In human speech one of the key aspects is its Formant structure. Formants are the resonances of the vocal tract, and as such they have a close relation to the vocal tract geometry. In other words Formant frequency may be refers to the spectral peak of the sound spectrum. Generally, for perception and discrimination three formants namely First (F1), Second (F2) and Third (F3) are considered. In this chapter, the formant frequency of Assamese vowels and words of typical structure i.e. CV, CVC, VCV are estimated which can be helpful for developing Assamese Automatic Speech recognition (ASR) system. It is observed that a significant variation is present in the formant frequencies with respect to gender for both vowel phoneme set as well as the word set. The investigations have shown that the range of formant frequency is maximum in case of isolated vowels, but when the vowels are placed in the nucleus of a structure like CV, VC or CVC, the formant frequency decreases.
5.1 INTRODUCTION

Formants frequencies are known as the distinguishing frequency components of the human speech. With the maximum energy concentration during the vowel utterances, it refers to the specific resonance frequencies of vocal tract. In other words, Formants are defined as the spectral peak of the sound spectrum [10]. Generally, three formant frequencies (F1, F2 and F3) are considered for perception and discrimination [13]. It is considered that one of the key aspects of a speech signal is its formant structure. Resonances of the vocal tract are Formants, and because of this they have a close relation to the vocal tract geometry. The shape of vocal tract is a function of the tongue, the lips, the jaw and the velum. The spectral properties of male, female and child too have differences in a number of ways. The difference between their average vocal tract lengths is one prominent difference. The vocal tract length of females is about 10% shorter compared to that of males, while for children, it is even shorter (up to 10%) than that of females [12]. According to the linear acoustic theory of speech production, it implies that towards the high end of the spectrum, all the formants in male speech undergo a vocal tract length -dependent scaling. Consequently, it is also implies that one has to warp the children spectra toward the lower end in order to match it with the adults speech.

Formants are associated with peaks in the smoothed power spectrum of speech [15]. The estimation of these resonance peaks and bandwidths is important in many applications [12, 13]. For example, to identify several phonetic events in speaker verification, the numbers of formants present in several selected frequency ranges are used. It plays an important role in the design of hearing aids [14]. Unfortunately, it is
very difficult to extract formant frequencies from the speech signals. However, several studies indicate that there exist approximately linear relationships between formant frequencies and other spectral representations [15]. Over the past few decades, methods such as the short-term Fourier transform [12], peak-picking on cepstrally smoothed spectra, and linear predictive coding (LPC) [11, 15] have been used to identify the location of formants in a speech signal. Among these different formant estimation techniques, LPC based methods have received considerable attention since it avoids the time bandwidth problems of the Fourier transform and is straightforward to implement [15]. In this scheme, the vocal-tract system is considered as an autoregressive (AR) model, the parameters of which are obtained through an LPC analysis of speech segments [18]. The formant frequencies or the resonance peaks are determined from these AR model parameters. Since the voiced speech is of a quasi-periodic in nature, the peaks of linear prediction spectral estimation are highly influenced by the frequency of pitch harmonics or we can say fundamental frequency, F0. The vocal fold structure and function influenced the acoustic characteristics of the source, and also depend on vocal tract shape and size highly influenced the acoustic characteristics of the filter. In High -pitched speech, wide spacing of harmonics makes very difficult while doing such estimations. In order to study whether the acoustic characteristics of either the vocal tract or during the study of the vocal fold the resonance frequencies of the vocal tract must be estimated accurately. Consequently, researchers long have attempted numerous modifications to the basic formulations of linear prediction analysis [11]. Since the source for vowels is quasi-periodic puffs of airflow through the vocal folds vibrating at a certain fundamental frequency, vowels are the largest phoneme group because the. Each vowel phoneme has a different vocal tract configuration. Various studies indicated
that the first two formant frequencies measured in the steady-state part of a vowel play an important role in its characterization [14]. However, the formants of the same vowel uttered by different speakers, speaking in different contexts, with different speaking rates and at different stress patterns, exhibit a lot of variability in their formant structure. In reality, sometimes some speech sounds contain a particular peak associated with one formant or may be with a pair, and sometimes a formant also may be so weak that it causes no peak in the spectrum as a consequence of weak excitation. Therefore, it is practically impossible to locate such formants exactly, which can cause all higher-frequency formants may be wrongly labeled, ultimately disastrous effects on the recognition. In such cases alternative labeling must be produced, and any uncertainties that cannot be resolved in other ways must be resolved within the Recognition algorithm [18]. In this Chapter, we have studied the results for the formant frequencies of eight Assamese vowel phonemes as well as the different word structure estimated from the 10th order Linear Predictive Coding (LPC) spectrum envelop. The first three consecutive peaks are generally known as the first three formants. Our results suggested that in frequency domain the distance between the first three formants varies appreciably. Particularly it was seen from our study that, for female speakers, the average distance between adjacent formants is generally much bigger than that of the average distance between adjacent formants for male speakers. We have also seen that the first three formants are distinctly different for all the eight Assamese vowel phonemes. The investigations from this study have shown that the range of formant frequency is maximum while vowels are in isolated form, but when the vowels are placed in the nucleus of a syllable structure like CV, VC or CVC, the formant frequency decreases. Further, it has been also seen that the location of the first three formants is
very different for male with the female informants, suggesting an efficient way to distinguish the gender of the speaker.

In the following section, the method of LPC analysis, specially, the autocorrelation method is presented in details.

### 5.2 LINEAR PREDICTIVE CODING METHOD

Linear prediction coding (LPC) method predicts the output of a linear system based on its input $x_n$ and previous outputs $s_{n-1}, s_{n-2}, \ldots s_{n-p}$. Mathematically

$$\hat{S}_n = \sum_{k=1}^{p} a_k s_{n-k} + \sum_{k=0}^{N} b_k x_{n-k}$$

\[ (5.1) \]

Where $\hat{S}_n$ is refers to the estimate or prediction of $S_n$. The problem is now to determine the constants $a_k$ and $b_k$ respectively in such a way that $\hat{S}_n$ approximates the real output $S_n$ accurately.

Based on the general expression which is given in Equation (5.1), various models have been proposed. In autoregressive model (AR), the output $\hat{S}_n$ is predicted by using only previous outputs and the current input [18]. This implies that $b_k = 0$ for $k > 0$, and hence the problem/aim is to find only $a_k$ and $b_0$ which corresponds to an all-pole filter. In moving average model (MA), the prediction is based only on the input which gives $a_k = 0$ corresponding to an FIR filter [14]. In autoregressive moving average model (ARMA), the general expression (5.1) is considered which corresponds to a general recursive filter [20].
Out of these three different models of LPC, due to several reasons the AR-model is usually preferred in speech processing. One of the striking reasons is because it is very easy to compute while determining the parameters $a_k$. In the following section with mathematical details the AR model is presented.

**5.3 ANALYSIS BASED ON AUTOREGRESSIVE (AR) MODEL**

In the AR model, Equation (5.1) changes to

$$\hat{S}_n = \sum_{k=1}^{p} a_k S_{n-k} \quad \ldots \quad (5.2)$$

Where $a_k$ is the predictor coefficients and $p$ is known as the order of the predictor. The problem is to determine the parameters $a_k$ so that $\hat{S}_n$ would be as close as possible with the recorded speech $S_n$ in some frame of the signal. In the LPC analysis, the speech signal is assumed to be generated from an all-pole source and each speech sample is estimated based on a linear combination of its $p$ previous samples [1]. The overall vocal tract filters which can be represented by a $p$-th order AR system with the help of a transfer function.

Here we would like to present the based on AR model, mathematical details of LPC analysis. If $S(z)$ is the $z$-transform of the speech signal $s(n)$ in time domain, then the transfer function of $H(z)$ is the ratio of output $S(z)$ and input $X(z)$, and is modeled by an all-pole filter is given by-

$$H(z) = \frac{S(z)}{X(z)} = \frac{b_0}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{b_0}{\sum_{k=1}^{p} a_k z^{-k}} \quad \ldots \quad (5.3)$$
Where $b_0$ is known as the gain factor, $a_k$ is the AR parameter of the filter, and $p_k = r_k e^{j\omega_k}$ are the poles with magnitude $r_k$ and angle $\omega_k$ respectively. It is assumed that during a short interval of time frame, a given speech signal is generally stationary. Therefore, within a frame, $H(z)$ can be modeled with a constant coefficients. A pair of complex conjugate poles is required, while to model each formant. Formant frequency ($F_k$) and bandwidth ($B_k$) both respectively can be computed using complex pole $p_k = r_k e^{j\omega_k}$ and sampling frequency $F_s$ as given below-

\[
F_k = \frac{F_s}{2\pi} \omega_k \\
B_k = -\frac{F_s}{\pi} \ln r_k
\]

\[\text{.... (5.4)}\]

Once the prediction polynomial which is $S(z)$ has been calculated, the formant parameters are formulated either by “peak-picking” on the filter response curve or by solving the roots of the equation $X(z) = 0$ respectively. Each pair of complex roots is used to find out the corresponding bandwidth and formant frequency.

After filtering the speech signal which is given by $X(z)$, can obtain an output $e_n$ which is known as error or residual signal given by the following way-

\[
e_n = S_n - \hat{S}_n = S_n - \sum_{k=1}^{p} a_k S_{n-k}
\]

\[\text{.... (5.5)}\]

Thus, mathematically, the aim/problem is to find the predictor coefficients $a_k$ which in return can minimizes the error $e_n$ in some sense over the desired range of samples.
5.3.1 Autocorrelation Method

In practice, the sum over n in Equation (5.5) is finite because of the finiteness of the signal but it is also can be noted that the frame is infinitely long and only few samples have nonzero value. Thus, we have a windowed only those signal $S_n$ where only a finite number of samples are nonzero which can be expressed as

$$ S_n = \begin{cases} 
\text{Some sampled signal, } 0 \leq n \leq N-1 \\
0, \quad n < 0 \text{ and } n \geq N
\end{cases}$$

...(5.6)

It is useful to minimize the total square error (E) which is given by

$$ E = \sum_n e_n^2 = (S_n - \hat{S}_n)^2 $$

...(5.7)

Now, using Equation (5.5) in Equation (5.7) and considering the nature of the windowed signal which is $S_n$ is again defined in general for all time $-\infty < n < \infty$, together with the assuming that $a_0 = 1$, we get

$$ E = \sum_{n=-\infty}^{\infty} (S_n - \sum_{k=1}^{p} a_k S_{n-k})^2 = \sum_{n=-\infty}^{\infty} (\sum_{k=0}^{p} a_k S_{n-k})^2 $$

...(5.8)

Now, the total squared error which can be minimized by setting the partial derivative of E with respect to each $a_i (1 \leq i \leq p)$, to zero as given:

$$ \frac{\partial E}{\partial a_i} = 2 \sum_{n=-\infty}^{\infty} (\sum_{k=0}^{p} a_k S_{n-k} \frac{\partial}{\partial a_i} a_k S_{n-k}) = 2 \sum_{k=0}^{p} a_k \sum_{n=-\infty}^{\infty} (S_{n-k} S_{n-i}) = 0, $$
This can be written as given below

\[ \sum_{k=0}^{p} \alpha_k r_{k,i} = 0 \]  \hspace{1cm} \text{.... (5.9)}

Where,

\[ r_{k,i} = \sum_{n=-\infty}^{\infty} (S_{n+k} S_{n-i}) = \sum_{n=-\infty}^{\infty} (S_{n+i} S_{n}) = \sum_{n=-\infty}^{\infty} S_n S_{n-(k-i)} \]  \hspace{1cm} \text{.... (5.10)}

As here in \textbf{Equation (5.10)}, it has been seen, the term \( r_{k,i} \) depends only on value of \( (k-i) \), and hence it can be written by one variable autocorrelation function \( r_{k-i} = r_{k,i} \). Thus, using the property of symmetry of \( r_i \) which is, \( r_i = r_{-i} \) and \( a_0 = 1 \), the condition for minimization of the total squared error given by the \textbf{Equation. (5.9)} can be written as follows

\[ \sum_{k=1}^{p} \alpha_k r_{k,i} = -r_i \]  \hspace{1cm} \text{.... (5.11)}

Where, \( r_i \) is the autocorrelation function of \( S_n \) with \( i = 1, 2, \ldots, p \). \textbf{Equation (5.11)} is called as the normal equation. Putting \textbf{Equation (5.11)} in \textbf{Equation (5.8)}, one get the minimum total square error as given below

\[ E_p = \sum_{k=0}^{p} \alpha_k r_k \]  \hspace{1cm} \text{.... (5.12)}

We note that the autocorrelation functions \( r_i \) and \( r_{k,i} \), both respectively depends on the summation limits. It has seen that in \textbf{Equation (5.10)}, the limits of \( n \) are from \(-\infty\) to \( \infty \).
Chapter 5  |  Formant Estimation Using LPC Approach

+∞, and for n < 0 and for n ≥ N, Sₙ is zero as given in Equation (5.6). In other words, truncation of Sₙ to zero beyond the N sample window demands the expression as in Equation (5.13)

\[ r_{k,i} = \sum_{n=-\infty}^{\infty} (S_{n+k}S_{n+i}) = \sum_{n=0}^{N-1-(i-k)} (S_{n+i+k}S_{n}) \]  

.... (5.13)

Which is equivalent to the error minimization over the interval of 0 ≤ n ≤ N + p - 1.

Here we would like to point out that an alternative to the autocorrelation method, which is namely, the covariance method chooses the summation limits in Equation (5.10) from p to N - 1, and thus all samples(speech) Sₙ are used during the computation of the covariance matrix. In other words, covariance method uses no explicit window on the speech sample where as the autocorrelation method examines N windowed speech samples, thereby imposes a distortion on the spectral estimation [1]. This distortion can be avoided by the covariance method. Therefore, with the extent as well as the type of smoothing dependent on the window shape and duration, the windowed speech spectrum is a smoothed version of the original. Thus, in the covariance approach

\[ \sum_{k=1}^{p} a_k c_{k,i} = -c_i \]

Where,

\[ c_{k,i} = \sum_{n=p}^{N-1} (S_{n+k}S_{n+i}) \]  

.... (5.14)

Given the autocorrelation function, one can calculate the AR parameters by solving a set of linear equations which is represented in the matrix form as in Equation (5.15)
These $p$ linear equations to be solved are best viewed in compact matrix form as $Rx = y$, where $R$ is the $p \times p$ matrix, $x$ is a column vector of LPC coefficients [7]. While solving the LPC vector, it needs inversion of the $R$ matrix and multiplication of the resulting $p \times p$ matrix with the $y$ vector. It is very difficult to compute the matrix inversion. But, the autocorrelation matrix $R$ has two special properties. First it is symmetric with respect to interchange of its indices, namely, $R_{ij} = R_{ji}$, and secondly it is Toeplitz i.e., all elements along with each diagonal are same. Because of these reasons or properties makes the autocorrelation approach computationally much simpler than the covariance method. Inversion of autocorrelation matrix requires $2p$ storage locations and $O(p^2)$ math operations, whereas the covariance matrix involves $p^2/2$ storage locations and $O(p^3)$ operations [1,7]. In addition, since the autocorrelation matrix $R$ satisfies the above two properties, it allow us to apply a more efficient Levinson-Durbin recursive procedure [18].

**5.3.2 Levinson-Durbin Recursion**

The basic idea of Levinson-Durbin method is to solve the matrix equation $Rx = y$ in steps, that is, by increasing the length of the vector $x$ and by calculating a new solution based on the previous solution [1]. The optimal coefficients satisfy Equation (5.12), namely
Levinson-Durbin recursion starts with the condition that \( r_0 = E_0 \), which may be the error of the zero\(^{th}\) degree predictor. Let us consider the case where \( p = 2 \). The matrix becomes like this Equation (5.17)

\[
\begin{pmatrix}
  r_0 & r_1 & r_2 \\
  r_1 & r_0 & r_1 \\
  r_2 & r_1 & r_0 \\
\end{pmatrix}
\begin{pmatrix}
  a_1 \\
  a_2 \\
  1 \\
\end{pmatrix}
= 
\begin{pmatrix}
  E^2 \\
  0 \\
  0 \\
\end{pmatrix}
\]

\[\text{Equation (5.17)}\]

The structure of matrix \( R \) yields

\[
\begin{pmatrix}
  r_0 & r_1 & r_2 \\
  r_1 & r_0 & r_1 \\
  r_2 & r_1 & r_0 \\
\end{pmatrix}
\begin{pmatrix}
  a_2 \\
  a_1 \\
  1 \\
\end{pmatrix}
= 
\begin{pmatrix}
  0 \\
  0 \\
  E^2 \\
\end{pmatrix}
\]

\[\text{Equation (5.18)}\]

Thus, from the Equation (5.17) and Equation (5.18), we have seen a nice property of the matrix that the matrix equations are still satisfied when the coefficient vector and the result vector are twisted upside down. Now let us try to use the following kind of solution to a bigger group of equations

\[
\begin{pmatrix}
  r_0 & r_1 & r_2 & r_3 \\
  r_1 & r_0 & r_1 & r_2 \\
  r_2 & r_1 & r_0 & r_1 \\
  r_3 & r_2 & r_1 & r_0 \\
\end{pmatrix}
\begin{pmatrix}
  1 \\
  a_1 + k_3 a_2 \\
  a_2 + k_3 a_1 \\
  k_3 \\
\end{pmatrix}
= 
\begin{pmatrix}
  E^2 + k_3 q \\
  0 \\
  0 \\
  k_3 E^2 + q \\
\end{pmatrix}
\]
To be a solution of above problem, it must require that all the elements, in the vector on the right side are equal to zero, except the first one. Thus,

\[ q + k_3 E_2 = 0, \quad \text{and} \quad k_3 q + E_2 = 0 \]

This yield

\[ E_3 = E_2 (1-k_3^2), \quad \text{with} \quad k_3 = \frac{1}{E_2} \sum_{i=0}^{2} a_i r_{3+i} \]

Thus in general,

\[ E_n = E_{n-1} (1-k_n^2), \quad \text{with} \quad k_n = \frac{1}{E_{n-1}} \sum_{i=0}^{n-1} a_i r_{n+i} \]

The prediction error which is \( E_n \) for the \( n^{th} \) degree filter is a square error, and because of this it can never be a negative, i.e., \( E_n > 0 \). This yields a condition on the reflection coefficients which is \( k_n \) from Equation (5.19) as

\[ |k_n| \leq 1 \]

This condition can be exactly related to acoustic tube models. Under this condition of \( k_n \), all the roots of \( X(z) \) will be inside or on the unit circle in the \( z \)-plane. This in turn returns a mathematically guarantees a stable LPC synthesis filter \( H(z) \). Thus, to obtain the LPC coefficients, the Levinson-Durbin algorithm can be written as given below

\[ E_0 = r_0 \]

... (5.20)
In this method spectral accuracy increases as the sampling rate $N$ increases. The autocorrelation method of linear prediction analysis assumes zero data outside of a short-time windowed speech segment and thus typically does not result in an exact solution, even when the data follows an all-pole model [7].

The order $p$ of the all-pole LPC model, it represents the number of poles present in the model and reflects the amount of detail in the LPC spectral estimate of the input speech signal. In real, the choice of $p$ directly influences a compromise of spectral accuracy, computation time and memory. It is considered that the all-pole model ignores spectral zeros and it assumes an infinitely-long stationary speech sound. Hence, it is a risk to assign a large number of poles ($p$) for modeling the expected number of formants. In that case, poles are used by the model which is again to handle non-formant effects in the windowed spectrum. This is, one of the weaknesses present in the all-pole model. The non-formant effects are deriving mostly from the vocal-tract excitation which is from both glottal and fricative respectively and also from lip radiation. In addition to this, spectral zero regularly present in nasalized sounds, which have more resonances than vowels [9]. Keeping all this in mind and to account for windowing effects, in the present study, we take 10 poles for 16000 Hz sampled speech.
5.4 RESULT ANALYSIS OF VOWEL DATABASE

The Vowel database as mentioned in Chapter 3 consists of 8 vowels and 20 speakers with equal numbers of male and female informants. All speakers are well educated at least graduate, spoke Assamese as their regular or native language, and their ages ranges between 25 to 35 years old. Speakers are chosen from across the state with different dialects namely, from Goalpara (South-West region), Kokrajhar (North-West region), Kamrup (Central region), Lakhimpur (North-East region), and Dibrugarh (North-East region) district. Each speaker uttered all the eight Assamese vowel phonemes in isolation. Each spoken phoneme is again repeated up to 10 times. Thus, our vowel database consists of 1600 phonemes.

We use the following Matlab program to construct the LPC spectral envelop for each sound file.

```matlab
x= wavread('f06w9_06.wav');
Fs=16000; %sampling frequency
P=fft(x); %Fast Fourier Transform of the signal
Y=abs(P);
Freq=1:8000; %Nyquist frequency
plot(Freq,8*log10(Y(Freq)));
hold;
a=lpc(x,10);
ak=a(2);
bk=a(3);
b=1;
[h,w]=freqz(b,a,100);
w=w*Fs/(2*pi); % w is assigned as formant frequency
plot(w,10*log10(abs(0.4*h)));
```

Now, using the above algorithm, the Formant Frequencies of Assamese Vowels are calculated for both Male and Female informants. The speech data which is 16 kHz sampling frequency is used in this study. Only native speakers of Assamese language
are selected. The quality of voice (speech) is tested by the phonetician. The duration for recording was up to 20 hours. **Table 5.1** depicts the values of the formant frequency for eight Assamese vowels corresponding to Assamese male and female informants.

**Table 5.1:** First three formant frequencies (Hz) of eight Assamese vowels corresponding to Assamese male and female informants.

<table>
<thead>
<tr>
<th>vowel</th>
<th>/o/</th>
<th>/a/</th>
<th>/ɔ/</th>
<th>/i/</th>
<th>/ɛ/</th>
<th>/e/</th>
<th>/o/</th>
<th>/u/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>550</td>
<td>930</td>
<td>730</td>
<td>430</td>
<td>530</td>
<td>510</td>
<td>490</td>
<td>450</td>
</tr>
<tr>
<td>F2</td>
<td>1080</td>
<td>1550</td>
<td>1330</td>
<td>730</td>
<td>1050</td>
<td>970</td>
<td>820</td>
<td>1100</td>
</tr>
<tr>
<td>F3</td>
<td>2850</td>
<td>3240</td>
<td>2770</td>
<td>3330</td>
<td>2930</td>
<td>3060</td>
<td>3350</td>
<td>2760</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>490</td>
<td>750</td>
<td>650</td>
<td>340</td>
<td>470</td>
<td>480</td>
<td>350</td>
<td>340</td>
</tr>
<tr>
<td>F2</td>
<td>960</td>
<td>1350</td>
<td>990</td>
<td>620</td>
<td>810</td>
<td>810</td>
<td>770</td>
<td>920</td>
</tr>
<tr>
<td>F3</td>
<td>2440</td>
<td>2550</td>
<td>2450</td>
<td>3080</td>
<td>2610</td>
<td>2620</td>
<td>3300</td>
<td>2320</td>
</tr>
</tbody>
</table>

We have seen in **Table 5.1** that the first three formants which are F1, F2 and F3 respectively are distinctly different for all eight Assamese vowel databases. Since different vowels have their corresponding formants at some characteristic places, the corresponding spectrum can distinguish each vowel from others. For example the result of spectra of Assamese vowel phoneme /i/ and /a/ for one female and one male speaker is given in the **Figure 5.1, Figure 5.2, Figure 5.3 and Figure 5.4** respectively.
Figure 5.1: Formant frequency of the Assamese vowel /i/ for a female speaker

Figure 5.2: Formant frequency of the Assamese vowel /i/ for a male speaker
Figure 5.3: Formant frequency of the Assamese vowel /a/ আ for a female speaker

Figure 5.4: Formant frequency of the Assamese vowel /a/ আ for a male speaker
In Table 5.2, it provides the range of average formant frequencies. It suggests that the distance between the first three formants varies appreciably in frequency domain. In particular, the average distance between adjacent formants for females is generally much bigger than the average distance between adjacent formants for males.

Table 5.2: Range of average first three formant frequencies (Hz) of eight Assamese vowels corresponding to Assamese male and female informants.

<table>
<thead>
<tr>
<th>vowel</th>
<th>/ɒ/</th>
<th>/ɑ/</th>
<th>/ɔ/</th>
<th>/ɪ/</th>
<th>/ɛ/</th>
<th>/e/</th>
<th>/o/</th>
<th>/u/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>450-</td>
<td>800-</td>
<td>670-</td>
<td>350-</td>
<td>380-</td>
<td>310-</td>
<td>350-</td>
<td>350-</td>
</tr>
<tr>
<td></td>
<td>700</td>
<td>1100</td>
<td>850</td>
<td>500</td>
<td>590</td>
<td>570</td>
<td>550</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>770-</td>
<td>1000</td>
<td>940-</td>
<td>550-</td>
<td>460-</td>
<td>420-</td>
<td>550-</td>
<td>560-</td>
</tr>
<tr>
<td></td>
<td>1350</td>
<td>1650</td>
<td>1550</td>
<td>1050</td>
<td>1100</td>
<td>1020</td>
<td>900</td>
<td>1190</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1440</td>
<td>2200</td>
<td>1980</td>
<td>2400</td>
<td>2300</td>
<td>2500</td>
<td>2500</td>
<td>2100</td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>3400</td>
<td>2850</td>
<td>3400</td>
<td>3200</td>
<td>3000</td>
<td>3400</td>
<td>2900</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>400-</td>
<td>650-</td>
<td>600-</td>
<td>220-</td>
<td>320-</td>
<td>290-</td>
<td>300-</td>
<td>320-</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>850</td>
<td>800</td>
<td>460</td>
<td>550</td>
<td>550</td>
<td>510</td>
<td>470</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>700-</td>
<td>900-</td>
<td>760-</td>
<td>500-</td>
<td>480-</td>
<td>460-</td>
<td>500-</td>
<td>520-</td>
</tr>
<tr>
<td></td>
<td>1200</td>
<td>1700</td>
<td>1100</td>
<td>920</td>
<td>1070</td>
<td>980</td>
<td>900</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2060</td>
<td>2000</td>
<td>1780</td>
<td>2200</td>
<td>2180</td>
<td>2200</td>
<td>2450</td>
<td>2020</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>2800</td>
<td>2500</td>
<td>3200</td>
<td>3060</td>
<td>3100</td>
<td>3400</td>
<td>2750</td>
</tr>
</tbody>
</table>
The first formant F1 is always associated with the changes occur in the mouth opening. Sounds or speech which requires a small mouth opening have low-frequency first formant which is F1 and those requires a wide mouth opening always have a high first formant frequency (F1). For example, the vowel /i/ requires a small mouth opening while comparing with that of to the vowel /a/. The F1 of the vowel /i/ for male informants is found to be about 220-460 Hz, whereas the F1 of the vowel /a/ for male informants is about 650-850 Hz. The location of the first two formants which are F1 and F2 respectively plays a significant role in determining vowel identity, although it is also true that the formants still differ from speaker to speaker. For instance, in Figure (5.4) and Figure (5.2) the first three formants for the vowel /a/ in case of a male speaker are at 750 Hz, 1350 Hz, and 2550 Hz, while the corresponding formants for /i/ for male speaker are at 340 Hz, 620 Hz, and 3080 Hz respectively. We have also seen from the Table 5.2, it has been seen that the location of the first three formants is very different for male and female informants. From the Figure 5.5, Figure 5.6 and Figure 5.7, it can be concluded that location of formants plays a great role in while determining gender of a speaker or informants. From the Figure 5.5, Figure 5.6 and Figure 5.7 it is seen that Females have higher formant frequencies than males for all the vowels. It concluded with an efficient way to distinguish the gender of the speaker as well as it helps to differentiate the phonemes (vowels). It is also suggested that in frequency domain the distance between the first three formants varies appreciably. It was observed that the average distance between adjacent formants for males is generally much smaller than the average distance between adjacent formants for females. Our study reveals that, although formant frequencies are associated with phonetic identity, there present a significant variation in these frequencies among different people. In fact, the vocal tract
length always tends to scale these formant frequencies which are leading to very different average values for adult males, adult females, and children.

**Variation of F1 with different VOWELS for male and female informants**

![Figure 5.5: Variation of F1 in Assamese Vowels](image)

**Variation of F2 with different VOWELS for male and female informants**

![Figure 5.6: Variation of F2 in Assamese Vowels](image)
Variation of F3 with different VOWELS for male and female informants

![Graph showing variation of F3 in Assamese Vowels](image)

Figure 5.7: Variation of F3 in Assamese Vowels

5.5 RESULT ANALYSIS OF ISOLATED WORD DATABASE

In this study we have collected the speech samples which are phonetically rich as well as which are frequently used in our day today life. The words have selected having the three words structure i.e. CV, CVC and VCV respectively. Each speaker uttered all the selected words in isolation. Table 5.3, Table 5.4 and Table 5.5 shows the examples of words in Assamese, their corresponding words in IPA format, in Assamese script and the meanings of the words. For the sake of compactness only five examples from each word structure have been mention in the following tables.
### Table 5.3: Word List of CV type

<table>
<thead>
<tr>
<th>Word (IPA)</th>
<th>Assamese</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ma/</td>
<td>মা</td>
<td>Mother</td>
</tr>
<tr>
<td>/kʰa/</td>
<td>খা</td>
<td>Eat</td>
</tr>
<tr>
<td>/nao/</td>
<td>নাও</td>
<td>Boat</td>
</tr>
<tr>
<td>/nɔ/</td>
<td>না</td>
<td>Nine</td>
</tr>
<tr>
<td>/ga/</td>
<td>গা</td>
<td>Body</td>
</tr>
</tbody>
</table>

### Table 5.4: Word List of CVC type

<table>
<thead>
<tr>
<th>Word (IPA)</th>
<th>Assamese</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>/nas/</td>
<td>নাচ</td>
<td>Dance</td>
</tr>
<tr>
<td>/gan/</td>
<td>গান</td>
<td>Song</td>
</tr>
<tr>
<td>/nak/</td>
<td>নাক</td>
<td>Nose</td>
</tr>
<tr>
<td>/gɔs/</td>
<td>গছ</td>
<td>Tree</td>
</tr>
<tr>
<td>/dʰɔn/</td>
<td>ধন</td>
<td>Money</td>
</tr>
</tbody>
</table>
Table 5.5: Word List of VC type

<table>
<thead>
<tr>
<th>WORD LIST</th>
<th>VC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word (IPA)</strong></td>
<td><strong>Assamese</strong></td>
</tr>
<tr>
<td>/am/</td>
<td>আম</td>
</tr>
<tr>
<td>/ah/</td>
<td>আহ</td>
</tr>
<tr>
<td>/ak/</td>
<td>আক</td>
</tr>
<tr>
<td>/ek/</td>
<td>এক</td>
</tr>
<tr>
<td>/azi/</td>
<td>আজি</td>
</tr>
</tbody>
</table>

We use the following Matlab program to construct the LPC spectral envelop for each sound file.

```matlab
x=wavread('f06w9_06.wav');
Fs=16000; %sampling frequency
P=fft(x); %Fast Fourier Transform of the signal
Y=abs(P);
Freq=1:4000; %Nyquist frequency
plot(Freq,8*log10(Y(Freq)));
hold;
a=lpc(x,10);  
ak=a(2);  
bk=a(3);  
b=1;  
[h,w]=freqz(b,a,100);  
w=w*Fs/(2*pi); % w is assigned as formant frequency  
plot(w,10*log10(abs(0.4*h)));
```
Now, using the above algorithm, the Formant Frequencies of Assamese words are estimated for both Male and Female informants. The speech data which was at 16 kHz sampling frequency is used in this study. The quality of speech/voice is tested by the Assamese language phonetician. The recording duration was around up to 20 hours. 

Table 5.6, Table 5.7 and Table 5.8 depicts the values of the formant frequency for Assamese words of different structure (CV, CVC and VC) corresponding to male and female informants.

Table 5.6: First three formant frequencies (Hz) of Assamese words (CV) corresponding to Assamese male and female informants.

<table>
<thead>
<tr>
<th>CV</th>
<th>/ ma / মা</th>
<th>/ kʰa/ খা</th>
<th>/ nao/ নাও</th>
<th>/ nɔ / ন</th>
<th>/ ga / গা</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>F1</td>
<td>420</td>
<td>440</td>
<td>370</td>
<td>320</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>1080</td>
<td>1020</td>
<td>860</td>
<td>800</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td>2070</td>
<td>2520</td>
<td>1900</td>
<td>1700</td>
</tr>
<tr>
<td>Male</td>
<td>F1</td>
<td>400</td>
<td>390</td>
<td>340</td>
<td>280</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>970</td>
<td>920</td>
<td>770</td>
<td>720</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td>1900</td>
<td>1850</td>
<td>1850</td>
<td>1650</td>
</tr>
</tbody>
</table>
Table 5.7: First three formant frequencies (Hz) of Assamese words (CVC) corresponding to Assamese male and female informants.

<table>
<thead>
<tr>
<th>CVC</th>
<th>/nas/নাচ</th>
<th>/gan/গান</th>
<th>/nak/নাক</th>
<th>/gɔs/গছ</th>
<th>/dʰen/ধন</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>F1</td>
<td>410</td>
<td>400</td>
<td>410</td>
<td>330</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>830</td>
<td>820</td>
<td>980</td>
<td>720</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td>1750</td>
<td>1980</td>
<td>2030</td>
<td>1670</td>
</tr>
<tr>
<td>Male</td>
<td>F1</td>
<td>370</td>
<td>350</td>
<td>380</td>
<td>290</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>800</td>
<td>770</td>
<td>690</td>
<td>680</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td>1680</td>
<td>1860</td>
<td>2010</td>
<td>1600</td>
</tr>
</tbody>
</table>

Table 5.8: First three formant frequencies of Assamese words (VC) corresponding to Assamese male and female informants.

<table>
<thead>
<tr>
<th>VC</th>
<th>/am/আম</th>
<th>/ ah/আহ</th>
<th>/ ak/আক</th>
<th>/ ek/এক</th>
<th>/ azi/আজি</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>F1</td>
<td>650</td>
<td>710</td>
<td>650</td>
<td>430</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>920</td>
<td>1200</td>
<td>930</td>
<td>670</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td>1900</td>
<td>2420</td>
<td>1970</td>
<td>1200</td>
</tr>
<tr>
<td>Male</td>
<td>F1</td>
<td>320</td>
<td>340</td>
<td>480</td>
<td>280</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>570</td>
<td>840</td>
<td>770</td>
<td>470</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td>1880</td>
<td>2120</td>
<td>1440</td>
<td>1130</td>
</tr>
</tbody>
</table>
Chapter 5 | Formant Estimation Using LPC Approach

A vowel is defined as a voiced sound, in which the air issues in continuous stream through the pharynx and mouth, there being no obstruction and no narrowing such as would cause audible friction [5]. And vowel sounds are known as the most important in any language for recognition process. To build a good recognizer, successful recognition of vowel sound is obvious. The performance of a good recognizer depends on the correct recognition of the vowel sounds. Vowel sounds are generally long in duration and hence they are spectrally well defined. But vowel sounds varies depending on the placement of it in the word. The Table 5.1 and Table (5.6, 5.7, and 5.8) have shown that the range of formant frequency is maximum in case of isolated vowels, but when the vowels are placed in the nucleus of a structure like CV, VC or CVC, the formant frequency decreases. For instance, the first three formants for the vowel /a/ in case of a male speaker are at 750 Hz, 1350 Hz and 2550 Hz, but the first three formants for the word consist of vowel sound /a/ after or before some consonant goes down. This is shown in the Table 5.6 that in case of a male speaker the word /kʰa/ is at 390 Hz, 920 Hz, and 1850 Hz. The resultant spectra of Assamese word of each type CV, VC, CVC for one female and one male speaker is given in the Figure 5.8, Figure 5.9, Figure 5.10, Figure 5.11, Figure 5.12 and 5.13 respectively. Again from the Figure 5.14, Figure 5.15, Figure 5.16, Figure 5.17, Figure 5.18, Figure 5.19, Figure 5.20, Figure 5.21 and Figure 5.22 it has been concluded that Female speakers has higher formant frequencies than male speakers for all the words.
Figure 5.8: Formant frequency of the Assamese word /kʰa/ “eat” for a female speaker

Figure 5.9: Formant frequency of the Assamese word /kʰa/ “eat” for a male speaker
Chapter 5 | Formant Estimation Using LPC Approach

Figure 5.10: Formant frequency of the Assamese word VC /azi/ আজি

“Today “ for a female speaker

Figure 5.11: Formant frequency of the Assamese word VC /azi/ আজি

“Today “ for a male speaker
Chapter 5 | Formant Estimation Using LPC Approach

Figure 5.12: Formant frequency of the Assamese word CVC /nak/ নাক “Nose” for a female speaker

Figure 5.13: Formant frequency of the Assamese word CVC /nak/ নাক “Nose” for a male speaker
Variation of F1 with different words of CV structure for male and female informants

![Graph of F1 variation with different words of CV structure for male and female informants](image1)

Figure 5.14: Variation of F1 for CV words

Variation of F2 with different words of CV structure for male and female informants

![Graph of F2 variation with different words of CV structure for male and female informants](image2)

Figure 5.15: Variation of F2 for CV words
Variation of F3 with different words of CV structure for male and female informants

Figure 5.16: Variation of F3 for CV words

Variation of F1 with different words of CVC structure for male and female informants

Figure 5.17: Variation of F1 for CVC words
Variation of F2 with different words of CVC structure for male and female informants

![Figure 5.18: Variation of F2 for CVC words](image)

Variation of F3 with different words of CVC structure for male and female informants

![Figure 5.19: Variation of F3 for CVC words](image)
Variation of F1 with different words of VC structure for male and female informants

![Figure 5.20: Variation of F1 for VC words](image)

Variation of F2 with different words of VC structure for male and female informants

![Figure 5.21: Variation of F2 for VC words](image)
Variation of F3 with different words of VC structure for male and female informants

![Figure 5.22: Variation of F3 for VC words](image)

5.6 SUMMARY OF THE CHAPTER

In this chapter, we perform an LPC analysis, which gives output of voiced speech sound as an all-pole filter which in response to a simple sequence of excitation of pulses. LPC analysis is based on the characterization of all-pole digital filters. It is the most common and efficient technique for low-bit-rate speech coding and is a very important tool in speech analysis, recognition and synthesis too. The popularity of LPC comes from its compact and precise representation of the speech spectral magnitude and it is very simple in computation. LPC is used to estimate \( F_0 \), vocal tract area functions, and the formant frequencies and bandwidths. The formants are known as the resonance peaks in the corresponding LPC spectra. The frequencies of the first three formants contain sufficient information for the recognition of vowels as well as other word structures. For each of the vowel phoneme and word speech of both male and female informants, we extract the first three formants by locating the first three consecutive peaks in spectrally
smoothed log spectra. We have seen that the first three formants are distinctly different for all the eight Assamese vowel phonemes. Since different vowels possess their formants at some characteristic places, the spectrum can distinguish vowels from each other. The first formant F1 is directly associated with changes in the mouth opening. Sounds which require a small mouth opening have low-frequency first formant which is F1 and those who require a wide mouth opening have high formant frequency (F1). So as when wide opening vowel like /a/ comes its first formant found at higher frequency as compared to words have small mouth opening vowel. Vowel sounds varies depending on the placement of it in the word. The research study have shown that the range of formant frequency is always maximum when they are in case of isolated vowels, but when the vowels are placed in the nucleus of a structure like CV, VC or CVC, the formant frequency decreases. It is also seen in this chapter that formant estimation is used as an efficient way to distinguish the sex of the speaker. It is also observed from the formant studies is that the average distance between adjacent formants for males is generally much smaller than the average distance between adjacent formants for females. Finally, we would like to conclude this chapter by reporting that the speaker variability caused by its accent is one of the most important and critical issue for automatic speech recognition, speaker recognition, pronunciation modeling, and as well as language learning. Accent can be defined as the patterns of pronunciation features which characterize an individual speech belonging to a particular native language group. When learning or speaking a second Language, spontaneously the speaker will carry these types of accent patterns of his native language into the new language. Therefore, many accent traits may be present in his/her speech. Recently, accent analysis based on formant frequencies has been performed on various languages.
This motivates us to carry the analysis of Assamese language based on accent based estimation of Formant frequency in the near future.