CHAPTER 8

BODO AND RABHA WORD RECOGNITION

8.1 INTRODUCTION

Recognition of the word(s) after word segmentation is an important task. The Bodo and Rabha words which are segmented through the boundary detection technique (Chapter-7) are now used as input to a recognizer which detects the presence of a key word in the sentence. The recognizer system is trained to recognize a set of keywords, i.e. a vocabulary, based on application. The recognizer is trained for each speaker separately. As the input word fed to the recognizer, the system decides, whether the input word is a keyword or not, which ultimately amounts to word recognition task.

8.2 FEATURE EXTRACTION

The purpose of feature extraction in speech recognition is to transform speech signal into a set of vectors, which are essential for speech recognition while discarding unreliable parts. The transfer vector is called feature vector. In the present study, Linear Predictive Coding (LPC), based on Cepstral Coefficients and phonetically important parameters are used as feature vectors. The input segmented word is first digitized at a 22.05KHz sampling rate with 16 bit resolution. It is then pre emphasized using a first order digital filter (Hamming window) with a pre emphasis factor $a=0.96$. The signal
is then blocked into 50 frames and each frame is multiplied by a Hamming window. A 10th order Linear Predictive Coding (LPC) analysis is performed on the data. 20 cepstral coefficients are generated from each frame.

8.3 CLUSTERING AND TRAINING VECTORS

To reduce the volume of the input data, the clustering technique has been used. The main concerned in the clustering process is to organize the data into sensible groups. Clustering makes it possible to look at properties of whole clusters instead of individual objects. The concept mostly utilizes geometric principles, where the samples are interpreted as points in a d-dimensional Euclidian space, and clustering is made according to the distances between points (usually, points which are close to each other will be allocated to the same cluster. Clustering algorithms are based on some initial assumptions in order to define partitioning of a data set. Most methods behave differently depending on the features of the data set and the initial assumptions for defining groups. Therefore, in most applications, the resulting clustering scheme requires some sort of evaluation regarding its validity. Evaluating and assessing the results of a clustering algorithm is the main object of cluster validity (Halkidi, M et al., 2002). The acoustic vectors extracted from input speech of each speaker provide a set of training vectors for that speaker so as to build a speaker-specific VQ codebook for each speaker using those training vectors. There is a well-know algorithm, viz K-means algorithm, for clustering a set of L training
vectors into a set of \( M \) codebook vectors. The algorithm is formally implemented by the following recursive procedure: (Rabiner & Jung, 1993)

Figure-8.1: Flow diagram of binary split codebook.

\[
\frac{D' - D}{D} < \varepsilon
\]
1) Design initially 1(first)-vector codebook; this is the centroid of the entire set of training vectors (hence, no iteration is required here).

2) Double the size of the codebook by splitting each current codebook \( y_n \) according to the rule

\[
\begin{align*}
y^+_n &= y_n(1 + \varepsilon) \\
y^-_n &= y_n(1 - \varepsilon)
\end{align*}
\]

where \( n \) varies from 1 to the current size of the codebook, and \( \varepsilon \) is a splitting parameter ( \( \varepsilon = 0.01 \) is chosen).

3) **Nearest-Neighbor Search**: for each training vector, find the codeword in the current codebook that is closest, and assign that vector to the corresponding cell by using K-means algorithm.

4) **Centroid Update**: update the codeword in each cell using the centroid of the training vectors assigned to that cell.

5) **Iteration I**: repeat steps 3 and 4 until the average distance falls below a preset threshold.

6) **Iteration II**: repeat steps 2, 3 and 4 until a codebook size of \( M \) is designed.

### 8.3.1 K-MEANS ALGORITHM

Among the different levels and techniques used in speech recognition tasks, the clustering methods have traditionally been used in order to find emerging patterns from data sets with unknown properties. In the present study, K-means clustering algorithms (Duda, R. O., Hart, P. E. and Stork, D. G, 2000) have been used and tested for both the Bodo and Rabha words. K-means clustering is an elementary but very popular approximate method that can
be used to simplify and accelerate the convergence. The K-means algorithm partitions or group objects based on their attributes (features) into K-number of groups, where K is the positive integer. The algorithm first chooses K cluster-centroids among the T feature vectors. Then, each feature vector is assigned to the nearest centroid, and the new centroids are calculated. This procedure is continued until a stopping criterion is met, that is, the mean square error between the feature vectors and the cluster centroids is below a certain threshold or there is no more change in the Cluster-center assignment. The algorithm is stated as below:

i) Set the number of clusters (K);
ii) Initialize cluster centroids ((\mu_1, \mu_2, \ldots, \mu_k);
iii) Classify the samples according to the nearest \mu_k;
iv) Recompute (\mu_1, \mu_2, \ldots, \mu_k) until there is no significant change.

8.3.2 SPEAKER DATABASE

In any speaker or speech recognition task, the creation of speech data base is a must. Thus, the first step in the speaker/speech recognition task is to build a speaker-database, C_{database} = \{C_1, C_2, \ldots, C_N\} consisting of N codebooks, one for each speaker in the database. This is done by first converting the raw input signal into a sequence of feature vectors X={x_1, \ldots, x_T}. These feature vectors are clustered into a set of M code words C={c_1, c_2, \ldots, c_M}. The set of code words is called a codebook. The clustering is done following a suitable clustering-algorithm.
8.3.3 SPEAKER MATCHING

In the recognition phase, an unknown speaker, represented by a sequence of feature vectors \( \{x_1, \ldots, x_T\} \), is compared with the database in the codebook. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion (Molgard and Jorgensen 2005) is identified as the speaker of the input utterance, as given in equation (8.2)

\[
s(X, C_j) = \frac{1}{T} \sum_{t=1}^{T} d(x_t, C_{\text{min}}^{j,t})
\]

(8.2)

Where, \( C_{\text{min}}^{j,t} \) denotes the nearest codeword \( x_t \) in the codebook \( C_j \) and \( d(.) \) is the Euclidean distance. Thus, each feature vector in the sequence \( X \) is compared with all the codebooks, and the codebook with the minimized average distance is chosen to be the best for the recognition task.

8.4 PERFORMANCE MEASUREMENT

To measure the effectiveness of the procedure used, the performance is measured as follows.

Efficiency = \( \frac{\text{No. of words correctly detected}}{\text{No. of words present in the test data file}} \) \hspace{1cm} (8.3)
8.5 RESULT AND CONCLUSION

A set of 50 sentences for Bodo and 50 sentences for Rabha languages were used for the word reorganization purpose. Each of the sentences were segmented and the words were given as input to the recognizer. The result have been shown in the Table-(8.1 & 8.2)

Table-8.1 Result of Bodo word recognition.

<table>
<thead>
<tr>
<th>Input word structure</th>
<th>Efficiency( in percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>84</td>
</tr>
<tr>
<td>CV</td>
<td>73</td>
</tr>
<tr>
<td>VC</td>
<td>82</td>
</tr>
<tr>
<td>CVC</td>
<td>72</td>
</tr>
</tbody>
</table>

Table-8.2 Result of Rabha word recognition.

<table>
<thead>
<tr>
<th>Input word structure</th>
<th>Efficiency( in percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>81</td>
</tr>
<tr>
<td>CV</td>
<td>69</td>
</tr>
<tr>
<td>VC</td>
<td>79</td>
</tr>
<tr>
<td>CVC</td>
<td>66</td>
</tr>
</tbody>
</table>

In the present investigation, it is found that the expected error of the recognition process go high when the word was started with a structure like CV, CVC etc. i.e. when the word started with a consonant, if the consonant phonemes were of the type of plosive, the word recognition was found to be less efficient. In case of Rabha word recognition, the efficiency of recognition further decreases in comparison to the Bodo words. It is found that due to the absence of the supra segmental feature of the Rabha words i.e. tone, the efficiency of recognition was affected.
CHAPTER 9

SUMMARY AND SCOPE FOR FUTURE WORK

9.1 SUMMARY OF WORK DONE

The north-eastern part of India comprises of the eight states—Assam, Arunachal Pradesh, Manipur, Meghalaya, Mizoram, Nagaland, Tripura and Sikkim. Among these eight states, the state of Assam has the highest population, which is 2.59% of the total population of India. The linguistic scenario of the north-eastern region of India is a mixture of Indo-Aryan and Sino-Tibetan family of languages. In Assam, Assamese, Bodo, Rabha, Hajong, Mech, Lalung, Tipra, Chutia and Moran etc. are the different linguistic components. Among these, Assamese is the main lingua franca (Link language). Bodo and Rabha are the next link languages, both population wise and area wise. The indigenous people living through-out the region represents a mixed tradition—both linguistically and culturally. Among these three major link languages of Assam, the Bodo language and the Rabha language are yet to be studied on proper scientific platform and so as to develop an Information Technology (IT) based transparent role models for these languages. To develop such a model, it is imperative to have a detail of study the characteristic features of these two languages as far as their—phonological, morphological, syntactic and semantic aspects are concerned.
Among the nine chapters embodied in this thesis –

In the chapter 1 outlines the introduction and advancement of speech researches, and the objective and methodology of the present study.

In chapter 2 the origin and history of Bodo and Rabha languages have been discussed along with the phonetic characteristics of both the Bodo and Rabha languages.

In chapter 3 the anatomy of human speech production system is described.

Chapter 4 deals with the measure of formant frequency characteristics of different Bodo and Rabha phonemes. It is observed in the present study that the Cepstral coefficients seem to play an important role in identifying the sex of the speaker with respect to Bodo and Rabha language. Specifically, the frame wise analysis of the cepstral coefficients of the vowels /i/ and /e/ show clear distinction of male and female speaker in few selected frames. This is a common finding for both the languages. The analysis of the formant characteristics (F1, F2 and F3) of vowels and some selected word sets show that when vowels are at the nucleus of a syllable or word, the formant frequencies of the word shifted to a lower value. It is observed that the second formant frequency (F2) increases significantly in case of a syllable structure like CV, VC, CVC etc. with respect to the vowels /o/, /a/, /u/ and /w/. But, in case /i/ or /e/ the F2 frequency decreases in terms of amplitude. Thus, /i/ and /e/ vowels, when at the nucleus could be used more significantly for word recognition.
In Chapter 5, the intonation pattern of the Bodo and Rabha sentences are carefully observed. The variation of the pitch (Fo) frequency for vowels is found to be rising for both Bodo and Rabha vowels. But, the utterance of the vowel /w/ in case of Bodo is distinctively different with respect to other vowels of Bodo and Rabha language irrespective of sex. Thus, the vowel /w/ could be effectively used for relative identification of native Rabha & Bodo speakers. Being a tonal language the overall intonation pattern of the Bodo sentences show a remarkable difference of pitch frequency at the beginning and at the end of a sentence. In case of Rabha sentence the variation of pitch value across a sentence is found nominal. This may account that Rabha language have gradually lost its tonal character (originally it was a tonal one) with respect to the socio-cultural environment. But, it is found that stress on phonemes and a general intonational tone still play an important role in diversifying the meaning of the same sentence.

In Chapter 6 an introduction to the prosody variation of the Bodo and Rabha languages has been made. An intense observation has been made to observe, how the meaning of the same sentence changes with the change of prosody of the sentence. The parameter used for the purpose is the pitch frequency variation. The prominent pitch pattern variation of the same sentence have been noticed and graphically depicted in this chapter.

In the Chapter 7 of the thesis, some issues related to developing a language-independent word boundary detection technique have been investigated. The pitch variation across a word in a sentence have been studied and the differences of the last pitch value of a word and the first pitch value of
next word of a sentence, have been applied for word segmentation. Although the word segmentation technique used in the present study does not furnish with cent percent accurate results, yet it has been able to provide us with at least 70% accuracy for both Bodo and Rabha languages.

In Chapter 8, recognition of the segmented word has been done. The LPC coefficients that have been extracted from the selected word sets are subjected to the vector quantization (VQ) and K-means algorithm for pattern matching followed by recognition have been used. The efficiency of recognition is found approximately 80% in case of Bodo and 73% for Rabha language.

9.2 SCOPE FOR FUTURE WORK

Knowledge is unlimited. There is enough scope for extending the present study. The present study shows the urgent necessity of further in-depth investigation of the two major languages of this region. Taking the present study as base, new ideas, techniques are expected to be incorporated to further strengthen the base for the development of languages of this region. It is believed that a role model would enhance the connectivity and integrity of the people of this region, encompassing all the social and cultural issues, which would definitely change the linguistic scenario of this region. To achieve all these objectives, a common platform to develop such an integrated transparent role-model of the different ethnic languages of India and North-East in general is a must. The present aim considered is only an attempt which is to be realized more broadly considering all the fundamental issues of any language concern.