# Connected Digit Recognition System in Bodo Language

[1]Aniruddha Deka, [2] Dr. Manoj Kumar Deka
[1]Reserach Scholar, [2]Assistant Professor
Department of Computer Science and Technology
[1]Bodoland University, Assam, India

***Abstract*— Bodo language is a tonal as well as Semitic language belonging to North-East of India and has a limited number of speakers in Assam. It comes under the Assam-Burmese group of languages. It is said to have branched off from the Tibeto-Burman family of languages and is spoken by Bodo people of Nepal and Bangladesh, apart from North-eastern India. Infect, the Bodo language happens to be amongst the official languages of the Indian state of Assam and is one of the 22 languages recognized by the eighth schedule of the Indian constitution. But in the field of speech processing a limited number of work has been done till date. Some work has been done in the field of isolated word. Connected word recognition is one area where no work has been done so far for Bodo language. In this paper, an effort has been made to build automatic speech recognizer to recognize Bodo digit from 0-9 by using acoustic modeling approach on HTK 3. 4. 1 speech engine. We have collected Bodo digit from 10 different speakers with all most all possible variation. In order to achieve all the possible tones for all the digits a permutation has been performed with the digits from 0 to 9 and then they are recorded. Overall recognition accuracy has been found to be 82. 12% at connected word model**

***Index Terms*—Bodo Digit, IPA Symbol, HMM, HTK**

---

## I. INTRODUCTION

The ability to listen spoken words, identify various sounds and recognize them as words of some known language present in it is known as speech recognition. In the field of computer it is defined as the ability of computer systems to accept spoken words in audio format - such as wav or raw - and then generate its content in text format. [1] This speech recognition comprise of two types. Isolated word recognition and connected word recognition. In this paper we mainly focus on to build a connected word recognizer.

Typically there are two-stages in speech recognition. It initially determines the phonemes location and their waveform characteristics using feature extraction process. Next it uses pattern recognition techniques to identify the phonemes, and maps these phonemes into words. In connected word based speech recognition system, the continuous speech signal is partition into equally spaced units of 10 to 20 msec, called frames. Then system estimates the pitch period and formant frequencies for each frame. The power spectrum of speech signal makes it easier to identify the locations of vowels, consonants, and noise. The final result of the feature extraction is a feature vector, which gives a set of 15 to 20 numbers that best represent a frame. The spectral characteristics such as formant frequencies and pitch period determined during the process of feature extraction helps to provide other clues to the location of boundaries. The phoneme recognition stage classifies each segment as a particular phoneme. In the final stage of recognition maps the phonetic sequence into words. In this stage phonetic dictionary, listing the phonetic spelling of all words that the speech engine is designed to understand, and a language model, listing the probabilities of specific sequences of words are required. [2]

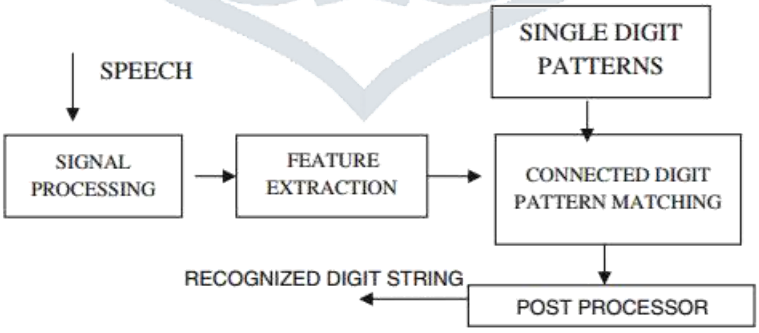


Fig 1: Block diagram of Connected Digit recognition Method

This paper represents an overview of our approach. Section two explains different research on speech recognition in Indian languages so far. In section three we give an approach to design a connected digit recognizer. In section four we evaluate the system performance and finally we conclude with some results and some observations.

## II. BACKGROUND STUDY

### *A. Speech recognition in other Indian languages*

We have studied different speech recognition system in following Indian languages. [3]

**Hindi language**:N. Rajput M. Kumar and A. Verma of IBM research lab, has developed a large-vocabulary continuous speech recognition system for Hindi. They developed a Hindi Speech Recognition system which has been trained on 40 hours of audio data that is trained with 3 million words. For a vocabulary size of 65000 words, the system gives a word accuracy of 75% .

**Telugu language:** Sunitha .K.V.N and Kalyani .N have built a speech recognition system that uses syllable as the basic unit. For training they have used 300 words and for testing they recorded 100 new words and 80% of the words were recognized correctly.

**Tamil language:** Vimala C. and V. Radha presented a speaker independent isolated speech recognition system for Tamil language. The experiments furnish high-quality word accuracy of 88% for trained and test utterances spoken by the speakers.

**Malayalam language:** A small vocabulary speech recognizer has been developed by Anuj Mohamed and K.N Ramachandran Nair using Hidden Markov Models. The system has produced 94.67% word accuracy.

**Punjabi language:** Ravinder et al. developed a speaker dependent real time, Isolated and connected word recognition systems for Punjabi language using acoustic template matching technique. It was designed for medium sized dictionary Vector quantization was used to transform signal parameters to codebook indices and Dynamic time warping techniques was used for finding the lowest distance path through the matrix, with some modification to noise and word detection algorithms. Accuracy of this ASR was just 61%.

**Assamese language:** In 2010, M. P. Sarma and K.K Sarma worked for the development of numeral speech recognition system for Assamese language. Gender and mood variations were given consideration during the recording of speech signals of 10 numeral digits at 8 KHz in mono channel mode. In 2011, M. P. Sarma and K.K Sarma have proposed the design of an optimal feature extraction block and ANN based architecture for speech recognition.

**Bengali language:** Neural network approach has been proposed by M. R. Hassan for Bengali phoneme recognition. A Bengali speech recognizer is built by training the HTK toolkit that can recognize any word in the dictionary. After acoustic analysis of speech signal, the words are recognized. Technically this work presents training the toolkit and builds a segmented speech recognizer of Bengali.

**Oriya Language:** Mohanty and Swain have made such effort for Oriya language. Mohanty and Swain have come forward to apply the benefit of automatic speech recognition systems to society by developing an isolated speech recognizer for Oriya language.

**Gujrati language:** A technique for fast bootstrapping of initial phone models of a Gujarati language is presented by Himangshu N. Patel. The training data for the Gujarati language is aligned using an existing speech recognition engine for English language. This aligned data is used to obtain the initial acoustic models for the phones of the Gujarati language. Speech recognition of Gujarati Language is presented by Patel Pravin and Harikrishna Jethva . Neural network was used for developing the system.

### *B. Speech recognition in Bodo language*

Bodo is one of the major Tibeto-Burman tone languages, spoken in many parts of the North-Eastern states of India as well as in parts of West Bengal and Nepal. The fact that it is now included in the eighth schedule of the Indian constitution, demonstrates its socio-political importance. According to recent Indian government census Bodo is spoken by almost a million speakers. There are no records indicating the origin of Bodo language. However, it is known to be a branch of the Sino-Tibetan family of language. A highpoint in the history of the Bodo language is the socio-political movement that was launched by local Bodo organizations, from 1913 onwards. Before 1953, the Bodo language had no standard form of writing. [4] Although, Roman script and Assamese script were used in the past, recently, Bodos adopted the Devanagiri script. According to some scholars, the Bodo language had a script of its own called 'Deodhai'. Bodo is one among the tonal languages of the world. There are two clearly distinguishable kinds of tone in Bodo; these are Low and High.

M.K.Deka has proposed an approach for Speech Recognition using LPCC (Linear Predictive Cepstral Coefficient) and MLP (Multilayer Perceptron) based Artificial Neural Network with respect to Assamese and Bodo Language [3]. A new simplified approach has been made for the design and implementation of a noise robust speech recognition using Multilayer Perceptron (MLP) based Artificial Neural Network and LPC-Cepstral Coefficient. Cepstral matrices obtained via Linear Prediction Coefficient are chosen as the eligible features. Here, MLP neural network based transformation method is studied for environmental mismatch compensation.

Utpal Bhattacharjee investigates the problems faced by tonal languages like Bodo during recognition process. The performance of speech recognition system degrades considerably when the recognizers are used to recognize the tonal words. Two approaches have been investigated in this paper for this purpose. In the first approach attempt has been made to develop a feature level solution to the problem of tonal word recognition. In the second approach, a model level solution has been suggested. Experiments were carried out to find the relative merits and demerits of both the methods. [5]

Following table shows the IPA symbol with Bodo phone set.

| **Bodo** | अ | आ | इ | उ | ए | ओ | ख | ग | ङ | ज | थ | द | न | फ | ब | म | य | र | ऱ | स | ह | व |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Label** | A | Aa | I | U | E | O | Kh | G | Ng | J | Th | D | N | Ph | B | M | Y | R | L | S | H | W |
| **IPA** | /a/ | /a:/ | /ɪ/,/i/ | /u/, /ʊ/ | /e/ | /o/ | /$k^h$/ | /g/ | /ŋ/ | /dʒ/ | /$t^h$/ | /d/ | /n/ | /$p^h$/ | /b/ | /m/ | / j/ | /r/ | /l/ | /s/ | / ɦ/ | /y/ |

Table 1: IPA symbol with Bodo phone set

Following table shows the Bodo digit:

| English | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Bodo | Lathikho | Nai | Se | Tham | Broi | Ba | Da | Sni | Dian | Gu |

Table 2: Bodo Digit pronunciation

## III. DESIGN APPROACH OF THE SYSTEM

In this section, the experimental work on digit recognition from a connected word speech corpus for Bodo language will be presented. Wave surfer is used for data recording, praat is used for data analysis, Mat lab is used for preprocessing and Hidden Markov Toolkit (HTK) is used for feature extraction, training and recognition steps.

***A. Data Collection:*** We have collected Bodo digit from 10 Bodo people out of which 5 recordings are of male and five are of female speakers. For most of the Digit recognition systems, the data recorded are of isolated digits i.e. the speaker speaks a digit and pause and again he/she speaks the next digit and takes a pause and so on. But in our method, in order to train the system more appropriately, we have used the permutation to generate random numbers from 0 to 9. Through a shell script all the numbers are shuffled and finally we get a sequence of 10 numbers from 0 to 9. This way a total of 1000 lines have been generated and out of which first 100 lines are used to record for the $1^{st}$ person, next 100 lines are used to record for the $2^{nd}$ person and so on and finally we have collected 1000 lines or words for 10 people (10*100 lines). In our method, $1^{st}$ 10 digits are recorded in continuous mode and then there is a pause. Again the speaker speaks the next 10 digits in continuous mode and so on. Now since a digit, spoken after another digit in a continuous mode might have various possible pitches, so this way we are able to get all the possible variation of the speech for a particular digit when spoken after another digit without a pause. But we make it sure that after recording we have to get the equal occurrence of each digit.

The implemented system is trained for 10 distinct Bodo language digits which have been used in various combination modes. The data is recorded with the help of a unidirectional microphone using a recording tool wave surfer in .wav format. The .wav files recorded are saved as HTK transcription. The sampling rate used for recording is 16 kHz. Speakers recorded the data and each speaker recorded total of 100 lines of digits in which each line contains all the numbers from 0 to 9. So the 100 distinct set of numbers resulted in 100 samples of 4 distinct male speakers and 4 distinct female speakers, files making a total of 800 (100*8) files. A labeling tool wave surfer is used to label the speech waveforms. The labeled files are used in acoustic model generation phase of the system.

***B. Acoustic Analysis***

The speech recognition tools cannot process directly on speech waveforms. These have to be represented in a more compact and efficient way. This step is called acoustical analysis. The original waveform is converted into a series of acoustical vectors. Mel Frequency Cepstral Coefficient (MFCC) technique has been used for feature extraction. [6]The purpose of feature extraction is to covert the speech waveform to some parametric representation. The computation steps of MFCC include:

a. Framing: The property of the speech signals change in every few millisecond, because of which speech should b analyzed in small duration frame. In this step the reemphasized speech signal is blocked into frames with a length of 25ms.
b. Windowing: It is an important factor to process signal correctly. There are two competing factors used to determine windowing. One factor smooth the discontinuity at the window boundary and other is used to not disturb the selected points of the waveform. In our approach we use hamming window, which shrinks the values of the signal toward zero at the window boundaries, avoiding discontinuities.
c. Extracting: It gives a compact representation of the s spectral properties of the frame. A vector of acoustical coefficients is extracted from each windowed frame.

### *C. Methodology for system design*

We propose our method in htk environment. We have used htk automated commend to evaluated our results.

Step1. Data preparation

The raw speech data is in the form of wave files. This needs to be converted to MFCC speech vectors. This is a form of spectral analysis of the raw waveform, and can be performed by using the tool HCopy. The MFC file can be viewed by using the HTK tool HList. A transcript is needed. This needs to be converted to Unicode, and then to label files. Label files are of the format

<Start> <end> <label>

Where start and end correspond to the beginning and end of the part of the waveform to which this label is being assigned. In case of word level it will be the whole file and thus 0 and -1.

In case of phone level recognition, a phonetic dictionary will be needed to convert the words into phones. Then the start and end values will be that of the various component phones.

Step2. Creating monophones

A prototype HMM model must be created, which must then be re-estimated using the data from the speech files. Silence models must be included. The prototype HMM model used is usually a 3 state left-to-right model with no skips. It has five states, but the first and last ones are dummy states which are used for continuity. The vector size used for the HMM models is usually 39, because it has been found to work well empirically. This prototype model is first initialized using HInit or HCompV and re-estimated by using HRest or HERest. We are using HInit and HRest to recognize the words. HInit initializes the HMM model based on one of the speech recordings. HRest is used to re estimate the parameters of the HMM model based on the other speech recordings in the training set

Step3. Creating Tied State triphones

The final step of model building is to create context dependant triphones from the monophones. The set of triphones is created by cloning the monophones and re-estimating. Similar acoustic states of these triphones are tied to ensure that all state distributions can be robustly estimated.

Step4. Recognition and evaluation

The test data can be recognized and the recognizer's performance can be evaluated. HVite and HResults are the HTK tools used for this.

## IV. PERFORMANCE ANALYSIS

We have evaluated different experiments and obtained an accuracy of 87.24 in training mode and 82.12 % in testing mode. In the experiment we have used 700 files to train the system and remaining 300 files are used to test the system.

| Mode | Word Accuracy | Number of deletion | Number of substitution | Number of Insertion |
|---|---|---|---|---|
| Training | 87.24 | 3 | 2 | 0 |
| Testing | 82.12 | 7 | 8 | 2 |

Table3: Performance evaluation of the system with training and testing data

## V. Conclusion

We have designed a digit recognizer in Bodo language for connected word and investigate the accuracy of the system using HTK. It is completely HMM based and by data is collected in normal environment. By considering the digit recognition analysis, the greatest accuracy was encountered in the case of digit 9, 1 and 2 while the least accuracy is encountered in the case of digit 6. The main cause for such a variation may be attributed to the tokens themselves; 6 is a monosyllabic Bodo digit which is pronounced as" /d/ द "which is short with lower amplitude than the other Bodo digits.

## REFERENCES


[1] Neema Mishra,Urmila Shrawankar,Dr. V. M Thakare ,"An overview of Hindi speech recognition", Proceedings of the International Conference, "Computational Systems and Communication Technology", 5th May 2010, Tamilnadu.

[2] Atif Zafar, J. Marc Overhage, and Clement J. McDonald, "Continuous Speech Recognition for Clinicians", journal of American medical informatics association.

[3] Cini Kurian, "A Survey on Speech Recogntion in Indian Languages", International Journal of Computer Science and Information Technologies, Vol. 5 (5), 2014, 6169-6175

[4] Priyankoo Sarma, "Some aspects of tonal phonology of Bodo", Ph. D. Dissertation: Central Institute of English and Foreign Languages, Center for Linguistics and Contemporary English, 2004

[5] Utpal Bhattacharjee, "Recognition of the Tonal Words of BODO Language", International Journal of Recent Technology And Engineering (IJRTE) ISSN: 2277-3878, Volume-1, Issue-6, January 2013114.

[6] http://htk.eng.cam.ac.uk/