5.1: Introduction

Acoustic-phonetics is the broad underpinning properties of all speech recognition work. A raw acoustic signal is a sequence of one-dimensional magnitudes corresponding to instantaneous measurements of air pressure, sampled at a rate between 8 and 44 kHz. In this form the signal is not susceptible to available statistical modelling techniques. Each sample is meaningful only in relation to others, and the time resolution afforded by the rapid sampling rate, the generating process is radically non-Markovian. Strong correlations exist across large numbers of samples, and the information of the signal lies in the waveform sketched out by this intricate web of non-local relationships. Transforming this one-dimensional sequence into a more usable representation is essential to the success of any modelling task. The information in thousands of correlated samples must be extracted and gathered into larger quanta whose properties can be non-trivially examined and compared; only from such substantive chunks can we hope to infer class-specific properties with any degree of generality. The kinds of sound categories that are interesting to humans are defined by very high level features. The relevant level of abstraction must be made explicit in the data, in the sense of being within the discriminative power of the model class in use, before
modelling can succeed. This makes clustering substantially more challenging than classification, and it would be unrealistic to expect comparable performance on the two tasks. However, under the assumption that class identities are unattainable or unreliable, we have no choice but to extract and emphasize the meaningful properties of the signal as selectively as possible.

In this chapter at first acoustic-phonetic properties of Assamese and Bodo phonemes with spectral representations are studied. Some of the useful and salient prosodic features, such as temporal energy, pitch variation and phoneme duration, as obtained in the present study have also been presented. Then this chapter describes how to extract information from a speech signal, which means creating feature vectors from the speech signal. A wide range of possibilities exist for parametrically representing a speech signal and its content.

5.2: Spectral Representation

Spectral features are usually referred as the characteristics related to the energy distribution of a speech sound in frequency domain. They also form the basis of analysis, synthesis and recognition of individual speech sound. The following sections represent a general picture of similarities and dissimilarities of the basic phonetic units of Assamese and Bodo vowels.
5.2.1: Formant Frequency

Formant frequency is an important indicator to describe the vowel like sounds in acoustic terms. It refers to specific resonant frequencies of vocal tract which have the greatest energy concentration. It is generally agreed that first three formant frequencies, F1, F2 and F3 are informative for vowel perception and discrimination [83].

The algorithm used to detect the formant frequencies is based on the mathematical model originally proposed by Welling et al [83] which is summarized below:

Based on digitized resonator technique, the entire frequency range is divided into a fixed number of segments, each segment representing a formant frequency. A second order resonator for each segment k with a specific boundary is defined. A predictor polynomial defined as the Fourier Transform of the corresponding second-order predictor is given by [83]:

\[ A_k(e^{jw}) = 1 - \alpha_k e^{-j\omega} - \beta_k e^{-j2\omega} \]  (5.1)

\( \alpha_k \) and \( \beta_k \) are the real valued prediction coefficients. Therefore, from (5.1)

\[ |A_k(e^{j\omega})|^2 = 1 + \alpha_k^2 + \beta_k^2 - 2\alpha_k \cos \omega (1 - \beta_k^2) - 2\beta_k \cos(2\omega) \]  (5.2)

\[ = (1 + \beta_k^2 + \alpha_k^2 + \frac{\alpha_k^2(1 - \beta_k^2)}{4\beta_k^2} - 4\beta_k \cos \omega + \frac{\alpha_k(1 - \beta_k)}{4\beta_k} \]  (5.3)

The parameter \( \beta_k \) determines the bandwidth of the resonator defined as negative logarithm of \( (-\beta_k)[A_k(e^{j\omega})]^2 \). The formant frequency is given by

\[ \varphi_f = \arccos \left( \frac{-\alpha_k(1 - \beta_k)}{4\beta_k} \right) \]  (5.4)

57
The beginning point and the end point of segment $k$ are denoted by $\omega_{k-1}$ and $\omega_k$ respectively. Using the predictor polynomial the prediction error is described as

$$E(\omega_{k-1}, \omega_k | \alpha_k, \beta_k) = \frac{1}{\pi} \int_{\omega_{k-1}}^{\omega_k} \left| S(e^{j\omega}) \right|^2 \left| A_k(e^{j\omega}) \right|^2 d\omega$$  

--- (5.5)

where $\left| S(e^{j\omega}) \right|^2$ denotes the short-time power density spectrum of the speech signal. Using (5.2) the predictor error can be represented as

$$E(\omega_{k-1}, \omega_k | \alpha_k, \beta_k) = (1 + \alpha_k^2 + \beta_k^1) r_k(0) - 2 \alpha_k (1 - \beta_k) r_k(1) - 2 \beta_k r_k(2)$$  

--- (5.6)

where $r_k(v)$ are the autocorrelation coefficients for segment $k$ for $v = 0, 1, 2$.

$$r_k(v) = r_{\omega_{k-1}, \omega_k}(v)$$

$$= \frac{1}{\pi} \int_{\omega_{k-1}}^{\omega_k} \left| S(e^{j\omega}) \right|^2 \cos(v \omega) d\omega$$  

--- (5.7)

By minimizing the prediction error as given by (5.6) with respect to $\alpha_k$ and $\beta_k$, the optimal predictor coefficients are expressed as [84]:

$$\alpha_k^{opt} = \frac{r_k(0) r_k(1) - r_k(1) r_k(2)}{r_k(0)^2 - r_k(1)^2}$$

$$\beta_k^{opt} = \frac{r_k(0) r_k(2) - r_k(1)^2}{r_k(0)^2 - r_k(1)^2}$$  

--- (5.8)

The value of the minimum predictor error is given by

$$E_{\text{min}}(\omega_{k-1}, \omega_k) = \min_{\alpha_k, \beta_k} E(\omega_{k-1}, \omega_k | \alpha_k, \beta_k)$$

$$= r_k(0) - \alpha_k^{opt} r_k(1) - \beta_k^{opt} r_k(2)$$  

--- (5.9)
Thus, from the minimization requirement, we obtain the constraint that the zeroes of the complex predictor polynomial $A(z)$ with complex $z$ must lie inside the unit circle [85]. Using these identities, the minimum requirement can be expressed in terms of predictor coefficient $a_k, \beta_k$ [86] as follows:

\[
\beta_k + a_k < 1 \\
\beta_k - a_k < 1 \\
|\beta_k| < 1
\]

These constraints result in a triangular region in the $(a_k, \beta_k)$-plane. In order to model a pole, i.e., a true second-order resonator, the conventional approach requires that the zeroes of the predictor polynomial should form a conjugate complex pair [86]. This requirement is fulfilled by the condition that

\[
\alpha_k^2 + 4\beta_k < 0
\]

which can be tightened further by combining the previous constraints to the new constraints [83]

It is thus evident that $|\cos \omega| < 1$, thus, the value of $\alpha_k$ and $\beta_k$ can be defined as

\[
\begin{align*}
\alpha_k &< 2 \\
-1 &< \beta_k < -\frac{\alpha_k^2}{4}
\end{align*}
\]

This constraint results in a parabolic boundary line. In case of such conjugate complex pair, the pole frequency is given by the equation

\[
\cos \theta = \frac{\alpha_k}{2\sqrt{(-\beta_k)}}
\]
In the approach presented by welling et al [83] the resonant frequency is the frequency at which the predictor polynomial attains its minimum. The resonance frequency for the $k^{th}$ segment is given by the equation

$$\cos \varphi_k = -\frac{\alpha_k (1 - \beta_k)}{4\beta_k} \quad \text{--- (3.12)}$$

From the inequality $|\cos \varphi_k| < 1$, we can obtain the following constraints for $\alpha_k$ and $\beta_k$

$$\begin{align*}
|\alpha_k| < 2 \\
-1 < \beta_k < -\frac{|\alpha_k|}{4 - |\alpha_k|}
\end{align*} \quad \text{--- (5.13)}$$

Plotting the corresponding boundary line in the $(\alpha, \beta)$ plane it is evident that these constraints are tighter than the constraints for a pole solution.

Thus resonance condition always implies the pole condition. The two frequencies converge to the same value if the damping of the pole approaches to zero, which is given by $\beta_k \to (-1)$.

Fig.5.1 (a) and Fig.5.1 (b) represent the plot of Formant estimation of eight Assamese vowels for male and female utterances respectively and Fig. 5.2 (a) and Fig.5.2 (b) represent the plot of Formant estimation of six Bodo vowels for male and female utterances respectively.

It has been notice that due to the varying vocal tract dimension and along with other factors, formants may vary considerably from speaker to speaker. Based on statistical measurements [87], the variation of F1 verses F2 in different syllabic structures are shown in the Table (5.1) for Assamese vowels for male and female informants respectively and Table (5.2) for Bodo vowels for male
and female informants. Their contour plot is given in the Fig.5.3 (a) and Fig.5.3 (b) for Assamese male and female and Fig.5.4 (a) and Fig.5.4 (b) for Bodo male and female respectively.

The sonorant consonants (semi-vowels) have vowel like spectral features. Their typical formant frequencies are shown in the Table 5.3(a) for Assamese and Table 5.3(b) for Bodo. For other consonants the formant frequencies are insignificant and it cannot be used as a measure for representing and separating the characteristics of the phoneme.

In the present study of the formant frequency, a remarkable shift in formant frequencies has been noticed between the same vowel utterances of male and female informants. Thus shift in formant frequencies can be considered as a parameter to identify the gender of the speaker. Further, it has been observed that formant position of common Assamese and Bodo phonemes differ significantly. This shift in format position can be used to identify the linguistic origin of the speaker.
Fig. 5.1 (a): Formant estimation of eight Assamese vowels for male utterances
Fig. 5.1 (b): Formant estimation of eight Assamese vowels for female utterances.
Fig. 5.2 (a): Formant estimation of six Bodo vowels for male utterances
Fig. 5.2 (b): Formant estimation of six Bodo vowels for female utterances.
Table (5.1): Range and variation of Formant frequencies of eight Assamese vowels

<table>
<thead>
<tr>
<th>Vowel</th>
<th>( F_1 )</th>
<th>( F_2 )</th>
<th>( F_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>[o]</td>
<td>0.54-0.70 (0.16)</td>
<td>0.2-0.4 (0.2)</td>
<td>0.8-1.3 (0.5)</td>
</tr>
<tr>
<td>[a]</td>
<td>0.7-1.0 (0.30)</td>
<td>0.3-0.5 (0.2)</td>
<td>1.25-1.45 (0.20)</td>
</tr>
<tr>
<td>[i]</td>
<td>0.15-0.3 (0.15)</td>
<td>0.2-0.3 (0.1)</td>
<td>0.2-0.45 (0.25)</td>
</tr>
<tr>
<td>[e]</td>
<td>0.3-0.5 (0.2)</td>
<td>0.22-0.85 (0.63)</td>
<td>1.6-2.2 (0.25)</td>
</tr>
<tr>
<td>[e]</td>
<td>0.5-0.7 (0.2)</td>
<td>0.2-0.9 (0.7)</td>
<td>1.5-2.0 (0.5)</td>
</tr>
<tr>
<td>[o]</td>
<td>0.2-0.5 (0.3)</td>
<td>0.20-0.50 (0.3)</td>
<td>0.3-1.0 (0.7)</td>
</tr>
<tr>
<td>[u]</td>
<td>0.1-0.3 (0.2)</td>
<td>0.22-0.32 (0.1)</td>
<td>0.5-1.0 (0.5)</td>
</tr>
<tr>
<td>[v]</td>
<td>0.2-0.7 (0.5)</td>
<td>0.3-0.6 (0.3)</td>
<td>0.8-1.5 (0.7)</td>
</tr>
</tbody>
</table>
Table (5.2): Range and variation of Formant frequencies of six Bodo vowels

<table>
<thead>
<tr>
<th>Vowel</th>
<th>RANGE AND VARIATION OF FORMANT FREQUENCY (KHZ)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F₁</td>
</tr>
<tr>
<td></td>
<td>Male</td>
</tr>
<tr>
<td>[a]</td>
<td>0.60-0.75</td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
</tr>
<tr>
<td>[e]</td>
<td>0.45-0.65</td>
</tr>
<tr>
<td></td>
<td>(0.2)</td>
</tr>
<tr>
<td>[i]</td>
<td>0.20-0.40</td>
</tr>
<tr>
<td></td>
<td>(0.20)</td>
</tr>
<tr>
<td>[o]</td>
<td>0.5-0.65</td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
</tr>
<tr>
<td>[u]</td>
<td>0.32-0.45</td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
</tr>
<tr>
<td>[ɪ]</td>
<td>0.35-0.45</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
</tr>
</tbody>
</table>
Fig. 5.3 (a): $F_1$-$F_2$ plot for male utterances of eight Assamese vowels
Fig. 5.3 (b): $F_1$-$F_2$ plot for female utterances of eight Assamese vowels.
Fig. 5.4 (a): $F_1$ - $F_2$ plot for male utterances of six Bodo Vowels
Fig. 5.4 (b): $F_1$-$F_2$ plot for female utterances of six Bodo Vowels
Table 5.3 (a): Formant frequencies of Assamese semi-vowels in Hz (for both male and Female informants)

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>603.8</td>
<td>949.9</td>
<td>2862.0</td>
</tr>
<tr>
<td>Female</td>
<td>672.4</td>
<td>967.3</td>
<td>2439.2</td>
</tr>
</tbody>
</table>

Table 5.3 (b): Formant frequencies of Bodo semi-vowels in Hz (for both male and female informants)

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>414.3</td>
<td>529.4</td>
<td>1244.1</td>
</tr>
<tr>
<td>Female</td>
<td>471.1</td>
<td>530.6</td>
<td>1120.3</td>
</tr>
</tbody>
</table>

72
5.2.2: Spectrogram

A spectrogram is a representation of the frequency content of a signal, changing over time. Speech waveform consists of sequence of different events. The time variation corresponds to highly fluctuating spectral characteristics over time. A single Fourier Transform of the entire acoustic signal cannot capture the time varying frequency contents for all the harmonics present. In order to capture the time varying nature of the speech signal, short-time Fourier Transform (STFT) consisting of a separate Fourier Transform on the piece of the waveform under a sliding window is used. The sliding window is represented by \( w[n, \tau] \), where \( \tau \) is the position of the window centre and \( n \) is the number of sample per window.

The Fourier Transform of the windowed speech waveform, i.e., STFT is given by \([85]\)

\[
X(\omega, \tau) = \sum_{n=-\infty}^{\infty} x[n, \tau] \exp[-j\omega n] \quad (5.14)
\]

where \( x[n, \tau] = w[n, \tau] x[n] \) represents the windowed speech segments as a function of the window centre at time \( \tau \). The spectrogram is a graphical display of the magnitude of time-varying spectral characteristics and is given by

\[
S(\omega, \tau) = |x(\omega, \tau)|^2 \quad (5.15)
\]

The difference between narrowband and wideband spectrogram is the length of the window \( w[n, \tau] \). The narrowband spectrogram gives good spectral resolution while the wideband spectrogram gives good temporal resolution.
For voiced speech, the output of a linear time-invariant system with impulse response \( h[n] \) and with a glottal flow input given by convolution of the glottal flow over one cycle \( g[n] \), with the impulse train, is given by

\[
p[n] = \sum_{k=-\infty}^{\infty} \delta[n - kP] \quad --- (5.16)
\]

In the windowed speech waveform the result can be expressed as

\[
x[n, \tau] = w[n, \tau] \{(p(n) * g(n)) * h(n)\}
\]

\[
= w[n, \tau](p[n] * \tilde{h}[n]) \quad --- (5.17)
\]

where the glottal waveform over a cycle and vocal tract impulse response are lumped into \( \tilde{h}[n] = g[n] * h[n] \).

Using Multiplication and Convolution theorem, the Fourier Transform of the speech segment is given by [85]

\[
X(\omega, \tau) = \frac{1}{p} W(\omega, \tau) \otimes \left[ H(\omega)G(\omega) \sum_{k=-\infty}^{\infty} \delta(\omega - \omega_k) \right]
\]

\[
= \frac{1}{p} \sum_{k=-\infty}^{\infty} H(\omega_k)G(\omega_k)W(\omega - \omega_k, \tau)
\]

\[
= \frac{1}{p} \sum_{k=-\infty}^{\infty} \tilde{H}(\omega_k)W(\omega - \omega_k, \tau) \quad --- (5.18)
\]

where \( \tilde{H}(\omega_k) = H(\omega_k)G(\omega_k) \) and \( \omega_k = \frac{2\pi k}{P} \) and \( \frac{2\pi}{P} \) is the fundamental frequency.

Therefore, the spectrogram of \( x[n] \) can be expressed as
The characteristics of the consonants can be better represented and separate from each other by spectrogram analysis [88]. The spectrogram of the Assamese and Bodo consonants are given in the Fig.(5.5) and Fig.(5.6) respectively.

From the spectrograms of Assamese and Bodo consonants, it has been noticed that in both Assamese and Bodo languages, nasal consonants \([n]\) and \([\eta]\) are similar in their spectral nature and difficult to isolate. Among the fricatives \([s]\) and \([z]\) are very much similar in spectral nature. The stop consonants are transient phonemes and their acoustical features depend on the context.
Fig. 5.5 (a): Spectrogram of Assamese consonants
Fig. 5.5 (b): Spectrogram of Assamese consonants (Cont...)
Fig. 5.5 (c): Spectrogram of Assamese consonants (Contd...)

![Spectrogram of Assamese consonants](image)

Fig. 5.5 (c): Spectrogram of Assamese consonants (Contd...)
Fig. 5.6 (a): Spectrogram of Bodo consonants
Fig. 5.6 (b): Spectrogram of Bodo consonants (cont...)

[diagram showing spectrograms of different Bodo consonants: [b], [s], [z], [r], [h], [l]]
5.3: Pitch

Pitch is defined as the fundamental frequency of quasi-stationary speech signal. The general problem of fundamental frequency estimation is to take a portion of signal and to find the dominant frequency of repetition. Thus, the difficulties arise are (i) all signals are not periodic, (ii) those are periodic may be changing in fundamental frequency over the time of interest, (iii) signals may be contaminated with noise, even with periodic signals of other fundamental frequencies, (iv) signals that are periodic with interval $T$ are also periodic with interval $2T$, $3T$ etc, so we need to find the smallest periodic interval or the highest fundamental frequency; and (v) even signals of constant fundamental frequency may be changing in other ways over the interval of interest.

A means to estimate fundamental frequency from the waveform directly is to use autocorrelation. The mathematical model based on which the fundamental frequencies are estimated is given below [89]:

A discrete-time short-time sequence is given by

$$S_n[m] = S[m]w[n - m]$$

--- (5.20)

where $w[n]$ is an analysis window of duration $N_w$. The short-time autocorrelation function $r_n[\tau]$ is defined by

$$r_n[\tau] = S_n[\tau] * S_n[-\tau]$$

$$= \sum_{m=-\infty}^{\infty} S_n[m]S_n[m + \tau]$$

--- (5.21)
when \( s[m] \) is periodic with period \( P \), \( r_n[x] \) contains peak at or near the pitch period, \( P \). For unvoiced sound no clear peak occurs near an expected pitch period. Location of the peak in the pitch period range provides a measure of pitch estimation and voicing decision.

The above correlation pitch estimator can be obtained more formally by minimizing, over possible pitch periods \( (P>0) \), the error criterion given by:

\[
E[P] = \sum_{m=-\infty}^{\infty} (S_n[m] - S_n[m + P])^2
\]

minimizing \( E[P] \) with respect to \( P \) yields

\[
\hat{P} = \max_p \left( \sum_{m=-\infty}^{\infty} S_n[m] S_n[m + P] \right)
\]

where \( P > \varepsilon \), i.e., \( P \) is sufficiently far from zero. This alternate view of autocorrelation pitch estimation is used to detecting the pitch of Assamese and Bodo vowels. Table 5.5 (a) shows average pitch of Assamese vowels and Table 5.5 (b) shows the average pitch of Bodo vowels.

Table 5.5 (a): Pitch of 8 Assamese vowels (Average of 100 utterances each)

<table>
<thead>
<tr>
<th></th>
<th>[a]</th>
<th>[i]</th>
<th>[e]</th>
<th>[ɛ]</th>
<th>[o]</th>
<th>[u]</th>
<th>[ɔ]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male</strong></td>
<td>250.42</td>
<td>242.42</td>
<td>533.33</td>
<td>666.67</td>
<td>219.72</td>
<td>727.27</td>
<td>618.7</td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td>276.18</td>
<td>284.17</td>
<td>535.29</td>
<td>680.23</td>
<td>250.0</td>
<td>750.0</td>
<td>690.4</td>
</tr>
</tbody>
</table>
Table 5.5 (b): Pitch of 6 Bodo vowels (Average of 100 utterances each)

<table>
<thead>
<tr>
<th></th>
<th>[a]</th>
<th>[e]</th>
<th>[i]</th>
<th>[o]</th>
<th>[u]</th>
<th>[w]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>119.45</td>
<td>727.27</td>
<td>615.39</td>
<td>123.08</td>
<td>530.76</td>
<td>727.28</td>
</tr>
<tr>
<td>Female</td>
<td>129.03</td>
<td>727.27</td>
<td>666.67</td>
<td>126.98</td>
<td>571.43</td>
<td>803.71</td>
</tr>
</tbody>
</table>

Pitch movement can be considered as an important characteristic for detecting the tone. It has been observed that in both the Assamese and Bodo languages pitch level at the beginning and at the end of a syllable and its movement (raising or falling) are most useful property for the identification of the tone, specially in case of Bodo language.

5.4: Temporal Energy

Temporal energy of the spectra due to a phoneme can be taken as an important representation of the characteristics of the phoneme. The energy of the sequence \( x(n) \) can be written as [90]

\[
\varepsilon_n = \sum_{-\infty}^{\infty} |x(n)|^2 \quad \text{--- (5.24)}
\]

\[
= \frac{1}{2\pi} \int_{-\pi}^{\pi} |X(e^{j\omega})|^2 d\omega
\]

\[
= \frac{\pi}{\pi} |X(e^{j\omega})|^2 d\omega \quad \text{--- (5.25)}
\]
for real sequence using even symmetry. From equation (5.25) the energy density spectrum of $x(n)$ can be expressed as:

$$\Phi_s(\omega) = \frac{\Delta |X(e^{j\omega})|^2}{\pi} \quad --- (5.26)$$

Then the energy of $x(n)$ in the $[\omega_1, \omega_2]$ band is given by

$$\int_{\omega_1}^{\omega_2} \Phi_s(\omega) d\omega, \quad 0 < \omega_1 < \omega_2 < \pi \quad --- (5.27)$$

Using the method stated above, the graphical representation of Energy Density Spectrum of Assamese syllables is presented in the Fig.5.7 (a) and Fig. 5.7 (b). Similarly, the graphical representation of the Energy Density Spectrum of Bodo syllables is presented in the Fig.5.8 (a) and Fig.5.8 (b). Short-term energy is useful in differentiating voiced sound from unvoiced sound and speech signals from the background noises. It has been noticed that the energy of the consonants sounds, in both the Assamese and Bodo languages are very low compared to the vowels and the energy distribution pattern of a syllable is completely dominated by the energy distribution pattern of the vowels.
Fig. 5.7 (a): Energy Spectral Density of Assamese word [o]

Fig. 5.7(b): Energy Spectral Density of Assamese word [olu]
Fig. 5.8 (a): Energy Spectral Density of Bodo vowel [o]

Fig. 5.8 (b): Energy Spectral Density of Assamese word [zo]

5.5: Duration

Duration of sound segments at various phonetic levels is important characteristics for representing continuous speech. Some general observations based on the present study have been detailed below:

1. Vowels are longer in duration than consonants, true for both the languages.
2. Voiced stop consonants are of very short duration.

3. Unvoiced fricatives have longer duration.

Again, it has been noticed that the duration of the vowel varies significantly in various syllabic structures. Typically, the duration of a vowel in CV syllable is longer than CVC structure. For example, duration of the vowel [s] in ‘bra’ (boy) is approximately 100 msec. whereas that in the word ‘la’ (to take) is nearly 230 msec.

Further, the duration of the sound segment seems playing important role in identifying the phonetic unit, particularly in the continuous speech. It has been observed in the present study that vowels in both the Assamese and Bodo languages are longer in duration than the consonants. Further, it has been observed that duration can not be considered as a discriminating feature for the identification of the gender of the speaker.

5.6: Feature Analysis

Spectral information is assumed to characterize the structure and basic properties of the utterance represented by the speech signal. In other words, the goal of such algorithms is to find parameters of the speech signals that can be somehow computed or estimated through processing of the signal waveform. Such parameters are termed as features. The purpose of feature extraction in speech recognition is to transform speech signals into a set of vectors essential for speech recognition while discarding unreliable parts. The transformed vector is called a feature vector. The speech is supposed to contain the sufficient amount of information needed for classification itself. Speech is the
result of interaction among different components of the vocal apparatus. The aim is to somehow track this interaction and the contribution from each component, allowing us to make decisions about a given speech signal. The amount of data involved in using the speech signal directly is huge, suggesting that the speech signal contains a lot of redundant information. Thus, using the speech signal directly wastes computational power and possibly fails to provide adequate information, in the worst case.

It is well-known that one of the key measurements used in speech processing is the short-term spectrum [91]. Such measure consists of some kind of local spectral estimate, typically measured over a relatively short region of speech (e.g., 20-30ms). It has been shown to be useful for a range of speech applications (speech coding, speech recognition).

The basic idea is generally to capture the time-varying spectral envelope of the speech with desirable reduction the effects of pitch on this estimate (it is irrelevant information to the discrimination). Therefore, in speech applications, the short-term spectral based algorithms are usually designed to estimate a spectral envelope that has a reduced influence from the pitch harmonics in voiced speech [92].

5.6.1: Objectives of an Optimal Feature-set

The objectives for an optimal feature-set are [93]:

**Improved Performance:** This is the major goal for designing feature-sets. Unfortunately the relationship between the feature-sets and their recognition performance can only be determined after all the training and testing cycles are
complete. This requires a substantial amount of computer resources and time. A less computationally costly and simple method of accurately predicting the performance of a feature-set has not been found. Thus the whole expensive process has to be repeated each time a modification is made to the front-end system.

**Noise Robustness:** The second important goal is to achieve noise robustness. Traditionally this has been a motivation for many signal-model designs. In real application there are many sources of noise from channels, microphone differences and background noise. Other sources of noise are breath noise from talkers at the beginning and especially at the end of utterances.

**Talker Independence:** In talker-independent systems it is important that the features capture useful semantic information and treat any talker dependent features as noise. For example, features that include pitch will not perform well in a talker-independent environment; this is ideal for a language like English, but may not be optimal for a tonal language. To achieve this talker-independence, most feature-set algorithms include some special explicit method to remove the pitch. However, if the features are to be used for talker identification, verification or characterization pitch would be helpful.

**Temporal Correlation:** Ideal features in a speech recognition system should be able to capture spectral dynamics. Speech is not a stationary process, although it has been successfully modelled as stationary on a short time basis of about 20 to 40 milliseconds. The stationary assumption is best for vocalic sounds (vowels, glides, etc); hence, the performance of most recognizers is best
for vocalic sounds than for rapidly-changing consonants such as stops. Temporal correlation is important for features to be optimal.

5.6.2: Review of Feature Measurement

Signal parameterization may be defined as a process of converting a sequence of time-domain speech samples via a well defined algorithm to a new parametric space. The resulting features are multi-dimensional vectors [94]. The goal of parameterization is to find features that enhance recognition probability of the sampled speech utterance. Feature measurement is basically a data-reduction technique whereby a large number of data points (in this case samples of the speech waveform recorded at an appropriate sampling rate) are transformed into a smaller set of features which are equivalent in the sense that, they faithfully describe the salient properties of the acoustic waveform. For speech signals, data-reduction rates from 10 to 100 are generally practical. For representing speech signals, a number of different feature sets have been proposed ranging from simple sets such as energy and zero crossing rates (usually in selected frequency bands), to complex, “complete” representations such as the short-time spectrum, linear-predictive coding (LPC), and the homomorphic model [95]. There have been many attempts to find optimal features [96-103]. Some have used purely signal processing models such as wavelets and time-frequency distributions, while other researchers have chosen perceptually meaningful models or mathematically tractable algorithms.
Early automatic speech recognition systems used threshold logic on analog voltages from acoustic waveforms [4, 104, 105] integrated over a long period of time to enable recognition of words or short utterances in talker-dependent environments. The next generation of recognizers used analog filterbanks for simple vocabularies like vowels and digits, and some moderately complex vocabularies [8, 12, 13]. During the 1970’s and 1980’s most signal processing front ends were based on Linear Predictive Coding (LPC) processing [101]. This provided a computationally inexpensive method that was mathematically tractable and modelled voiced sounds extremely well [106]. The use of LPC has since continued, but there is a discrete Fourier transform (DFT)-based processing. Some researchers, however, believe that LPC will remain dominating [1]; they argue that optimal performance for a recognition system lies in a statistical modelling rather than in the feature sets.

This section discusses different feature measurement approaches. Most recognition systems use a combination of these to define their signal processing.

a) Waveform Approaches

This type of signal processing is known as waveform or time-domain method. Proponents of this method point to the purity of the data as compared to transform-domain methods where data may be corrupted [107]. For example, in many DFT-based models, the phase of the signal is lost. The features used are typically the amplitude of the signal, zero-crossing rates, level crossing rates, durations and energy [1]. Sheikhzadeh and Deng [108] have used a new
approach that is purported to have some potential to compete with transform-domain methods. They have designed a waveform-based recognition system that combines a time-varying autoregressive filter with the HMM in a hidden filter model (HFM) first suggested by Poritz. Each speech sample serves as a one dimensional observation vector. They found recognition performances between 87% and 90.4% on a simple task of talker-dependent discrete-utterances. Santhoshkumar et al [109], have used the energy envelop and zero-crossing rate for digit recognition and achieved 80% recognition performance, a very low score by current standards. The reason so few researchers have studied waveform approaches is, apart from their claimed advantage of data purity, the lack of proof that they can perform any better than or even as well as other methods. They also have a high data rate. The evidence that the ear itself is performing some kind of signal transformation does not boost confidence in waveform approaches. So far, applications of waveform approaches have been on small tasks and appear to be viable no longer.

b) Transform Domain Approaches

Experimental digital signal processing advances have made these approaches more advantageous than waveform techniques. They may require less data for processing and may be developed to preserve the important information and to discard useless and/or redundant information. Their disadvantage is that some useful information may be lost in the transformation process. Among these transform methods are filterbanks, discrete Fourier
transform (DFT), linear predictive coding (LPC) and other orthogonal transforms such as the Karhunen-Loeve and Hadamard transforms [110, 111].

Linear predictive coding used to be the dominant transform domain method for speech recognition. It was introduced in the early 1970s. It also ignited activity in vector quantization (VQ) research [112]. LPC modelling provides a good model for voiced speech signals, is analytically tractable and, more importantly, it works reasonably well in recognition applications. The linear-predictive coding method provides a robust, reliable, and accurate method for estimating the parameters that characterize the linear time-varying system [84, 95, 106].

Filterbanks [113] were the first transform methods to be used [101], and were mostly applied using analog circuits. In his pioneering work on speech coding, Dudley [114,115] of Bell Telephone Laboratories used analog bandpass filter banks to process speech. Also Davis, Biddulph, and Balashek [4] built a talker-dependent system for isolated digit recognition using analog filter banks. Without digital computers this was the best method to process speech. However most implementations today use digital filters. Filterbanks models are based on the so called Place theory [101], which states that at any time there is a group of nerves in the ear that are only sensitive frequencies in a speech signal, which can be represented by a centre of some bandpass filter. Secondly the theory of just noticeable differences (JND) from experiments in human perception has shown that frequencies of a complex sound within a
certain bandwidth of some nominal frequency can not be individually identified. When one of the components falls out of this critical bandwidth, which is usually 10-20% of the centre frequency it can be distinguished. Filterbanks continue to be used in some large vocabulary systems but they are generally similar to or outperformed by other transform methods such as LPC [116, 117, 118].

c) Auditory-Based Spectral Analysis Models

The motivation for this kind of processing is the ability of the human ear to deal with many kinds of noises (microphone, channels and room reverberations) [1, 119, 99], and to have a high recognition performance [120, 121, 122, 123] even under conditions of phonetic variability. Studies on the auditory system have demonstrated the principles by which sounds are encoded in the auditory nerve, but little is known beyond the encoding. Current research in auditory modelling is mostly devoted to the simulation of the auditory periphery. Several models have been designed to simulate either the direct firing activity in the cochlea or are related to the representation of the cochlear output. Among the popular direct models that have been explored are the Ensemble Interval Histogram (EIH) model developed by Ghitza [124, 125], the Payton model [126, 99], and the cochlear auditory model designed by Lyon [127].

The EIH model focuses on the mechanical motion of the basilar membrane and simulates the activity of the auditory nerves. The Payton auditory model focuses on the effect of the acoustic pressure-wave signal on
the eardrum. The model attempts to approximate the three sections of the hearing mechanism: the middle ear, basilar membrane and the hair-cell sections that include the properties of synaptic connections to auditory nerve fibres. The Lyon cochlear model differs from the others in that its objective is not to accurately model the internal structure of the ear, but only to approximate the information contained in the auditory nerve. It converts a speech signal into a set of features representing information sent to the brain.

Another popular auditory-based model is designed using autoregressive models, and it is called Perceptual Linear Prediction (PLP). The PLP speech analysis technique [122] estimates an all-pole autoregressive model of an auditory-like short-time speech spectrum. It has been shown to efficiently suppress talker-dependent components in a speech signal, and it also uses fewer coefficients than LPC and DFT. The LPC model estimates an all-pole autoregressive model of the short-time spectrum but there is no modification performed on the spectrum to simulate the hearing mechanism. The PLP auditory-like spectrum is obtained by integrating the short-time power spectrum of a speech signal over a simulated range of masking nerves. Some of these auditory models have been applied in large-vocabulary speech recognition systems at IBM [128, 19] and at MIT [129]. Besides the four auditory models discussed so far, there are many others such as the Patterson and Seneff's [99, 130]. The models discussed are different in how they model the hearing system. The Lyon model is only concerned about the input-output relationship, the hearing system is treated like a black box. Whereas, the
Payton and EIH models do approximate the internal workings of this complex system. The PLP on the other hand tries to improve on the LPC. Finding exactly how the ear processes auditory information might not necessarily improve recognition performance, without fully understanding the role of the brain in hearing. It still remains to be seen how far the auditory-based models will contribute to automatic recognition systems.

**d) Homomorphic Signal Processing**

The currently most popular basis for signal processing is homomorphic or cepstral processing. Many interesting sound features are not frequency-specific, a particular word, for instance, can be spoken at a broad range of different pitches. A means of extracting the pitch-invariant properties of a signal, widely used in speech recognition [131], is an extension of spectral analysis known as cepstral analysis. Noll [132] was the first to apply and extend the cepstral notions to speech in his research on pitch detection. However, his ideas were based on earlier work on seismic signals performed by Bogert, Healey and Tukey [133]. Tukey is credited with the term cepstral, which is a play on spectral. Homomorphic signal processing is based on a linear model of speech production, in which speech is viewed as a convolution of a periodic excitation (pitch) and time-varying vocal tract area linear filter model. The calculation takes the inverse FFT of the logarithm of the spectrogram. The cepstrum can be computed without having to transform the data into the frequency domain using Linear Prediction (LPC) coefficients [101].
Cepstral analysis itself, however, is less domain specific. In nontrivial acoustic domains, there is variability in spectral shape, the resonance characteristics of the sound, independent of pitch. If clustering is to focus on these properties, it is essential that they be represented in a way that is relatively pitch-invariant. Cepstral analysis is by far the most commonly used feature extraction model [101, 131]. The motivation for performing analysis in the cepstral domain is two-fold. First, the speech signal contains two key pieces of information: the excitation and the vocal tract shape. Recognition systems typically operate on the latter while speaker identification systems model the former. Cepstral analysis, a form of homomorphic signal processing, provides a mechanism for separating out these two components of the signal. Second, cepstral processing is attractive because it gives the ability to improve noise robustness via very simple and inexpensive mechanisms (such as cepstral mean normalization [134]).

Cepstral analysis, in its ability to deconvolute resonant features from harmonic ones, is general enough to be potentially useful across a broad range of clustering problems. Whether or not it should be employed, and which cepstral coefficients should be retained—low ones for phonetic-like features, higher ones for harmonic information—can only be decided in the context of a particular task. Finally, the cepstral domain does indeed satisfy several requirements of a “good” feature space [101] for speech processing.

For recognition systems, the motivation for choosing one feature set over another is often complex and highly dependent on constraints imposed on
the system (e.g., cost, speed, response time, computational complexity, etc.)[135]. Three of the most important of these criteria are:

1) Computation time
2) Storage
3) Ease of implementation.

Of course the ultimate criterion is overall system performance (i.e., accuracy with which the recognition task is performed). However, this criterion is a complicated function of all system variables.

In the present investigation, LPC-based cepstral coefficients and phonetically important parameters are used as feature vectors.

5.6.3: Feature Extractor

A block diagram for feature extraction is shown in Fig. (5.9).

Speech Signal

Low Pass Filter → Digitization → End-point Detection

Frame Windowing

Frame Blocking → Pre-emphasis

LPC/Cepstral Analysis → Cepstral Weighting

Pitch-Related Feature Detection → Feature Combination

Feature Vector

Fig. 5.9: Block Diagram of Feature Extractor
The basic steps in the feature extraction include the following:

(i) **Low Pass Filter**

A speech signal is first lowpass filtered to prevent the aliasing effect in sampling and remove parts of speech spectrum which are not important and may contain noise.

(ii) **Digitization**

In the digitization procedure, the analog speech waveform is sampled at 8 kHz and quantized by a 16 bit analog-to-digital converter for further digital processing of speech signals.

(iii) **End-point Detection**

The endpoint detection eliminates unnecessary silence parts at the beginning and end of an utterance.

(iv) **Pre-emphasis**

Voiced Speech has a high frequency roll off of -12db/octave while radiation at the lips may be approximated by a 6db/octave spectral lift, resulting in a combined spectral tilt of -6db/octave. It is desirable to have a constant dynamic range across the entire frequency spectrum [136] and the speech is therefore processed to give a 6db/octave lift. Pre-emphasis filter attempt to offset the attenuation and hopefully improve the efficiency of analysis [95].

This spectral shaping is usually accomplished by using a finite impulse response (FIR) filter $H(z) = \sum_{n=0}^{N} a(n)z^{-n}$.
Normally a simple pre-emphasis digital filter with one coefficient is used $H(z) = 1 + az^{-1}$.

The value of $a$ ranges in the set $[-0.4, -1.00]$, adjusting from a full differentiator ($a = -1.00$) to very lossy for ($a = -0.4$). The most common range is $[-0.99, -0.94]$ the slightly lossy/differentiator range. The pre-emphasis filter is intended to boost the signal spectrum and has been shown, in some cases, to improve recognition performance. Also the higher frequencies are amplified by the pre-emphasis filter and more importantly the DC bias is reduced. This amplification of higher frequencies is usually not a problem since they will be compressed to some degree or eliminated depending on the design of the signal processing front-end. The transfer function of the filter used in the experiment is $H(z) = 1 + 0.96z^{-1}$. The frequency response of the filter is shown in Fig. 5.10.

(v) Frame Blocking

The speech signal is continuous when observed over long periods. While over periods of 20-30 ms the signal is, to a reasonable approximation,
stationary. The signal is stationary over this time due to physiological limitations of the speech articulators — the various organs involved in speech production are unable to move fast enough to change their output in a shorter time span [136]. The pre-emphasised speech is therefore parameterised in overlapping blocks as shown by Fig. 5.11 and the signal in each block is assumed to be stationary. For block-processing, consecutive speech samples of 30 ms are taken as a single frame. Frames are overlapped by 20 ms to produce a frame rate of 10 ms.

Fig. 5.11: Frame Blocking

\( w(n) = 0.54 - 0.46 \cos \left( \frac{2 \pi n}{N - 1} \right), 0 \leq n \leq N - 1 \) ---- (5.28)

(vi) Frame Windowing

Frame blocking introduces artifacts into the frequency response of the signal due to the sharp discontinuities at the beginning and end of each frame. A Hamming window, as shown in Fig. 5.12, which tapers at its edges rather than having a sharp discontinuity, introduces fewer artifacts and is therefore used. Speech samples in each frame are multiplied by Hamming window which is given by
where $N$ is the number of samples in a block. This windowing reduces the undesirable effect of Gibbs phenomenon which occurs in the theory of Fourier series [137].

![Fig.5.12: Weighting function of Hamming window](image)

(vii) LPC Cepstral Coefficients

The method of computing the LPC coefficients are based on the assumption that a speech sample at time $n$, $s(n)$ can be approximated by a linear combination of the past $p$ speech samples as given by equation (5.29)[84].

$$S(n) \approx a_1s(n-1) + a_2s(n-2) + \ldots + a_ps(n-p) \quad (5.29)$$

where $a_1, a_2, \ldots, a_p$ are constant coefficients.

The equation can be further transformed by including an excitation term $Gu(n)$ to:

$$S(n) = \sum_{i=1}^{p} a_is(n-i) + Gu(n) \quad (5.30)$$
where $G$ is the gain and $u(n)$ is the normalized excitation. The transformation of equation (5.30) to $z$-domain is given by equation (5.31).

$$S(z) = \sum_{n=1}^{\infty} a_i z^{-i} s(z) + Gu(z) \quad \text{--- (5.31)}$$

and the corresponding transfer function $H(z)$ is described as

$$H(z) = \frac{S(z)}{Gu(z)}$$

$$= \frac{1}{1 - \sum_{n=1}^{\infty} a_i z^{-i}} = \frac{1}{A(z)} \quad \text{--- (5.32)}$$

This corresponds to the transfer function of digital time varying filter. It is important to note the higher order of the model; the better is the representation of the spoken sound. A linear predictor with coefficient $a_i$ is defined with the polynomial $P(z)$ as

$$P(z) = \sum_{k=1}^{\infty} a_k z^{-k} \quad \text{--- (5.33)}$$

where output is given by

$$\hat{s}(n) = \sum_{k=1}^{\infty} a_k s(n - k) \quad \text{--- (5.34)}$$

The predictor error $e(n)$ is defined as

$$e(n) = s(n) - \hat{s}(n)$$

$$= s(n) - \sum_{k=1}^{\infty} a_k s(n - k) \quad \text{--- (5.35)}$$

with error transfer function

$$A(z) = \frac{E(z)}{S(z)} = 1 - \sum_{k=1}^{\infty} a_k z^{-k} \quad \text{--- (5.36)}$$
It is now required to obtain the set of coefficients, $a_k$, that minimizes the prediction error in a short segment of speech. The mean short time predictor error per frame is given as

$$E_n = \sum_{m} e_n^2(m) = \left[ s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right]^2$$ \hspace{1cm} \text{(5.37)}

where $s_n(m)$ is the segment of speech selected in the neighbourhood of a sample $s_n(m) = s(m+n)$. The value of the coefficients $a_k$ that minimize the error $E_n$ can be obtained considering $\frac{dE_n}{da_i} = 0, i = 1, 2, 3, \ldots \ldots, p$, as given in equation (5.38).

$$\sum_{m} s_n(m-i)s_n(m) = \sum_{k=1}^{p} a'_k \sum_{m} s_n(m-i)s_n(m-k), 1 \leq i \leq p$$ \hspace{1cm} \text{(5.38)}

where $a'_k$ are the values of $a_k$ that minimizes $E_n$. Defining $\Phi_n(i,k) = \sum_{m} s_n(m-i)s_n(m-k)$, the equation (5.38) can be rewritten as

$$\sum_{k=1}^{p} a_k \Phi_n(i,k) = \Phi_n(i,0), i = 1, 2, \ldots, p$$ \hspace{1cm} \text{(5.39)}

This is a system of $p$ equations with $p$ variables. The choice of is a compromise between modelling accuracy and computation time. In general, one pair of poles is required to model each of the formants, plus residual 4-6 poles to model possible zeros and general spectral trends in the signal [138]. $p$ is generally therefore between 10-15. The equations can be solved to find $a_k$ coefficients for the segment $s_n(m)$. Thus $E_n$ can be represented as

$$E_n = \sum_{m} s_n^2(m) - \sum_{k=1}^{p} a_k \sum_{m} s_n(m)s_n(m-k)$$ \hspace{1cm} \text{(5.40)}
and in compact form, \( E_n \) further reduced to equation (5.41)

\[
E_n = \Phi(0,0) - \sum_{k=1}^{p} a_k \Phi_n(0,k) \quad \quad (5.41)
\]

Now, the values \( \Phi_n(i, k) \) have to be obtained for \( 1 \leq i \leq p \) and \( 1 \leq k \leq p \) and \( a_k \) coefficients are obtained by solving equation (5.39). The solution of equation (5.41) is obtained following the autocorrelation method.

In the autocorrelation method, the segments \( s_n(m) = 0 \) are considered outside the interval \( 1 \leq m \leq N \) and \( s_n \) is described by the equation

\[
s_n(m) = s_n(m + n)w(m) \quad \text{in the interval, where } w(m) \text{ is a finite length window.}
\]

If \( s_n(m) \) differs from zero for the interval \( 1 \leq m \leq N \), then the corresponding prediction error \( e_n(m) \), for a linear predictor of order \( p \), will be different from zero in the interval \( 1 \leq m \leq N + p \), which is given by

\[
E_n = \sum_{n=0}^{N+p} e_n^2(m) \quad \quad (5.42)
\]

So this method gives comparatively higher prediction error at the beginning and at the end of the specified interval which might be due to the prediction of null samples at the extremes. For this reason, every segment should be subjected to the windowing process (Hamming window in our case) for reducing the border values. Considering \( s_n(m) \) as null outside the interval \( 1 \leq m \leq N \), the error function \( \Phi_n(i, k) \), can be expanded as

\[
\Phi_n(i, k) = \sum_{m=1}^{N+p} s_n(m-i)s_n(m-k), 1 \leq i \leq p, 1 \leq k \leq p \quad \quad (5.43)
\]

which can be further rewrite as
\[ \Phi_n(i,k) = \sum_{m=1}^{N-1} s_n(m)s_n(m+i-k), 1 \leq i \leq p, 1 \leq k \leq p \quad \text{--- (5.44)} \]

In this case, \( \Phi_n(i,k) \) is related to the short-time autocorrelation function value \( R_n(i,k) \) as

\[ \Phi_n(i,k) = R_n(i,k) \quad \text{--- (5.45)} \]

where \( R_n(k) = \sum_{m=1}^{N-k} s_n(m)s_n(m+k) \), is a pair function. Thus

\[ \Phi_n(i,k) = R_n(|i+1-k|), 1 \leq i \leq p, 1 \leq k \leq p \quad \text{--- (5.46)} \]

and therefore,

\[ \sum_{k=1}^{p} a_k R_n(|i+1-k|) = R_n(i), 1 \leq i \leq p \quad \text{--- (5.47)} \]

By analogy, the square of the prediction error can be expanded as

\[ E_n = R_n(0) = \sum_{k=1}^{p} a_k R_n(k) \quad \text{--- (5.48)} \]

The Levinson-Durbin recursion algorithm is used to solve equation (5.47). The complete algorithm is described as follows.

Taking \( E^1 = R(1) \), the following sets of equation is solved recursively for \( i = 2, 3, \ldots, p \)

\[ R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j) \]

\[ k_i = \frac{1}{E^{(i-1)}} \quad \text{--- (5.49)} \]

\[ a_j^{(i)} = k \quad \text{--- (5.50)} \]

\[ a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j-1}^{(i-1)}, 1 \leq j \leq i-1 \quad \text{--- (5.51)} \]

\[ E^{(i)} = (1-k_i^2)E^{(i-1)} \quad \text{--- (5.52)} \]
Thus we have
\[ a_j = a_j^{(p)}, 1 \leq j \leq p \]  

--- (5.53)

Instead of using directly the LPC coefficients as feature vectors, cepstral coefficients based on LPC analysis are usually used because of their superior recognition performances [139]. The LPC-based cepstral coefficients can be derived as follows.

\[ c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}, 1 \leq k \leq p \]  

--- (5.54)

\[ c_m = \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}, m > p \]  

--- (5.55)

where \( c_k \) is an LPC-based cepstral coefficient. It has been established that LPC-based cepstral coefficients produce better recognition results when they are appropriately weighted [140,141]. The weighting function is given as

\[ w(m) = 1 + \frac{Q}{2} \sin \left( \frac{\pi m}{Q} \right), 1 \leq m \leq Q \]  

--- (5.56)

where \( Q \) is the order of cepstral coefficients. In addition to the cepstral coefficients, their time-derivative approximations are used as feature vectors to account for the dynamic characteristic of speech signal. The time derivative is approximated by a linear regression coefficient over a finite window, which is defined as [142], [143]

\[ \Delta \hat{c}_j(m) = \left[ \sum_{k=k}^{k} \hat{c}_{j-k}(m) \right] G, 1 \leq m \leq Q \]  

--- (5.57)
where $\hat{c}_i(m)$ is the $m^{th}$ weighted cepstral coefficient at time $l$ and $G$ is a constant used to make the variances of the derivative terms equal to those with the original cepstral coefficients.

In the present investigation, the following typical values are used: $N=240$, $P=12$, $Q=12$, $K=2$ and $G=0.316$ [143]. The weighted cepstral and the corresponding time-derivatives are concatenated, which resulted in a 24-dimensional observation vector [143], given by

$$U(t) = [\hat{c}_1 \Delta \hat{c}_1 \Delta^2 \hat{c}_1 \Delta^3 \hat{c}_1 \Delta^6 \hat{c}_1 \Delta^9 \hat{c}_1 \Delta^{12} \hat{c}_1]$$  \hspace{1cm} (5.58)

In practice, to reduce the computational cost, some of the less useful cepstral features can be discarded. After discarding the less useful cepstral features, the following features set with only 13 elements are considered for the recognition of basic phoneme of the present study.

$$U(t) = [\hat{c}_1 \Delta \hat{c}_1 \Delta^2 \hat{c}_1 \Delta^3 \hat{c}_1 \Delta^6 \hat{c}_1 \Delta^{12} \hat{c}_1]$$  \hspace{1cm} (5.59)

(viii) Pitch Related Features Detection

Bodo language like other Sino-Tibetan languages is a tonal language. A language is said to be tonal if a word with same phonetic structure but different tone convey different meaning. It has been observed that if pitch related feature is used, the performance of the system degrades considerably for non-tonal languages like Assamese, as considered in the present study. Therefore, in the present study, a new approach has been made to combine features with pitch related features. Again, pitch related features are normalized to work efficiently with tonal as well as non-tonal languages.
Through a pitch detector algorithm the pitch related acoustic features are extracted—including frame energy, the probability of voicing and pitch period. The same window size and frame rates are used to make the extracted pitch features more consistent with the original coefficients based features.

Thus the signal $s(n)$, is first divided into frames. For each frame decisions are made for: (a) speech vs. non-speech, (b) voiced vs. unvoiced and (c) the pitch period. The basic features of the algorithm are as described below:

First to discriminate between speech and non-speech, the signal energy level is computed using autocorrelation and it is then compared with fixed threshold. Cepstral coefficients are computed. In cepstrum domain, first peak ($R_0$) is zero cepstral coefficient, which is partly depends on the frame energy. In voiced the second peak ($R_1$) is present showing the energy of $F_0$. For unvoiced frame, no predominate 2nd peak is present. Therefore, the ratio of $R_1$ against $R_0$ denoted by $R_c$ is compared with a fixed threshold $t$. If $R_c$ is longer than $t$, the frame is classified as voiced and the position of $R_1$ is the pitch period.

For the features to be useful for speech recognition, it is better to make soft decision instead of hard decision for both speech silence differentiation and voiced/unvoiced differentiation. By using autocorrelation value $e$ as a feature, we can estimate the conditional distribution $Pr (e|\text{non-speech})$ and $Pr (e|\text{speech})$ empirically using non-parametric estimation techniques (such as histogram). By using Bayes rule and empirical estimation of $Pr (\text{speech})$ and $Pr (\text{non-speech})$, we can estimate the probability, $Pr (\text{speech}|e)$, for each frame.
A similar modification is also applied for voicing decision. In the voicing decision, first the phonetic alignment is obtained. The distribution of $Pr(R_c|\text{voiced, speech})$ and $Pr(R_c|\text{unvoiced, speech})$ is estimated using Discriminative Learning Algorithm [144,145]. Then instead of comparing $R_c$ with a fixed threshold, we compute the probability $Pr(\text{voiced} | R_c, \text{speech})$, which can be computed using Bayes rule are expressed as given in equation (5.60).

$$Pr(V | R_c, S) = \frac{Pr(R_c | V, S) \cdot Pr(V, S)}{Pr(R_c | V, S) \cdot Pr(V, S) + Pr(R_c | U, S) \cdot Pr(U, S)}$$

--- (5.60)

where $V=\text{Voiced}$, $S=\text{Speech}$, $N=\text{Non-speech}$, $R_c = \frac{R_l}{R_o}$.

Finally, the probability of voicing is computed by the following equation

$$Pr(V) = (1 - Pr(N) \cdot Pr(V | R_c, S))$$

--- (5.61)

The algorithm stated above generates three pitch related features for each frame, namely, the transfer energy $E_n(t)$, the probability of voicing $Pr(V)$ and the pitch period. For using these features in real speech recognition application we are to normalize these parameters as described in the following paragraphs.

The energy of the voicing region is higher than that in unvoiced region and so it is intuitively a useful feature. However, the energy can be affected by loudness which is irrelevant to phoneme identity. In the present study, we use the transformed energy $E_n(t)$, which is given by

$$E_n(t) = \frac{E(t) - E_{\text{channel}}}{E_{\text{max}} - E_{\text{channel}}}$$

--- (5.62)
where \( E_n(t), E_{\text{channel}} \) and \( E_{\text{max}} \) are energy at frame \( t \), average energy in the silence period and maximum energy across the whole utterance respectively. In our study, we consider two type of transformation of \( E_n(t) \) which are given by \( \log(E_n(t)) \) and \( \Delta \log(E_n(t)) \).

Pitch period or \( F_0 \) is the most important feature because it directly related to tone. However, as the pitch period is only defined in the voiced region, depending on the pitch extraction algorithm, it is sometimes set to 0(zero) during unvoiced and silence region. This problem is similar to the problem of probability of voicing that can have zero variance if a hard 0/1 decision is made during feature extraction. Different solutions have been proposed to deal with this problem [146]. In the present investigation, it has been observed that pitch period of unvoiced frame are self sustainable by itself and no spectral treatment is required. Therefore, the pitch period is normalized using average pitch of a sentence as described in equation (5.63).

\[
F_n(t) = \frac{F_0}{\bar{F}_0}
\]

\[--- (5.63)\]

Since tone is actually a segmental feature, modelling the pitch per frame may not be sufficient in determining the tone pattern as derivatives are the normal approaches for modelling frame dependency, therefore, the first order and the second order derivatives of the normalized pitch period, i.e., \( \Delta F_n(t) \) and \( \Delta^2 F_n(t) \) has been considered.
Further, it has been observed that pitch period itself contains voiced/unvoiced information; therefore, we discard the probability of voicing feature parameter. Thus, the pitch related feature vector for frame $t$ is given by

$$U_p(t) = \log(E_n(t)), \Delta \log(E_n(t)), F_n(t), \Delta F_n(t), \Delta^2 F_n(t)$$

--- (5.64)

Combining these pitch related features with the cepstral features, we get a feature vector with 18 elements, which are suitable for recognizing tonal as well as non-tonal languages. The feature vector is finally expressed as

$$U_p(t) = [\log(E_n(t)) \Delta \log(E_n(t)) F_n(t) \Delta F_n(t) \Delta^2 F_n(t) \hat{c}_1 \hat{c}_2 \hat{c}_3 \ldots \hat{c}_k \Delta \hat{c}_2 \Delta \hat{c}_3 \Delta \hat{c}_4 \Delta \hat{c}_5 \Delta \hat{c}_6]$$

--- (5.65)

The plot of the cepstral coefficients of the Bodo vowel [a] extracted from the 45$^{th}$ frame, which is a central frame of the utterance, has been depicted in the Fig. 5.13. Fig. 5.14 and Fig. 5.15 present the plot for weighted cepstral coefficients and their first order derivative for the same frame. Fig.5.16, Fig. 5.17 and Fig. 5.18 depicts the pitch period, its first order and second order derivative for each and every frame of the Bodo vowel [a]. Fig. 5.19 and Fig. 5.20 shows the log energy and differential log energy for all the frames of the Bodo vowel [a]

112
Fig. 5.13: Cepstral coefficient extracted from the 45th frame of the utterance of Bodo vowel [a]
Fig. 5.14: Weighted Cepstral coefficient extracted from the 45th frame of the utterance of Bodo vowel [a]
Fig. 5.15: First order derivative weighted cepstral coefficient extracted from the 45th frame of the utterance of Bodo vowel [a]
Fig. 5.16: Pitch period for the utterance of Bodo vowel [a]
Fig. 5.17: Derivative of the pitch period for the utterance of Bodo vowel [a]
Fig. 5.18: Second order derivative of the pitch period for the utterance of Bodo vowel [a]
Fig. 5.19: Log energy for the utterance of Bodo vowel [a]
Fig. 5.20: Differential log energy for the utterance of Bodo vowel [a]