MATERIALS AND METHODS

This chapter discusses the materials and methods that have been adopted in the study conducted. Broadly, it comprise of theoretical concepts, materials and methods, mathematically derived formulae that have been conceptualised in the area of noise reduction as well as in speech enhancement. Further, this chapter introduce the various tools used and experiments designed and performed in the study. Besides this, Data sampling, Collection of noise from noisy speech and Simulation of Data, Instrumentation and Analysis techniques that have been adopted for the study is also inferred in this chapter. Various filtering techniques used for characterizing the noise are also discussed in this chapter. To discuss about the noisy speech production, the idea of sound waves interaction is introduced as below.

When two or more sound waves are superimposed, they add to and subtract from each other depends on phase differences between the two waves. If their peaks and troughs are perfectly in phase, they reinforce each other, resulting in a waveform that has higher amplitude than either individual waveform. This is illustrated in figure 3.1.

If the peaks and troughs of two waveforms are perfectly out of phase, they cancel each other out, resulting cancellation of intensity as shown in figure 3.2.

In most cases, waves are out of phase in varying amounts, resulting in a combined waveform that is more complex than individual waveforms as shown in figure 3.3. A complex waveform that represents music, voice, noise, and other sounds, combines the waveforms from each
sound together. Because of its unique physical structure, a single source can create extremely complex waves.

![Diagram](image1)

Fig 3.1. In-phase waves reinforce each other.

![Diagram](image2)

Fig 3.2. Out-of-phase waves cancel each other.

![Diagram](image3)

Fig 3.3 Two simple waves combine to create a complex wave.

Here, superimposition of two signals takes place to produce a complex wave. When a signal which is undesirable or unwanted, interacts with the signal of our interest, then it can impart undesirable effects in the resultant complex wave. This in turn degrades the quality of the signal producing a noisy signal. The mathematical background of superimposition of signals is explained with the help of Linear Time-Invariant theory (LTI).
3.1 Linear Time-Invariant theory (LTI)

The fundamental result in LTI theory is that any LTI system can be characterized by a single function called the system's impulse response. The output of the system is the convolution of the input to the system with the system's impulse response. This method of analysis is called the time domain point-of-view. The same result is true for discrete-time linear shift-invariant systems, in which signals are discrete-time samples, and convolution is defined on sequences.

For a linear and time invariant system; Linearity means that the relationship between the input and the output of the system satisfies the superimposition property. If the input to the system is the sum of two component signals:

\[ x(t) = (a) x_1(t) + (b) x_2(t) \ldots \ldots 3.1 \]

Then the output of the system will be

\[ y(t) = (a) y_1(t) + (b) y_2(t) \ldots \ldots 3.2 \]

where, \( y_n(t) \) is the output resulting from the input \( x_n(t) \) and \( a \ & b \) are any constants.

It can be shown that, given this superimposition property, the scaling property follows for any rational scalar quantity. If the output due to input \( x(t) \) is \( y(t) \), then the output due to input \( cx(t) \) is \( cy(t) \), where \( c \) is a constant.

Then, formally, a linear system is a system which exhibits the following property: if the input of the system is
\[ x(t) = \sum_{n} c_n x_n(t) \] \[ y(t) = \sum_{n} c_n y_n(t) \]

Then the output of the system will be

\[ y(t) = \sum_{n} c_n y_n(t) \]

For any constants \( C_n \) and where each \( y_n(t) \) is the output resulting from the sole input \( x_n(t) \).

Time invariance means that whether we apply an input to the system now or \( T \) seconds from now, the output will be identical, except for a time delay of the \( T \) seconds. If the output due to input \( x(t) \) is \( y(t) \), then the output due to input \( x(t - T) \) is \( y(t - T) \). More specifically, an input affected by a time delay should affect a corresponding time delay in the output; hence it is time-invariant.

Equivalently, any LTI system can be characterized in the frequency domain by the system's transfer function, which is the Laplace transform of the system's impulse response (or Z transform in the case of discrete-time systems). As a result of the properties of these transforms, the output of the system in the frequency domain is the product of the transfer function and the transform of the input. In other words, convolution in the time domain is equivalent to multiplication in the frequency domain. Figure 3.4 shows the relationship between time domain and frequency domain representations.

![Fig-3.4. Relationship between the time domain and the frequency domain](image-url)

\[ x(t) \quad h(t) \quad y(t) = h(t) * x(t) \]

\[ X(s) \quad H(s) \quad Y(s) = H(s) \cdot X(s) \]
3.2 Impulse response

Impulse response of a system is the output of the system when no input signal is fed into the system. In signal processing, the approaches are analogue and digital; accordingly which the impulse response of a system is explained herewith.

3.2.1 Analogue or Continuous-time systems

Consider a time-varying system whose impulse response is a two-dimensional function and the condition of time invariance helps to reduce it to one dimension. Suppose the input signal is \( x(t) \) where its index set is the real line, i.e., \( t \in \mathbb{R} \). The linear operator \( \mathcal{H} \) represents the system operating on the input signal. The appropriate operator for this index set is a two-dimensional function \( h(t_1, t_2) \) where \( t_1, t_2 \in \mathbb{R} \).

Since \( \mathcal{H} \) is a linear operator, action of the system on the input signal \( x(t) \) is a linear transformation represented by the following superimposition integral

\[
y(t_1) = \int_{-\infty}^{\infty} h(t_1, t_2) x(t_2) \, dt_2. \quad \text{...............3.5}
\]

If the linear operator \( \mathcal{H} \) is also time-invariant, then

\[
h(t_1, t_2) = h(t_1 + \tau, t_2 + \tau) \quad \forall \, \tau \in \mathbb{R}. \quad \text{...............3.6}
\]

Let then \( \tau = -t_2 \), it follows that

\[
h(t_1, t_2) = h(t_1 - t_2, 0) \quad \text{...............3.7}
\]
Neglecting the zero second argument to \( h(t_1, t_2) \) for brevity of notation so that the superimposition integral now becomes the familiar convolution integral used in filtering

\[
y(t_1) = \int_{-\infty}^{\infty} h(t_1 - t_2) x(t_2) \, dt_2 = (h \ast x)(t_1). \quad \text{………3.8}
\]

Thus, the convolution integral represents the effect of a linear, time-invariant system on any input function. If we input a Dirac delta function to this system, the result of the LTI transformation is known as the impulse response because the delta function is an ideal impulse.

\[
(h \ast \delta)(t) = \int_{-\infty}^{\infty} h(t - \tau) \delta(\tau) \, d\tau = h(t), \quad \text{………3.9}
\]

(by the shifting property of the delta function).

Note that \( h(t) = h(t,0) \) (with \( t = t_1 - t_2 \)) so that \( h(t) \) is the impulse response of the system.

### 3.2.2 Discrete-time systems

Consider the impulse response of time varying system is a two dimensional function and the condition of time-invariance is used to reduce it to one dimension. Suppose the input signal is \( x[n] \) where its index set is the integers, i.e., \( n \in Z \). The linear operator \( \mathcal{H} \) represents the system operating on the input signal. The appropriate operator for this \( h[n_1,n_2] \) where \( n_1,n_2 \in Z \) index set is a two-dimensional function

Since \( \mathcal{H} \) is a linear operator, the action of the system on the input signal \( x[n] \) is a linear transformation represented by the following superimposition sum
\[ y[n_1] = \sum_{n_2=-\infty}^{\infty} h[n_1, n_2] x[n_2], \quad \cdots \cdots \cdots 3.10 \]

If the linear operator \( \mathcal{H} \) is also time-invariant, then

\[ h[n_1, n_2] = h[n_1 + m, n_2 + m] \quad \forall m \in \mathbb{Z}. \quad \cdots \cdots \cdots 3.11 \]

Let \( m = -n_2 \) then it follows that

\[ h[n_1, n_2] = h[n_1 - n_2, 0] \quad \cdots \cdots \cdots 3.12 \]

Similarly, neglecting the zero second argument to \( h[n_1, n_2] \) for brevity of notation so that the superimposition integral now becomes the familiar convolution sum used in filtering

\[ y[n_1] = \sum_{n_2=-\infty}^{\infty} h[n_1 - n_2] x[n_2] = (h \ast x)[n_1]. \quad \cdots \cdots \cdots 3.13 \]

Thus, the convolution sum represents the effect of a linear, time-invariant system on any input function. If we input a discrete delta function to this system, the result of the LTI transformation is known as the impulse response because the delta function is an ideal impulse.

\[ (h \ast \delta)[n] = \sum_{m=-\infty}^{\infty} h[n - m] \delta[m] = h[n], \quad \cdots \cdots \cdots 3.14 \]
(by the sifting property of the delta function). Note that $h[n] = h[n_1-n_2,0]$

where, $n = n_1-n_2$, so that $h[n]$ is the impulse response of the system.

### 3.3 Noise Reduction

The goal of the noise reduction is to reduce the noise level without distorting the speech, thus reduce the stress on the listener and - ideally - increase intelligibility. There are many different ways to perform the noise reduction. Basically, the solutions can be split in two classes: mono and multi sensor (microphone) systems, where multi sensor systems exploit the spatial properties of speech and noise and a mono sensor system usually rely on the temporal characteristics. A fundamental requirement for most mono sensor systems is that speech and noise are additive and results of uncorrelated statistical processes, and that the spectral characteristics of the noise changes markedly slower than those of the speech.

#### 3.3.1 Noise Reduction by Spectral Weighting

In spectral weighting different spectral regions of the mixed signal of speech and noise are attenuated with different factors to attain a signal which contains less noise than the original one. Besides requiring a minimal distortion of the original speech, it is also important that the residual noise, i.e. the noise remaining in the processed signal, does not sound unnatural.

Spectral weighting is usually performed in a transformed domain, e.g. the frequency domain. A common transform is the Fourier transform which provides an equidistant frequency solution. If we denote the speech with $s(k)$, the noise with $n(k)$, and the microphone signal $x(k)$, we have

$$x(k) = s(k) + n(k)$$

Taking the Fourier transform of this equation leads to
The actual spectral weighting is now performed by multiplying the spectrum \( X(f) \) with a real weighting function \( H(f) \geq 0 \). Here, \( H(f) \) is a weighting function or weighting rule. The result \( Y(f) \) is then,

\[
Y(f) = H(f) X(f) \quad \text{3.17}
\]

and the output signal \( y(k) \) of the system is obtained by transforming \( Y(f) \) back into the time domain.

Due to the short-time stationarity of speech, the processing is done on a frame-by-frame basis. A basic system for this is shown in figure 3.5 below. In addition to the functions shown, the other necessary functions are framing, windowing, zero padding, overlap-and-add. Because \( H(f) \) is a real function, only the magnitude of \( X(f) \) is changed and phase is retained for the reconstruction.

![Fig-3.5. The principle of spectral weighting.](image)

The weighting function \( H(f) \) is usually a function of the magnitude spectra \(|Y(f)|\) and \(|N(f)|\), or of the power spectral densities \( R_y(f) \) and \( R_n(f) \). Thus, to calculate \( H(f) \) some estimate of the noise which should be reduced is necessary.
3.3.1.1 Spectral Weighting Rules

In this section a few weighting rules for noise reduction are discussed.

a) Spectral Subtraction

One of the first weighting rules proposed for noise reduction was the spectral subtraction [6]. One version of it is the magnitude spectral subtraction. It means that an estimate $|N^*(f)|$ of the noise magnitude spectrum is subtracted from the instantaneous input magnitude spectrum $|X(f)|$ such that

$$|Y(f)| = |X(f)| - |N^*(f)| \ldots \ldots \ldots \ldots \ldots \ldots 3.18$$

Written as a weighting rule we have

$$H(f) = 1 - (|N^*(f)/|X(f)|) \ldots \ldots \ldots \ldots \ldots \ldots 3.19$$

To prevent $Y(f)$ from being negative, $|N^*(f)|$ must not be greater than $|X(f)|$.

Although the noise level is reduced by the spectral subtraction, the disadvantage is that an unnatural sounding residual noise still remains there. It can be explained by the statistical nature of the noise [6]. For example, consider some frequency of the instantaneous spectrum which does not contain any speech, $X(f) = N(f)$. The effect of too small a noise estimate, $|N^*(f)| < |N(f)|$, is a remaining excitation at this frequency. On the other hand, if the noise is estimated to be higher than it actually is, the result will be zero due to the necessary bounding, $Y(f) = 0$. The result is short sinusoids randomly distributed over time and frequency, which remain in the processed signal. This kind of noise is called musical noise.
b) The Wiener Filter

The Wiener filter rule is derived from the optimal filter theory [20]. It is based on minimizing the mean squared error between the speech $S(f)$ and the estimate $Y(f)$,

$$
E\{ |S(f) - Y(f)|^2 \} = \min, \text{...............3.20}
$$

and leads to the solution

$$
H(f) = \frac{Rs(f)}{Rx(f)} = \frac{Rs(f)}{(Rs(f) + Rn(f))}, \text{........3.21}
$$

where we have assumed that speech and noise are statistically uncorrelated, Rs(f) is the PSD of the speech, Rx(f) the PSD of the input signal and Rn(f) the PSD of the noise. The PSD of the speech can be estimated from $Rs(f) = Rx(f) - Rn(f)$. The result when using the Wiener rule above also suffers from musical noise. However, the Wiener rule can be implemented in other ways which help to reduce the amount of musical noise, [46].

c) The Weighting Rules of Ephraim & Malah

Ephraim & Malah [47] & [42] proposed two weighting rules which were derived by taking the distribution of the magnitude spectrum of the speech into account. In comparison to the spectral subtraction and the standard Wiener rule, the weighting rules proposed by Ephraim and Malah drastically reduce the musical noise in the processed signal. However, depending on how strong the noise is reduced, the residual noise still sounds unnatural.
d) **Psychoacoustically Motivated Weighting Rules**

When dealing with signals intended for a human listener, the properties of the human auditory system should be taken into account. In auditory perception the effect called auditory masking occurs when a weak signal is inaudible if a stronger signal is present at the same or a nearby frequency. The weak signal is masked by the stronger one [48].

When the speech $s(k)$ masks the additive noise $n(k)$, the noise in such situations are not audible, although it is present in the total signal $x(k) = s(k) + n(k)$. This effect is applied in audio codecs. The masked threshold defines the input level of the noise which is not audible. In the following it is denoted as $R_t(f)$.

There are several possible approaches that uses the advantage of the effect of auditory masking for noise reduction as well. One of the most apparent one [49] is to estimate the speech using weighting rules $H(f)$ (discussed above), then estimate the masked threshold from this result. At frequencies where the estimated noise lies below the masked threshold, a new weighing rule $H'(f)$ is set to 1, otherwise $H'(f)$ takes the value of $H(f)$: 1, if $R_n(f)<R_t(f)$

\[
H'(f) = \begin{cases} 
H(f), & \text{otherwise} 
\end{cases}
\]

Then $Y(f) = H'(f)X(f)$ is used to obtain the final result.

Although this usage of the masked threshold may lead to a lower distortion of the speech (the speech remains unchanged at frequencies where the noise is masked), still there remain an unnaturally sounding residual noise in parts of the spectrum where only a weak or no speech is present.
e) **A Novel Psychoacoustically Motivated Weighting Rule**

A new weighting rule has been developed with the aim of preserving the characteristics of the background noise while minimizing the speech distortion [5050] & [51]. Important is that a complete removal of noise is obtained; instead a low level of natural sounding background noise remain in the signal. The desired noise attenuation is defined by a factor $Z$.

Basically, the weighting rule is defined by choosing the weighting factor at a certain frequency such that the power of the difference between the desired residual noise level, defined by $Z N(f)$, and the actual noise level after processing, $H(f) N(f)$, is masked by the speech. The effect is that the “distortion” of the desired residual noise is masked by the speech. The resulting weighting rule is called HIND from Inaudible Noise Distortion. With $R_t(f)$ as the masked threshold, $R_n(f)$ as the noise PSD, it can be written as

$$HIND(f) = \sqrt{R_t(f) / R_n(f) + Z} \ldots \ldots \ldots 3.23$$

From the above equation, if no speech is present, then the masked threshold is zero, and $HIND(f) = Z$. The output signal is identical to the input signal multiplied by a scalar factor. In other words, the background noise characteristics are preserved during the process.

### 3.3.2 Potential drawbacks of the noise reduction system

Audible distortions may be still present in the enhanced speech signal, mainly due to the fact that some components exhibit strong fluctuations in noisy areas of the short-term input spectrum when the local SNR comes low. As a result of the Short-Term Suppression factor principle, the residual noise will be composed of sinusoidal components with random
frequencies in each short-term frame. This phenomenon is at the origin [52] of the “musical noise”. Furthermore, the Short-Term Suppression rule may also introduce distortions on some components of the speech signal. It has been observed (Scalart and Benamar, 1995) that these spectral imperfections may have a strong impact on the coder performance and result in audible distortions in the reconstructed speech signals after coding/decoding operations, even if the subjective quality of the enhanced speech signal seems quite good. These distortions can be interpreted as spectral mismatches between the enhanced and the clean speech signals which may disturb the analysis operations (Short-Term and Long-Term Predictors) in the full rate Global System of Mobile (GSM) coder/decoder.

### 3.4 Speech Enhancement

Speech Enhancement includes, suppression/reduction of noise contained in the recording and removing or reducing the other unwanted effects in the signal, thus to improve the quality of the speech signal. Speech Enhancement depends on good signal processing technique, human perceptual factor. Speech quality and intelligibility are dependent on short term spectral amplitude and insensitive to spectral phase. Fig-3.6 and Fig-3.7. shows noisy speech production and its enhancement.

The first category of speech enhancement algorithms is based on the short-time spectral estimation such as the spectral subtraction and Wiener filtering techniques. The algorithms in the second category are comb filtering and adaptive noise cancellation techniques which use the quasi-periodic nature of the speech signal. The third category contains algorithms that are based on the statistical modelling of the speech signal and use Hidden Markov Model (HMM) or expectation and maximization (EM) for speech enhancement. Another speech enhancement approach is the signal subspace (SS) method introduced by Dendrinos et al. in 1991 [53] and further improvement in this approach was made by Ephraim & Van Trees in 1995 [54]. There have been many such approaches
that improved the quality of the speech in noisy recordings. They include well known algorithms like soft-decision estimation based on Maximum Likelihood estimation by MC Aulay and Malpass in 1980, followed by Yang in 1993 and Brancaccio and Pelaez in 1993; or Minimum Mean-Square Error (MMSE) estimation [47] of the amplitude of enhanced speech signal.

![Diagram of a typical speech enhancement system](image1)

recover $s(n)$ from $y(n) = s(n) + d(n)$

Fig-3.6. Diagram of a typical speech enhancement system

![Diagrammatic representation of the short-time spectral magnitude enhancement](image2)

Fig-3.7. Diagrammatic representation of the short-time spectral magnitude enhancement

Some of the important methods developed for speech enhancement namely, Spectral Subtraction technique, Wiener filtering, Comb filtering, Cepstrum Mean Normalization (CMN), Relative Spectral processing of speech (RASTA), Short-Time Spectral Attenuation (STSA),
Deconvolution method, Normalized Least Mean Square (NLMS) and Artificial Neural Networks (ANN) are discussed herewith.

3.4.1 Spectral Subtraction technique

A classical noise reduction approach for speech enhancement and robust recognition is the spectral subtraction method that was first proposed by Boll [6]. The basic idea is to restore the magnitude spectrum or power spectrum of a signal observed in additive noise through subtraction of an estimate of the average noise spectrum from the noisy signal spectrum. Assuming that the noise is a stationary or a slowly varying process, the noise spectrum is estimated or updated during the periods when the speech signal is absent. The advantage of the spectral subtraction method is its simplicity and effectiveness.

The spectral subtraction method assumes that the noise and speech are uncorrelated and additive in the time domain. In that case, the power spectrum of the noisy signal is the sum of the noise and the speech spectra. The method also assumes that the noise characteristics change slowly relative to those of speech signals, so that the noise spectrum estimated during a non-speech period can be used for suppressing the noise contaminating the speech.

3.4.1.1 Principle of the basic spectral subtraction method

If we assume that \( y(n) \), the discrete noisy input signal, is composed of the clean speech signal \( s(n) \) and the uncorrelated additive noise signal \( d(n) \), then we can represent it as:

\[
y(n) = s(n) + d(n) \quad \text{.................3.24}
\]

Processing is done on a frame-by-frame basis. Analysis of overlapping frames of the noisy signal is implemented by using the Discrete Fourier Transform (DFT) preceded by a Hamming window. The power spectrum of the noisy signal can be written as:
\[ |Y(k)|^2 \approx |S(k)|^2 + |D(k)|^2 \quad \ldots \ldots \ldots \ldots \ldots 3.25 \]

where, the DFT of \( Y(k) \) is given by:

\[
Y(k) = \sum_{n=0}^{N-1} y(n) e^{-j \frac{2\pi kn}{N}} = |Y(k)| e^{j \varphi(k)} \quad \ldots \ldots \ldots \ldots \ldots 3.26
\]

where \( \varphi(k) \) is the phase of the noise-corrupted signal, i.e. the phase of \( Y(k) \).

Since the noise spectrum \( D(k) \) cannot be directly obtained, a time-average of the power spectrum \( \hat{D}(k) \) is calculated during a period of silence. Assuming that noise is uncorrelated with the speech signal, an estimate of the modified speech spectrum can be given as:

\[
|\hat{S}(k)|^2 = |Y(k)|^2 - |\hat{D}(k)|^2 \quad \ldots \ldots \ldots \ldots \ldots 3.27
\]

From equation 3.27 it can be seen that the subtraction process involves the subtraction of an averaged estimate of the noise from the instantaneous speech spectrum. Due to the error in computing the noise spectrum, we may have some negative values in the modified spectrum. These values are set to zero. This process is called half-wave rectification. With half-wave rectification the modified spectrum can be written as:

\[
|\hat{S}(k)|^2 = \begin{cases} 
|\hat{S}(k)|^2 & \text{if } |\hat{S}(k)|^2 > 0 \\
0 & \text{else}
\end{cases} \quad \ldots \ldots \ldots \ldots \ldots 3.28
\]
The modified spectrum of equation 3.28 is combined with the phase information from the noise corrupted signal to reconstruct the time signal by using the Inverse Discrete Fourier Transform (IDFT) in conjunction with the OLA method.

\[
\hat{s}(n) = \text{IDFT} \left( \hat{S}(k) | e^{j\phi(k)} \right) \quad \text{………………3.29}
\]

The noise suppression can also be implemented as a time-varying filtering process by rewriting the spectral subtraction method as:

\[
S'(k) = H(k) \cdot Y(k) \quad \text{………………3.30}
\]

where \( H(k) \) is a gain function represented by:

\[
H(k) = \sqrt{1 - \frac{|\hat{D}(k)|^2}{|Y(k)|^2}} \quad \text{………………3.31}
\]

\[
H(k) = \sqrt{\frac{|Y(k)|^2 - |\hat{D}(k)|^2}{|Y(k)|^2}} \quad \text{………………3.32}
\]

In this case, the modified spectrum is obtained by applying a time varying weight \( H(k) \) to each frequency component. From equation 3.31, it can be deduced that the frequency dependent gain is a function of the noisy signal-to-noise ratio (NSNR) of each of the frequency components. The enhanced time signal is synthesized as given in equation
3.29, using the original noisy phase portion. The enhanced signal has largely reduced noise levels compared to the original noise corrupted signal resulting in a better SNR and improved speech quality.

The major calculation in the spectral subtraction method is the discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT) which can be efficiently implemented using the Fast Fourier transform (FFT) algorithms. It is also proved by various researchers that the spectral subtraction method can improve the signal-to-noise ratio (SNR) and word recognition accuracy (WRA) under different SNR conditions namely, Nolazco and Young in 1994 [55] & Le Bouquin 1996.

The method is simple and efficient for stationary or slowly varying wide-band additive noise. The performance of a noise subtraction system also relies on noise/speech classification decision. A misclassification may result in misestimation of noise model and thus a degradation of speech estimate. The decision can be based on modeling log-energy histograms as a bimodal distribution. The subtraction may result in negative power spectrum values which are set to non-negative thresholds. This non-linear operation produces residual noise commonly called musical noise. This artifact is due to the residual noise in the enhanced speech. Subtraction techniques cannot be performed in the logarithmic spectrum domain where noise, even uncorrelated with the signal in the time domain, becomes signal-dependent. The main problem in spectral subtraction is the processing distortions caused by the random variations of the noise spectrum.

3.4.2 Wiener filtering

The Wiener filter, a filter named after Norbert Wiener who introduced the optimal estimation theory, separates signals based on their frequency spectra. Wiener filter is an optimum filter in the mean-square error sense. In Wiener filtering, an assumption is made for known a
signal and noise spectrum which gives better performance in enhancement. The gain of the filter at each frequency is determined by the relative amount of signal and noise at that frequency. Also imposing constraints from speech production model and speech characteristics produce better signal spectrum estimation and hence improve performance.

Theoretically, Wiener filtering minimizes the variance of the difference between noisy speech and clean speech, assuming additive signal and noise. In frequency domain, the filter is derived as the ratio of power spectrum of clean speech over the sum of the power spectrum of speech and that of noise.

The Wiener filter obtains a linear estimate of the original speech signal while minimising the mean-squared error between the original and enhanced signal (= MMSE estimate). In Wiener filtering, both the speech and the noise signal are considered as realisations of stationary stochastic processes. It is assumed that the second order statistics of the speech and noise signals are a priori known. The filter characteristic is then given by

\[ W(f) = \frac{\Phi_s(f)}{\Phi_x(f)} \]  \hspace{1cm} 3.32

with \( \Phi_s(f) \) and \( \Phi_x(f) \) the power spectra obtained via the Fourier transform of the correlation functions of the clean and noisy signal respectively. Since, in practice, the speech and noise signals are non-stationary and their statistics are not exactly known, the filter should be derived by estimating the short-time power spectrum of the noisy speech \( P_x(f) \), and the long-term average of the noise power spectrum \( P_n(f) \). In this case, the Wiener filter is given by

\[ W(f) = \frac{P_s(f) - P_n(f)}{P_s(f)} \]  \hspace{1cm} 3.33
Apart from some minor differences, Wiener filtering can be seen as a time-domain implementation of PSS. The similarity between both becomes apparent by rewriting the spectral subtraction formula, except for the non-linear noise flooring and overestimation operations, as a filter operation. However, it is noted that the Wiener filter yields Minimum Mean Square Error (MMSE) estimate of the short-time Fourier transform whereas PSS obtains an MMSE estimate of the short-time spectral amplitude without changing the phase. In practice, both algorithms behave very similar [47]. The Wiener filter obtains a least \( \ell^2 \) squares estimate of under stationarity assumptions of speech and noise. The construction of the Wiener filter requires an estimate of the power spectrum of the clean speech and the noise:

\[
W(f) = \frac{\hat{s}_m(f)}{\hat{s}_m(f) + \hat{\Phi}_m(f)} \quad \text{.................3.34}
\]

Most of the Wiener filtering-based algorithms are iterative since an estimate of clean speech power spectrum is required in the formulation. The frequency response, represented by \( H[f] \), is determined by the frequency spectra of the noise, \( N[f] \), and the signal, \( S[f] \). Only the magnitudes are important; all of the phases are zero.

\[
H[f] = \frac{S[f]^2}{S[f]^2 + N[f]^2} \quad \text{.................3.35}
\]

A transfer function of the Wiener filter, \( \text{wiener} H \), is expressed in terms of the power spectrum of clean speech and the power spectrum of noise as:

\[
H_{\text{wiener}}(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_n(\omega)} \quad \text{.................3.36}
\]
But, the noncasual Wiener filter in Equation 3.36 cannot be applied directly to estimate the clean speech since speech cannot be assumed to be stationary and the power spectrum $P_s(\omega)$ cannot be assumed known. Therefore, an adaptive Wiener filter implementation can be used to approximate the above filter as:

$$H_{\text{wiener}}(\omega) = \frac{E[|S(\omega)|^2]}{E[|S(\omega)|^2 + E[|D(\omega)|^2]} \quad \text{------------3.37}$$

Since the power spectrum of clean speech is also not known, the power spectrum of the noisy speech signal is used instead, and the above Equation 3.33 can be written using Equation 3.34 as follows:

$$H_{\text{wiener}}(\omega) = \frac{E[|Y(\omega)|^2] - E[|D(\omega)|^2]}{E[|Y(\omega)|^2]} \quad \text{------------3.38}$$

Direct application of Wiener filtering to noise cancellation for noisy speech is limited by the non-stationarity of speech signals. HMMs automatically divide speech into quasi-stationary segments corresponding to the model states. This property of HMMs is exploited [56] to implement noise canceling filters within a speech recognition process. Under this scheme, in each state for each filter channel, an optimum FIR Wiener filter is attached as an additional parameter of the model. During recognition, a filtered estimate of clean speech is calculated on the basis of a sequence of noisy input vectors. The estimate is used to compute the output probability of the state. In the frequency domain, Wiener filtering corresponds to a multiplication of the signal spectrum and the filter frequency response. In the cepstral domain, this is equivalent to an additive operation due to the logarithmic scaling involved. An HMM state-dependent Wiener filter was implemented to additively correct cepstral observation vectors prior to the
calculation of the output probability for that state developed by V.L. Beattie and S.J. Young [56].

Cepstral correction by Wiener filtering was also used in a template-based dynamic programming recognition system, where the filters were frame-dependent [57]. For cepstral-time features modeled by a mixture Wiener filters produce an optimal signal-to-noise ratio and matched filters are optimal for target detection. In the Wiener filtering algorithm, the estimator is obtained by finding the optimal MMSE estimates of the complex Fourier transform coefficients.

3.4.3 Comb filtering

If the period of noisy voiced speech can be determined, then comb filters can be applied in the frequency domain to reduce the noise level [58]. Comb filtering multiplies in the frequency domain, the observation signal by a sequence of Dirac functions whose interval is the period of the speech signal. In the time domain, this operation is equivalent to averaging the signal waveform over several periods. A real speech is a quasi-periodic signal with frequency modulation in the period. The resulting effect is that, compared to lower harmonics, the spectral peaks of the higher harmonics are more broadened and lowered in amplitude. To correct the non-stationarity in the periodicity, comb filters can be applied to time-domain dewarped speech signal. The clean estimate of the speech can be obtained by re-warping the comb filtered speech to re-introduce the quasi-periodicity of the original speech [59] & [60]. The effect of comb filtering on the SNR of speech in different noise was evaluated [61] & [62]. While it improves the SNR ratio, over a wide range of SNR ratios the intelligibility is in fact reduced by the comb filtering. A speech signal can be represented as a sum of the harmonics of its fundamental frequency component. Under the assumption that the contaminating noise is Gaussian with known variance and that the amplitudes of the harmonics are independent, it has been shown [63] that the optimum estimation of speech may be considered to be a combination of comb filtering and Wiener filtering. Comb filtering makes the assumption that noise is additive and short-time stationary. Also, an accurate determination of the period of noisy speech is critical for
the success of comb filtering. The filter is not applicable to speech segments with fast transitions, voiced fricatives, as well as unvoiced speech.

### 3.4.4 Cepstrum Mean Normalization (CMN)

Atal in 1974 has developed an effective technique to filter the slowly varying background noise by using Cepstrum Mean Normalization (CMN) technique [17]. It is the simplest and efficient way to remove slow variations in the background noise which removes the mean from cepstral vectors [17], [64] & [65]. This method compensates for the difference of bias on cepstrum coefficients (CC) between training data and test data by subtracting the mean value of the cepstrum coefficients CC calculated from a certain amount of given speech data. Such adaptation data are not phonetically balanced, which makes it difficult to get an accurate mean value for the cepstrum coefficients CC. Cepstrum mean normalization is an effective method for recognizing distorted telephone speech. It is observed [66] that CMN performs similarly as Relative Spectral processing (RASTA), does not degrade baseline system and improves the recognition accuracy when a new microphone is used.

### 3.4.5 Relative Spectral processing of speech (RASTA)

RASTA - Relative Spectral processing of speech is based on the technique involving suppression of constant additive offsets in every log spectral component of the short-term spectrum. The technique involves filtering trajectories of short term power spectrum of 8 KHz sampled speech. The short term power spectrum is estimated using a 256 point FFT with Hamming window and a frame step of 64 samples. For 256 point FFT we must process 129 unique spectral trajectories. The cubic-root compressed spectral trajectories are given as inputs to a filter bank. The filters are designed on training data to optimally map the noisy spectral trajectories to clean spectral trajectories.
Both linear filters and neural networks have been used to perform the mapping. The linear filtering utilizes 21 tap noncausal optimal FIR filter. The neural network architecture consists of a feed forward structure with time delay inputs to cover a 180ms time window (total of 21x129 inputs). Sparse connectivity is used such that each spectral channel feeds to 6 neurons with a 4 adjacent channel overlap.

RASTA operates in the logarithmic domain and noise which is additive in the logarithmic domain such as slow-changing communication channel characteristics can be efficiently reduced. But noises that are additive in the time domain and therefore signal-dependent in the log domain cannot be removed. Estimates of clean magnitude spectrum are combined with original noisy phase in an overlap-add-synthesis approach to reconstruct time-domain speech signal.

3.4.6 Signal subspace method (SS)

The idea to perform subspace-based signal estimation was originally proposed by Tufts et.al in 1982 [67]. In their work, the signal estimation is actually based on a modified Singular Value Decomposition (SVD) of data matrices. Later on, [68] presented a general framework for recovering signals from noisy observations. The idea behind subspace methods is to project the noisy signal into two subspaces: the signal-plus-noise subspace, or simply signal subspace (since the signal dominates this subspace), and the noise subspace. The noise subspace contains signals from the noise process only; hence an estimate of the clean signal can be made by removing or nullifying the components of the signal in the noise subspace and retaining only the components of the signal in the signal subspace. Or in other words, the enhancement is performed by removing the noise subspace and estimating the clean speech from the remaining signal-plus-noise subspace.

The decomposition of the space into two subspaces can be done using either the SVD [69] & [70] or the Eigen Value Decomposition (EVD). In this approach, a nonparametric linear
estimate of the unknown clean-speech signal is obtained based on a decomposition of the observed noisy signal into mutually orthogonal signal and noise subspaces. This decomposition is possible under the assumption of a low-rank linear model for speech and uncorrelated additive (white) noise interference. Under these conditions, the energy of less correlated noise spreads over the whole observation space while the energy of the correlated speech components is concentrated in a subspace thereof. Also, the signal subspace can be recovered consistently from the noisy data. The basic idea is to model the colored noise by an Autoregressive (AR) process. A modified covariance method is used to estimate the Autoregressive AR parameters of the colored noise process and a preprocessing filter is based on the estimated AR parameters. The noisy speech signal filtered by the preprocessing filter is then enhanced by the Energy-constrained Signal Subspace (ECSS) method and an inverse filter is used in the post-processing stage in order to remove the distortion to the speech signal caused by the preprocessing filter.

Noise reduction is obtained by nullifying the noise subspace and by removing the noise contribution in the signal subspace. It is assumed that the original signal exhibits some well-defined properties or obeys a certain model. Signal enhancement is then obtained by mapping the observed signal onto the space of signals that possess the same structure as the clean signal. This theory forms the basis for all subspace-based noise reduction algorithms. A first and indispensable step towards noise reduction is obtained by nullifying the noise subspace (least squares (LS) estimator) [71]. However, for improved noise reduction also the noise contribution in the (signal + noise) subspace should be suppressed or controlled, which is achieved by all other estimators. Of particular interest is the Minimum Variance (MV) estimation, which gives the best linear estimate of the clean data, given the rank p of the clean signal and the variance of the white noise [72] & [73]. Later on, a subspace-based speech enhancement with noise shaping was proposed in [74]. Based on the observation that signal distortion and residual noise cannot
be minimised simultaneously, two new linear estimators are designed—Time Domain Constrained (TDC) and Spectral Domain Constrained (SDC)—that keep the level of the residual noise below a chosen threshold while minimising signal distortion. Parameters of the algorithm, control the trade-off between residual noise and signal distortion. In subspace based speech enhancement with true perceptual noise shaping, the residual noise is shaped according to an estimate of the clean signal masking threshold, as discussed in some papers [74], [75] & [76]. Although basic subspace-based speech enhancement is developed for dealing with white noise distortions, it can easily be extended to remove general coloured noise provided that the noise covariance matrix is known (or can be estimated) [77] & [78]. A detailed theoretical analysis of the underlying principles of subspace filtering can be found in [72], [74] & [80]. The excellent noise reduction capabilities of subspace filtering techniques are confirmed by several studies, both with the basic LS estimate [71] and with the more advanced optimisation criteria [74], [78] & [79]. Especially for the MV and SDC estimators, a speech quality improvement that outperforms the spectral subtraction approach is revealed by listening tests.

It was found that the signal-subspace method helped to improve the Word Recognition Accuracy (WRA) by a certain amount under different SNR levels. However, the SS method did not work well for the low-energy speech segments in continuous speech such as consonants. It was found that the Energy-constrained Signal Subspace (ECSS) method is very effective for recovering the low-energy segments in continuous speech which in turn results significant improvement in recognition accuracy on the enhanced speech.

### 3.4.7 Short-Time Spectral Attenuation (STSA)

The Short-Term Spectral Amplitude (STSA) of speech has been exploited successfully in the development of various speech enhancement algorithms. The basic idea is to use the STSA of the noisy speech input and recover an estimate of the clean STSA by removing the part contributed by the additive noise. A general representation of the technique is given in Figure
3.8. The input to the system is the noise-corrupted signal $y(n)$. While there are many methods for the analysis-synthesis processing, the Short-Term Fourier Transform (STFT) of the signal with Over Lap and Add (OLA) is the most commonly used method. The spectral amplitude $|Y(k)|$ of the noisy input signal $y(n)$ is modified using a correction factor. Usually this correction factor could be the spectral amplitude of the estimated noise signal $d(n)$, measured during periods of silence/non-activity in the speech signal or obtained from a reference channel (dual-channel method). The correction is obtained by subtracting the spectral amplitude of the noise signal from that of the noisy speech input. Hence, these methods are also referred to as subtractive-type algorithms. If the noise is assumed to be uncorrelated with the speech signal, then the corrected amplitude can be considered as an estimate $|S^{*}(k)|$ of the original clean speech signal $s(n)$. The unprocessed phase of the noisy input signal is used to synthesize the enhanced speech signal under the assumption that the human ear is not able to perceive the distortions in the phase of the speech signal.

Spectral subtraction is a well-known noise reduction method based on the STSA estimation technique. The basic power spectral subtraction technique, as proposed by Boll [6], is popular due to its simple underlying concept and its effectiveness in enhancing speech degraded by additive noise. The basic principle of the spectral subtraction method is to subtract the magnitude spectrum of noise from that of the noisy speech. The noise is assumed to be uncorrelated and additive to the speech signal. An estimate of the noise signal is measured during silence or non-speech activity in the signal. Since Boll [6] first proposed this method, several variations and enhancements have been made to the techniques to overcome some inherent drawbacks in the method. Short-Time Spectral Attenuation (STSA) speech enhancement algorithms are ineffective in the presence of highly non-stationary noise due to difficulties in the accurate estimation of the local noise spectrum.
3.4.8 Deconvolution method

In convolution, mixing of two signals occurs, but in deconvolution, instead of mixing two signals they are isolated. This is useful for analysing the characteristics of the input signal and the impulse response when only given the output of the system. Although the modulation envelope in deconvolution method has been claimed to be position independent, it reduces the intelligibility of the speech signal. Deconvolution is the process of filtering a signal to compensate for an undesired convolution. The goal of deconvolution is to recreate the signal as it existed before the convolution took place. This usually requires the characteristics of the convolution (i.e., the impulse or frequency response) to be known.

Deconvolution is difficult to handle in the time domain, but quite effective in the frequency domain. Each sinusoid that composes the original signal can be changed in amplitude and/or phase as it passes through the undesired convolution. To extract the original signal, the deconvolution filter must undo these amplitude and phase changes. For example, if the convolution changes a sinusoid's amplitude by 0.5 with a 30 degree phase shift, the deconvolution filter must amplify the sinusoid by 2.0 with a -30 degree phase change. Cole.D, Moody.M, and Sreedharan.S, in 1997 shown that the waveform deconvolution (inversion) approach is impractical in any situation where either source or receiver is mobile, due to variation of room impulse response with position [23].

3.4.9 Normalized Least Mean Square (NLMS)

Least mean squares (LMS) algorithms are used in adaptive filters to find the filter coefficients that relate to producing the least mean squares of the error signal (difference between the desired and the actual signal). It is a stochastic gradient descent method in that the filter is only adapted based on the error at the current time. It was invented in 1960 by Bernard Widrow and Ted Hoff. The idea behind LMS filters is to use the method of steepest descent to find a coefficient vector which minimizes a cost function. Normalized Least Mean Square
(NLMS) also gets it performance degraded significantly in the presence of high levels of background noise [24].

3.4.10 Artificial Neural Networks (ANN)

ANN- Multilayer Perceptrons (MLPs) using Back Propagation (BP) algorithm in ANN with time delay and Adaptive Networks has been used for Noise Reduction and echo cancellation. More generally, arbitrarily complex transformations can be achieved using an ANN such as the multi-layer feed-forward perceptrons. ANNs have been used for noise reduction [81] & [82] particularly for vector sequence mapping (mapping of a noisy magnitude spectrum to a clean magnitude spectrum) [83]. ANNs gives better results when compared to a linear transformation [83]. The noise reduction can be performed on time signal [84]. Some results show that while SNR was significantly improved, there was no significant improvement in recognition scores [85]. ANN with radial basis function (RBF) has also been used to map noisy observation to clean feature and results similar to that of a probabilistic mapping approach have been reported. Noisy to clean speech mapping by ANN has been used in combination with a linked predictive neural network (LPNN) speech recognizer [86]. This scheme is extended to an integration of ANN and LPNN (where parameters in the two components are initialized by separate trainings and then optimized through a joint training. Due to their universal function approximation capability, very simple ANNs can be used to obtain reasonable results. However, since no parametric models on noise or on speech are involved, the performance may be highly dependent on the noise level and type of the training data. The main disadvantage with this method is that the network requires noisy speech and the corresponding clear speech for it’s training which is not simultaneously available in many practical situations.
3.5 Filters

Filters have two uses: signal separation and signal restoration. Signal separation is needed when a signal has been contaminated with interference or noise. A filter might be used to separate the signals so that they can be individually analysed. Signal restoration is used when a signal has been distorted in some way. To reduce or eliminate various noise and distortion sounds from an audiotape, forensic audio specialists use variety of filters. If filtering of noise has not been carried out appropriately; corresponding features of noise may also be extracted together with the feature of the actual speech signal during the feature extraction process [16]. Thus, the desired parametric representation carries a high amount of error rate.

3.5.1 Analogue Filters

In the earlier decades, tape enhancement techniques traditionally relied upon exclusively on analogue instruments such as high-pass, low-pass, notch filters, graphic and parametric equalizers, comb filters and compressor-limiters. An analogue filter uses analogue electronic circuits made up from components such as resistors, capacitors and op-amps to produce the required filtering effect. The band pass filters, for example, were often effective in reducing specific tones or noises within the speech frequency range such as tape hiss, which could be eliminated with the appropriate high-pass filter.

3.5.2 Digital Filters

The introduction of the digital filter has significantly increased the expert’s ability at improving speech intelligibility while removing room reverberation effects. These filters are capable of accomplishing adaptive filtering, analysis-synthesis, spectral subtraction and deconvolution. These filters are effective in reducing pure tones, noise, and convolutional effects caused from room acoustics and reverberation. A digital filter uses a digital processor to perform numerical calculations on sampled values of the signal. The processor may be a general-purpose
computer such as a PC, or a specialised Digital Signal Processor (DSP) chip. In a digital filter, the signal is represented by a sequence of numbers, rather than a voltage or current.

3.5.2.1 Advantages of using digital filters

Some of the main advantages of digital filters over analogue filters are discussed herewith. A digital filter is programmable, i.e. its operation is determined by a programme stored in the processor's memory. This means that the digital filter can easily be changed without affecting the circuitry (hardware) whereas an analogue filter can only be changed by redesigning the filter circuit.

Digital filters are easily designed, tested and implemented on a general-purpose computer or workstation. The characteristics of analogue filter circuits (particularly those containing active components) are subject to drift and are dependent on temperature. Digital filters do not suffer from these problems. Due to this, they are extremely stable with respect to both time and temperature.

Unlike their analogue counterparts, digital filters can handle low frequency signals accurately. As the speed of DSP technology continues to increase, digital filters are being applied to high frequency signals in the radio frequency (RF) domain, which in the past was the exclusive preserve of analogue technology.

Digital filters are more versatile in their ability to process signals in a variety of ways; this includes the ability of some types of digital filter to adapt to changes in the characteristics of the signal. Fast DSP processors can handle complex combinations of filters in parallel or cascade (series), making the hardware requirements relatively simple and compact in comparison with the equivalent analogue circuitry.

Analogue filters are effective primarily when the noise sources are basically stationary, but they also reduce a portion of the desired voice signal and thus were inefficient in eliminating wide-band noise such as room reverberation effects.
Due to above mentioned factors, an Audio Forensic Expert rely on various digital filters that can help in enhancing the speech portion in a noisy speech recording.

3.5.2.2 Parameters of Digital Filters

Order of a digital filter is the number of previous inputs (stored in the processor's memory) used to calculate the current output. Following are few examples of digital filters.

- Zero order: \( y_n = a_0 x_n \)
- First order: \( y_n = a_0 x_n + a_1 x_{n-1} \)
- Second order: \( y_n = a_0 x_n + a_1 x_{n-1} + a_2 x_{n-2} \)

Similar expressions can be developed for filters of any order. The constants \( a_0, a_1, a_2, \ldots \) appearing in these expressions are called the filter coefficients. It is the values of these coefficients that determine the characteristics of a particular filter.

The important time domain parameters of Digital Filters are Risetime, Overshoot and Phase linearity. Risetime is the number of samples quote between the 10% and 90% amplitude levels. Signal jumps before settling down, this is termed as Overshoot. Phase linearity refers to the symmetry of the upper half of the step response with the lower half.

The important Frequency domain parameters of Digital Filters are Pass-band, Stop-band and Transition-band. Pass-band refers to those frequencies that are passed. Stop-band contains those frequencies that are blocked. Band which is in between Pass-band & stop-band is called Transition band.

3.5.2.3 Operation of digital filters

In this section, basic theory of the operation of digital filters is discussed. Suppose the “raw” signal which is to be digitally filtered is in the form of a voltage waveform described by the function
where \( t \) is time.

This signal is sampled at time intervals \( h \) (the sampling interval). The sampled value at time \( t = ih \) is \( x_i = x(ih) \).

Thus the digital values transferred from the ADC to the processor can be represented by the sequence \( x_0, x_1, x_2, x_3, \ldots, x_n \) corresponding to the values of the signal waveform at \( t = 0, h, 2h, 3h, \ldots \) and \( t = 0 \) is the instant at which sampling begins.

At time \( t = nh \) (where \( n \) is some positive integer), the values available to the processor, stored in memory, are \( x_0, x_1, x_2, x_3, \ldots, x_n \). The digital output from the processor to the DAC consists of the sequence of values \( y_0, y_1, y_2, y_3, \ldots, y_n \).

In general, the value of \( y_n \) is calculated from the values \( x_0, x_1, x_2, x_3, \ldots, x_n \). The way in which the \( y \)'s are calculated from the \( x \)'s determines the filtering action of the digital filter.

Every linear filter has an impulse response, a step response and a frequency response. Each of these responses contains complete information about the filter, but in a different form. Impulse response is given a special name as filter kernel.

### 3.5.2.4 Classification

Two types of Digital Filters are there, recursive or Infinite Impulse Response (IIR) filter & non-recursive or Finite Impulse Response (FIR) filter.

These terms refer to the differing “impulse responses” of the two types of filter. The impulse response of a digital filter is the output sequence from the filter when a unit impulse is applied at its input. (A unit impulse is a very simple input sequence consisting of a single value of 1 at time \( t = 0 \), followed by zeros at all subsequent sampling instants).
An FIR filter is one whose impulse response is of finite duration. An IIR filter is one whose impulse response theoretically continues forever because the recursive (previous output) terms feed back energy into the filter input and keeps it working. The term IIR is not accurate because the actual impulse responses of almost all IIR filters reduce virtually to zero in a finite time.

### 3.5.2.4.1 Finite Impulse Response (FIR) filter

If the current output \( y_n \) is calculated from the current and previous input values \( (x_n, x_{n-1}, x_{n-2},...) \). This type of filter is said to be non-recursive or Finite Impulse Response (FIR) filter. It uses a finite impulse response function. They are also called convolutive filters which can be constructed by convolving the input with filter function. As the design methods are linear, there is no feedback between the structures.

**Advantages of FIR filter**

FIR filters have exact linear phase. They are always stable. The design employed methods are generally linear. FIR filters can be realized efficiently in hardware.

### 3.5.2.4.2 Infinite Impulse Response (IIR) filter

An Infinite Impulse Response (IIR) filter is one which in addition to input values also uses previous output values. The previous input values, are stored in the processor's memory. They have an impulse response function which is non-zero over an infinite length of time.

They are also called recursive filter. The word recursive literally means “running back”. Recursive filters re-use one or more output(s) of the filter as inputs. This feedback results in an unending impulse response characterized by exponentially growing, decaying, or sinusoidal signal output components. The expression for a recursive filter, therefore, contains not only terms involving the input values \( (x_n, x_{n-1}, \)
xn-2,...) but also terms in yn-1, yn-2,... Since there are previous output terms in the filter expression as well as input terms, it is anticipated that recursive filters require more calculations to be performed. In fact, the reverse is usually the case: to achieve a given frequency response characteristic using a recursive filter generally requires a much lower order filter and therefore fewer terms to be evaluated by the processor than the equivalent nonrecursive filter.

**Advantages of IIR filter**

IIR filter requires few parameters than FIR filters. It uses less memory requirements and posses lower computational complexity.

### 3.5.3 Detection of signals in noise

#### 3.5.3.1 Detection theory

Detection theory, or signal detection theory, is a means to quantify the ability to discern between signal and noise. According to the theory, there are a number of psychological determiners for how to detect a signal and where to keep threshold levels. Experience, expectations, physiological state (e.g. fatigue) and other factors affect thresholds.

In the detection of signals in noise, the aim is to determine if the observation consists of noise alone or if it contains an information-bearing signal. The “Golden rule” of enhancement is that no audio signals are removed or attenuated that decrease the speech intelligibility, even slightly [87]. The noisy observation \( y(m) \) can be modelled as

\[
y(m) = b(m) * x(m) + n(m) \quad \cdots \cdots \cdots \cdots \cdots 3.40
\]

where, \( x(m) \) is the signal to be detected, \( n(m) \) is the noise and \( b(m) \) is a binary-valued state sequence, such that \( b(m) = 1 \) indicates the presence of the signal \( x(m) \) and \( b(m) = 0 \) indicates the presence of noise.
0 indicates that the signal is absent. If the signal \( x(m) \) has a known shape, then a correlator or a matched filter can be used to detect the signal.

The impulse response \( h(m) \) of the matched filter for detection of a signal \( x(m) \) is the time reversed \( x(m) \) given by

\[
h(m) = x(N-1-m) \quad 0 < m < N-1 \quad \cdots \quad 3.41
\]

where, \( N \) is the length of \( x(m) \).

The output of the matched filter is given by,

\[
Z(m) = \sum_{m=0}^{N-1} n(m-k) y(m) \quad \cdots \quad 3.42
\]

The matched filter output is compared to a threshold and a binary decision is made as,

\[
b(m) = \begin{cases} 
1 & \text{If } Z(m) \geq \text{Threshold} \\
0 & \text{Otherwise}
\end{cases} \quad \cdots \quad 3.43
\]

where \( b(m) \) is an estimate of \( b(m) \), and it may be erroneous in particular if the signal to noise ratio is low.

Figure-3.9 shows the configuration of a matched filter followed by a threshold comparator for detecting signals in noise.
3.6 Methods

3.6.1 Collection of Speech Samples and Sampling

In the study, noisy speech samples, clear speech samples and simulated noisy speech samples are collected from the exhibits received in case examination in the laboratory. Clear speech or speech signal under studio conditions of high quality has been recorded in the laboratory. Simulated samples have been prepared from the noise speech samples and recorded clear speech samples. All the samples are sampled for duration of 20 seconds as standard for this study.

3.6.1.1 Noisy speech samples

One hundred and fifty noisy speech samples received in actual crime cases at Central Forensic Science Laboratory, Chandigarh, were collected for the study. The samples were classified into direct, telephonic (land line) and mobile phone conversation based upon their mode of recording (hereafter referred as Noisy Speech-I, Noisy Speech-II & Noisy Speech-III respectively). Such categorization is being done for the purpose of identifying the sources of the noise as well. The samples collected are used for identifying the speaker by comparing with corresponding specimen samples as well as for recognizing the speech in the context.
3.6.1.2 Speech samples recorded in Ideal condition

Specimen voices recorded were also collected from the actual crime cases received in the laboratory. This has been carried out for comparison purpose for Speaker Identification from the corresponding collected questioned noisy speech samples. In order to keep a Reference Speech Signal (hereafter referred as RSS) for the second phase of experiment—Speech Recognition, speech samples have been recorded in studio condition in the laboratory.

3.6.1.3 Simulated Speech Samples

Noise portion from the actual case exemplars for each samples is collected and kept as Reference Noise Signal (hereafter referred as RNS). This RNS for each noisy speech sample is mixed with RSS for preparation of simulated noisy samples in the laboratory. The simulated speech samples are used for conducting the experiment for Speech Recognition as well as for Speaker Identification purposes.

3.6.2 Analogue Signal

The signal, which may assume any value in a continuous range, is called analogue signal. Analogue signals are real-valued signals which are defined at every instant of time over a continuous domain, such as an interval; or a union of intervals. Analogue signal can have value for any instant of time and it may have an infinite number of different instantaneous values. They are continuous-amplitude signals: taking any value from a continuous region, a number of possible values. Analogue signals are both continuous-time and continuous-amplitude.
### 3.6.3 Digital Signal

The Latin word “digit” means “finger”, standing for a simple aid for counting. The signal, which makes abrupt transitions between two well-defined levels, is called digital signal. The property of digital signal is that it is defined only for a limited number of values and for certain instances. Digital signal have values at specific interval of time and hence they are discrete. Basic form of analogue and digital signals are represented in Fig-3.9.

![Digital Signal Diagram](image)

Fig-3.9 The normal representation of analogue and digital signal.

### 3.6.4 Digitization

In analogue signal processing signal and noise cannot be separated completely. It is a characteristic of the analogue signal processing, that with each step of processing, more unwanted signals (noise, harmonics, etc.) are added to the original signal. Digital signal
processing does not produce any of the above mentioned changes to the original signals at all. Because of the fact that the quality, accuracy, speed and flexibility of processing a signal in digital domain is far better as compared to analogue domain, it is always preferred to process a signal in digital domain.

![Figure: 3.11 Comparison of analogue and digital signal processing](image)

To convert an analogue signal to a digital signal, two operations are required: 1) The analogue signal must be evaluated at certain instances. 2) The detected analogue values must be converted to binary digital values. For each of the two operations a certain device will be required. Each evaluation and conversion of an analogue value, thus each digital sample of the analogue signal, will produce a certain amount of data. If a certain time, of an analogue signal is to be converted to a digital signal, the amount of binary data produced depends on: - The number of samples has taken during the period. - The number of bits used to represent one sample. The total number of bits produced is the product of samples and bits per sample.

In an ADC, the input analogue voltage can have any value in a range. But the digital output can have only 2N discrete values for N-bit ADC. Therefore, the whole range of
analogue voltage is required to be represented suitably in $2N$ intervals and each interval then corresponds to a digital output.

When audio is recorded in computer, the sound card starts the recording process and specifies what sample rate and bit depth to use. Through Line In or Microphone In ports, the sound card receives analogue audio and digitally samples it at the specified rate. Computer stores each sample in sequence until recording is stopped. When a file is played, the process happens in reverse. Computer sends a series of digital samples to the sound card. The card reconstructs the original waveform and sends it as an analogue signal through Line Out ports to speakers.

To summarize, the process of digitizing audio starts with a pressure wave in the air. A microphone converts this pressure wave into voltage changes. A sound card converts these voltage changes into digital samples. After analogue sound becomes digital audio, it is possible to record, edit, process, and mix the audio. A computer based analogue to digital converter is shown here (Fig-3.11.).

![Fig-3.11 Analogue-to-digital converter](image)
3.6.5 Methodology

Different kinds of noise associated with each mode of recording is identified and grouped according to their basic features. For this, non-speech regions containing noise have been collected from the sample and are mixed with Reference Speech Signal (RSS) recorded in ideal condition for preparation of simulated noisy samples to enable a comparative study of systematic noise filtering. Both the original noisy speech and simulated noisy speech is subjected to various filtering techniques. The improvements produced by various filtering techniques were studied in a statistical way. A comparative study of systematic noise filtering by comparing the distortion level of the filtered simulated signal with RSS is carried out. This helps in deciding degree of quality of the speech signal retrieved for Speech Recognition and Speaker Identification in noisy conditions. This is carried out by comparing the distortion level and SNR of the filtered speech signal with RSS. From the resultant signal, speech is reclaimed to a maximum extent by nullifying the effect of embedded noise with the help of specific filter(s). This in turn helps to decide the degree of quality of speech signal retrieved for Speaker Identification under noisy conditions in terms of SNR and perceptual features.

Thus, various groups of noise embedded with the speech samples are subjected for suitable filtering technique(s) for efficient noise reduction. The SNR of simulated speech samples before and after applying the filters are also compared and studied. Then the effect of filter application on SNR and speaker dependent features of speech samples before and after the application of filters are also studied.

\[
\text{SNR(dB)} = 10 \log_{10}(P_{\text{signal}}/P_{\text{noise}}) = 20 \log_{10}(A_{\text{signal}}/A_{\text{noise}}) \quad \ldots \ldots \ldots \text{3.44}
\]
As perceptual features are very much important in the case of Forensic Speaker Identification, listening tests are conducted to ensure that the perceptual features of the original noisy speech are preserved while applying filters. For this purpose, 13 listeners of age group of 25-30 were trained with the original noisy speech until they are familiar to follow the perceptual features of each speaker. The filtered speech samples are then subjected to critical listening by the trained listeners. Finally, an opinion is made based on the method of Mean Opinion Score (MOS); the most widely accepted method for speech quality evaluation and a simplest subjective measure for the assessment that gives an overall opinion of the performance. The standard and possible set of score for MOS is presented in the Table-3.1.

<table>
<thead>
<tr>
<th>Score</th>
<th>Opinion</th>
<th>Impairment scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Unsatisfactory</td>
<td>Annoying and unacceptabe</td>
</tr>
<tr>
<td>2</td>
<td>Poor</td>
<td>Annoying</td>
</tr>
<tr>
<td>3</td>
<td>Fair</td>
<td>Slightly annoying</td>
</tr>
<tr>
<td>4</td>
<td>Good</td>
<td>Perceptible but acceptable</td>
</tr>
<tr>
<td>5</td>
<td>Excellent</td>
<td>Imperceptible</td>
</tr>
</tbody>
</table>

Table-3.1- Mean Opinion Score (MOS) definition

A decision on the processed samples is made by estimating SNR with the help of various analyses conducted with Computerized Speech Laboratory (CSL) and Mean Opinion Score (MOS) test.

3.6.6 Instrumentation

3.6.6.1 Digitization of noisy speech samples

Sony model number- HCD-V515, HI FI audio system MHC- V777 is used to play the recorded tape and the output is fed into the computer through a connector. Digitization of noisy speech samples are conducted with the help of sound blaster card of
creative audigy make installed on the computer along with the Adobe audition software (version 1.0.). The noisy speech samples in analogue mode are converted in PCM format with 16bit quantization (unsigned) and 22050Hz sampling rate.

3.6.6.2 Recording speech in studio situation

Reference Speech Signal (RSS) is recorded directly in digital form using hardware and software of Computerised Speech Laboratory (CSL). A unidirectional microphone of Shurie make is used in the process of recording the Reference Speech Signal (RSS).

3.6.6.3 Simulation of noisy speech samples

With the help of Adobe audition software (version 1.0.) the noise only portion collected from the noisy speech samples and the recorded speech in studio condition are mixing together to produce simulated noisy speech samples. This is carried out by superimposing the RNS of each noisy speech sample and RSS to produce corresponding simulated noisy speech samples.

3.6.6.4 Computerized Speech Laboratory (CSL)

Computerized Speech Laboratory (CSL) 4300B is used for the recording and analysis of the speech samples. It is a windows (Operating System) software that is used for the acquisition, acoustic analysis, display and playback of speech signals.

Signal acquisition, storing speech to disk memory, graphical and numerical display of speech parameters, audio output, signal editing are the main operations of CSL with options like spectrographic analysis, pitch contour analysis, LPC Analysis, Cepstrum analysis, FFT and Energy contour analysis etc. can be performed with CSL.
3.6.7 Experiment

As a preliminary test, the classified digital noisy speech samples are subjected for critical listening. This is done in order to understand the perceptual characteristics of noise embedded with the speech. A specific duration of noise only portion is selected and is kept as Reference Noise Signal (RNS) for the particular sample throughout the experiment. The sample containing more than one type of noise, are segregated. From these segregated samples with each having unique kind of noise structure, a specific duration of noise only portion are selected. The selected noise only part which is prominent in the sample is kept as Reference Noise Signal (RNS) for the particular sample throughout the experiment. This is used to carry out the study exclusively on noise, its structure and its effect when embedded on speech. Thereafter the degree of distortion is compared as the factor of applying different filter(s).

3.6.7.1 Filters used

The classified noisy speech exemplars are subjected to undergo application of various filtering techniques. Filters are applied upon original noisy speech samples and simulated samples in a twofold manner. The filtering is carried out for Speech Recognition and Speaker Identification both separately. The filters applied for Speech Recognition are FFT filter, Noise Reduction, Noise Gate, Notch filter, Band-pass, Butterworth filter, Digital Equalizer and Parametric Equalizer. For Speaker Identification filters applied are Noise Reduction, Noise Gate, Notch filter, Band-pass and Butterworth filter. Such unique Identification of Software/Filters is necessary to preserve the Speaker-Specific Information in the case of Speaker Identification. For such reason corresponding Filters is chosen keeping the Speaker Dependent Characteristics unaltered at any level of the signal. The outcome of each filter upon noisy speech samples is studied and analysed.
Speech is reclaimed to a maximum extent by nullifying the effect of embedded noisy signal from the resultant signal by the application of filters. The degree of distortion upon speech signal due to embedded noise is also studied by comparing the results obtained by the application of filters.

3.6.7.2 Analysis of Speech Exemplars

Various analyses are performed using Computerized Speech Laboratory. In the experiment, Blackman windowing technique is used with a frame length of 20 milliseconds.

Both time domain and frequency domain analysis, namely, Energy contour analysis, FFT analysis, LTA analysis and LPC analysis are carried out in order to study the effects of embedded noisy signal upon speech signal for original noisy speech and simulated noisy speech before and after applying various filters.

The Energy contour analysis is carried out with a Sampling rate of 22050Hz, Frame Length of 20 (msec), total 1013 samples with no Smoothing Level. The Blackman window weighing technique is used for the analysis.

While taking the FFT power spectrum, the sampling rate is kept at 22050 with 512 FFT size and zero pre-emphasis level. Blackman window weighting technique with no smoothing level and 257 samples is used for the study.

During the LTA analysis, sampling rate is kept at 22050 with 512 FFT size, zero pre-emphasis level. Using Blackman window weighting of no smoothing level and 257 samples, Long Term Average (LTA) for original noisy speech samples and simulated noisy speech samples were calculated.

Formant values were extracted from LPC spectra. The sampling rate is kept at 22050 with Frame Length of 20 msec with 221 points. Autocorrelation analysis method is
used with filter order of 12. Blackman window weighting technique with pre-emphasis of 0.90 for 201 samples is used for the analysis.

Based on the inference from the subsequent study conducted upon the results of various analyses, appropriate Filters for each class of noise associated with Forensic speech samples is identified and discussed for their efficiency in enhancing the speech for Speech Recognition and Speaker Identification.

3.6.7.3 Statistical study

Statistical study is conducted over the results produced in terms of Signal to Noise Ratio (SNR) by the various filtering techniques and the subsequent analysis by CSL. Thus the improvement produced by various filters upon original noisy speech and simulated noisy speech samples is studied. The percentage of improvement by each filters are calculated and compared for both original noisy speech and simulated noisy speech samples. Based on this study characterization of noise and their classification are performed. Thus, a systematic characterization of noise has been achieved by which it is possible to classify the same.