CHAPTER 4

Acoustic Representation of Bodo Phonemes and Words of Different Types

4.1 Introduction

A model is defined as the representation of essential structure of some object or event in the real world. An ideal model of a speech signal is that which most accurately reflects different aspect of the speech production system. It has been observed by many scientist that speech can be represented as the output of a linear, time-varying system, where, as the time changes, its properties changes very slowly. This leads to the idea that a sort segment of speech can be modeled as the output of a stationary or quasi stationary wave. The speech signal can also be modeled as the output of a Linear Time-invariant system excited as a quasi-periodic impulse train [55].

Since the excitation and the impulse response of a linear time-invariant system are combined by convolution technique, so it is very difficult to separate the different parts of the convolution. So, as far as speech analysis is considered, homomorphic systems and the cepstrum are considered as some of the useful tools [55].
To study the basic components of the speech signal, the individual effect of speech production organs are to be separated. Homomorphic Analysis [56] is a useful method in this case.

4.2 Homomorphic Filtering

Homomorphic filtering is a class of non-linear signal processing technique, based on a generalization of the principal of superposition that defines linear system. A.M.Noll (1967), first applied the homomorphic analysis to speech processing which led to the success of cepstral pitch detection. In the following section a homomorphic analysis of speech signal in the form of LPC (Linear Predictive Coding) model, has been illustrated and its application in the present analysis of Bodo speech signals have been described [63].

4.2.1 Cepstral Analysis - A Special Case of Homomorphic Filtering

Homomorphic filtering is a generalized technique involving -

A nonlinear mapping to a different domain, linear filters are applied followed by mapping back to the original domain.

Consider the transformation defined by

\[ Y(n) = L[x(n)], \quad \ldots\ldots(4.1) \]

Where, \( x(n) \) = signal under study

If \( L \) is a linear system, it will satisfy the principle of superposition. i.e.

\[ L[x_1(n) + x_2(n)] = L[x_1(n)] + L[x_1(n)] + L[x_2(n)] \quad \ldots\ldots (4.2) \]
By analogy, we define a class of systems that obey a generalized principle of superposition where addition is replaced by convolution. i.e.

\[ H[x(n)] = H[x_1(n) \ast x_2(n)] \]

\[ = H[x_1(n)] \ast H[x_2(n)] \] \hspace{1cm} \text{(4.3)}

Systems which possess this property are known as *homomorphic systems* for convolution, and can be depicted as shown below

An important property of homomorphic systems is that they can be represented as a cascade of three homomorphic systems:

- The first system takes inputs combined by convolution and transforms them into an additive combination of the corresponding outputs, i.e.

\[ D_\ast[x(n)] = D_\ast[x_1(n) \ast x_2(n)] \]

\[ = D_\ast[x_1(n)] + D_\ast[x_2(n)] \]

\[ = \mathcal{X}_1(n) + \mathcal{X}_2(n) \]

\[ = \mathcal{X}(n) \hspace{1cm} \text{(4.4)} \]

Where

- \( \mathcal{X}_1(n), \mathcal{X}_2(n) \) = cepstra of the signals \( x_1(n) \) and \( x_2(n) \) respectively.

\[ D = \text{Discrete delta} \]
• The second system is a conventional linear system that obeys the principle of superposition.

• The third system is the inverse of the first system: it transforms signals combined by addition into signals combined by convolution, i.e.

\[ D^{-1}_r[y(n)] = D^{-1}_r[\overline{y}_1(n) + \overline{y}_2(n)] = D^{-1}_r[y_1(n)] * D^{-1}_r[y_2(n)] = y_1(n) * y_2(n) = y(n) \quad \cdots (4.5) \]

The above expression is important because the design of such system reduces to the design of a linear system.

### 4.3 Speech Model

The human speech production system can be modeled using a simple structure: the lung which generates the air or energy to excite the vocal tract and are represented by a white noise source. The acoustic path composed comprising the larynx, vocal cords, pharynx, mouth and nose, can be modeled by a time-varying digital filter.
Figure-4.1 shows a block diagram of human speech production system.

4.4 Recording Setup used for Speech Research

The accuracy and perfectness of the result of any research depends on the environment in which the research is conducted. It is very much important to follow a recording standard, which is globally standardized. In the current research the following setup has been employed to maintain the globally accepted standard:

- **Microphone Shure**: Model no. SM63,
- **Frequency Response**: 80 Hz to 20 KHz
- 5 channel audio mixer 20 Hz to 20 KHz, model UNISOUN, UB-100e
- Sound card: Sound blaster live 5.1, creative make
- Distance between the speaker and microphone: 8 inch

The block diagram of the recording setup is shown the figure 4.3.
4.5 Linear Predictive Coding (LPC)

4.5.1 Introduction

Linear Predictive Coding (LPC) is defined as a digital method for encoding an analog signal in which a particular value is predicted by a linear function of the previous values of the signal. It was first proposed as a method for encoding human speech by the Department of Defense (DoD), United States. Human speech is produced in the vocal tract which comprises of the throat, the mouth and the nasal cavity that can be approximated as a variable diameter tube. The linear predictive coding (LPC) model is based on a mathematical approximation of the vocal tract represented by this tube of a varying diameter. At a particular time, $t$, the speech sample $s(t)$ is represented as a linear sum of the $p$ previous samples. The most important aspect of LPC is the linear predictive filter which allows the value of the next sample to be determined by a linear combination of previous samples [58].

It has been observed that speech production occurs through slow anatomical movements. The frequency of human speech production ranges from around 300Hz
to 3400Hz. Speech coding is a technique associated with the reduction of the amount of information needed to represent a speech signal. Even if there is some loss of information during the encoding process, but still it is undetectable to the human ear, it is said to be based on lossy algorithm (loss of information). It has been observed that a conversation process consists more than 50% of silence. If this percentage of conversation can be avoided, than there is a great save in the amount of bandwidth requirement. Moreover it has also been observed that, mechanically there is a high correlation between the adjacent samples.

Linear Predictive Coding (LPC) is one of the methods of compression that models the process of speech production. Specifically, LPC models this process as a linear sum of earlier samples using a digital filter inputting an excitement signal. An alternate explanation is that linear prediction filters attempt to predict future values of the input signal based on past signals. “LPC” - models speech as an autoregressive process, and sends the parameters of the process as opposed to sending the speech itself [59].

The basic problem of the LPC system is to determine the formants from the speech signal. The solution of this problem is a difference equation, which expresses each sample of the signal as a linear combination of previous samples. Such an equation is called a linear predictor i.e. Linear Predictive Coding. The coefficients of the difference equation (the prediction coefficients) characterize the formants. Therefore, the LPC system needs to estimate these coefficients. The
estimation is made by minimizing the Mean Square Error (MSE) between the predicted signal and the actual signal.

The basic idea behind the LPC model is that a given speech sample \( S(n) \) at time \( n \), can be approximated as a linear combination of the past \( p \) speech samples [60] such that

\[
s(n) \approx a_1 s(n - 1) + a_2 s(n - 2) + \ldots + a_p s(n - p) \quad \ldots (4.6)
\]

Where the coefficients \( a_1, a_2, \ldots, a_p \) are assumed to be constants over the speech analysis frame. The equation (4.6) can be converted to an equality by including an excitation term \( G(u(n)) \),

\[
s(n) = G(u(n)) + \sum_{i=1}^{p} a_i s(n - i) \quad \ldots (4.7)
\]

Where, \( u(n) \) is normalized excitation and \( G \) is the gain of excitation.

Expressing equation 4.7 in Z domain we get the relation:

\[
S(z) = G(u(z)) + \sum_{i=1}^{p} a_i z^{-i} s(z) \quad \ldots (4.8)
\]

leading to the transfer function:

\[
H(z) = \frac{s(z)}{G(u(z))} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \quad \ldots (4.9)
\]
based on our knowledge that the actual excitation function for speech is essentially either voiced speech sounds or an unvoiced sound.

4.5.2 LPC Analysis

The relation between \( s(n) \) and \( u(n) \) is defined as (based on the speech production model Figure-4.1)

\[
s(n) = Gu + \sum_{i=1}^{p} a_k s(n) \quad \quad \ldots (4.10)
\]

Where, \( p \) = order of predictor (Assumed as 10 in the present study)

We consider the linear combination of past speech samples as the estimate \( \hat{s}(n) \) defined as,

\[
\hat{s}(n) = \sum_{k=1}^{p} a_k s(-k) \quad \quad \ldots (4.11)
\]

The predictor error, \( e(n) \), is defined as,

\[
e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n - k) \quad \ldots (4.12)
\]

And the error transfer function is,

\[
A(z) = \frac{E_z}{S_z}
\]
One of the main problems of linear prediction analysis is to determine the set of predictor coefficients $a_k$, directly from the speech signal so that the speech properties of the digital filter match those of the speech waveform within the analysis window.

To set up the equations that must be solved to determine the predictor coefficients, we define the short-term speech and error segments at time $n$ as,

$$s_n(m) = s(n + m) \quad \ldots \ldots (4.14)$$

$$e_n(m) = e(n + m) \quad \ldots \ldots (4.15)$$

and tried to minimize the mean square error signal at time $n$,

$$E_n = e_n^2(m) \quad \ldots \ldots (4.16)$$

using equation (4.14) & (4.15) we can write

$$E_n = \left[ \sum_m s_n(m) - \sum_{l=1}^{p} a_k s_n(m - k) \right]^2 \quad \ldots \ldots (4.17)$$

Where $s_n(m)$ is the segment of speech selected from the neighborhood of a sample $s_n(m) = s(m + n)$. The value of the coefficient $a_k$ that minimize the error $E_n$, can be obtained by considering,

$$\frac{\partial E_n}{\partial a_k} = 0, \quad k = 1, 2, 3 \ldots p \quad \ldots \ldots (4.18)$$
Giving,

\[ \sum_m s_n(m-i)s_n(m) \]
\[ = \sum_{i=1}^{p} a'_{k} \sum_m s_n(m-i)s_n(m-k), 1 \leq i \leq p \quad \ldots \quad \ldots \quad (4.19) \]

where, \( a'_{k} \) are the values of \( a_{k} \) that minimizes \( E_n \).

The terms \( \sum_m s_n(m-i)s_n(m-k) \) are related to the short term covariance of \( s_n(m) \) i.e.,

\[ \phi(i, k) = \sum_m s_n(m-i)s_n(m-k) \quad \ldots \quad (4.20) \]

which can be expressed in compact notation as,

\[ \phi_n(i, 0) = \sum_{k=1}^{p} a_k \phi_n(i, k) \quad \ldots \quad (4.21) \]

which describe a set of \( p \) equations. It is readily shown that the minimum mean-square error, \( E_n \), can be defined as:

\[ E_n = \sum_m s_n^2(m) - \sum_{i=1}^{p} a_k \sum_m s_n(m)s_n(m-k) \quad \ldots \quad (4.22) \]

Thus the compact form of the above equation is

\[ E_n = \phi_n(0, 0) - \sum_{k=1}^{p} a_k \phi_n(0, k) \quad \ldots \quad (4.23) \]

Thus the minimum **Mean-Squared Error (MSE)** consists of a fixed term \( \phi_n(0,0) \) and is dependent on the predictor coefficients. To solve Equation (4.21)
for the optimum coefficients $\tilde{\alpha}_k$, we have to compute $\varphi_n(i,k)$, for $1 \leq i \leq p$ and $0 \leq k \leq p$, and then solve the resulting set of $p$ simultaneous equations. A method to solve these equations and compute the coefficients is the autocorrelation method.

4.5.3 The Autocorrelation Method

In this method, the segment $s_n(m) = 0$ are considered outside the interval $1 \leq m \leq N$, and $s_n$ is describe by the equation

$$s_n(m) = s_n(m + n)w(m)$$

Where, $w(m)$, is identically zero outside the rang $0 \leq m \leq N$, Thus, the speech sample for minimization can be expressed as

$$s_n(m) = \begin{cases} s(m + n), & 0 \leq m \leq N - 1 \\ 0 & \text{otherwise} \end{cases}$$

.....(4.24)

Thus, if $s_n(m)$ differs from zero to 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 can be expressed as

$$E_n = \sum_{m=0}^{N+p} e_n^2(m)$$

.....(4.25)

and $\varphi_n(i,k)$ can be expressed as,

$$\varphi_n(i,k) = \sum_{m=0}^{N+p} s_n(m - i)s_n(m - k)$$

.....(4.26)
Where, \( 1 \leq i \leq p \) \text{ and } \( 0 \leq k \leq p \)

Or

\[
\phi_n(i, k) = \sum_{m=0}^{N-(i-k)} s_n(m)s_n(m + i - k) \quad \ldots \quad (4.27)
\]

where \( 1 \leq i \leq p \) \text{ and } \( 0 \leq k \leq p \)

since equation (4.27) is only a function of \( (i - k) \), rather than the two independent variables \( i \) and \( k \), the covariance function, \( \phi_n(i, k) \), reduces to the simple autocorrelation function,

\[
\phi_n(i, k) = r_n(i - k)
\]

\[
= \sum_{m=0}^{N-(i-k)} s_n(m)s_n(m + i - k) \quad \ldots \quad (4.28)
\]

Since the autocorrelation function is symmetric, i.e. \( r_n(-k) = r_n(k) \), the LPC equations can be expressed as

\[
\sum_{k=1}^{p} r(|i - k|)\tilde{a}_k = r_n(i), \quad \text{for } 1 \leq i \leq p \quad \ldots \quad (4.29)
\]

and can be expressed in the matrix form, known as the Yule-Walker equations [7] as follows:
The \((p \times p)\) matrix of autocorrelation value is a **Toeplitz matrix** (all diagonal elements are equal) and hence can be solved efficiently through several methods. One of the numerical methods is called the **Levinson-Durbin algorithm**.

### 4.5.4 The Levinson-Durbin Algorithm

A numerical solution of \(P\) equations with \(P\) unknowns, as needed in solving the Yule Walker equations, requires about \(P^3\) multiply-add operations [61]. The Levinson-Durbin algorithm solves the Yule-Walker equations in approximately \(P^2\) multiply-add operations. The algorithm is as follows:

1. Define \(S_0 = r(0)\)
2. For \(i = 0, 1, 2, \ldots, p - 1\) then
   
   \[
   \rho_{i+1} = \frac{r(i + 1) + \sum_{k=1}^{i} a_{i,k} r(i + 1 - k)}{s_i}
   \]
   
   a. \(s_i (i + 1) = s_i (1 - \rho_i (i + 1))^2\)
   
   b. \(a_{i+1,k} = a_{i,k} - \rho_{i+1} a_{i,i+1-k} \quad 1 \leq k \leq i\)
   
   c. \(a_{i+1,i+1} = -\rho_{i+1}\)

The coefficient \(a_{i+1,i+1}\) recursively computed by this algorithm, correspond to the LPC coefficients, \(a_k\), in the equation (4.10)
4.5.5 The LPC- Cepstral Coefficient

In the present study, LPC-based cepstral coefficients and phonetically important parameters are used as feature vectors. Cepstral weighted feature vector is obtained for each frame by block processing of continuous speech signals. The analog speech waveform is then sampled and quantized. To spectrally flatten the signal, the speech signal has been subjected to the pre-emphasis procedure through a first order digital filter whose transfer function has been given by

$$H(z) = 1 - az^{-1}, \text{for } 0 \leq a \leq 1.0 \quad \ldots \ldots (4.31)$$

Consecutive speech signal are taken as a single frame. To reduce the undesired effect of Gibbs phenomenon, the frames are multiplied by a windows function (Hamming window), which is given by [62],

$$w(n) = 0.54 - 0.46 \cos \left( \frac{2\pi n}{N - 1} \right), \text{for } 0 \leq n \leq N, \quad \ldots \ldots (4.32)$$

where $N$ is the number of sample in a block. Now, each frame of the windowed signal is next auto correlated to give

$$r_f(m) = \sum_{n=0}^{N-m} \tilde{x}_f(n)\tilde{x}_f(n + m), \text{ for } m = 0, 1, 2, \ldots p \quad \ldots (4.33)$$

Where the highest auto correlated value, $p$, is the order of the LPC analysis.
4.5.6 LPC Parameter Conversion to Cepstral Coefficient

The LPC cepstral coefficients, which are a set of values that have been found to be more robust, reliable feature set for speech recognition than the LPC coefficients. These coefficients are obtained recursively as follows,

\[ c = \ln(\sigma^2), \text{where } \sigma^2 \text{ is the gain term in the LPC model.} \]

\[ c_m = a_m + \sum_{k=1}^{p} \left( \frac{k}{m} \right) c_{m-k}a_k, \text{ for } 1 \leq m \leq p, \quad \ldots (4.34) \]

\[ c_m = \sum_{k=1}^{p} \left( \frac{k}{m} \right) c_{m-k}a_k, \text{ for } m > p, \quad \ldots (4.35) \]

Equation (4.35) shows the computation of cepstral coefficients \( C_{p+1}, C_{p+2}, \ldots, C_p \). Generally, \( q > p \) is taken for cepstral representation.

Figure-4.2 Block diagram of an LPC front-end processor
4.6 Determination of Mel-Frequency Cepstral Co-efficient (MFCC)

The mel-cepstrum is a widely used parameter for speech recognition [55]. There are several methods that have been used to obtain the Mel-Frequency Cepstral Coefficient (MFCC). In the present study, The MFCCs i.e. frequency transformed cepstral coefficients have been calculated from the LPC coefficient using recursion formulae. The LPC coefficients are first transformed to the cepstral-coefficients, and the cepstral coefficients transformed to the mel-frequency cepstral coefficients by using the recursion formula [73].

In the speech processing technology, The Mel-Frequency Capstrum (MFC), is referred as the representation of short term power spectrum of a sound signal, based on a linear cosine transform of a log power spectrum on a non linear mel- scale of frequency. Mel-Frequency Cepstral Co-efficient (MFCCs) are coefficient, which collectively make up an MFC. They are derived from a type of cepstral representation of a speech sound. The frequency band are equally spaced on the mel scale in the MFC that approximates the human auditory system’s response more closely than the linearly spaced frequency bands used in the normal spectrum. Such frequency wrapping can allow for better representation of sound.

MFCC’s are based on the known variation of the human ear’s critical bandwidths with frequency. The two types of filters used in the MFCC technique are linearly spaced filters and logarithmically spaced filters. The speech signal is expressed in the Mel frequency scale for determining the phonetically important characteristics of speech. The scale has a linear frequency spacing between 1000Hz
and a logarithmic spacing above 1000Hz. The normal speech waveform may vary from time to time depending on the physical condition of the speaker's vocal cord. Rather than the speech waveforms themselves, MFCCs are less susceptible to the said variations [8][79]. The following formula is used to compute the mels for a given frequency $f$, in Hz [84]:

$$\text{mel}(f) = 2595 \times \log_{10}(1 + \frac{f}{700})$$

.... (4.36)

The mel cepstrum is a useful and widely used parameter for speech recognition[8]. There are several methods that have been used to obtain Mel-Frequency Cepstral Co-efficients. One of the common methods of deriving MFCCs are as follows [78]:

1. Take the Fourier transform of a signal.
2. Map the powers of the spectrum obtained above into the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete Cosine Transform (DCT) of the list of mel log powers, as if, it were a signal.
5. The MFCC's are the amplitudes of the resulting spectrum.

As the mel spectrum coefficients are real numbers, they may be converted to the time domain using the Discrete Cosine Transform (DTC). The MFCCs is calculated using the equation (4.37) [84, 73]:

66
\[
C_n = \sum_{n=1}^{K} (\log S_K) \left[ n \left( K - \frac{1}{2} \frac{\pi}{K} \right) \right]
\]

where \( n = 1, 2, ... K \)

Here \( K \) represents the number of mel cepstrum coefficient, which has been taken as 12 in the present study. \( C_0 \) is excluded from the DCT as it represents the mean value of the input signal which carries little speaker specific information. For each speech frame of about 20ms with overlap, a set of mel-frequency cepstrum coefficients is computed. This set of coefficients is called an acoustic vector. To represent and recognize the speech characteristic of the speaker.

In the present study, by using the equation (4.37), 32 mel-frequency cepstral coefficients have been calculated for the utterances of the Bodo vowels corresponding to the male and female informants.

In the present study, by using the equation (4.37), 20 Mel-Frequency Cepstral Co-efficient (MFCC) have been calculated for the utterances of Bodo vowels and words of CV, VC, CVC type.
Figure 4.4: MFCC comparison of Male and Female in vowel /e/
Figure 4.5: MFCC comparison of Male and Female in CV type word

Bu-HH

Bu-LL

69
Figure 4.6: MFCC comparison of Male and Female in CVC type word
Figure 4.7: MFCC comparison of Male and Female in VC type word

ER-HH

ER-LL
The plot of the MFCCs corresponding to the Male and Female utterances of Bodo vowel \( \text{e} \) is depicted in figure (4.4), and the plot of MFCCs corresponding to the Male and Female utterance of CV, VC, and CVC type of word is depicted in the figure (4.5), (4.6), and (4.7) respectively.

4.7 Observation

In the present study, the Cepstral analysis of Bodo vowels and words of type CV, VC, CVC is carried out with respect to Mel-Frequency Cepstral Co-efficient in case of all vowels shows a remarkable distinction between Male and Female informants. So, MFCC can be considered as a better practice for speaker and speech verification in Bodo language.

Similarly in case of CV type of words, it has been observed that in case of high tone, there is a remarkable change in MFCC, with respect to Male and Female informants, where as in case of lone tone, there is not much difference in MFCC, with respect to Male and Female informants. Similar characteristics were also observed in case of VC type of words also.

In case of CVC type of words, it is observed that the MFCC are almost constant in case of Male informants where as there is slight differences observed in case of Female informants.