CHAPTER-III

ACOUSTICAL REPRESENTATION OF ASSAMESE AND BODO PHONEMES

This chapter is devoted to represent the acoustical features of Assamese and Bodo phonemes. 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.

3.1 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.

3.1.1 FORMANT FREQUENCY

Format 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, $F_1$, $F_2$ and $F_3$ are informative for vowel perception and discrimination[128].
The algorithm used to detect the formant frequencies is based on the mathematical model originally proposed by Welling et al [128] 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 [128]:

\[
A_k(e^{j\omega}) = 1 - \alpha_k e^{j\omega} - \beta_k e^{-j\omega} \tag{3.1}
\]

\(\alpha_k\) and \(\beta_k\) are the real valued prediction coefficients. Therefore, from (3.1)

\[
|A_k(e^{j\omega})|^2 = 1 + \alpha_k^2 + \beta_k^2 - 2\alpha_k(1 - \beta_k)\cos\omega - 2\beta_k\cos(2\omega) \tag{3.2}
\]

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

\[
\tag{3.3}
\]

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

\[
\varphi_f = \arccos\left[\frac{-\alpha_k(1 - \beta_k)}{4\beta_k}\right] \tag{3.4}
\]

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

\[
E(\omega_k - 1, \omega_k | \alpha_k, \beta_k) = \frac{1}{\pi} \int_{\omega_{k-1}}^{\omega_k} \left| S(e^{j\omega}) \right|^2 \left| A_k(e^{j\omega}) \right|^2 d\omega \tag{3.5}
\]

where \(|S(e^{j\omega})|^2\) denotes the short-time power density spectrum of the speech signal. Using (3.2) the predictor error can be represented as

46
\[
E(\omega_{k-1}, \omega_k | \alpha_k, \beta_k) = (1 + \alpha_k^2 + \beta_k^2) r_k(0) - 2\alpha_k(1 - \beta_k) r_k(1) - 2\beta_k r_k(2)
\]

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

\[
r_k(\nu) = \omega_{k-1, \omega_k}(\nu)
= \frac{1}{\pi} \int S(e^{j\omega})^2 \cos(\nu \omega) d\omega
\]

--- (3.7)

By minimizing the prediction error as given by (3.6) with respect to \( \alpha_k \) and \( \beta_k \), the optimal predictor coefficients are expressed as [77]:

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

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

--- (3.8)

The value of the minimum predictor error is given by

\[
E_{\text{min}}(\omega_{k-1}, \omega_k) = \min_{\alpha_k, \beta_k} E(\omega_{k-1}, \omega_k | \alpha_k, \beta_k)
= r_k(0) - \alpha_k^{opt} r_k(1) - \beta_k^{opt} r_k(2)
\]

--- (3.9)

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 [54]. Using these identities, the minimum requirement can be expressed in terms of predictor coefficient \( \alpha_k, \beta_k \) [20] as follows:

\[
\beta_k + \alpha_k < 1
\]

\[
\beta_k - \alpha_k < 1
\]

\[
| \beta_k | < 1
\]
These constraints result in a triangular region in the \((\alpha_k, \beta_k)\) 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 [20]. 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 [128]

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

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

--- (3.10)

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

\[
\cos \theta = \frac{\alpha_k}{2\sqrt{(-\beta_k)}}
\]  

--- (3.11)

In the approach presented by welling et al [128] the resonant frequency is the frequency at which the predictor polynomial attains it minimum. The resonance frequency for the \(k^{th}\) segment is given by the equation

\[
\cos \varphi_k = -\frac{\alpha_k(1 - \beta_k)}{4\beta_k}
\]  

--- (3.12)

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

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

--- (3.13)
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 $\beta_k \rightarrow (-1)$.

Fig-3.1 (a) and Fig - 3.1 (b) represent the plot of Formant estimation of eight Assamese vowels for male and female utterances respectively and Fig - 3.2 (a) and Fig - 3.2(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 [135], the variation of $F_1$ verses $F_2$ in different syllabic structures are shown in the Table - 3.1 (a) and Table - 3.1 (b) for Assamese vowels for male and female informants respectively and Table - 3.2 (a) and Table - 3.2 (b) for Bodo vowels for male and female informants. Their contour plot is given in the Fig-3.3 (a) and Fig-3.3 (b) for Assamese male and female and Fig-3.4 (a) and Fig-3.4(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 – 3.3(a) for Assamese and Table – 3.3(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-3.1 (a): Formant estimation of eight Assamese vowels for male utterances
Fig. 3.1 (b): Formant estimation of eight Assamese vowels for female utterances
Fig - 3.2 (a): Formant estimation of six Bodo vowels for male utterances
Fig - 3.2 (b): Formant estimation of six Bodo vowels for female utterances
Table 3.1: Range and variation of Formant frequencies of eight Assamese 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>[ɔ]</td>
<td>0.54-0.70 (0.16)</td>
</tr>
<tr>
<td>[a]</td>
<td>0.7-1.0 (0.30)</td>
</tr>
<tr>
<td>[ɪ]</td>
<td>0.15-0.3 (0.15)</td>
</tr>
<tr>
<td>[ɛ]</td>
<td>0.3-0.5 (0.2)</td>
</tr>
<tr>
<td>[ɛ]</td>
<td>0.5-0.7 (0.2)</td>
</tr>
<tr>
<td>[o]</td>
<td>0.2-0.5 (0.3)</td>
</tr>
<tr>
<td>[u]</td>
<td>0.1-0.3 (0.2)</td>
</tr>
<tr>
<td>[ʊ]</td>
<td>0.2-0.7 (0.5)</td>
</tr>
</tbody>
</table>
Table 3.2: Range and variation of Formant frequencies of six Bodo vowels

<table>
<thead>
<tr>
<th>Vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RANGE AND VARIATION OF FORMANT FREQUENCY (KHZ)</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>[a]</td>
</tr>
<tr>
<td>[e]</td>
</tr>
<tr>
<td>[i]</td>
</tr>
<tr>
<td>[r]</td>
</tr>
<tr>
<td>[u]</td>
</tr>
<tr>
<td>[m]</td>
</tr>
</tbody>
</table>
Fig-3.3 (a): $F_1$-$F_2$ plot for male utterances of eight Assamese vowels
Fig-3.3 (b): $F_1$-$F_2$ plot for female utterances of eight Assamese vowels
Fig -3.4 (a): F₁ - F₂ plot for male utterances of six Bodo Vowels
Fig-3.4 (b): F1-F2 plot for female utterances of six Bodo vowels

Fig-3.4 (b): F1-F2 plot for female utterances of six Bodo vowels
Table – 3.3 (a): Formant frequencies of Assamese semi-vowels in Hz (for both male and female)

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

Table – 3.3 (b): Formant frequencies of Bodo semi-vowels in Hz (for both male and female)

<table>
<thead>
<tr>
<th></th>
<th>[w]</th>
<th>[j]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>F1</td>
<td>414.3</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>529.4</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td>1244.1</td>
</tr>
<tr>
<td>Female</td>
<td>F1</td>
<td>471.1</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>530.6</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td>1120.3</td>
</tr>
</tbody>
</table>
3.1.2 SPECTROGRAM

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 \( w[n, r] \), where \( r \) 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 \[ X(\omega, r) = \sum_{n=-\infty}^{\infty} x[n, r] \exp(-j\omega n) \]  
\[ --- (3.14) \]

where \( x[n, r] = w[n, r]x[n] \) represents the windowed speech segments as a function of the window centre at time \( r \). The spectrogram is a graphical display of the magnitude of time-varying spectral characteristics and is given by \[ S(\omega, r) = |x(\omega, r)|^2 \]  
\[ --- (3.15) \]

The difference between narrowband and wideband spectrogram is the length of the window \( w[n, r] \). 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 \[ p[n] = \sum_{k=-\infty}^{\infty} \delta[n-kP] \]  
\[ --- (3.16) \]
In the windowed speech waveform the result can be expressed as
\[ x[n, \tau] = w[n, \tau] \{(p(n) * g(n)) * h(n)\} \]
\[ = w[n, \tau](p[n] * \tilde{h}[n]) \]
--- (3.17)
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 [54]
\[
X(\omega, \tau) = \frac{1}{p} W(\omega, \tau) \otimes \left[ H(\omega)G(\omega) \sum_{k=-\infty}^{\infty} \delta(\omega - \omega_k) \right]
\]
\[
= \frac{1}{p} \sum_{k=-\infty}^{\infty} H(\omega_k)G(\omega_k)W(\omega - \omega_k, \tau)
\]
\[
= \frac{1}{p} \sum_{k=-\infty}^{\infty} \tilde{H}(\omega_k)W(\omega - \omega_k, \tau)
\]
--- (3.18)
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
\[
S(\omega, \tau) = \frac{1}{p^2} \left| \sum_{k=-\infty}^{\infty} \tilde{H}(\omega_k)W(\omega - \omega_k, \tau) \right|^2
\]
--- (3.19)

The characteristics of the consonants can be better represented and separate from each other by spectrogram analysis [27]. The spectrogram of the Assamese and Bodo consonants are given in the Fig-3.5 and Fig-3.6 respectively.
From the spectrograms of Assamese and Bodo consonants, it has been noticed that in both Assamese and Bodo languages, nasal consonants [n] and [ŋ] 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.

![Spectrogram of Assamese consonants](image)

**Fig – 3.5 (a): Spectrogram of Assamese consonants**
Fig - 3.5 (b): Spectrogram of Assamese consonants (Cont...)

[k]

[kh]

[g]

[gh]

[s]

[z]

[x]

[h]
Fig – 3.5 (c): Spectrogram of Assamese consonants (cont...)
Fig – 3.6 (a): Spectrogram of Bodo consonants
Fig - 3.6 (b): Spectrogram of Bodo consonants (cont...)

[Fig showing spectrograms for Bodo consonants: [p], [s], [z], [r], [h], [l]]
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 [93]:

A discrete-time short-time sequence is given by

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

--- (3.20)

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

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

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

--- (3.21)

when $s[m]$ is periodic with period $P$, $r_n[\tau]$ 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 \( P > \epsilon \), 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-
3.5 (a) shows average pitch of Assamese vowels and Table -3.5 (b) shows the average
pitch of Bodo vowels.

**Table -3.5 (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>[o]</th>
<th>[u]</th>
<th>[ə]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</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>
</tr>
<tr>
<td>Female</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>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>690.4</td>
<td>444.44</td>
</tr>
</tbody>
</table>
Table 3.5 (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.

3.3 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 [44]

$$E_n = \sum_{-\infty}^{\infty} |x(n)|^2$$

--- (3.24)

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} |X(e^{jn})|^2 \, dw$$

$$= \int_{0}^{\pi} |X(e^{jn})|^2 \, dw$$

--- (3.25)
for real sequence using even symmetry. From equation (3.25) the energy density spectrum of $x(n)$ can be expressed as:

$$
\Phi_x(\omega) = \frac{\Delta}{\pi} \left| X(e^{j\omega}) \right|^2
$$

--- (3.26)

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

$$
\frac{\omega_2}{\omega_1} \int_{\omega_1}^{\omega_2} \Phi_x(\omega) d\omega, \quad 0 < \omega_1 < \omega_2 < \pi
$$

--- (3.27)

Using the method stated above, the graphical representation of Energy Density Spectrum of Assamese syllables is presented in the Fig-3.7 (a) and Fig - 3.7 (b). Similarly, the graphical representation of the Energy Density Spectrum of Bodo syllables is presented in the Fig-3.8 (a) and Fig-3.8 (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 - 3.7 (a): Energy Spectral Density of Assamese word [o]

Fig - 3.7(b): Energy Spectral Density of Assamese word [blu]
Fig - 3.8 (a): Energy Spectral Density of Bodo vowel [o]

Fig - 3.8(b): Energy Spectral Density of Assamese word [zo]
3.4 DURATION

Duration of sound segments at various phonetic levels are 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 [ɔ] in 'lōra' (boy) is approximately 100 msec. whereas that in the word ‘lā’ (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 then 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.