Chapter 2
SPEECH ANALYSIS AND SYNTHESIS

2.1 INTRODUCTION:

Speech signal analysis is used to characterize the spectral information of an input speech signal. Speech signal analysis [52-53] techniques are employed in a variety of systems, including voice recognition and digital speech compression. Popular method of analyzing speech signals uses Linear Predictive Coding (LPC) [43-45]. Usually each sample of a digital speech signal is represented as a combination of an innovation sample and a weighted series of past speech samples in Linear Predictive Coding. The series coefficients, or weights, are referred to as LPC coefficients. Real-Time LPC analysis of speech signals is a computationally burdensome process.

Many voice recognition devices currently use LPC speech analysis techniques to generate useful spectral information about an input speech signal. In voice recognition, LPC techniques are employed to create observation vectors, which are used by voice recognizers. These observation vectors are compared or matched to stored model vectors in order to recognize the input speech. Voice recognition systems have been utilized in various industries, including telephony and consumer electronics. For example, mobile telephones may employ voice recognition to allow "hands free" dialing, or voice dialing. For the speech signal analysis, first the windowing technique is used.

2.2 WINDOWING:

One may choose to use some kind of windowing function in the linear analysis to smooth the estimated power spectrum and to avoid abrupt transitions in frequency response between adjacent frames. Spectral smoothing techniques are used to avoid distinct peaks in the spectrum, which will result in poles near the unit circle. The effect of multiplying the input with a finite-length window [55] is equal to convolving the power spectrum with the frequency response of the window. This causes the side-
lobes in the frequency response of the window to have an averaging effect on the signal power spectrum. The use of 160-sample frames in the linear analysis would be equal to windowing the input with a 160-point rectangular window.

In the analysis, a 160-point Hamming window is generally used, which has better frequency properties than the rectangular window. The effect of this is to produce a weighted average of the input, where the 160 samples in the center of the Hanning window correspond to the frame being processed, i.e. the last sub-frame of the preceding frame and the first one of the next frame are also included in the analysis. This alleviates the effect of abrupt transitions in the frequency properties of adjacent frames.

There are two important windows which are used frequently for windowing techniques. They are discussed below:

### 2.2.1 Hamming and Hanning Windows:

The window function is expressed as

\[
W_2(n) = W_1(n)\left(\beta_1 - 2\beta_2 \cos(\alpha n)\right), \quad \alpha = \frac{2\pi}{L - 1}
\]

\[
W_2(e^{i\omega}) = \beta_1 W_1(e^{i\omega}) - \beta_2 W_1(e^{i(\omega+\alpha)}) - \beta_2 W_1(e^{i(\omega-\alpha)})
\]

Where \( L \) represents the width, in samples, of a discrete-time and Pulling out the exponential terms, and simplifying we get

\[
e^{i\left(\frac{\omega(L-1)}{2}\right)} \left[ \beta_1 \frac{\sin\left(\frac{\omega L}{2}\right)}{\sin\left(\frac{\omega}{2}\right)} + \beta_2 \frac{\sin\left(\frac{(\omega + \alpha)L}{2}\right)}{\sin\left(\frac{(\omega + \alpha)}{2}\right)} + \beta_2 \frac{\sin\left(\frac{(\omega - \alpha)L}{2}\right)}{\sin\left(\frac{(\omega - \alpha)}{2}\right)} \right]
\]

\[
\left(2.3\right)
\]
The first term in the sum is $180^\circ$ out of phase with the second and third terms for every side lobe, so it is possible to make the side lobes of $\omega_2(e^{j\omega})$ arbitrarily close to zero by choosing the correct $\beta_1$ and $\beta_2$. Some typical choices are:

- For Hanning window: $\beta_1=0.5$, $\beta_2=0.25$.
- For Hamming window: $\beta_1=0.54$, $\beta_2=0.23$.

However with these constants, the first side-lobe amplitude is less than 1% of the main lobe amplitude that is; the first side lobe is down more than 40 dB.

### 2.2.2 Characteristics of the Hamming Window:

- Main lobe: $W_2(e^0) = 0.54L$  \hspace{1cm} (2.4)
- First null: $W_2(e^{\frac{2\pi}{L}}) = 0$  \hspace{1cm} (2.5)
- First side lobe: $20\log_{10}\left(\frac{|W_2(e^{\frac{2\pi}{L}})|}{|W_2(e^0)|}\right) \approx -40$ dB  \hspace{1cm} (2.6)

### 2.2.3 Window Length:

Optimum window length can be chosen as follows:

- a) To resolve Fundamental frequency $f_0$ (may be for pitch detection), the main lobe width ($2 * W_c$) should not exceed $f_0$ or $f_s$. For example, for hamming window, 
  \[
  \frac{8\pi}{L} \leq \frac{2\pi f_0}{f_s}
  \]  \hspace{1cm} (2.7)

- b) To resolve the formants, main lobe should be wide enough to “smooth together” the pitch harmonics, but not so wide that it blurs the formant peaks.

- c) For rapidly changing events (e.g. stop release), $L$ should be as short as possible ($5$-$10$ms max).

After windowing technique then LPC analysis is implemented. The LPC order and gain are the two main factors important for analysis.
2.3 CHOOSING THE LPC ORDER:

For an excellent filter used in LPC [3]:
- Pass band ripple should be tending to zero.
- Stop band attenuation should tend to infinite
- Transition band of Roll off

First two things can be taken care by proper windowing and that is the reason Hamming or Hanning window is generally used, which gives good roll off also. The numbers of LPC coefficients are decided by Roll off parameter and the relation is given by

\[
\text{No. of LPC coefficient (m) = } \frac{4}{\text{transition bandwidth (K Hz)}} \tag{2.8}
\]

And
\[
\text{Transition bandwidth = Sampling frequency } \times \text{ some fraction} \tag{2.9}
\]

The fraction can lie between 0-0.5. So for good roll off fraction it is 0.05 hence the number of LPC coefficient (m) comes to \(\frac{4}{(8 \text{ K Hz}) \times 0.05} = 10\). Less fraction value increases the number of LPC coefficient which will increase the computation time. So it is a tradeoff between quality and computation time.

2.4 CHOOSING THE LPC GAIN:

The LPC excitation is defined by the formula

\[
s(n) = \sum_{k=1}^{p} a_p(k)s(n - k) + Gu(n) \tag{2.10}
\]

The LPC error is defined by the formula

\[
e(n) = s(n) - \sum_{k=1}^{p} \left(a_p(k)s(n - k)\right) + Gu(n) \tag{2.11}
\]

Defining
\[
\sum_{k=1}^{N-1} u^2(n) = 1 \quad (2.12)
\]

Then
\[
G^2 = \frac{\sum e^2(n)}{\sum u^2(n)} \quad (2.13)
\]

### 2.5 Calculation of LPC Coefficients:

The LPC coefficients [3] are determined by minimizing the mean energy of the residual signal using classical least squares method. This is given by

\[
\varepsilon_p = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} \left[ s(n) + \sum_{k=1}^{p} a_p(k)s(n-k) \right]^2 \quad (2.14)
\]

Windowing will determine the range of the coefficients either the speech or the residual signal, leading to the autocorrelation or covariance method, respectively. The autocorrelation method is generally preferred as the performance is more efficient in terms of computations than the covariance method. As a result and the synthesis filter is always stable.

### 2.6 Quantization of LPC Coefficients:

In low bit rate speech coding, to encode spectral envelope, the LPC coefficients are widely used. In forward based LPC coders [3], the LPC coefficients are calculated from the original speech input and these are quantized and further it is transmitted frame-wise.

The overall bit rate depends on the transmission of these coefficients, thus it is important to quantize the LPC coefficients as few bits as possible without introducing excessive spectral distortion and with reasonable complexity, a very important requirement is that the all-pole synthesis filter remains stable after quantization.
2.7 INTERPOLATION OF THE LPC COEFFICIENTS:

LPC analysis [4] in speech coding systems is carried out through a frame-by-frame basis with a new set of parameters computed, quantized and transmitted at frame intervals of 20 to 30ms. Changes in LPC parameters with adjacent frames are due to slow update of frames, which may produce undesired transients or clicks in the reconstructed signal.

To overcome this problem, interpolation of LPC parameters is used at the receiver to get smooth variations in their values. Usually, interpolation is done linearly, at equally spaced time instants called sub frames, four sub frames are generally used.

All-pole synthesis filter can become unstable if the interpolation is not done directly on the LPC coefficients. In fact, stability issues in the interpolation are very similar to those encountered in quantization interpolation on the reflection coefficients; log area ratios; inverse sine coefficients and LSF parameters always produce stable filters. It is better to use interpolation of the same LPC representation that was used for quantization and it is known that LSF representation. It was proved to be the best interpolation performance.

2.8 LINEAR PREDICTIVE CODING:

The correlations in the speech signal are of two types: 1) Correlations over time lags of less than 2 ms, the so-called short-term correlations and 2) Correlations resulting from the periodicity of the speech signal, which are observed over time lags of 2 ms or more and which are called the long-term correlations. The envelope of the power spectrum is determined by the short-term correlations and structure of the power spectrum is determined by long-term correlations.

However these correlations can be interpreted and it is generally considered to be useful to extract and encode these correlations as a first step in encoding the speech signal. The LPC analysis captures the short-term correlations and describes them in the form of LPC parameters.
Generally the rate at which the vocal tract changes its shape are limited and it has been found that an update rate for the LPC parameters of about 50 Hz for coding purposes. Linear prediction [4] is the basic fundamental technique for removing redundancy in a signal. Linear prediction is used to estimate the value of the current speech sample based on a linear combination of past speech samples.

Let \( s(n) \) be the sequence of speech samples and \( a_k \) be the \( k^{th} \) predictor coefficient in a predictor of order ‘\( p \)’. The estimated speech sequence \( \hat{s}(n) \) is given by

\[
\hat{s}(n) = a_1 s(n - 1) + a_2 s(n - 2) \ldots \ldots + a_p s(n - p)
\]

\[
= \sum_{k=1}^{p} a_k s(n - k) \tag{2.15}
\]

The prediction error \( e(n) \) is found from

\[
e(n) = s(n) - \hat{s}(n) \tag{2.16}
\]

Linear prediction can be more effective for the speech signal if the time-varying filter is adapted. Redesigning of the filter is necessary in order to get the time-varying characteristics of the speech per frame due to the time-varying vocal tract shape associated with successive distinct sounds. Speech coders are generally designed with a frame size of the order of 20 ms (corresponding to 160 samples at an 8 k Hz sampling rate).

### 2.9 LINE SPECTRAL FREQUENCIES:

A 10\(^{th}\)-order LPC analysis[4] results in an all-pole filter with 10 poles whose transfer function is denoted by \( H(z) = 1/A(z) \) in which

\[
A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_{10} z^{-10} \tag{2.17}
\]
and \([a_1, a_2, \ldots a_{10}]\) are the LPC coefficients. LSF parameters can be represented with LPC coefficients which are related to the zeros of a function of the polynomial \(A(z)\). The LSF parameters are represented by

\[
l = [l_1, l_2, \ldots l_{10}]^T
\]

and they are in fact a scaled version of the angular frequencies known as Line Spectral Pairs which are located between 0 and \(\frac{1}{4}\). The ordering property of the LSF parameters states that these parameters are ordered and bounded within a range, i.e.,

\[
0 < l_1 < l_2 < \ldots < l_{10} < 0.5
\]

The ordering property of LSF parameters [5] encapsulates a large portion of their intra frame dependencies. If the property is successfully employed in the quantizer structure then it can increase the performance of the quantizer.

The following are the properties of LSF parameters:

1) All zeros of LSP polynomials are on the unit circle.
2) Zeros of \(P(z)\) and \(Q(z)\) are interlaced with each other.
3) The minimum phase property of \(A(z)\) can be easily preserved if the first two properties are intact after quantization.

2.10 SPEECH ANALYSIS FILTER:

Linear Predictive Coding is most efficient form of coding technique and it is used in different speech processing applications for representing the envelope of the short-term power spectrum of speech. In LPC analysis [52] [53] of order ‘\(p\)’ the current speech sample \(s(n)\) is predicted by a linear combination of \(p\) past samples \(k\), \(\hat{s}(n)\)

\[
\hat{s}(n) = \sum_{k=1}^{p} a_p(k). s(n - k)
\]
Where $\hat{s}(n)$ is the predictor signal and $\{a_1, \ldots, a_p\}$ are the LPC coefficients. The residual signal $e(n)$ is derived by subtracting $\hat{s}(n)$ from $s(n)$ and the reduced variance is given below

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_p(k) \cdot s(n-k) \quad (2.21)$$

By applying the Z-transform to the equation (2.21) which gives rise to

$$E(z) = A_p(z) \cdot s(z) \quad (2.22)$$

Where $s(z)$ and $E(z)$ are the transforms of the speech signal and the residual signal respectively, and $A_p(z)$ is the LPC analysis filter of order ‘$p’

$$A_p(z) = 1 - \sum_{k=1}^{p} a_p(k) z^{-k} \quad (2.23)$$

The short-term correlation of the input speech signal is removed by giving an output $E(z)$ with more or less flat spectrum.

After implementation of analysis filter, the quantization techniques are implemented and the speech signal is to be brought from the quantized signal at the receiver and so the quantized signal is to be synthesized to get the speech signal.

### 2.11 SPEECH SYNTHESIS FILTER:

The short-term power spectral envelope of the speech signal can be modeled by the all-pole synthesis filter

$$H_p(Z) = \frac{1}{A_p(Z)} = \frac{1}{1 - \sum_{k=1}^{p} a_p(k) \cdot Z^{-k}} \quad (2.24)$$
The above equation is the basis for the LPC analysis [6] model. On the other hand, the LPC synthesis model consists of an excitation source $E(z)$, which provides input to the spectral shaping filter $H_p(Z)$, which will give way the synthesized output speech $S(z)$:

$$S(z) = H_p(z).E(z) \quad (2.25)$$

By following certain principles $E(Z)$ and $H_p(Z)$ are chosen, so that $S(Z)$ is as close as possible to the original speech.

The synthesis of speech signal can be represented by the following diagram:

![Diagram for synthesis of speech signal](Figure 2.1 Diagram for synthesis of speech signal)

To identify the sound whether it is voiced or unvoiced, the LPC analysis [52-54] of each frame can act as decision-making process. The impulse train is used to represent voiced signal, with nonzero taps occurring for every pitch period.

To determine the correct pitch period / frequency, pitch-detecting algorithm is used. The pitch period can be estimated using autocorrelation function. However, if the frame is unvoiced, then white noise is used to represent it and a pitch period of $T=0$ is transmitted.
Therefore, either white noise or impulse train becomes the excitation of the LPC synthesis filter [6]. Hence it is important to emphasize on pitch, gain and coefficient parameters that will be varying with time and from one frame to another.

The above model is often called the LPC Model. This model speaks about the digital filter (called the LPC filter) whose input is either a train of impulses or a white noise sequence and the output is a digital speech signal. The relationship between the physical and the mathematical model is given below:

\[
\begin{align*}
\text{Vocal Tract} & \iff H(z) \text{ (LPC Filter)} \\
\text{Air} & \iff u(n) \text{ (Innovations)} \\
\text{Vocal Cord Vibration} & \iff V \text{ (Voiced)} \\
\text{Vocal Cord Vibration Period} & \iff T \text{ (Pitch Period)} \\
\text{Fricatives and Plosives} & \iff UV \text{ (unvoiced)} \\
\text{Air volume} & \iff G \text{ (gain)}
\end{align*}
\]