6.1 Introduction

There are a variety of speech enhancement techniques and most of them have been developed with attention to noise reduction considering normal individuals. The noise reduction techniques estimate speech noise ratio and use a mathematical criteria depending on this ratio. These algorithms improve the quality of speech but fail to improve intelligibility like Boll [8], Elberling [81] and Jamieson [82], Levitt [83]. With the filtering methods using Wiener filter or Kalman filter in the improved speech residual or musical noise is still observed. The musical noise is the result of random harmonics present during short-time frames [84][34][71][85]. Among Tsoukalas [86], Virag [87], Nandkumar and Hansen [88], Nandkumar and Hansen suggested speech improvement considering auditory masking properties. Nandkumar and Hansen have proposed a method to estimate noise and noise free speech based on MAP estimation. But the scheme did not give much attention on noise masking threshold. Sarikaya and Hansen suggested estimation of Noise Masking Threshold (NMT) using a different approach. Some researchers attempted to minimize the musical noise effect with the help of human auditory system [89-94]. The human ear has masking properties in the sense that the noise components above NMT are irritating while
those below the threshold are unimportant perceptually [66]. Now in speech quality improvement one can concentrate on removing the audible noise components. Lu and Wang [92-94] developed a linear estimator to suppress residual noise. In order to achieve musical noise suppression, the essential thing is to estimate a priori SNR accurately. Hu and Loizou [90] have given a scheme of a priori SNR estimation.

Using multi-band Wiener filter followed by harmonic regeneration a good improvement in quality of speech is observed. However, there is still some musical noise present. In order to further improve the quality of speech, a new scheme has been suggested by the author based on the work of [70]. In the new scheme a post processing is applied with a varying NMT for each critical band. As suggested by [95-96] between 1 kHz to 4 kHz there is frequency spectrum that contains disturbing musical noise. Accordingly in the present work the musical noise is observed in the critical bands 9 to 18. Some researchers [87] used a fixed offset for relative threshold, while [95][97-98] used a variable one. For better NMT estimation the author has proposed a scheme merging these two. Using this, NMT constrained perceptual weighting filter is designed and used for eliminating musical noise.

6.2 Estimation of the NMT

For estimation of noise the periods of pauses in speech are used. However, the actual noise amplitude would differ from the estimated values. This difference appears as a residual noise. The musical noise
is responsible for undesired tones. The masking of one sound with another could be either frequency masking or simultaneous masking if both sounds occur at same time. When the two appear with a delay it is known as temporal masking. In the present work the author has considered simultaneous masking. In simultaneous masking when a strong signal masks weak signal the lateral will be in audible. The modelling is done via NMT. Hence all spectral components below the NMT will be inaudible. In modelling, the NMT is obtained considering the details in [66]. The estimation of modified NMT and its offset are shown schematically in Fig.6.1.

Fig.6.1 Block diagram for estimation of NMT
6.2.1 Estimation of Critical Band Power Spectrum

A frequency band where two sounds are perceptually heard as one sound will be called a critical band, when the energy of the single sound equals the sum of energies of the constituent two sounds. The human ear sub-divides the sounds into critical bands.

The first step to find NMT is to estimate Bark scale power spectral density.

The output of multi-band Wiener filter is estimated clean speech. The DFT coefficients of this speech are used and for each critical band the energies are added. The Bark scale power spectral density is given by

\[
B(k) = \sum_{w(k)} |\hat{X}(m,k)|^2 \quad k = 1 \text{ to } K. \tag{6.2.1}
\]

where, \(k\) is critical band number and \(K\) is total number of critical bands, which depends on the sampling frequency and 18 critical bands are needed for sampling frequency 8KHz. \(w(k)\) is the frequency index and is given by:

\[
\frac{w_l}{w_s/N} \leq w(k) \leq \frac{w_h}{w_s/N} \tag{6.2.2}
\]

where \(w(k)\) is frequency index and is given by:

\[
\frac{w_l}{w_s/N} \leq w(k) \leq \frac{w_h}{w_s/N} \tag{6.2.3}
\]

where \(w_s\) is sampling frequency, \(N\) is frame size and \(w_l\) and \(w_h\) are lower and upper frequency limits of \(k^{th}\) critical band.
6.2.1.1 Convolving with a Spreading Function

With the aid of spreading function \([70]\) given by

\[
SF(k) = 15.81 + 7.5(k + 0.474) - 17.5\sqrt{1 + (k + 0.474)^2} \quad [dB] \quad (6.2.4)
\]

The interaction of critical bands and mask between them is modelled. Convolution of spreading function and bark spectrum yields critical band spectrum \(C(k)\) as

\[
C(k) = SF(k) * B(k) \quad (6.2.5)
\]

6.2.2 Calculation of Tonality Coefficients

The presence and absence of musical tones can be detected using tonality coefficient evaluated by method of \([66]\). Using denoised speech a reference signal \(\hat{X}_r(m,k)\) is constructed, which is subdivided into critical bands. The Spectral Flatness Measure (SFM) is formed from

\[
SFM_{dB} = 10\log_{10}\left[\frac{G_m}{A_m}\right]. \quad (6.2.6)
\]

Where \(G_m\) is geometric mean and \(A_m\) is arithmetic mean of the power spectrum. The tonality coefficient \(\alpha\) is estimated from

\[
\varphi_t = \min\left[1, \frac{SFM_{dB}}{SFM_{dB_{\text{max}}}}\right]. \quad (6.2.7)
\]

Using \(-60dB\) for maximum for \(SFM_{\text{max}}\) we get one or zero for \(\alpha\) for tone like signal and narrow band noise respectively. The tonality coefficients \(\varphi_{dk}\) and \(\varphi_{rk}\) for denoised speech and for reference signal are formed in each critical band. The difference of tonality coefficients \(\Delta \varphi_k = \varphi_{dk} - \varphi_{rk}\) is calculated for each critical band \(k\). Selecting \(\tau_k'\) threshold, which is nearly 0.06 for all critical bands. It is inferred that
residual noise will be present in \( k \)th critical band if \( \varphi_{dk} > \varphi_k \) and further it will be audible if \( \Delta \varphi_k > \tau'_k \).

The estimation of the NMT using tonality coefficient given in [66], [70] is modified in the present work in the sense that \( \varphi_m \) is replaced by \( \varphi_r \) if \( \Delta \varphi_k \geq \tau_k \).

### 6.2.3 Relative Threshold Offset

The NMTs are \( T_{nk} \) for noise or tone like masking signal and \( T_{nk} \) for the masked signal.

**Noise masked by tone:**

\[
T_{nk} = C(k) - 14.5 - k \quad [dB]
\]  \hspace{1cm} (6.2.8)

**Noise masking a tone:**

\[
T_{nk} = C(k) - 5.5 \quad [dB]
\]  \hspace{1cm} (6.2.9)

The offset threshold \( O(k) \) is formed from [66]

\[
O(k) = \alpha(14.5 + k) + (1 - \varphi)5.5 \quad [dB].
\]  \hspace{1cm} (6.2.10)

In order to detect musical noise in critical band 9 to 18, while calculating \( O(k) \) critical band tonality coefficient is used. To estimate modified relative threshold offset \( M(k) \), a Boolean flag based on \( \Delta \varphi_k \) is used

\[
M(k) = \begin{cases} 
1 & \text{if } \Delta \varphi_k \geq \tau'_k \\
0 & \text{otherwise}
\end{cases}
\]  \hspace{1cm} (6.2.11)

When \( M(k) = 1 \), \( O(k) \) will be nearly unity. To have a better estimation it is proposed by author to replace \( k \)th critical sub-band tonality coefficient with that of reference signal. Accordingly
\[ O_M(k) = \varphi_{mk}(14.5 + k) + (1 - \varphi)5.5 \text{ [dB]} \]  

(6.2.12)

where

\[ \varphi_{mk} = \begin{cases} 
\varphi_{dk}, & \text{for } M(k) = 1 \\
\varphi_{rk}, & \text{for } M(k) = 0 
\end{cases} \]  

(6.2.13)

Generally noisy speech signal to be processed only is available. So preliminary estimation of thresholds using (6.2.10) and/or (6.2.12) also get affected. Fixed relative threshold offset is merged with modified offset resulting in final modified threshold offset \( O'_M(k) \)

\[ O'_M(k) = \beta O_M(k) + (1 - \beta)O(k) . \]  

(6.2.14)

Where \( \beta \) is a weighting constant and used as 0.95. Using this \( O'_M(k) \) and critical band spectrum given in (6.2.5), the NMT is computed from

\[ T(k) = 10^{[\log_{10}[C(k)] - [O(k)/10]]} \]  

(6.2.15)

### 6.2.4 Renormalization and Comparison with the Absolute Threshold of Hearing

To reduce the effect convolving the \( B(k) \) with the spreading function renormalization is done. A renormalized spreading function that increases the estimate of energy in each band is used. For renormalization each \( T(k) \) is multiplied by the inverse of the energy gain in each sub-band. Finally the renormalized NMT is compared with the absolute threshold of hearing [66]. Then the modified NMT is estimated using

\[ T_f(k) = \max(T_{abs}(k), T(k)) \]  

(6.2.16)

where \( T_{abs}(k) \) is the absolute threshold of hearing.
6.3 Constrained Perceptual Weighting Filter

In the last two decades, psychoacoustic models are used to improve perceptual aspects of speech signal. Getting this perceptual aspect of human ear is incorporated in the speech enhancement process. The perceptual characteristics of human auditory system depend on its masking properties which are used to design perceptually motivated speech enhancement systems. The main idea is to mask the distortion partially or totally. Based on this, several techniques have been proposed over past two decades [2-8]. Although perceptual methods perform better compared to the non-perceptual methods, still some residual noise is left over, due to lowering of NMT. The goal of this work is to present a novel approach considering variable NMT levels for speech enhancement thereby overcoming the drawbacks of the conventional perceptual speech enhancement algorithms.

6.3.1 Derivation of Constrained Perceptual Weighting Filter

In general noisy speech signal $y(n)$ appears as sum of clean speech $x(n)$ and additive noise $d(n)$ and noise is considered as uncorrelated with original speech signal.

$$y(n) = x(n) + d(n)$$  \hspace{1cm} (6.3.1)

By applying the Fourier transform to the noisy speech signal

$$Y(m,k) = X(m,k) + D(m,k)$$  \hspace{1cm} (6.3.2)

Where $m$ is frame index and $k$ is frequency bin index. $Y(m,k), X(m,k)$ and $D(m,k)$ are spectral components of $y(n), x(n)$ and
$d(n)$. Every frequency component of the original speech $\hat{X}(m,k)$ is estimated in basic speech enhancement methods

$$\hat{X}(m,k) = H(m,k)Y(m,k) \quad (6.3.3)$$

where $H(m,k)$: spectral weighting factor. The error signal is formulated by finding the difference between original speech and estimated speech signal, which is given by

$$e(m,k) = \hat{X}(m,k) - X(m,k) = (H(m,k)Y(m,k) - X(m,k)$$

$$= (H(m,k) - 1)X(m,k) + H(m,k)D(m,k) \approx r_s + r_d. \quad (6.3.4)$$

The first term in (6.3.4) is the speech distortion due to spectral weighting and the second term is residual noise.

Let

$$\varepsilon_s^2 = E[\|r_s\|^2] = E[\|X(m,k)\|^2]H(m,k) - 1)^2 \quad (6.3.5)$$

be the power of the speech distortion. Similarly, let

$$\varepsilon_d^2 = E[\|r_d\|^2] = E[\|D(m,k)\|^2]H^2(m,k) \quad (6.3.6)$$

denotes the power of the residual noise.

**6.3.1.1 Perceptual Constraint**

The linear spectral weighting estimator with perceptual constraint on the residual noise is given by

$$\min_{H(m,k)} \varepsilon_s^2$$

subject to: $\varepsilon_d^2 \leq T_F \quad (6.3.7)$

where $T_F$ is a NMT. Using the NMT and speech distortion to residual noise ratio, the weighting factor derived minimizes the speech
distortion over all linear filters which result in the permissible residual noise level $T_F$. The spectral weighting factor is estimated in lieu of (6.3.7) and using the Kuhn-Tucker necessary conditions for constrained minimization [99-100]. Then cost function $J$ is formulated based on speech distortion and residual noise power.

$$J = \varepsilon_s^2 + \mu [\varepsilon_D^2 - T_F]$$

(6.3.8)

Where $\mu$ is the Lagrangian multiplier and this value is zero if the level of residual noise is less than noise masking threshold. Substituting (6.3.5) and (6.3.6) in (6.3.8), the cost function becomes

$$J = E[X(m,k)] H(m,k) - 1 + \mu E[D(m,k)] H^2(m,k) - T_F$$

(6.3.9)

To minimize the cost function, first we differentiate (6.3.9) with respect to the weighting factor $H(m,k)$. Equate the result to zero. The optimized spectral weighting factor is

$$H(m,k) = \frac{E[X(m,k)]}{E[X(m,k)] + \mu E[D(m,k)]}$$

(6.3.10)

Now differentiate (6.3.9) with respect to $\mu$, and equate to zero. Then obtained spectral perceptual weighting factor is

$$H(m,k) = \frac{T_r}{\sqrt{E[D(m,k)]}} \quad 0 \leq H(m,k) \leq 1$$

(6.3.11)

Since (6.3.10) should be equal to (6.3.11), the relation can be arranged as

$$E[X(m,k)] + \mu E[D(m,k)] = E[X(m,k)] \frac{E[D(m,k)]}{T_r}$$

(6.3.12)

and the Lagrangian multiplier becomes
\[
\mu = \begin{cases} 
\mathbb{E}[X(m,k)] \left( \frac{\mathbb{E}[D(m,k)]}{T_f} \right)^2 \leq 1, & \text{if } \mathbb{E}[D(m,k)] \geq T_f \\
0, & \text{if } \mathbb{E}[D(m,k)] < T_f
\end{cases} \quad (6.3.13)
\]

The accuracy of NMT is essential. The wrong estimation of NMT will affect the performance of weighting factor. Estimating a lower bound of NMT prevents the underestimation of weighting factors.

**6.3.1.2 Robust Constraint**

In order that the speech distortion is less than residual noise

\[
\frac{\varepsilon_s^2}{\varepsilon_D^2} \leq 1 \quad (6.3.14)
\]

substituting (6.3.5), and (6.3.6) into (6.3.14), we have

\[
\frac{\mathbb{E}[X(m,k)]^2}{\mathbb{E}[D(m,k)]^2} \frac{(H(m,k) - 1)^2}{H^2(m,k)} \leq 1. \quad (6.3.15)
\]

Substituting weighting factor (6.3.10) into (6.3.15), it gives

\[
\mu. \frac{\mathbb{E}[D(m,k)]}{\mathbb{E}[X(m,k)]} \leq 1. \quad (6.3.16)
\]

Substituting the Lagrangian multiplier (6.3.13) into (6.3.16), we have

\[
\left( \frac{\mathbb{E}[X(m,k)]}{\mathbb{E}[D(m,k)]} \right)^2 \left( \frac{\mathbb{E}[D(m,k)]}{T_f} \right)^2 \left( \mathbb{E}[X(m,k)] \right)^2 \leq 1 \quad (6.3.17)
\]

and

\[
\mathbb{E}[X(m,k)] \left( \frac{\mathbb{E}[D(m,k)]}{\sqrt{T_f}} - \sqrt{T_f} \right)^2 - \mathbb{E}[D(m,k)]^2 \leq 0. \quad (6.3.18)
\]

The minimum value of NMT is
\[
T_{\text{min}} = \frac{\sqrt{E[D(m,k)]^2} \cdot E[X(m,k)]^2} {\left( \sqrt{E[X(m,k)]^2} + \sqrt{E[D(m,k)]^2} \right)} .
\] (6.3.19)

### 6.3.2 Modified Perceptual Constraint

The perceptual weighting filter gain function \( H_m(m,k) \) will be optimum if the short-term spectral energy associated with the speech distortion is minimum with a condition

\[
\min_{\varepsilon_s^2}
\]

Subject to the constraint \( \varepsilon_s^2 + \varepsilon_t^2 \leq T_F . \) (6.3.20)

In this work, the perceptual gain factor is estimated in lieu of (6.3.20) and the Kuhn-Tucker conditions for constrained minimization [100]. Then cost function \( J \) is found based on speech distortion and residual noise spectral energy

\[
J = \varepsilon_s^2 + \mu [\varepsilon_s^2 + \varepsilon_t^2 - T_F] .
\] (6.3.21)

Substituting (6.3.5) and (6.3.6) in (6.3.21), the cost function becomes

\[
J = E[X(m,k)]^2 \left[ H_m(m,k) - 1 \right]^2 + \mu E[X(m,k)]^2 \left[ H_m(m,k) - 1 \right]^2 + E[D(m,k)]^2 \left[ H_m(m,k) - T_F \right] .
\] (6.3.22)

Now differentiate (6.3.22) with respect to \( H_m(m,k) \) and equate to zero. The perceptual weighting gain factor will be

\[
H_m(m,k) = \frac{E[X(k)]^2} {E[X(k)]^2 + E[D(k)]^2} \frac{\mu} {1 + \mu}
\]

\[
= \frac{\xi(m,k)} {\xi(m,k) + \beta}
\] (6.3.23)

where \( \xi(m,k) \) is the a priori SNR. \( \beta \) is equal to \( \mu / 1 + \mu \). For this \( H_m(m,k) \), the sum of speech distortion and residual noise is
Assuming that constraints given in (6.3.20) are satisfied with equality
\[
(H_m(m,k)-1)^2 E\left[|X(m,k)|^2\right] + H_m^2(m,k)E\left[|D(m,k)|^2\right]. \tag{6.3.24}
\]
and \(P = \frac{T_F}{E[|D(m,k)|^2]}\). (6.3.25) is a quadratic equation in \(H_m(m,k)\) and
will have two roots. Noting that \(\mu \geq 0\), the only possible root for optimum gain is
\[
H_m(m,k) = \frac{\xi(m,k) + \sqrt{\xi(m,k)(P-1)+P}}{\xi(m,k)+1}, P \leq 1 \tag{6.3.26}
\]
and
\[
\mu^{opt} = \frac{\xi(m,k) - \xi(m,k)\sqrt{\xi(m,k)(P-1)+P}}{(\xi(m,k)+1)\sqrt{\xi(m,k)(P-1)+P}}. \tag{6.3.27}
\]
By letting \(q = \xi(m,k)(P-1)+P\), it can be verified that \(\mu^{opt} \geq 0\) when \(P \leq 1\) and \(q \geq 0\). The gain \(H_m(m,k)\) in (6.3.26) is complex when \(q < 0\). In this case, the phase spectral components of noisy speech will get modified. However, in lieu of noise reduction techniques the phase of noisy speech signal does not change. Then the gain \(H_m(m,k)\) is used as in (6.3.11) when \(q < 0\) and when \(P > 1\), \(H_m(m,k) = 1\). This choice of \(H_m(m,k)\) is perceptually reasonable because \(P > 1\) implies \(E[|D(m,k)|^2] < T_F\). It can be seen from (6.3.24) that \(H_m(m,k) = 1\) will result in no signal distortion. Then \(H^2(m,k)E[|D(m,k)|^2] = E[|D(m,k)|] < T_F\). That is the energy of the residual
noise will be below noise masking threshold. Finally the proposed gain function is

\[
H^0_w(m,k) = \begin{cases} 
1 & P > 1 \\
H_w(m,k) & P \leq 1, q \geq 0 \\
H(m,k) & P \leq 1, q < 0 
\end{cases}
\] (6.3.28)

It can be verified from (6.3.11), (6.3.26) and (6.3.28) that proposed gain is bounded by \(0 < H^0_w(m,k) \leq 1\). In this work this modified constraint perceptual weighting filter is attached as a second stage with multi-band Wiener filter with harmonic regeneration to enhance the noise corrupted speech signal and to improve quality as well as intelligibility of the enhanced speech signal.

6.4 Simulation Results

In the experiments the speech data is obtained from the database NOIZEUS [61]. We have considered three different types noise: airport, car and train at 0, 5 and 10 dB SNRs. In all our experiments, initially we have estimated noise using the first initial 120 msec. segment because it contains only noise. Then this initial noise estimate is updated. For the proposed method, we choose frame size as 160 samples with 50% overlapping. To compare the performance of the multi-band Wiener filter with harmonic regeneration and combined with constrained perceptual weighting filter, performance measures considered are (i) the average segmental SNR and (ii) PESQ. Details of these measures are given in section 2.6. It is well known that the speech distortion accurately represented by average segmental SNR rather than the overall SNR. It is stated that higher value of the
average segmental SNR correspond to weaker speech distortion. The higher PESQ score indicates better perceived quality of the processed signal. The performance of the proposed technique is compared with multi-band Wiener filter and multi-band Wiener filter with harmonic regeneration.

The obtained output average segmental SNR values are given in Table 6.1. From the tabulated values it can be observed that for input SNR values < 10 dB improvement in segmental SNR values with constrained perceptual weighting filter is better compared to other methods in all noise environments. Fig.6.2 illustrates graphical representation of comparison of enhanced speech signals as a function of output average segmental SNR values. Table 6.2 illustrates the comparison of PESQ values of enhanced speech signals obtained with various methods. It is observed that with constrained perceptual weighting filter, improvement in PESQ values is in the range 0.2 to 0.5. Fig.6.3 represents comparison among enhanced speech signals in terms of PESQ values. From these results it is observed that multi-band Wiener filter with harmonic regeneration
<table>
<thead>
<tr>
<th>Type of noise and SNR (dB)</th>
<th>Wiener Filter (WF) [34]</th>
<th>Multi-band Wiener Filter (MWF)</th>
<th>MWF with Harmonic Regeneration (MWFHR)</th>
<th>MWFHR with Constrained Perceptual Weighting Filter (MWFHRCPWF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airport-0</td>
<td>-4.06</td>
<td>-2.39</td>
<td>-1.76</td>
<td>-0.89</td>
</tr>
<tr>
<td>Airport-5</td>
<td>-2.23</td>
<td>0.67</td>
<td>0.87</td>
<td>0.43</td>
</tr>
<tr>
<td>Airport-10</td>
<td>-0.68</td>
<td>0.43</td>
<td>0.46</td>
<td>1.09</td>
</tr>
<tr>
<td>Airport-15</td>
<td>0.77</td>
<td>3.13</td>
<td>3.43</td>
<td>3.65</td>
</tr>
<tr>
<td>Babble-0</td>
<td>-4.40</td>
<td>-1.91</td>
<td>-1.14</td>
<td>-0.32</td>
</tr>
<tr>
<td>Babble-5</td>
<td>-1.80</td>
<td>0.05</td>
<td>0.36</td>
<td>0.42</td>
</tr>
<tr>
<td>Babble-10</td>
<td>-0.16</td>
<td>2.36</td>
<td>2.48</td>
<td>2.74</td>
</tr>
<tr>
<td>Babble-15</td>
<td>0.76</td>
<td>3.06</td>
<td>3.97</td>
<td>4.15</td>
</tr>
<tr>
<td>Car-0</td>
<td>-4.48</td>
<td>-1.02</td>
<td>-1.28</td>
<td>-0.85</td>
</tr>
<tr>
<td>Car-5</td>
<td>-1.90</td>
<td>1.69</td>
<td>1.75</td>
<td>1.80</td>
</tr>
<tr>
<td>Car-10</td>
<td>-0.08</td>
<td>2.40</td>
<td>2.60</td>
<td>2.78</td>
</tr>
<tr>
<td>Car-15</td>
<td>0.75</td>
<td>2.71</td>
<td>2.95</td>
<td>3.81</td>
</tr>
<tr>
<td>Street-0</td>
<td>-3.16</td>
<td>-1.97</td>
<td>-1.39</td>
<td>-0.08</td>
</tr>
<tr>
<td>Street-5</td>
<td>-2.28</td>
<td>-0.29</td>
<td>-0.11</td>
<td>-0.13</td>
</tr>
<tr>
<td>Street-10</td>
<td>-0.31</td>
<td>2.42</td>
<td>2.54</td>
<td>2.70</td>
</tr>
<tr>
<td>Street-15</td>
<td>0.58</td>
<td>2.48</td>
<td>2.88</td>
<td>3.42</td>
</tr>
<tr>
<td>Train-0</td>
<td>-3.45</td>
<td>-2.13</td>
<td>-1.77</td>
<td>-1.15</td>
</tr>
<tr>
<td>Train-5</td>
<td>-1.75</td>
<td>0.93</td>
<td>1.90</td>
<td>2.11</td>
</tr>
<tr>
<td>Train-10</td>
<td>-0.81</td>
<td>1.69</td>
<td>2.86</td>
<td>2.98</td>
</tr>
<tr>
<td>Train-15</td>
<td>0.49</td>
<td>2.57</td>
<td>3.57</td>
<td>3.75</td>
</tr>
<tr>
<td>Restaurant-0</td>
<td>-5.11</td>
<td>-3.44</td>
<td>-3.05</td>
<td>-2.45</td>
</tr>
<tr>
<td>Restaurant-5</td>
<td>-2.72</td>
<td>-0.15</td>
<td>0.21</td>
<td>0.43</td>
</tr>
<tr>
<td>Restaurant-10</td>
<td>-0.66</td>
<td>1.28</td>
<td>1.56</td>
<td>1.98</td>
</tr>
<tr>
<td>Restaurant-15</td>
<td>0.49</td>
<td>2.47</td>
<td>2.68</td>
<td>3.15</td>
</tr>
<tr>
<td>Station-0</td>
<td>-3.72</td>
<td>0.51</td>
<td>0.89</td>
<td>0.78</td>
</tr>
<tr>
<td>Station-5</td>
<td>-1.83</td>
<td>1.18</td>
<td>1.57</td>
<td>1.65</td>
</tr>
<tr>
<td>Station-10</td>
<td>-0.36</td>
<td>2.39</td>
<td>2.89</td>
<td>2.95</td>
</tr>
<tr>
<td>Station-15</td>
<td>0.76</td>
<td>2.86</td>
<td>3.52</td>
<td>3.66</td>
</tr>
</tbody>
</table>
Fig. 6.2 Comparison of enhanced speech signals in terms of output average segmental SNR values for different noises (a) babble (b) car (c) restaurant noise (d) station (e) street and (f) train and constrained perceptual weighting filter gives better performance compared to other methods. More accurate information about the residual noise and speech distortion can be obtained from time frequency distribution of speech signals rather than time domain waveforms. Comparing spectrograms for each of the method, it is confirmed that there is a reduction of the musical noise and speech distortion. Fig.6.4 represents the spectrograms of (i) clean speech signal, (ii) noisy signal and (iii) enhanced speech signals.
### Table 6.2 PESQ values of the enhanced speech signals

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>Input SNR (dB)</th>
<th>Noisy speech</th>
<th>Wiener Filter (WF) [34]</th>
<th>Multi-band Wiener Filter (MWF)</th>
<th>MWF with Harmonic Regeneration (MWFHR)</th>
<th>MWFHR with Constrained Perceptual Weighting Filter (MWFHRCP WF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babble</td>
<td>0</td>
<td>1.01</td>
<td>1.221</td>
<td>0.952</td>
<td>1.2</td>
<td>1.427</td>
</tr>
<tr>
<td>Babble</td>
<td>5</td>
<td>1.74</td>
<td>1.728</td>
<td>1.750</td>
<td>1.77</td>
<td>1.836</td>
</tr>
<tr>
<td>Babble</td>
<td>10</td>
<td>2.10</td>
<td>2.034</td>
<td>2.276</td>
<td>2.36</td>
<td>2.402</td>
</tr>
<tr>
<td>Babble</td>
<td>15</td>
<td>2.61</td>
<td>2.127</td>
<td>2.609</td>
<td>2.65</td>
<td>2.718</td>
</tr>
<tr>
<td>Car</td>
<td>0</td>
<td>1.39</td>
<td>1.165</td>
<td>1.439</td>
<td>1.54</td>
<td>1.734</td>
</tr>
<tr>
<td>Car</td>
<td>5</td>
<td>1.82</td>
<td>1.694</td>
<td>1.697</td>
<td>1.76</td>
<td>2.107</td>
</tr>
<tr>
<td>Car</td>
<td>10</td>
<td>2.16</td>
<td>1.921</td>
<td>2.168</td>
<td>2.22</td>
<td>2.318</td>
</tr>
<tr>
<td>Car</td>
<td>15</td>
<td>2.51</td>
<td>2.265</td>
<td>2.645</td>
<td>2.98</td>
<td>3.127</td>
</tr>
<tr>
<td>Train</td>
<td>0</td>
<td>1.37</td>
<td>1.450</td>
<td>1.482</td>
<td>1.38</td>
<td>1.731</td>
</tr>
<tr>
<td>Train</td>
<td>5</td>
<td>1.72</td>
<td>1.680</td>
<td>1.715</td>
<td>1.86</td>
<td>2.133</td>
</tr>
<tr>
<td>Train</td>
<td>10</td>
<td>2.00</td>
<td>2.009</td>
<td>2.096</td>
<td>2.15</td>
<td>2.479</td>
</tr>
<tr>
<td>Train</td>
<td>15</td>
<td>2.39</td>
<td>2.040</td>
<td>2.032</td>
<td>2.35</td>
<td>2.714</td>
</tr>
<tr>
<td>Airport</td>
<td>0</td>
<td>1.58</td>
<td>1.472</td>
<td>1.561</td>
<td>1.67</td>
<td>1.759</td>
</tr>
<tr>
<td>Airport</td>
<td>5</td>
<td>2.03</td>
<td>1.492</td>
<td>1.769</td>
<td>1.98</td>
<td>2.242</td>
</tr>
<tr>
<td>Airport</td>
<td>10</td>
<td>2.37</td>
<td>2.025</td>
<td>2.413</td>
<td>2.39</td>
<td>2.538</td>
</tr>
<tr>
<td>Airport</td>
<td>15</td>
<td>2.46</td>
<td>2.249</td>
<td>2.579</td>
<td>2.56</td>
<td>2.715</td>
</tr>
<tr>
<td>Street</td>
<td>0</td>
<td>1.53</td>
<td>1.636</td>
<td>1.782</td>
<td>1.83</td>
<td>1.817</td>
</tr>
<tr>
<td>Street</td>
<td>5</td>
<td>1.65</td>
<td>1.679</td>
<td>1.857</td>
<td>1.95</td>
<td>1.968</td>
</tr>
<tr>
<td>Street</td>
<td>10</td>
<td>2.20</td>
<td>2.119</td>
<td>2.260</td>
<td>2.36</td>
<td>2.392</td>
</tr>
<tr>
<td>Street</td>
<td>15</td>
<td>2.54</td>
<td>2.380</td>
<td>2.573</td>
<td>2.59</td>
<td>2.683</td>
</tr>
</tbody>
</table>
Fig. 6.3 Comparison of enhanced speech signals in terms of PESQ values for different types of noises (a) babble (b) car (c) train (d) airport and (d) street
6.5 Conclusion

In this chapter speech enhancement using perceptual weighting filter is discussed. Intelligibility of the speech signal depends on perceptual characteristics of human auditory system. To improve quality and intelligibility of enhanced speech signal, a perceptual weighting filter is combined with multi-band Wiener filter with harmonic regeneration as post processing stage. Performance of this
method is compared with other methods in terms of average segmental SNR values and PESQ values. Simulation results confirm significant improvement with the method proposed by the author.

In the next chapter deals with enhancement of noisy speech signal in high frequency regions using unvoiced speech enhancement.