# Chapter 3

# Speech Enhancement and Detection Techniques: Transform Domain

This chapter describes techniques for additive noise removal which are transform domain methods and based mostly on short time Fourier transform (STFT). The discrete short time Fourier transform is used as transformation tool in most techniques used at present [1-2, 4]. These methods are based on the analysis-modify-synthesis approach. They use fixed analysis window length (usually 20-25ms) and frame based processing. They are based on the fact that human speech perception is not sensitive to spectral phase but the clean spectral amplitude must be properly extracted from the noisy speech to have acceptable quality speech at output and hence they are called short time spectral amplitude or attenuation (STSA) based methods [3]. The phase of noisy speech is preserved in the enhanced speech. The synthesis is mostly done using overlap-add method. They have been one of the well-known and well investigated techniques for additive noise reduction. Also they have less computation complexity and easy implementations. The detailed mathematical expression for the transfer gain function for each method is described along with the terms used in the function. The relative pros and cons of all available methods as well as applications are mentioned. The chapter starts with the brief of analysis and synthesis procedures used in the methods. The other transformation used is discrete wavelet transform (DWT) and the techniques based on DWT are also described in brief here.

The performance evaluation of any algorithm is very important for comparisons. There are several objective and subjective measures are available to evaluate the speech enhancement algorithms. The objective measures are described in brief in this chapter.

## 3.1 Signal Processing Framework

This section discusses backbone signal processing theories utilized by STSA algorithms.

### 3.1.1. Short Time Fourier Transform (STFT) Analysis

The short time Fourier transform (STFT) is a time varying Fourier representation that reflects the time varying properties of the speech waveform. The short – time Fourier transform (STFT) is given by:

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega n} \qquad (3.1)$$

where $x(m)$ is the input signal , and $w(m)$ is the analysis window, which is time – reversed and shifted by $n$ samples as shown in figure 3.1. The STFT is a function of two variables: the discrete – time index, $n$, and the (continuous) frequency variables $\omega$. To obtain $X(n+1,\omega)$, slide the window by one sample, multiply it with $x(m)$, and compute the Fourier transform of the window signal. Continuing this will generate a set of STFTs for various values of $n$ until the

end of the signal$x(m)$ is reached.

A discrete version of the STFT is obtained by sampling the frequency variable $\omega$ at $N$ uniformly spaced frequencies, i.e., at $\omega_k = \frac{2\pi k}{N}, k = 0,1, \ldots\ldots, N-1$. The resulting discrete STFT is defined as:

$$X(n, \omega_k) \triangleq X(n,k) = \sum_{m=-\infty}^{\infty} x(m)\, w(n-m) e^{\frac{-j2\pi km}{N}} \tag{3.2}$$

The STFT $X(n,\omega)$ can be interpreted in two distinct ways, depending on how one treat the time ($n$) and frequency ($\omega$) variables. If $n$ is fixed, but $\omega$ varies, $X(n,\omega)$ can be viewed as the discrete time Fourier transform of the windowed sequence $x(n-m)w(m)$. As such, $X(n,\omega)$ have the same properties as the DTFT. If $\omega$ is fixed and the time index $n$ varies, a filtering interpretation emerges.

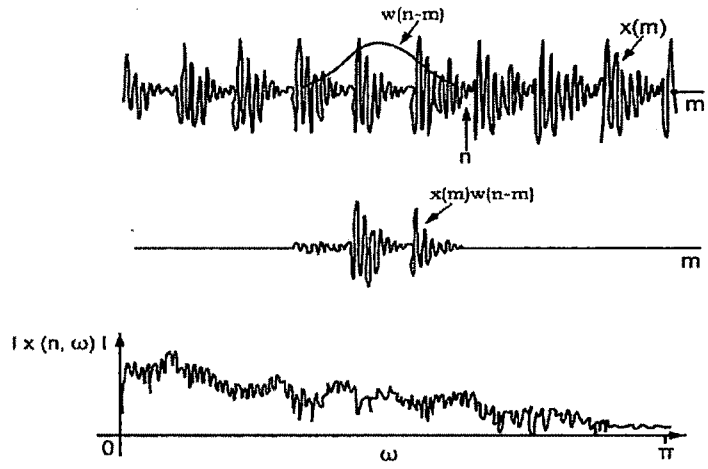


**Fig. 3.1 STFT of speech signal**

The STFT $X(n,\omega)$ is a two dimensional function of time $n$ and frequency $\omega$. In principle, $X(n, \omega)$ can be evaluated for each value of $n$; however, in practice $X(n, \omega)$ is decimated in time due partly to the heavy computational load involved and partly to the redundancy of information contained in consecutive values of $X(n,\omega)$ (e.g., between $X(n,\omega)$ and $X(n+1,\omega)$). Hence, in most practical applications $X(n,\omega)$ is not evaluated for every sample but for every $R$ sample, where R corresponds to the decimation factor, often express as a fraction of the window length. The sampling, in both time and frequency, has to be done in such a way that $x(n)$ can be recovered from $X(n,\omega)$ without aliasing.

Considering the sampling of $X(n,\omega)$ in time, from equation 3.2 it can be shown that bandwidth of the sequence $X(n,\omega_k)$ (along n, for a fixed frequency $\omega_k$) is less than or equal to the bandwidth of the analysis window $w(n)$. This suggests that $X(n,\omega_k)$ has to be sampled at twice the bandwidth of the window $w(n)$ to satisfy the Nyquist sampling criterion. For the $L$ – point Hamming window, which has an effective bandwidth of:

$$B = \frac{2F_s}{L} Hz \tag{3.3}$$

where $F_s$ is the sampling frequency. For this window,$X(n,\omega_k)$ has to be sampled in time at a minimum rate of 2B sample/sec $=\frac{4F_s}{L}$ sample /sec to avoid time aliasing. The corresponding sampling period is $\frac{L}{4F_s}$ sec or L/4 samples. This means that for an $L$ –point Hamming window $X(n,\omega_k)$ needs to evaluate at most every $L/4$ samples, corresponding to a minimum overlap of 75% between adjacent windows. This strict requirement on the minimum amount of overlap between adjacent windows can be relaxed if zeros are allowed in the window transform [5]. In speech enhancement application, it is quite common to use a 50% rather than 75% overlap between adjacent windows. This implies that $X(n,\omega_k)$ is evaluated every L/2 samples; that is, it is decimated by a factor of L/2, where L is the window length. As STFT $X(n,\omega_k)$(for fixed $n$) is the DTFT of the window sequence $w(m)x(n-m)$. Hence, to recover the windowed sequence $w(m)x(n-m)$ with no aliasing, it is required that the frequency variable $\omega$ be sampled at $N$ $(N \geq L)$ uniformly spaced frequencies, i.e., at $\omega_k = 2\pi k/N$ , $k = 0,1, \ldots \ldots N-1$.

### 3.1.2. Overlap Add Synthesis

The method for reconstructing $x(n)$ from its STFT is overlap add method, which is widely used in speech enhancement. Assuming the STFT $X(n,\omega)$ sampling in time every R samples as $X(rR,\omega)$, the overlap add method is given by following equation [5]:

$$y(n) = \sum_{r=-\infty}^{\infty} [\frac{1}{N} \sum_{k=0}^{N-1} X(rR,\omega_k) e^{j\omega_k n}] \tag{3.4}$$

The term in brackets is an IDFT yielding for each value of $r$ the sequence:

$$y_r(n) = x(n)w(rR-n) \tag{3.5}$$

Equation 3.4 can be expressed as:

$$y(n) = \sum_{r=-\infty}^{\infty} y_r(n) = x(n) \sum_{r=-\infty}^{\infty} w(rR - n) \tag{3.6}$$

From Equation 3.6 it can be seen that the signal $y(n)$ at time $n$ is obtained by summing all the sequences $y_r(n)$ that overlap at time $n$. Provided that the summation term in Equation 3.6 is constant for all $n$, we can recover $x(n)$ exactly (within a constant) as:

$$y(n) = C.x(n) \tag{3.7}$$

where $C$ is a constant. It can be shown that if $X(n, \omega)$ is sampled properly in time, i.e., $R$ is small enough to avoid time aliasing, and then C is equal to:

$$C = \sum_{r=-\infty}^{\infty} w(rR - n) = \frac{W(0)}{R} \tag{3.8}$$

independent of time $n$ [5]. Equation 3.7 and Equation 3.8 indicate that $x(n)$ can be reconstructed exactly (within a constant) by adding overlapping sections of the windowed sequences $y_r(n)$. The constraint imposed on the window is that it satisfies equation 3.8; that is, the sum of all analysis windows shifted by increments of R samples adds up to a constant. Furthermore, $R$ needs to be small enough to avoid time aliasing. With $R = {}^{L}/_{2}$ (i.e., 50% window overlap), which is most commonly used in speech enhancement, the signal $y(n)$ consists of two terms:

$$y(n) = x(n)w(R - n) + x(n)w(2R - n);\ 0 \leq n \leq R - 1 \tag{3.9}$$

Figure 3.2 shows how the overlap addition is implemented for an L-point Hamming window with 50% overlap ($R = L/2$). In the context of speech enhancement, the enhanced output signal in frame $t$ consists of the sum of the windowed signal [with $w(R - n)$] enhanced in the previous frame $(t - 1)$ and the windowed signal [with $w(2R - n)$] enhanced in the present frame $(t)$.

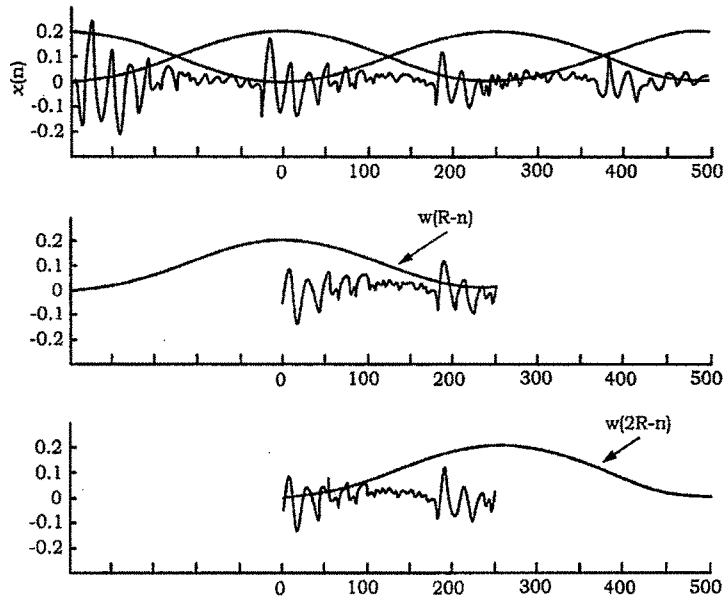


**Fig. 3.2 Overlap add synthesis with 50% overlap (L=500, R=L/2)**

Figure 3.3 shows the flow diagram of the analysis-modify-synthesis method, which can be used in any frequency domain speech enhancement algorithm. The L-point signal sequence needs to be padded with sufficient zeroes to avoid time aliasing. In the context of speech enhancement, the input signal $x(n)$ in figure 3.3 corresponds to the noisy signal and the output signal $y(n)$ to the enhanced signal.

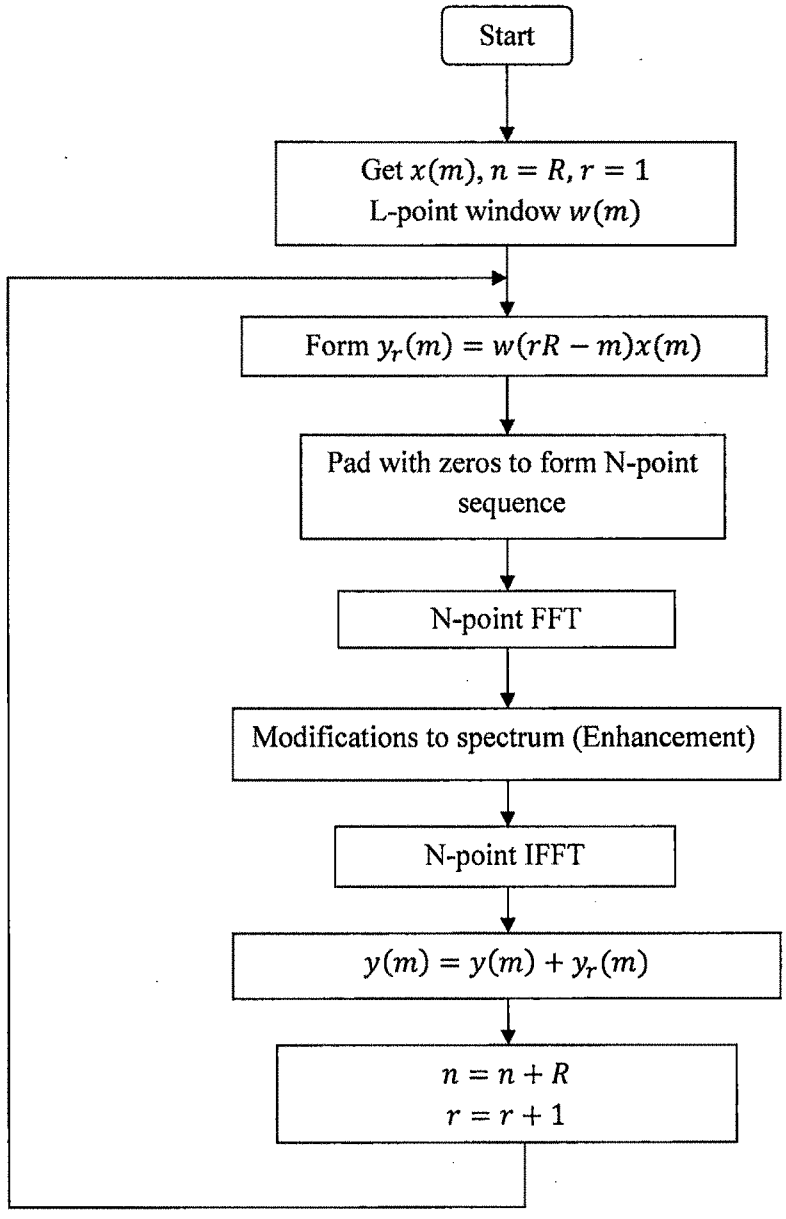


**Fig. 3.3 Flow chart of analysis-modify-synthesis method**

### 3.1.3. Spectrographic Analysis of Speech Signals

The two dimensional function $|X(n,\omega)|^2$ provides the spectrogram of the speech signal – a two dimensional graphical display of the power spectrum of speech as a function of time.

This is a widely used tool employed for studying the time varying spectral and temporal characteristic of speech. It is given by:

$$S(n,\omega) = |X(n,\omega)|^2 \tag{3.10}$$

The spectrogram describes the speech signal's relative energy concentration in frequency as a function of time and, as such, it reflects the time varying properties of the speech waveform. Frequency is plotted vertically on the spectrogram with time plotted horizontally. Amplitude, or loudness, is depicted by gray scale or color intensity. Color spectrograms represent the maximum intensity as red gradually decreasing through orange, yellow, green and blue (illustrated in figure 3.5).

Two kinds of spectrograms, narrow-band and wide-band, can be produced, depending on the window length used in the computation of $S(n,\omega)$. A long duration window (at least two pitch periods long) is typically used in the computation of the narrow-band spectrogram and a short window in the computation of the wide band spectrogram. The narrow-band spectrogram gives good frequency resolution but poor time resolution. The fine frequency resolution allows the individual harmonics of speech to be resolved. These harmonics appear as horizontal striations in the spectrogram (figure 3.5, top panel). The main drawback of using long windows is the possibility of temporally smearing short-duration segments of speech, such as the stop consonants. The wideband spectrogram uses short-duration windows (less than a pitch period) and gives good temporal resolution but poor frequency resolution. The main consequence of the poor frequency resolution is the smearing (in frequency) of individual harmonics in the speech spectrum, yielding only the spectral envelope of the spectrum (figure 3.5, bottom panel). The fundamental frequency (reciprocal of pitch period) range is about 60-150Hz for male speakers and 200-400Hz for females and children [5]. So the pitch period varies approximately 2-20ms. Therefore, in practice a compromise is made by setting a suitable practical value for window duration of 20-30ms. This way it is possible to accommodate a broad range of general speakers. These values are used throughout the research work. This also represents the harmonic structure of speech fairly correctly.

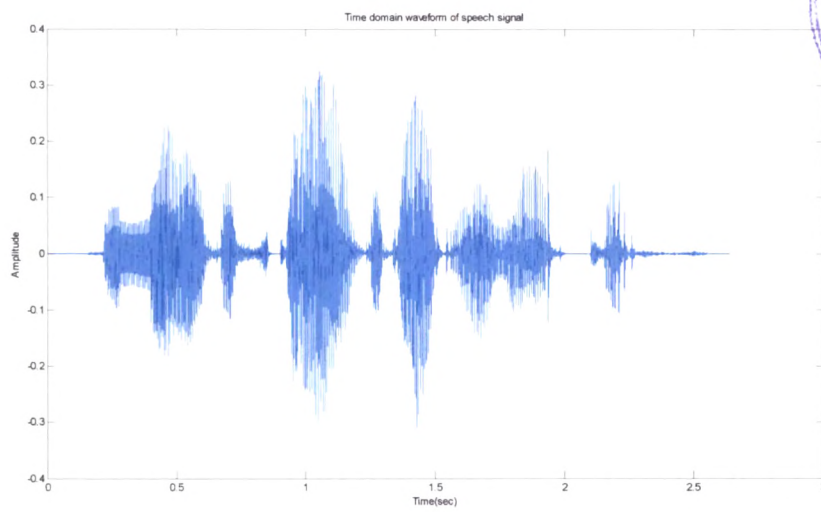


**Fig. 3.4 Time domain waveform of speech signal containing sentence 'He knew the skill of the great young actress'**

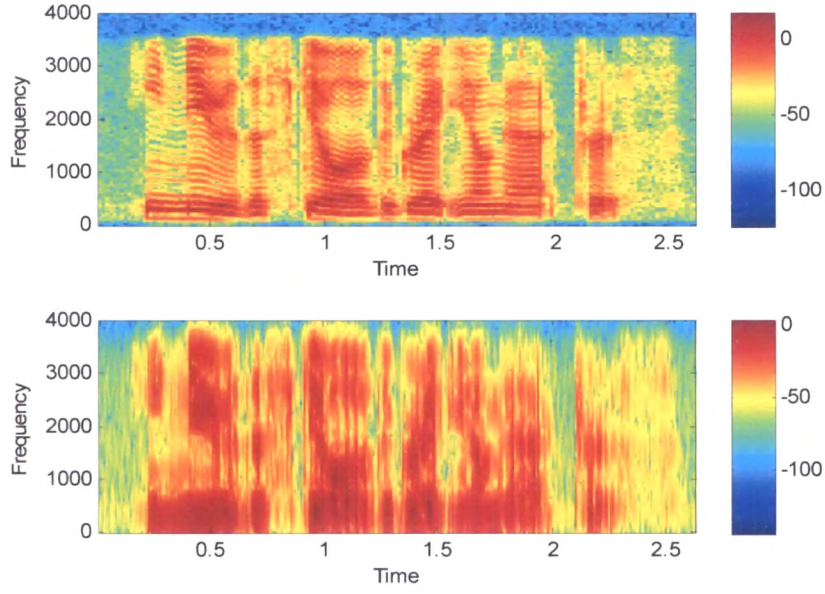


**Fig. 3.5 Narrowband (top panel) and wideband (bottom panel) spectrogram of the speech signal in figure 3.4**

## 3.2 Short Time Spectral Amplitude (STSA) Algorithms

Figure 3.6 specifies the various STSA algorithms along with their original proposer. STSA based approaches assume that noise is additive white noise and stationary for a frame and changes slowly in comparison with the speech. Most real environmental noise sources such as

vehicles, street noise, babble noise etc. are non-stationary and coloured in nature. Therefore complete noise cancellation is more complex as it is not possible to completely track such noises. However, using this assumption it is possible to achieve significant reduction in the background noise levels using simple techniques. The noise statistics are typically characterized during voice-inactivity regions between speech pauses using a voice activity detector (VAD). The VAD always becomes an integral part of any STSA based algorithm [3-4]. The operation and types of VAD are described in next section. Table 3.1 describes the list of symbols used in STSA methods description.

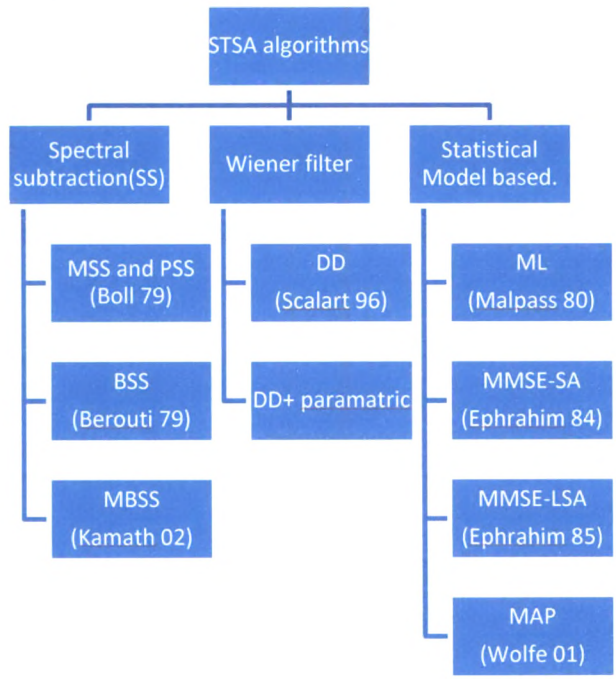


**Fig. 3.6 A chart showing various STSA algorithms**

| Symbol | Meaning |
|---|---|
| $y(n)$ | Degraded Speech signal |
| $x(n)$ | Clean speech signal |
| $d(n)$ | Additive noise |
| $\alpha$ | Over subtraction factor |
| $\beta$ | Spectral floor parameter |
| $p$ | Spectral power |
| $\eta$ | Smoothing constant |
| $K$ | Discrete frequency bin |
| $\delta$ | Tweaking factor |
| $\mu, v$ | Parameters of Wiener filter |
| $\xi(K)$ | *A priori* SNR at frequency bin K $= \frac{\lvert\hat{X}(K)\rvert^2}{\lvert\hat{D}(K)\rvert^2}$ |
| $\gamma(K)$ | *A posteriori* SNR at frequency bin K $= \frac{\lvert Y(K)\rvert^2}{\lvert\hat{D}(K)\rvert^2}$ |
| $i$ | Frequency band |
| $\phi_y(K)$ | Phase of signal y(n) at frequency bin K |
| $F_s$ | Sampling frequency |
| $f_i$ | Upper frequency in the $i^{th}$ frequency band |

**Table 3.1 List of symbols used in STSA algorithms**

## 3.3 Spectral Subtraction (SS) Methods

Spectral subtraction method was first proposed by S.F.Boll [7]. The basic principle of spectral subtraction is to subtract an estimate of the average noise spectrum from noisy speech magnitude spectrum. Degraded speech signal is modelled as

$$y(n) = x(n) + d(n) \tag{3.11}$$

Taking DFT of (1) gives

$$Y(K) = X(K) + D(K) \tag{3.12}$$

The estimate of $D(K)$ is obtained by using VAD and updated during non-speech or silence periods. For good initial estimate it requires initial silence period of around 0.2 seconds.

### 3.3.1 Magnitude and Power Spectral Subtraction (MSS and PSS)

From equation 3.12 taking only magnitude of spectrum we can write

$$|\hat{X}(K)| = \begin{cases} |Y(K)| - |\hat{D}(K)| \; if \; |Y(K)| > |\hat{D}(K)| \\ 0 \quad else \end{cases} \tag{3.13}$$

The half wave rectification process is only one of many ways of ensuring non-negative$|\hat{X}(K)|$. The original speech estimate is given by preserving the noisy speech phase $\emptyset_y(K)$. This is partly

motivated by the fact that phase that does not affect speech intelligibility [19], may affect speech quality to some degree.

$$\hat{X}(K) = [|Y(K)| - |\hat{D}(K)|]e^{j\emptyset_y(K)} \tag{3.14}$$

The preceding discussion of magnitude spectrum subtraction can be extended to power spectrum domain as

$$|\hat{X}(K)|^2 = \begin{cases} |Y(K)|^2 - |\hat{D}(K)|^2 & if\ |Y(K)|^2 > |\hat{D}(K)|^2 \\ 0 & else \end{cases} \tag{3.15}$$

The spectral power subtraction can be generalized [11] with an arbitrary spectral order $p$, called generalized spectral subtraction (GSS) and defined as

$$|\hat{X}(K)|^p = \begin{cases} |Y(K)|^p - |\hat{D}(K)|^p & if\ |Y(K)|^p > |\hat{D}(K)|^p \\ 0 & else \end{cases} \tag{3.16}$$

The general block diagram of spectral subtraction method is shown in figure 3.7.

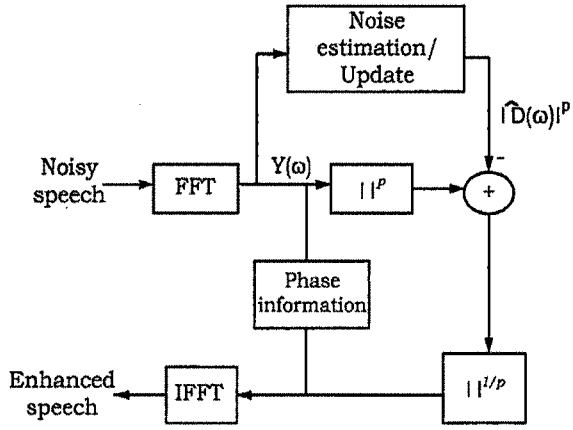


**Fig. 3.7 Block representation of general spectral subtraction method**

### 3.3.2 Berouti Spectral Subtraction (BSS)

The major problem of the basic spectral subtraction is that, the algorithm may itself introduce a synthetic noise, called musical noise. The half wave rectification is non-linear process and it creates small, isolated peaks in the spectrum occurring at random frequency locations in each frame. In time domain these peaks result in tones with randomly changing frequency from frame to frame. This musical noise is more disturbing to the listener than the original noise. Most researchers suggest that it is difficult to minimize musical noise without

affecting the speech signal. So there is always a trade-off between the amount of noise reduction and speech distortion.

Berouti *et al.* [8] proposed an important variation of the original method, which improves the noise reduction compare to the basic spectral subtraction. It introduces an over subtraction factor ($\alpha \geq 1$) and spectral floor parameter ($0 < \beta < 1$) ; and it is defined as

$$|\hat{X}(K)|^2 = \begin{cases} |Y(K)|^2 - \alpha|\hat{D}(K)|^2 \; if \; |Y(K)|^2 > (\alpha+\beta)|\hat{D}(K)|^2 \\ \beta|\hat{D}(K)|^2 \quad else \end{cases} \tag{3.17}$$

The parameter $\beta$ controls the amount of remaining residual noise and the amount of perceived musical noise. Large $\beta$ produces audible residual noise but small musical noise and vice versa. The parameter $\alpha$ affects the amount of speech spectral distortion caused by the subtraction in equation 3.17. Large values of $\alpha$ produce high speech distortion and vice versa [9]. The value of $\alpha$ should vary linearly with SNR in dB on per frame basis as

$$\alpha = \alpha_0 - s \times (SNR) \tag{3.18}$$

where $\alpha_0$ is the value of $\alpha$ at 0 dB SNR, $s$ is slope and SNR is estimated *a posteriori* frame SNR in dB. The optimized value of $\alpha_0$ is between 3 to 6 and that of $\beta$ is in the range of 0.02 to 0.06 for SNR$\leq 0dB$ and in the range of 0.005 to 0.02 for SNR$> 0dB$. Though usage of over subtraction of the noise spectrum and the introduction of a spectral floor serve to minimize residual noise and musical noise, musical noise is not completely avoided.

Equation 3.17 can be extended for general $p^{th}$ power as

$$|\hat{X}(K)|^p = \begin{cases} |Y(K)|^p - \alpha|\hat{D}(K)|^p & if \; |Y(K)|^p > (\alpha+\beta)|\hat{D}(K)|^p \\ \beta|\hat{D}(K)|^p \quad else \end{cases} \tag{3.19}$$

From this

$$H(K) = \frac{|\hat{X}(K)|}{|Y(K)|} = \begin{cases} \sqrt[p]{1 - \alpha\frac{|\hat{D}(K)|^p}{|Y(K)|^p}} = \frac{(\sqrt{\gamma(K)}^p - \alpha)^{\frac{1}{p}}}{\sqrt{\gamma(K)}} & if \; |\gamma(K)|^{p/2} > (\alpha+\beta) \\ \beta^{\frac{1}{p}} \frac{1}{\sqrt{\gamma(K)}} \quad else \end{cases} \tag{3.20}$$

In the context of linear system theory, $H(K)$ is known as the system's transfer function. In speech enhancement, $H(K)$ is referred to as the gain function, or suppression function. $H(K)$ in equation 3.18 is real and, in principle, is always positive, taking values in the range of $0 \leq H(K) \leq 1$. Negative values are sometimes obtained owing to inaccurate estimates of the noise

spectrum.$H(K)$ is called the suppression function because it provides the amount of suppression (or attenuation, as $0 \leq H(K) \leq 1$) applied to noisy power spectrum $|Y(K)|^2$ at a given frequency to obtain the enhanced power spectrum $|\hat{X}(K)|^2$. The shape of the suppression function is unique to a particular speech enhancement algorithm. For this reason, different algorithms are compared by comparing their corresponding suppressions functions.

### 3.3.3 Multiband Spectral Subtraction (MBSS)

This method proposed by S.D.Kamath [10] performs spectral subtraction with different over subtraction factor in different non-overlapped frequency bands. It is based on the fact that, in general, noise will not affect the speech signal uniformly over the whole spectrum. Some frequencies will be affected more adversely than others depending on the spectral characteristics of the noise. This can address the problem of colored noise reduction. The spectral subtraction rule in $i^{th}$ frequency band is given by

$$|\hat{X}_i(K)|^2 = \begin{cases} |\bar{Y}_i(K)|^2 - \alpha_i \delta_i |\hat{D}_i(K)|^2 \; if \; |Y_i(K)|^2 > \alpha_i \delta_i |\hat{D}_i(K)|^2 \\ \beta |\bar{Y}_i(K)|^2 \; else \end{cases} \quad for \; b_i \leq K \leq e_i \tag{3.21}$$

where the spectral floor parameter $\beta$ is set to 0.002. The over subtraction parameter in $i^{th}$ band is specified as

$$\alpha_i = \begin{cases} 4.75 & SNR_i < -5\,dB \\ 4 - \frac{3}{20}(SNR_i) & -5\,dB \leq SNR_i \leq 20\,dB \\ 1 & SNR_i > 20\,dB \end{cases} \tag{3.22}$$

where the band $SNR_i$ is given by:

$$SNR_i(dB) = 10log_{10}\left(\frac{\sum_{K=b_i}^{e_i} |\bar{Y}_i(K)|^2}{\sum_{K=b_i}^{e_i} |\hat{D}_i K)|^2}\right) \tag{3.23}$$

The additional over subtraction factor$\delta_i$; called tweaking factor provides additional degree of control in each frequency band. The values of this factor are empirically determined and set according to following equation. Usually 4-8 linearly spaced frequency bands are used.

$$\delta_i = \begin{cases} 1 & f_i < 1\,kHz \\ 2.5 & 1\,kHz \leq f_i \leq \frac{F_s}{2} - 2\,kHz \\ 1.5 & f_i > \frac{F_s}{2} - 2\,kHz \end{cases} \tag{3.24}$$

In the preceding equations $\bar{Y}_i(K)$ is the smoothed noisy spectrum of the $i^{th}$ frequency band estimated in the preprocessing stage. A weighted spectral average is taken over preceding and succeeding frames of speech as follows:

$$|\bar{Y}_j(K)| = \sum_{i=-M}^{M} W_i |Y_{j-i}(K)| \tag{3.25}$$

The number of frames $M$ is limited to 2 to prevent spectral smearing and weights $W_i$ = [0.09,0.25,0.32,0.25,0.09] set empirically. To further mask any remaining musical noise, a small amount of the noisy spectrum is introduced back to the enhanced spectrum as follows:

$$|\bar{\bar{X}}_i(K)|^2 = |\hat{X}_i(K)|^2 + 0.05\,|\bar{Y}_i(K)|^2 \tag{3.26}$$

where $|\bar{\bar{X}}_i(K)|^2$ is the newly enhanced power spectrum.

The block diagram of the multiband method proposed in [10] is shown in figure 3.8. The signal is first windowed and the magnitude spectrum is estimated using FFT. The noisy speech spectrum is then preprocessed to the noise and speech spectra are divided into N contiguous frequency bands and the over subtraction factors for each band are calculated. The individual frequency bands of the estimated noise spectrum are subtracted from the corresponding bands of the noisy speech spectrum. Lastly, the modified frequency bands are recombined and the enhanced signal is obtained by taking the IFFT of the enhanced spectrum using the noisy speech phase. The motivation behind the preprocessing stage is to reduce the variance of the spectral estimate and consequently reduce the residual noise. The preprocessing serves to precondition the input data to surmount the distortion caused by errors in the subtraction process. Hence, instead of directly using the power spectrum of the signal, a smoothed version of the power spectrum is used. Smoothing of the magnitude spectrum as per [7] was found to reduce the variance of the speech spectrum and contribute to speech quality improvement. However it is not reducing the residual noise [10].

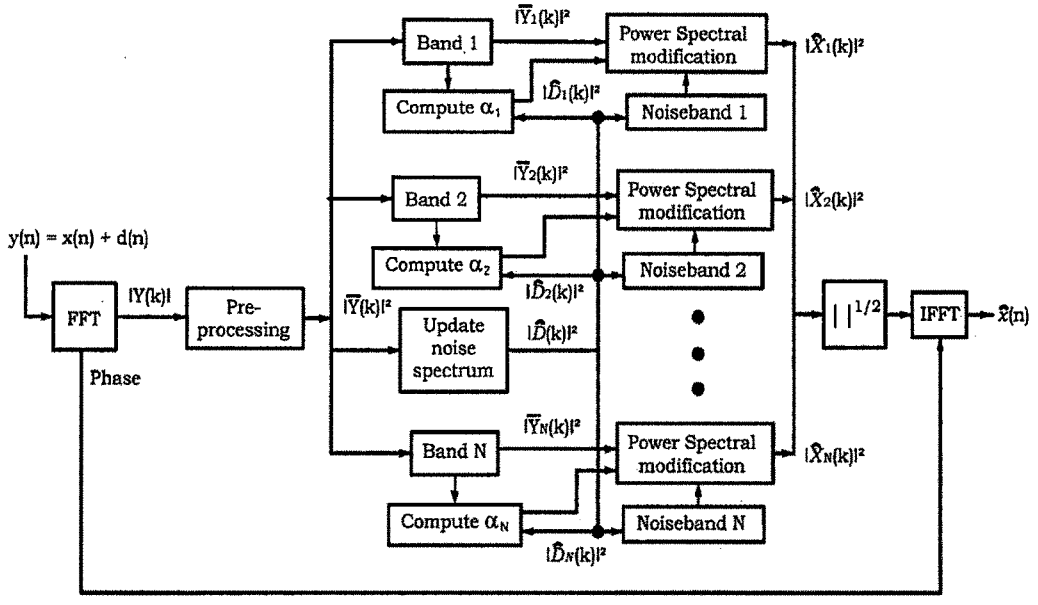


**Fig. 3.8 Block diagram of MBSS method**

## 3.4 Wiener Filtering Methods

The traditional Wiener filter used in most adaptive filtering and control applications can also be applied to speech enhancement. The Wiener filter is an optimal filter that minimizes the mean square error of a desired signal in time domain and assumes that the speech and noise are uncorrelated. In terms of our speech enhancement problem the Wiener filter is given by

$$|\hat{X}(K)| = \frac{\xi(K)}{1+\xi(K)}|Y(K)| \tag{3.27}$$

This filter is a function of *a priori* SNR.

### 3.4.1 Decision Direct (DD) Approach

The Wiener filter is non-causal and cannot be implemented in real time as it requires a prior knowledge of clean speech signal spectrum $|\hat{X}(K)|$. As a solution, Ephraim and Malah [13] proposed the decision directed rule to estimate this ratio and it is used by Scalart *et al.* [15] with Wiener filter. The decision direct rule for frame $t$ is given by

$$\xi^{(t)}(K) = \eta\frac{|\hat{X}^{(t-1)}(K)|^2}{|\hat{D}^{(t)}(K)|^2} + (1-\eta)max(\gamma^{(t)}(K)-1,0) \tag{3.28}$$

Where $0 \leq \eta \leq 1$ is smoothing constant and normally it is set to 0.98. In Wiener filter $0 \leq H(K) \leq 1$, and $H(K) \approx 0$ when $\xi(K) \to 0$ (i.e., at extremely low-SNR regions) and $H(K) \approx 1$ when $\xi(K) \to \infty$ (i.e., at extremely high-SNR regions). So, according to equation 3.26, the Wiener filter emphasizes portions of the spectrum where the SNR is high and attenuates portions of the spectrum where the SNR is low. This recursive relationship provides smoothness in the estimate of $\xi(K)$, and consequently can eliminate the musical noise [18]. Good performance was reported in [17] with the algorithm. Speech enhanced by the preceding algorithm had little speech distortion but had notable residual noise.

### 3.4.2 DD Approach with Parametric Wiener Filter

A more general Wiener filter gain function estimation was obtained by Lim and Oppenheim [6] and it is called parametric Wiener filter and it is given by

$$|\hat{X}(K)| = \left(\frac{\xi(K)}{\mu + \xi(K)}\right)^{v} |Y(K)| \tag{3.29}$$

By varying parameters $\mu$ and $v$ we can obtain different Wiener filters with different attenuation characteristics.

## 3.5 Statistical Model Based Methods

The Wiener filter is a linear estimator of the complex spectrum of the signal; an alternate approach is to use non-linear estimators of the magnitude spectrum only using various statistical model and optimization criteria. These estimators consider the probability density function (pdf) of speech and noise DFT coefficients explicitly into account and use Gaussian distribution. Various techniques of estimation theory can be applied to speech enhancement problem and mainly they fall in following categories.

### 3.5.1 Maximum Likelihood (ML) Approach

The ML approach is first applied to speech enhancement by McAulay and Malpass [12]. The magnitude and phase of clean signal are assumed to be unknown but deterministic. The pdf of noise Fourier transform coefficients is assumed to be zero-mean complex Gaussian. Based on this the ML estimation is given by

$$|\hat{X}(K)| = \frac{1}{2}\left(|Y(K)| + \sqrt{|Y(K)|^2 - |\hat{D}(K)|^2}\right) \tag{3.30}$$

Analysis reports that it provides smaller attenuation at lower SNRs compared to SS and Wiener filter methods and hence this method is not preferred as speech enhancement method.

### 3.5.2 Minimum Mean Square Error (MMSE) Approach

This method takes MMSE estimate of spectral amplitude rather than complex spectrum as in Wiener filter. The MMSE-SA optimization suggested by Ephrahim and Malah [13] is given by the equation:

$$|\hat{X}(K)| = \frac{\sqrt{\pi}\sqrt{v(K)}}{2}\frac{}{\gamma(K)}e^{-\frac{v(K)}{2}}\left[(1+v(K))I_0\left(\frac{v(K)}{2}\right)+v(K)I_1\left(\frac{v(K)}{2}\right)\right]|Y(K)|; \quad (3.31)$$

$$v(K) = \frac{\xi(K)}{1+\xi(K)}\gamma(K)$$

Here $I_0(.)$ And $I_1(.)$ Denote the modified Bessel functions of zero and first order. This estimation assumes that the speech and noise signal spectrum are statistically independent zero mean complex Gaussian random variables. The decision direct rule is used to estimate *a priori* SNR. Research shows that with the speech corrupted by an additive white noise; enhanced speech with this approach has colorless residual noise; that is, the residual noise produced by this method is not musical as in SS, Wiener filter and ML method. The speech distortion is also less compared to Wiener filter. The smoothing parameter $\eta$ controls the trade-off between speech distortion and residual noise. In summary, it is the smoothing behavior of the decision-directed approach in conjunction with the suppression rule that is responsible for reducing the musical noise effect in the MMSE algorithm. Using the method of Lagrange multipliers, the optimal solution for phase estimation can be shown to be

$$\exp(j\hat{\phi}_x) = \exp(j\phi_y) \quad (3.32)$$

That is, the noisy noise phase ($\phi_y$) is the optimal in the MMSE sense.

### 3.5.3 MMSE Log Spectral Amplitude (LSA) Approach

As a variant, Ephrahim and Malah [14] proposed MMSE log spectral amplitude (MMSE-LSA) estimator based on the fact that a distortion measure with the log spectral amplitudes is more suitable for speech processing. It minimizes the mean square error of the log amplitude spectra and the estimate of the clean speech is given by the equation:

$$|\hat{X}(K)| = \frac{\xi(K)}{1+\xi(K)}\exp\left(-\frac{1}{2}\int_{v(K)}^{\infty}\frac{e^{-t}}{t}dt\right)|Y(K)| \quad (3.33)$$

The integral in preceding equation is exponential integral and can be evaluated numerically. The exponential integral,$Ei(x)$, can be approximated as follows [11]:

$$Ei(x) = \int_{x}^{\infty} \frac{e^{-x}}{x} dx \approx \frac{e^{x}}{x} \sum_{k} \frac{k!}{x^{k}} \tag{3.34}$$

This method reduces the residual noise considerably without introducing much speech distortion.

### 3.5.4 Maximum a Posteriori (MAP) Approach

This method estimates clean speech spectral amplitude based on maximization of the *a posteriori* (MAP) pdf [16]. The MAP estimator is given by the equation

$$|\hat{X}(K)| = \frac{\xi(K) + \sqrt{\xi(K)^2 + 2(1+\xi(K))\frac{\xi(K)}{\gamma(K)}}}{2(1+\xi(K))} |Y(K)| \tag{3.35}$$

The MAP and MMSE estimates are nearly same for high *a priori* and *a posteriori* SNRs. The MAP phase estimate gives the noisy phase, which also happens to the MMSE phase estimate. Also, the MAP estimator gives simple computation compared to MMSE.

Table 3.2 summarizes the gain function (suppression function) of various STSA methods. In all the spectral subtraction methods the spectral floor can be set as per equation 3.18 with different values of parameters $\alpha, \beta$ and $p$. A noise pre-processor based on STSA has been developed by Motorola for enhanced variable rate codec (EVRC) being used in CDMA based telephone systems. In this pre-processor the input speech spectrum is divided into 16 non-uniform, non-overlapping bands similar to MBSS where input speech spectrum is divided into 3 bands. The speech is enhanced by using a gain function similar to MMSE based methods to each band. The VAD used to decide speech/silence frame and noise estimation is embedded within the algorithm. The sub-modules of EVRC noise pre-processor are optimized and highly inter-dependent.

| Sr. No. | Class | Method | Gain (suppression or attenuation) function H(K) | Remarks |
|---|---|---|---|---|
| 1 | Spectral subtraction | 1.MSS | $\frac{\sqrt{\gamma(K)}-1}{\sqrt{\gamma(K)}}$ | Simple<br>High Residual noise |
| | | 2.PSS | $\sqrt{\frac{\gamma(K)-1}{\gamma(K)}}$ | Simple<br>Musical noise artifact |
| | | 3.GSS | $\frac{(\sqrt{\gamma(K)}^{p}-1)^{1/p}}{\sqrt{\gamma(K)}}$ | Flexible<br>Musical and residual noise trade-off |
| | | 4.BSS | $\sqrt{\frac{\gamma(K)-\alpha}{\gamma(K)}}$ | Simple<br>Less musical noise<br>High Residual noise |
| 2 | Wiener | 1.Scalart | $\frac{\xi(K)}{1+\xi(K)}$ | Non-causal |
| | | 2.Para-metric | $\left(\frac{\xi(K)}{\mu+\xi(K)}\right)^{v}$ | Non-causal<br>but flexible |
| 3 | Statistical modeling | 1.ML | $0.5+0.5\sqrt{\frac{\gamma(K)-1}{\gamma(K)}}$ | Less attenuation<br>Not preferred<br>High musical noise |
| | | 2.MMSE -SA | $\frac{\sqrt{\pi}}{2}\frac{\sqrt{v(K)}}{\gamma(K)}e^{-\frac{v(K)}{2}}\left[(1+v(K))I_0\left(\frac{v(K)}{2}\right)+v(K)I_1\left(\frac{v(K)}{2}\right)\right]$ | Complicated<br>Less musical and residual noise but Speech distortion |
| | | 3.MMSE -LSA | $\frac{\xi(K)}{1+\xi(K)}\exp\left(-\frac{1}{2}\int_{v(K)}^{\infty}\frac{e^{-t}}{t}dt\right)$ | Complicated<br>Less musical and residual noise with less speech distortion |
| | | 4.MAP | $\frac{\xi(K)+\sqrt{\xi(K)^2+2(1+\xi(K))\frac{\xi(K)}{\gamma(K)}}}{2(1+\xi(K))}$ | Simple<br>Alternate to MMSE |

**Table 3.2 A summary of STSA methods**

## 3.6 Voice Activity Detection (VAD) and Noise Estimation

In speech communications, speech can be characterized as a discontinuous medium because of the pauses which are a unique feature compared to other multimedia signals, such as video, audio and data. The regions where voice information exists are classified as voice-active and the pauses between talk spurts are called voice-inactive or silence regions. An example illustrating active and inactive voice regions for a speech signal is shown in figure 3.9. A voice

activity detector (VAD) is an algorithm employed to detect the active and inactive regions of speech.

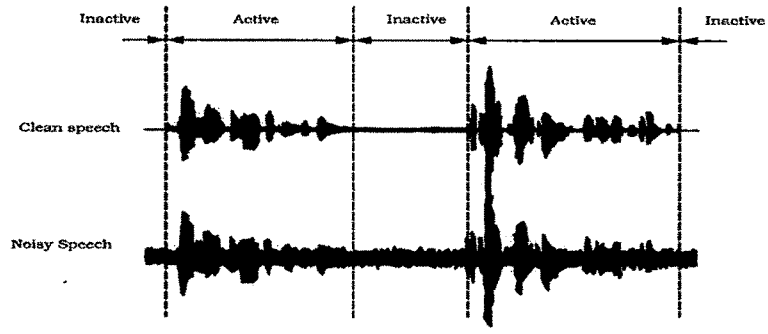


**Fig. 3.9 Voice active and inactive regions**

A practical speech enhancement system consists of two major components, the estimation of noise power spectrum, and the estimation of clean speech. The first part is performed along with voice activity detection (VAD) and second part uses output from first part and apply algorithm for clean speech estimation. Therefore, a critical component of any frequency domain enhancement algorithm is the estimation of the noise power spectrum [19]. The basic VAD and noise estimation operation is described in figure 3.10.

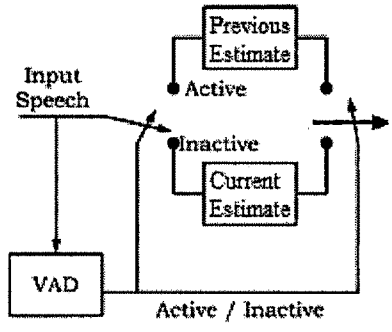


**Fig. 3.10 Block diagram of VAD and noise estimation**

The speech/silence detection finds out the frames of the noisy speech that contain only noise. Speech pauses or noise only, frames are essential to estimate noise. If the speech/silence detection is not accurate then speech echoes and residual noise tend to be present in the enhanced speech. Several methods are used for VAD, such as voiced/unvoiced classification used in ITU G.723.1, zero crossing method used in G.729, and spectral comparison used in both G.729 and

GSM vocoders in addition to different power thresholds variations. However they are suitable for clean speech only. For speech enhancement it is required to operate with noisy speech and hence the magnitude spectral distance VAD which is generic, simple and easy to integrate with speech enhancement algorithm is most common in applications. In [20] it is reported that this VAD is most suitable for real time implementation.

Let $|Y(K)|$ is the current frames magnitude spectrum, which is to be labeled as noise or speech, N is noise magnitude spectrum template (estimation), NC is noise counter which reflects the number of immediate previous noise frames, NM is noise margin and it is spectral distance threshold. Hangover counter is the number of noise segments after which the "Speech flag" resets (goes to zero). "Noise flag" is set to one if the segment is labeled as noise. Spectral distance is calculated by using following formula and based on this the decision is taken.

$$\begin{gathered} Spectral\ distance\ =\ log_{10}(|Y(K)|) - log_{10}(N) \\ If\ Spectral\ distance < NM: Noise\ flag = 1, NC = NC + 1 \\ Else:\ Noise\ flag = 0, NC = 0 \\ If\ NC > Hangover\ counter: Speech\ flag = 0 \\ Else: Speech\ flag = 1 \end{gathered} \tag{3.36}$$

## 3.7 Speech Enhancement Using Wavelet Transform

The STFT allows representing the signal in frequency domain through time windowing function. The window length determines a constant time and frequency resolution. Thus, a shorter time windowing is used in order to capture the transient behavior of a signal at the cost of frequency resolution. The nature of the speech signals is quasi-stationary; such signals cannot easily be analyzed by conventional transforms. So, an alternative mathematical tool- wavelet transform should be selected to extract the relevant time amplitude information from a signal. In this thesis, only some key equations and concepts of wavelet transform are stated, more rigorous mathematical treatment of this subject can be found in [21]. A continuous time wavelet transform (CWT) of signal $x(t)$ is defined as:

$$X_w(\tau, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) h_{\tau,a}{}^{*}(t) dt;\quad h_{\tau,a}(t) = \frac{1}{\sqrt{a}} h\left(\frac{t-\tau}{a}\right) \tag{3.37}$$

Here $a\ (Scaling\ factor), \tau(Time\ shift) \in R, a \neq 0$ and they are dilating and translating

coefficients, respectively. This multiplication of $\frac{1}{\sqrt{a}}$ is for energy normalization purposes so that the transformed signal will have the same energy at every scale. The analysis function $h(t)$, the so-called mother wavelet (basic or prototype wavelet), is scaled by $a$, so a wavelet analysis is often called a time-scale analysis rather than a time-frequency analysis. The wavelet transform decomposes the signal into different scales with different levels of resolution by dilating a single prototype function, the mother wavelet. Furthermore, a mother wavelet has to satisfy that it has a zero net area, which suggest that the transformation kernel of the wavelet transform is a compactly support function (localized in time), thereby offering the potential to capture the transients [21].

Calculating wavelet coefficients at every possible scale is a fair amount of work, and it generates an awful lot of data. It turns out, rather remarkably, that if scales and positions are based on powers of two; so-called *dyadic* scales and positions; then the analysis will be much more efficient and just as accurate. Such an analysis forms the discrete wavelet transform (DWT) of discrete time signal $x(n)$.

$$a = 2^m \quad and \quad \tau = n2^m;\ m, n \in N \tag{3.38}$$

$$X_w(n,m) = \sum_{p=-\infty}^{\infty} x(p) h_{n,m}{}^*(p) \tag{3.39}$$

The family of dilated mother wavelets of selected $a$ *and* $\tau$ constitute an orthonormal basis of $L^2(R)$. In addition, sampling of $X_w(\tau, a)$ in dyadic grid also called dyadic orthonormal wavelet transform. Due to the orthonormal properties, there is no information redundancy in the DWT. In addition, with this choice of $a$ *and* $\tau$, there exists the multi-resolution analysis (MRA) algorithm, which decomposes a signal into scales with different time and frequency resolution. MRA is designed to give good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies. The discrete time dyadic wavelet transform can be efficiently implemented by using filter banks. The filtering implementation of the forward transform is given by an iterative cascade of identical stages, each stage consisting of low pass and high pass decomposition of the signal followed by the 2 to 1 down-sampling. A similar iterative structure can be used for inverting the wavelet transform from the wavelet coefficients. Further details can be obtained from [21].

The differences between different mother wavelet functions (e.g., Haar, Daubechies, Coiflets, Symlet, Biorthogonal and etc.) consist in how scaling signals and the wavelets are defined. The choice of wavelet determines the final waveform shape; likewise, for Fourier transform, the decomposed waveforms are always sinusoid. To have a unique reconstructed signal from wavelet transform, it is needed to select the orthogonal wavelets to perform the transforms.

### 3.7.1 Thresholding of Wavelet Co-efficients for Speech Enhancement

One of the first wavelet-based methods de-noising was developed by Donoho and Johnstone [22-23]. It reduces noise by thresholding the wavelet coefficients so that only the coefficient with values above the threshold are retained. Signal energy is concentrated on a small number of wavelet coefficients in many signals; while wavelet coefficients of noise are spread over a wide number of coefficients. Appropriate thresholding of wavelet coefficients can lead to high noise reduction with low signal distortion. The general wavelet de-nosing procedure is as follows:

- Apply DWT to the noisy signal to produce the noisy wavelet coefficients to the level.
- Select appropriate threshold limit at each level and threshold method to best remove the noises.
- Inverse DWT of the thresholded wavelet coefficients to obtain a de-noised signal.

Performing the DWT of equation 3.11

$$Y_{j,k} = X_{j,k} + D_{j,k} \tag{3.40}$$

where $Y_{j,k}$ is the $k^{th}$ wavelet coefficient in the scale $j$. There are two common ways to threshold ($\lambda$) the resulting wavelet coefficients. The first is referred to as hard thresholding which sets the coefficients to zero whose absolute value is below the threshold.

$$\hat{Y}_{j,k} = \begin{cases} Y_{j,k} & if \ |Y_{j,k}| > \lambda \\ 0 & else. \end{cases} \tag{3.41}$$

Soft thresholding goes one step further and decreases the magnitude of the remaining coefficients by the threshold value

$$\hat{Y}_{j,k}^{soft} = sign\,(Y_{j,k})\, max(|Y_{j,k}| - \lambda, 0) \tag{3.42}$$

Hard thresholding maintains the scale of the signal but introduces ringing and artifacts after reconstruction due to a discontinuity in the wavelet coefficients. Soft thresholding eliminates this

discontinuity resulting in smoother signals slightly decreases the magnitude of the reconstructed signal.

Many methods for setting the threshold have been proposed. The most time-consuming way is to set the threshold limit on a case-by-case basis. The limit is selected such that satisfactory noise removal is achieved. For a Gaussian noise if orthogonal wavelet transform is applied to the noise signal, the transformed signal will preserve the Gaussian nature of the noise, which the histogram of the noise will be a symmetrical bell-shaped curve about its mean value. To obtain the threshold value for a signal of length $d$, the approach in [22] seeks to minimize the maximum error over all possible samples. This method assumes that $d(t)$ having some known standard deviation $\sigma$. The universal threshold is given by

$$\lambda_{uni} = \sigma\sqrt{2\log(d)} \tag{3.43}$$

is shown to be asymptotically optimal in the minimax sense when employed as a hard threshold with $\sigma = MAD/0.6745$, where MAD represents the absolute median estimated on the first scale $Y_{HH_1}$. Donoho and Johnstone [23] also proposed a more advanced strategy based on Stein's unbiased risk estimate (SURE). Here, soft thresholding is used because it is more mathematically tractable (i.e., continuous) and the clean signal is estimated as

$$\lambda_{SURE} = argmin_{0\le\lambda\le\sqrt{2log(d)}} SURE(\lambda, Y_j) \tag{3.44}$$

$$where;\ SURE(\lambda, Y_j) = \sigma^2 + \frac{1}{c}\sum_{k=1}^{c}[min(|Y_{j,k}|, \lambda]^2 - \frac{2\sigma^2}{c}\sum_{k=1}^{c} I(|Y_{j,k}| < \lambda)$$

Johnstone and Silverman [24] studied the correlated noise situation and proposed a "level-dependent" threshold

$$\lambda_j = \sigma_j\sqrt{2\log(d_j)} \tag{3.45}$$

with $\sigma = MAD/0.6745$, and $d_j$ is the number of samples in scale $j$.

During the past decade, the wavelet transforms have been applied to various research areas. Their applications include signal and image de-noising, compression, detection, and pattern recognition. To the best of knowledge, de-noising methods based on the wavelet thresholding have not been successfully applied to speech enhancement. The difficulties are simultaneously associated to the speech signal complexity and to the nature of the noise.

However, to improve the wavelet thresholding enhancement, following suggestions are proposed [25-27]:

- The use of the wavelet packet transform (WPT) instead of the wavelet transform,
- To extend the concept of the level-dependent threshold (Equation 3.45) to the WPT,
- The use of time-adapted threshold based on the speech waveform energy.

As a result the wavelet based techniques are ruled out here for further refinements. It is considered in next chapter only for comparison with the STSA based techniques.

## 3.8 Objective Quality Measures for Speech Enhancement Methods

Quality is one of many attributes of the speech signal. Quality is highly subjective in nature and it is difficult to evaluate reliably. This is partly because individual listeners have different internal standards of what constitutes "good" or "poor" quality, resulting in large variability in rating scores among listeners. Quality measures assess 'how' a speaker produces an utterance, and includes attributes such as "natural", "raspy", "hoarse", "scratchy", and so on. Quality possesses many dimensions, too many to enumerate. For practical purposes it is restricted to only a few dimensions of speech quality, depending on the application.

Intelligibility measures assess "what" the speaker said, i.e., the meaning or the content of the spoken words. Unlike quality, intelligibility is not subjective and can be easily measured by presenting to a group of listeners speech material (sentences, words, etc.) and asking them to identify the words spoken. Intelligibility is quantified by counting the number of words or phonemes identified correctly. The relationship between speech intelligibility and speech quality is not fully understood, and this is in part because no one has yet identified the acoustic correlates of quality and intelligibility [28]. A good speech enhancement algorithm needs to preserve or enhance not only speech intelligibility but also speech quality. This is based on the observation that it is possible for speech to be both highly intelligible and of poor quality. Also, although two different algorithms may produce equal word intelligibility scores, listeners may perceive the speech of one of the two algorithms as being more natural, pleasant, and acceptable. There is, therefore, the need to measure other attributes of the speech signal besides intelligibility. Reliable evaluation of speech quality is considered to be a much more challenging task than the task of evaluating speech intelligibility

Quality assessment of speech enhancement algorithms can be done using subjective listening tests or objective quality measures. Subjective listening tests uses mean opinion score

(MOS) to evaluate the performance of speech enhancement algorithms [17]. But they are time consuming, expensive, involve human subjects, not easily repeatable and rating is based on their overall perception (possess inherent variability in interpretation). A consistent listening environment is required and perceived distortion can vary with factors such as the playback volume and type of listening instrument used. For provisional investigations objective quality measures can be used. Objective evaluation involves a mathematical comparison of the original and processed speech signals. Objective measures quantify quality by measuring the numerical "distance" between the original and processed signals. Clearly, for the objective measures to be valid, it needs to correlate well with subjective listening tests, and for that reason much research has been focused on developing objective measures that model various aspects of the auditory system [29].

Objective measures of speech quality are implemented by first segmenting the speech signal into 10-30 ms frames and then computing a distortion measure between the original and processes signals. A single, global measure of speech distortion is computed by averaging the distortion measures of each speech frame. A large number of objective measures have been evaluated, particularly for speech coding applications. Reviews of objective measures can be found in [30]. The focus here is on a subset of those measures that have been found to useful for evaluation of speech enhancement algorithms [29].The STSA and wavelet based algorithms are compared using several objective measures and results are shown in chapter 4. In addition, the MOS subjective measure is also used to compare modified and proposed method with existing algorithms and it is described in chapter 6. A final comment on the quality of the enhanced speech can be made only after referring to both the objective measures and subjective test. Figure 3.11 illustrates the typical system setup.

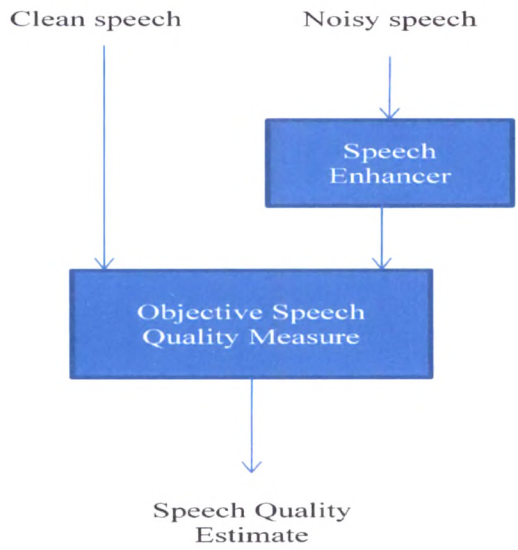


**Fig. 3.11 Objective speech quality measuring system**

Table 3.3 presents a brief summary of important objective measures used for speech quality assessments.

| Sr. No. | Objective measure | Mathematical relation | Terminology and significance |
|---|---|---|---|
| 1 | Segmental SNR (SSNR) [31] | $\frac{10}{M}\sum_{m=0}^{M-1} log_{10}\left(\frac{\sum_{n=Nm}^{Nm+N-1} x^2(n)}{\sum_{n=Nm}^{Nm+N-1}(x(n)-\hat{x}(n))^2}\right)$ | $x(n)$ is the original (clean) signal, $\hat{x}(n)$ is the enhanced signal, N is the frame length (typically chosen to be 15-20 ms), and M is the number of frames in the signal. It is based on the geometric mean of the SNRs across all frames of the speech signal. |
| 2 | Log Likelihood Ratio Distance (LLR) [4] | $d(a_x, \bar{a}_{\hat{x}}) = log\frac{\bar{a}_{\hat{x}}^T R_x \bar{a}_{\hat{x}}}{a_x^T R_x a_x}$ | $a_x^T = [1, -\alpha_x(1), -\alpha_x(2), \ldots., -\alpha_x(p)]$ are the LPC coefficient of the clean signal, $\bar{a}_{\hat{x}}^T = [1, -\alpha_{\hat{x}}(1), -\alpha_{\hat{x}}(2), \ldots., -\alpha_{\hat{x}}(p)]$ are the coefficients of the enhanced signals, and $R_x$ is the $(p+1)\times(p+1)$ autocorrelation matrix (Toeplitz) of the clean signal. It is based on the dissimilarity between all-pole models of the clean and enhanced signals. |
| 3 | Weighted Spectral Slope Distance (WSS) [32-34] | $d_{WSS}(C_x, \bar{C}_x) = \sum_{k=1}^{L} W(k)(S_x(k) - \bar{S}_{\hat{x}}(k))^2$<br>$S_x(k) = C_x(k+1) - C_x(k)$<br>$\bar{S}_{\hat{x}}(k) = \bar{C}_{\hat{x}}(k+1) - \bar{C}_{\hat{x}}(k)$ | $C_x(k)$ is clean and $C_{\hat{x}}(k)$ is enhanced critical-band spectra expressed in dB, $W(k)$ is weight for band $k$, L is the number of critical bands. It is based on phonetic distance. Thirty six overlapping filter of progressively larger bandwidths to estimate the smoothed short time speech spectrum every 12 ms are used. The filter bandwidths approximate auditory critical bands so as to give equal perceptual weight to each band. |
| 4 | Perceptual Evaluation of Speech Quality (PSEQ) [35] | The process is described by block diagram in figure 3.12. | It closely resembles to the subjective MOS measure. The range of the PESQ score is 0.5 to 4.5. |

**Table 3.3 Objective measures used for speech quality assessments**

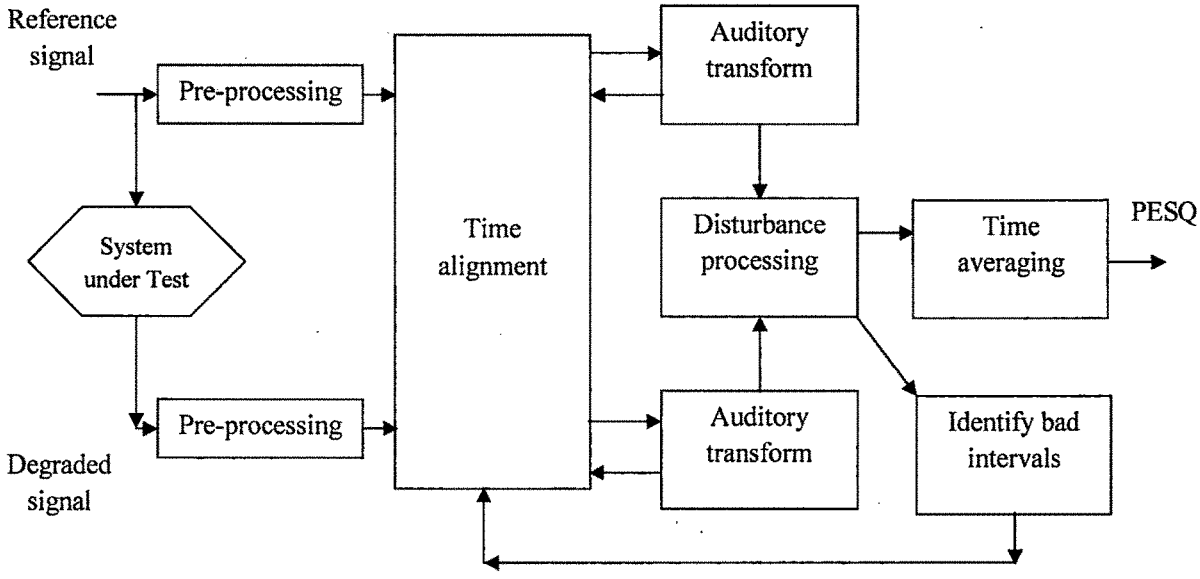


**Fig. 3.12 Block diagram of PESQ measure computation**

## 3.9 Summary

The transform domain techniques particularly STSA techniques are frequent in speech enhancement and they are discussed in detail. They are characterized by their gain function. The gain function requires computation of a *posteriori* and/or *a priori* SNR. The frame by frame processing using decision direct rule allows the computation of both SNRs. The gain function depicts the complexity of computation. The MMSE-STSA85 (LSA) method has complex gain function but provides good resistance against musical noise. The amount of speech distortion perceived is also reduced. So it is preferred in practical applications. The wavelet based transform domain techniques are also touched here. The de-noising is done by using thresholding of wavelet co-efficients. There is no optimized way for thresholding and hence they are still inferior in comparison to STSA techniques. The objective quality measures SSNR, LLR, WSS and PESQ are used to assess the effectiveness of speech enhancement algorithms. In next chapter the simulation and objective evaluation results of these techniques are presented.