat{\varphi}_s(\omega) \approx \varphi_s(\omega) \]  (7)
Therefore from equations (6) & (7), we can obtain the estimated speech frame:
\[
\hat{S}(\omega) = \hat{S}(\omega) e^{j\phi_y(\omega)} = \left[|Y(\omega)|^2 - |\hat{D}(\omega)|^2\right]^{1/2} e^{j\phi_y(\omega)} \tag{8}
\]
And a generalization for (8) is [1]:
\[
\hat{s}(n) = IDTFT\left\{\left[|Y(\omega)| - |\hat{D}(\omega)|\right]^{1/\gamma} e^{j\phi_y(\omega)}\right\} \tag{9}
\]
Where the power exponent “\(\gamma\)” can be chosen by optimization. The problem with this method is negative spectral components created after some subtractions. This problem is due to the error in estimation of the noise spectrum. So, these negative values must be modified. This problem can be eliminated by two methods [1]. One is half-wave rectification:
\[
|\hat{S}(\omega)|^2 = \begin{cases} |\hat{S}(\omega)|^2 & \text{if } |\hat{S}(\omega)| > 0 \\ 0 & \text{else} \end{cases} \tag{10}
\]
and the other is full-wave rectification:
\[
|\hat{S}(\omega)| = \text{abs}\left(|\hat{S}(\omega)|^2\right) \tag{11}
\]
These two modification methods introduce a new noise named “musical” noise which is the major limitation of spectral subtraction methods.

### III. GENERALIZED SPECTRAL SUBTRACTION (GSS)

As described, Brouti[4] proposed a generalized spectral subtraction for enhancement of speech corrupted by acoustic noise. In this method, spectral subtraction is done as the following equation:
\[
|\hat{S}(\omega)| = |Y(\omega)| - \alpha |\hat{D}(\omega)| \quad \alpha \geq 1 \tag{12}
\]
Where \(\alpha\) is over-subtraction factor and is determined as a function of input SNR, [2]:
\[
\alpha = \begin{cases} \alpha_0 + \frac{3}{4} & \text{SNR}_i \leq -5\text{db} \\ \alpha_0 - \frac{3}{20} & -5\text{db} \leq \text{SNR}_i \leq 20\text{db} \\ \alpha_0 - \frac{3}{3} & \text{SNR}_i \geq 20\text{db} \end{cases} \tag{13}
\]
The segmental SNR, of the \(i\) th noisy signal frame is calculated as:
\[
\text{SNR}_i = 10 \log_{10} \frac{\sum_{k=b_i}^{b_i} |Y_k|^2}{\sum_{k=b_i}^{b_i} |\hat{D}_k|^2} \tag{14}
\]
Where \(b_i\) and \(e_i\) are the beginning and ending frequency bins of the \(i\) th input frame.

Brouni[4] proposes for \(\alpha_0\) to be between 3 & 6. Experimental results have shown that \(\alpha_0 = 4\) is appropriate [2].

![Figure 1: Over-subtraction factor \(\alpha\) as a function of SNR with \(\alpha_0 = 4\).](image)

For negative spectral components, Brouni [4] proposes using spectral-floor factor as:
\[
|\hat{S}(\omega)| = \begin{cases} |\hat{S}(\omega)| & \text{if } |\hat{S}(\omega)| > \beta |\hat{D}(\omega)| \\ \beta |\hat{D}(\omega)| & \text{else} \end{cases} \tag{15}
\]
Where \(\beta\) is spectral-floor factor and \(\beta \ll 1\).

### IV. WAVELET TRANSFORM

Most single channel speech enhancement algorithms are applied in the frequency domain using Discrete Fourier Transform (DFT). The reason behind this, is that it is often easier to separate between noise and signal in the frequency domain than in the time domain. Wavelet transform has the advantage of using a variable window size 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 [10]. There are numerous types of wavelets to choose from. All of the wavelets are scaled version of the “mother wavelet”. The continuous wavelet transform is defined as [2]:
\[
CWT(s,t) = \int f(t)\psi_s(t) dt \tag{16}
\]
\[
\psi_s(t) = \frac{1}{\sqrt{s}} \Psi\left(\frac{t - \tau}{s}\right) \tag{17}
\]
Where \(s\) and \(\tau\) are real and \(\Psi(t)\) is the mother wavelet. The wavelets are contracted \((s<1)\) or dilated \((s>1)\) and moved over the signal to be analyzed by time shift \(\tau\). Contraction and dilation scale the frequency response to allow the set of wavelets to span the desired frequency ranges[2].
In 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 discrete wavelet transform, the decomposition process is shown in Figure 2.

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

After processing the wavelet coefficients, the processed components can be assembled back into the enhanced signal. This process is called reconstruction. The reconstruction process is shown in Figure 4.

V. SPECTRAL SUBTRACTION IN THE WAVELET DOMAIN

In this section we describe the idea of applying the spectral subtraction method to wavelet approximations and details coefficients. The first step is to apply Discrete Wavelet Transform (DWT) to the noisy signal frame, so the approximations and details coefficients are acquired. Also DWT is applied to the noise estimated from silence frames to acquire the estimated approximations and details coefficients of noise. In the next step, the GSS algorithm proposed by Brouti [4] has been improved and the improved algorithm is applied to both of approximations and details of noisy signal in parallel. The improved algorithm is described as:

\[
\begin{align*}
    &\text{if } |\hat{S}_{\text{app/dt}}(\omega)| > \beta |\hat{D}_{\text{app/dt}}(\omega)| \\
    &|\hat{S}_{\text{app/dt}}(\omega)| = |Y_{\text{app/dt}}(\omega)| - \eta \alpha_{\text{app/dt}} |\hat{D}_{\text{app/dt}}(\omega)| \\
    &\text{else} \\
    &|\hat{S}_{\text{app/dt}}(\omega)| = \beta |\hat{D}_{\text{app/dt}}(\omega)| \\
    &\text{end}
\end{align*}
\]

Where \( \hat{S}_{\text{app/dt}}(\omega) \) is the spectrum of enhanced approximations / details signal, \( Y_{\text{app/dt}}(\omega) \) is the spectrum of noisy approximations / details signal and \( \hat{D}_{\text{app/dt}}(\omega) \) is the spectrum of estimated noise approximations / details signal. The modified over-subtraction factor \( \alpha_{\text{app/dt}} \) is determined as:

\[
\alpha_{\text{app/dt}} = \begin{cases} 
\alpha_0 + 1 & \text{SNR}_{\text{app/dt,j}} \leq -5\text{db} \\
\alpha_0 - \frac{1}{5}\text{SNR}_{\text{app/dt,j}} & -5\text{db} \leq \text{SNR}_{\text{app/dt,j}} \leq 15\text{db} \\
\alpha_0 - 3 & \text{SNR}_{\text{app/dt,j}} \geq 15\text{db}
\end{cases}
\]

(18)

Brouti[4] proposes for \( \alpha_0 \) to be between 3 & 6. Our experimental results have shown that \( \alpha_0 = 4 \) is appropriate. The segmental SNR of the \( i \) th noisy approximations / details signal frame (SNR\(_{\text{app/dt,i}}\)) is calculated as:

\[
\text{SNR}_{\text{app/dt,i}} = 10 \log_{10} \frac{\sum_{k=0}^{e_i} |Y_{\text{app/dt,i}}(k)|^2}{\sum_{k=0}^{e_i} |\hat{D}_{\text{app/dt,i}}(k)|^2} 
\]

(19)

Where \( b_i \) and \( e_i \) are the beginning and ending frequency bins of the \( i \) th noisy approximations / details signal frame. We proposed a new factor \( \eta \) which is determined as:

\[
\eta = \begin{cases} 
0.5 & \text{if the speech frame is unvoiced} \\
1 & \text{else}
\end{cases}
\]

(20)

The voiced/unvoiced decision is determined by the algorithm proposed by Seok[11]. The unvoiced regions where the speech is active (not silence frames) are considered as “unvoiced speech frames”. So we need a voice activity detector (VAD). We propose the following algorithm in order to determine if the noisy frame is unvoiced speech frame:

\[
VADSNR = 10 \log_{10} \frac{\text{noisy frame energy}}{\text{estimated noise energy}}
\]

\[
\eta = 0.5 \quad \text{if } \{\text{input frame is unvoiced and } (VADSNR > 1)\}
\]

\[
\eta = 1 \quad \text{else}
\]

end
The spectral-floor factor has been determined as $\beta = 0.01$ through experiments. The power factor $\gamma$ is determined by optimization, so we varied it from 1 to 2 for corrupted speech by white Gaussian noise with global SNRs of -10db to 10db (by 5db steps). The SNRs of enhanced speech is depicted in Figure 5. Therefore, the best value was chosen $\gamma = 1.5$.

For removing the musical noise, the algorithm must be iterated. At the end of the algorithm, the enhanced signal frame is substituted for noisy signal frame and with new iteration, we will have new $SNR_{\text{app/dt,app}}$ and $a_{\text{app/dt}}$. Note that through new iteration, the old estimated noise spectrum is subtracted from old noisy signal spectrum magnitude by new $a_{\text{app/dt}}$ and old $\eta$.

Two times of iteration are enough to remove musical noise. At the end of this algorithm which is applied to both approximations and details signal in parallel, the spectrum of enhanced approximations and details signal is calculated as:

$$\hat{S}_{\text{app/dt}}(\omega) = \hat{S}_{\text{app/dt}}(\omega) e^{j\Phi_{\text{app/dt,app}}(\omega)}$$  \hspace{1cm} (21)

where $\Phi_{\text{app/dt,app}}$ (phase spectrum of enhanced approximations and details signal) is determined from equation (7):

$$\hat{\Phi}_{\text{app/dt,app}} \approx \Phi_{\text{app/dt,app}}$$  \hspace{1cm} (22)

The final step is to apply inverse DWT to reconstruct enhanced speech:

$$\hat{s}(n) = IDWT\{IDFT(\hat{S}_{\text{app/dt}}(\omega))\}$$  \hspace{1cm} (23)

VI. EXPERIMENTAL RESULTS

The proposed speech enhancement algorithm has been tested on the spoken English sentence which has been chosen from TIMIT database. The sentence is “Please shorten this skirt for Joyce” with the sampling rate of 16kHz and spoken by a male speaker. First, for optimization of power factor “$\gamma$”, we varied $\gamma$ from 1 to 2 for corrupted speech by white Gaussian noise with global SNR of -10db to 10db (by 5db steps) and the SNR of enhanced speech is depicted as a function of $\gamma$ in Figure 5.

The output enhanced speech was checked by hearing and observing the spectrogram and SNR improvement, then the best value for power factor was chosen $\gamma = 1.5$. With this value, the algorithm has been tested on the corrupted speech by white Gaussian noise. The wavelet which we used was “sym8”, but for most of wavelets, results don’t have considerable changes. Figures 6(a), 6(b) and 6(c) show the temporal results and Figures 7(a), 7(b) and 7(c) show the spectrograms of clean, noisy and enhanced speech by the proposed method.

![Figure 6. speech enhancement results:](image)

1. Clean speech
2. Noisy speech (SNR=5db)
3. Enhanced speech by proposed method (SNR=10.7db)

![Figure 7. speech enhancement spectrograms:](image)

1. Clean speech
2. Noisy speech (SNR=5db)
3. Enhanced speech by proposed method (SNR=10.7db)

The proposed spectral subtraction in the wavelet domain has been implemented to enhance the noisy speech with several SNRs on several speakers. Table 1 shows the enhancement performances.

![Figure 5. SNR of enhanced speech as a function of the power factor “$\gamma$”](image)
Table 1. SNR of speech enhancement by spectral subtraction in the wavelet domain for four speakers with different input SNRs

<table>
<thead>
<tr>
<th>Noisy SNR (db)</th>
<th>Male1</th>
<th>Male2</th>
<th>Female1</th>
<th>Female2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10</td>
<td>2.3</td>
<td>2.3</td>
<td>3.2</td>
<td>3.8</td>
</tr>
<tr>
<td>-5</td>
<td>4.9</td>
<td>4.9</td>
<td>5.5</td>
<td>6.4</td>
</tr>
<tr>
<td>0</td>
<td>7.5</td>
<td>8.1</td>
<td>8.7</td>
<td>9.1</td>
</tr>
<tr>
<td>5</td>
<td>10.7</td>
<td>11.8</td>
<td>11.7</td>
<td>12.1</td>
</tr>
<tr>
<td>10</td>
<td>14.1</td>
<td>15.6</td>
<td>14.9</td>
<td>15.4</td>
</tr>
<tr>
<td>15</td>
<td>18.1</td>
<td>19.5</td>
<td>18.3</td>
<td>18.8</td>
</tr>
<tr>
<td>20</td>
<td>22.3</td>
<td>23.4</td>
<td>22.4</td>
<td>22.5</td>
</tr>
</tbody>
</table>

VII. CONCLUSION

In the present study, the modified spectral subtraction algorithm is applied to wavelet approximations and details signal. The proposed method has shown that it can considerably enhance the noisy speech and remove the musical noise. In order to improve enhancement of unvoiced speech frames, we proposed a new factor that showed to work better relative to old algorithm. In order to remove the musical noise, we optimized the power factor in GSS algorithm and applied two iterations to the algorithm. Experimental results have shown considerable improvement in the signal-to-noise ratios as well as spectrograms.

REFERENCES