International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

N-Gram Language Model based Continuous Voiced Odia Digit Recognition

Prithviraj Mohanty, Ajit Kumar Nayak

Abstract*: With the enormous improvement in the area of signal processing, speech processing systems are creating a massive impact in recognizing the voices, controlling the commands and making as communication interfaces. A continuous speech recognition system is essential for voice identification hands free system used as a voice dialer, voice originated security systems and voice based automatic electronic machines. The proposed work suggests a finest speaker independent continuous voiced digit recognition for Odia language. The model integrates the concept of Mel Frequency Cepstral Coefficient (MFCC) and continuous density Hidden Markov Model (HMM), relating to speech parameterization and recognition respectively. The performance of the model is explored for different levels of HMM like word-level and phoneme-level. Further the model output is evaluated using different N-Gram approaches of the language model. Finally it is shown that the model using phoneme-level HMM with a tri-gram language model is superior to other methodologies.

***Index Terms*: MFCC, HMM, Phoneme-level, N-Gram, Language Model**

I. INTRODUCTION

The utmost common and efficient method of communication between the people is their native speech. In Today's world, handling electronic appliances like computers, mobile phones, tablets and laptops through speech, is quite challenging and enigmatic. Speech technologies assist machines to react correctly and consistently to human voices and deliver useful and valuable services as and when needed. Interacting computers using voice is faster and easier rather than manually using input devices like: keyboard and mouse. So, people will prefer such type of voice operated systems. Communication between the human lives is conquered by spoken language, therefore it is expected for people to presume voice interface systems functioned in their natural languages [1].

Automatic speech-to-text (STT) conversion on a machine involves transformation of spoken utterances into its corresponding text. With the progressive increase in STT conversion technology, people with physical disabilities like blindness can easily controlled, simultaneously access the purposes along with applications of computers and mobile phones [2].The performance of such system depends upon how accurately it converts the spoken voices to its corresponding text. Thus, a reliable and robust STT system is always desirable for recognizing correct words spoken by the user. Depending upon what types of voices the speech recognition systems have the ability to recognize, it can be divided in to several classes [3]. Broadly it can be separated into two categories: word level recognition and phoneme level recognition.

Word level recognition: This level of recognition deals with recognizing each word independently. An HMM model is to be created for every individual word present in the training set which is to be constructed from the utterances of different speakers. A training set needs to be prepared, when a new word is added to the vocabulary [4]. These HMM models are to be trained competently. If the vocabulary contains less number of words, then this level of recognition is suitable. Word level recognition skill can be measured with two phases: training phase and testing phase. In both the phases, the speech is initially recorded then pre-processed (noise elimination) and parameterized wave form is generated by applying the feature extraction technique. In training phase, using training corpus and the language model, different HMM models are generated. In testing phase using the models and classifier generated in training phase, the equivalent texts are produced.Fig.1 depicts the architecture for word level recognition.

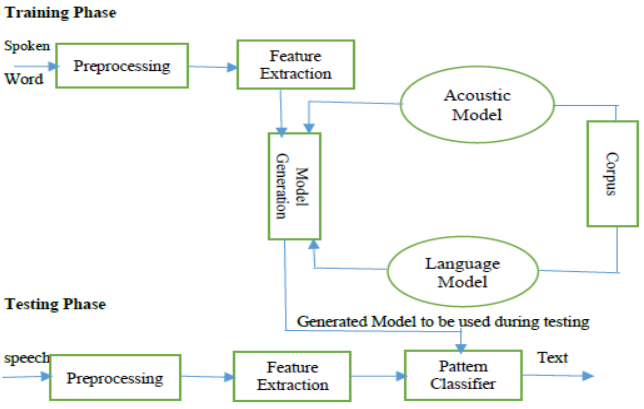


Fig.-1 Architecture for word level recognition

Revised Manuscript Received on 30 July 2019.
* Correspondence Author
Prithviraj Mohanty*, Department of CS&IT, ITER, S'O'A (Deemed to be) University, Bhubaneswar, India. **Email:** prithvirajmohanty@soa.ac.in
Ajit Kumar Nayak, Department of CS&IT, ITER, S'O'A (Deemed to be) University, Bhubaneswar, India. **Email:** ajitnayak@soa.ac.in

Phoneme level recognition: Phoneme is a discrete and distinctive unit of a language that can be used to differentiate between words [5].

Each individual word is segmented into phones which is to be recognized independently [4]. For example, the English digit FIVE is composed of three phones (FIVE=f +ay+ v). Similarly, the Odia digit ସାତ (sāta) =ସା+ତ (sa+ ta), is being made up of two phones. For each phone, there will be a corresponding HMM to recognize it. This type of model is more valuable for continuous speech recognition. Pronunciation of a single word may vary differently. So, a phonetic dictionary is always needed for this level of recognition.Fig.2 represents the architecture for phoneme level recognition.

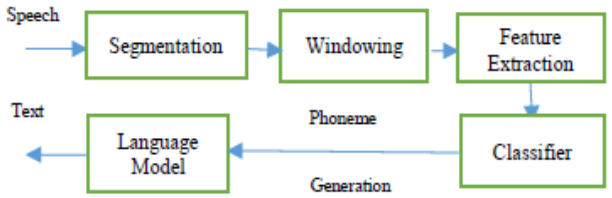


Fig.-2 Architecture for phoneme level recognition

The basic components of a statistical word recognition system are: Signal recoding and preprocessing, Acoustic Analysis, Acoustic Model, Pronunciation Model, Language Model and the Decoder [6]. First the voices are recorded using different recording environment and stored in *.wav* or any other supported format. In acoustical analysis, the physical aspects of spoken language is characterized by the technique of waveform analysis like: Fast Fourier Time (FFT) or Linear Predicting Coefficient (LPC) analysis, voice onset time (VOT) measurements, formant frequency measurements etc. [7]. An acoustic model grasps the data describing the acoustic feature of all the phonemes agreed by the system. The pronunciation dictionary always holds the set of words and their pronunciations using a common set of phonemes [8]. In the pronunciation dictionary, multiple entries can appear for a single word. It is the work of the linguist to verify the correct pronunciation of the words that should appear in the dictionary.The aim of language modelling is to approximate the probability distribution of several linguistic units, such as words, phrases and sentences. The decoder is the main component of the speech recognition system. It's work is to decode a sequence of speech signal to reveal what words are articulated [6],[9]. The main focus of the research work represented in this paper contains the different approaches of language model utilized to check the accuracy of the text generated.

Language Model: The naivest model that allocates probabilities to sequence of words and sentences are called language models. N-gram consists of N number of words, a 2-gram (bigram) is an arrangement of two consecutive words such as "What about", "please sign" or "ଭଲ ପିଲା (bhala pilā)", a 3-gram (or trigram) is a three word sequence of words like "What about you", "please sign it", or "ଭଲ ପିଲା ଅଟେ (bhala pilā aṭe)". The model, N-gram is used to compute the likelihood of the last word, provided with the sequences of n-1 words.

The application of language model is quiet necessary in the field of speech recognition, handwriting recognition, spelling correction, machine translation and augmentative communication [10]. The probability of a given word follows a sequence of words can be acquired from relative frequency count: considering a very large corpus , counting the number of times the same word follows the sequence of words and total number of that word sequences the corpus contains. As the language is innovative, new sentences are generated all the time, so, it is not possible to count entire sentences. The joint probability method for counting the entire word sequence is also not to be an effective method as it is required to find all possible word sequences. Hence, it requires a new method for guessing the probability of a word ***w*** given a prior history, or the probability of the whole word sequence **W**. An order of N words can be represented as $w_1 \ldots w_n$ or(w_1^n).The probabilities of entire word sequence Pr($w_1 \ldots w_n$) or Pr(w_1^n) can be computed easily by using the chain rule of probability. So, the final probability can be obtained as:

$$Pr(w_1^n) = Pr(w_1)Pr(w_2|w_1)Pr(w_3|w_1^2) \ldots . Pr(w_n|w_1^{n-1}) = \prod_{k=1}^{n} Pr(w_1|w_1^{k-1}) \quad (1)$$

The chain rule displays the link for calculating the conditional likelihood of a word, given previous words and calculating the combined probability of a given sequence. By multiplying the conditional probabilities, the joint probability of entire sequence of words can be estimated. It is difficult to calculate the actual occurrence of a word given a lengthy sequence of former words: $Pr(w_n|w_{n-1})$. The intuition behind the n-gram model is that, it is not required to compute the possibility of a word specified its complete history, rather we can estimate by just considering its previous words. For an instance, the bigram model estimates the likelihood of a word specified all the former words $Pr(w_n|w_1^{n-1})$ by simply considering the conditional likelihood of the former word $Pr(w_n|w_{n-1})$ .So, in bigram model, the following approximation is used to predict the conditional possibility of the next word.

$$Pr(w_n|w_1^{n-1}) \approx Pr(w_n|w_{n-1}) \quad (2)$$

Markov assumptions belongs to the class of probabilistic modelling which assumes the prediction regarding probability of various forthcoming units without observing a long back. So, it may be the generalized form of bigram approach (considers single word about past), trigram approach (considers two consecutive words about past) and n-gram approach (considers n-1 words regarding past) [10]. If the probability of a single word can be computed using bigram approach, then the probability of sequence of words can be determined by the following equation.

$$Pr(w_1^n) \approx \prod_{k=1}^{n} Pr(w_n|w_{n-1}) \quad (3)$$

Using the maximum likelihood estimation (MLE), the probabilities of bigram or n-gram approaches can be better assessed. The term MLE evaluates the factors relating to n-gram model by obtaining similar counts from the respective corpus. Then it has to be normalized, so that the counts should have the values between 0 to 1.The net amount of entire bigram counts, which begins with certain word w_{n-1}, that must be same with the unigram count considering the same word w_{n-1} :

$$Pr(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})} \quad (4)$$

Over the past few decades, a lot of research has been over and done with automatic speech recognition. Researchers are trying their best for developing ASR systems for their own languages. In India, considering different regional languages, ASR systems has been implemented for languages such as: Bengali, Telegu, Punjabi, Tamil, Malayalam, Kannada, Marathi, Hindi etc. Regional language like Odia is spoken in Odisha as well as in numerous areas of West Bengal, Bihar, Chhattisgarh, Jharkhand and Andhra Pradesh. Government has already approved that, for Odisha, Odia is to be treated as the first official language and for the state like Jharkhand, it should be the second official language. Even though considerable work can be found in many Indian languages, Odia language is not well explored yet by researchers in the perspective of speech recognition. Speech interface supports several valuable applications like spoken database querying for novice users, telephone directory assistance, office dictation devices, hands busy applications in banking and medical line, automatic voice translation into foreign languages, etc. Digit recognition is one of the common application for designing automatic voiced dialer system [11]. Continuous digit recognition system is important for physically challenged (particularly for blind people) or elderly people for making a telephonic conversation without physically dialing the numbers. An automatic Odia digit recognition system is required for Odia spoken people those are blind or can speak Odia digits but not able to recognize them. So, our work is contributed to develop such type of system. The novelty of this paper originates from the fact that we have developed a continuous voiced Odia digit recognition system which can recognize spoken digits for Odia language. In our proposed work, we used the acoustic phonetic, word level and phoneme level approach for recognizing voiced mobile numbers for Odia language. The rest of the paper is structured as follows. Section 2 describes the related work over speech recognition. Section 3 outlines the work flow of the proposed continuous voiced Odia digit recognition system. The comparison results of different approaches for continuous Odia digit recognition is presented in Section 4. Finally section 5 concludes the paper and describes the future directions.

II. RELATED WORK

As the speech/voice signal is non-stationary and holds multidimensional data, automatic speech/voice recognition turn out to be a complex task. Different speaker's ascent for the same word can vary, so extracting and manipulating textual data from speech signals becomes even more challenging. Individual utterances for the same word, also differ in length and other parameter depending upon the scenario and the context of the text. Various feature extraction methods are used in speech processing such as: MFCC, LPC, FFT, Discrete Wavelet Transform (DWT) and perceptual linear prediction (PLP) coefficients [5]. Due to the orthogonality behavior shown by MFCC coefficients and fits with the Mel filter bank which resembles in accordance with the human beings, is the prime concern of selecting as the parameter extraction technique. Also the 1st order and 2nd order derivatives of MFCC confirms its suitability for the task of speech analysis. In speech recognition, the models that uses MFCC including HMM model provides better outcomes compare to the other models. A lot of research work has been already proposed by different researchers for speech recognition. The research work related to isolated word recognition, continuous word recognition and digit recognition for different spoken languages with MFCC as parametric representation of the speech are main focus for our contemplation. The author Dua et al. proposed an ASR model considering isolated words for Punjabi language [9]. The said model has been developed by the help of HTK toolkit considering HMM model. Initially, the model is trained by considering 115 different words from Punjabi language accumulated from eight number of speakers, further it is tested over samples collected from six individuals and applied in real time scenario.The system is being tested in different environments: noisy and noiseless. The experimental outcome indicates the overall efficiency has to be 95.63% in class room (noiseless) environment and 94.08% in open space (noisy) environment. Medhi et.al suggested an ASR method for isolated Assamese words which uses the parameters like: Zero Crossing Rate (ZCR), MFCC and Short Time Energy (STE). These parameters has been extracted from the sounds belonging to Assamese words [12]. The model works well for both speaker dependent and speaker independent. The system comprises three stages: the recognition stage, the training stage and the testing stage. The database was constructed by considering the utterances of 20 speakers (10 male and 10 female). Hundred commonly used isolated Assamese words whose syllable deviates in the range 1 to 5 (mono-syllabic to penta-syllabic) was taken for recording and recognition purpose. The recognition/accuracy rate was found to be high for both the cases: speaker independent (93%) and speaker dependent (99%). An application of HMM based isolated Odia word recognition system for the visually impaired students in school and public education was proposed in [13]. A Mohamed and Ramachandran developed a small vocabulary, context independent and connectionist-statistical continuous speech recognition system. The system usages the Artificial Neural Network (ANN) to estimate the posterior probability, which can be used in state emission probability of the HMM by applying Bayes rule. The hybrid system produces 86.67 % word and 66.67% sentence recognition rate [14].

S Mohanty et.al developed a system for both speech recognition ("what was said?") and speaker identification ("who said?")

HMM based model is used for speech recognition where as speaker identification task was achieved using Support Vector Machine (SVM). Gamma tone cepstral coefficients and MFCC are extracted for speaker identification and speech recognition respectively [15]. Firoze et. al proposed a speech recognition system for Bengali language that is based upon fuzzy logic[16].They empirically revealed that, fuzzy logic consequences is a better response for Bengali speech which may contain ambiguous entities. Experimental results shown that the system based on fuzzy logic (86% accurate) and ANN (90% accurate) has better performance compared to a general HMM based recognizer which is of 73% accuracy.

Speech recognition for the digits in Hindi language was developed in[17].Authors have designed the system by considering both noisy and noiseless environment.Using eight speakers' audio data, the acoustic model was trained .The size of the vocabulary for recognition is 10 words/digits (0 to 9). HTK toolkit with HMM model is considered for performance evaluation of the system. The efficiency of the Hindi digit recognizer has been considered for phoneme level as well as word level. In [30], an isolated Kannada digit recognition model was developed by using HMM and MFCC. An optimal speaker independent continuous digit recognizer for Malayalam language was proposed by C. Kurian and K. Balakrishnan [18]. For speech parameterization, the system employs PLP coefficient technique whereas for recognition purpose HMM model is used. The training set contains the voices of 21 speakers with ages between 20 to 40 years. The voice has been recorded in office environment and individual speaker is requested to speak a continuous set of 20 digits. The system acquired efficiency level having 99.5 % for the untrained data. Isolated Odia digit recognition using HTK tool is represented in [19].An isolated, speaker independent digit recognition for Malayalam language and its various application was presented in [20]. For their system, MFCC is used for feature vector generation, while HMM is used as the technique for recognition purpose. Based on word level and HTK approach, Marathi digit recognition was proposed in [21].They created the corpus which comprises 800 sounds of 40 speakers of 20 male speakers and 20 female speakers. MFCC technique is used for extracting the acoustic features of the utterances present in the corpus. The outcome analysis of the system displays a recognition rate of 99.75% with 48.75% accuracy. In [22], for recognition of Bangla digits, the corpus was collected from the people of Bangladesh. MFCC technique has been utilized for feature extraction and for the task of recognition, HMM classifiers are used. The experimental outcomes shows for higher recognition rate (>= 95%) for the digits 0 to 5 and lower recognition rate (<= 90%) for the digits 6 to 9.In [23], the corpus was constructed by considering the utterances of ten digits (English) zero to nine. The digit recognition system mainly contains two parts, one is for feature extraction and other is for matching the features. For feature extraction, cepstrum technique is applied while for matching the features, vector quantization method is used.

A unique approach for constructing syllable-based continuous speech recognizers for Indian languages is proposed in [24]. Both in training and testing phase, the speech signal is automatically segmented into syllables. In order to build the syllable models for better recognition, syllable boundary information was considered. In training phase, a rule based segmentation algorithm was used which syllabifies the texts. The syllabified signal and text automatically generates the annotated data which was used for constructing isolated syllable models. During testing phase, the acoustic waveform is again automatically segmented using group delay technique. In [25], a novel method was proposed for aligning sequence of input frames with corresponding output labels. Based on sequence modeling and attention mechanism, a Recurrent Neural Network (RNN) was trained to do the above process. The system can be applied to Large Vocabulary Continuous Speech Recognition (LVCSR) by mixing the decoding RNN with an n-gram language model. The processing speed up for the system can be achieved by constraining selections done by attention mechanism**.** Wageesha et.al proposed an Interactive Voice Response (IVR) based Sinhala Speech Recognition System [26]. The system uses HMM to spot Sinhala digits and Sinhala songs name, which is set as a ring back tone**.** The voice based communication system has been represented in [27]. The system was constructed by assembling various subsystems as: an ASR engine, a TTS engine, a dialogue management subsystem and an interface with the bank database. The system is generally used for providing services for the banking system. In [28], an ASR system for mobile phone applications used in Punjabi language has been proposed. The system was tested with four different acoustic models like: context independent, context dependent untied, context dependent tied, and context dependent deleted interpolation models. It was observed that, context dependent untied models overtake others by having lower word error rate and better accuracy. An end-to-end acoustic modeling approach using convolutional neural networks (CNNs) for HMM based ASR was proposed in [29]. In their proposed acoustic modeling approach, the appropriate features and the classifier are mutually learned from the raw speech signals.

III. PROPOSED CONTINUOUS VOICED ODIA DIGIT RECOGNITION SYSTEM

The overall construction steps for continuous voiced Odia digit recognition system is depicted in Fig.3.

A. Data Preparation

1. Recording the Data: Recording of the voiced data may be done using a quiet environment or noisy environment. A headset microphone is used for recording purpose. The audacity (sound recording and editing software) is used setting the microphone volume level as 1.0. Recording for the mobile numbers are done with a normal voice. Utterances are taken by considering the sampling frequency of 16 KHz with 16 bits per sample. From different cities of Odisha, speakers are chosen. At the time of recording, a distance of 2 - 4 inches was retained between microphone and the speaker.

Smart mobile phones are also used for recording the utterances. For isolated digit recognition approach, the voices recorded continuously from different speakers are manually segmented using audacity and labelled with the corresponding word of that digit. The segmented files that are created, stored separately and also named based on the digit and name of the particular speaker's utterance. For our system, the training corpus is being considered with 100 mobile numbers recorded by 20 different speakers (13 male and 7 female) whose age's lies between 20 to 50 years. Each speaker is asked to speak 10 set of mobile numbers. So, a total of 10X20=200 sound files are recorded.

2. Sample Preprocessing: The input sound signal articulated by a speaker may contain some silence period, circumstantial noise and other inference signal along with the needful information. The preprocessing step is used to minimize the noise and removes the silence period which was present during recording of the speech. Using the energy of the signal, the starting point of a speech in the recorded Odia digit can be found. For our system, the silence and noise removal property present in audacity, is used for removal of the noises present in the recorded wave forms.

3. Creating the Transcription Files: Each sound file recorded earlier, should be labelled with its corresponding text.It is developed by considering its equivalent transcription files (*trans.txt*). There are two ways the transcription files will be developed using HTK Toolkit: Word Level Transcriptions (WLT) and Phone Level Transcriptions (PLT). The *trans.txt* file created can't be processed directly by the HTK toolkit. A Master Label File (MLF) has to be created. It is a single file containing a tag entry for each and every line in the *trans.txt* file. This is the easiest approach, which is used for our purpose. Using a Python script, a *words.mlf* is generated from our *trans.txt* file. The ***HLEd*** command of HTK is used to expand the WLT to PLT which will substitute each word with its corresponding phonemes. Hence, a new Phone Level Master Label File is resulted. This is created by considering each word in the MLF file, and looking for the phones in the dictionary which made that word. Using the above command twice, two files are generated: *phones0.mlf*, (no space ("sp") after each word phone group) and *phones1.mlf* (short pauses ("sp") after each phone word group) [32]. These files are required when we will use the phoneme level approach for recognition.

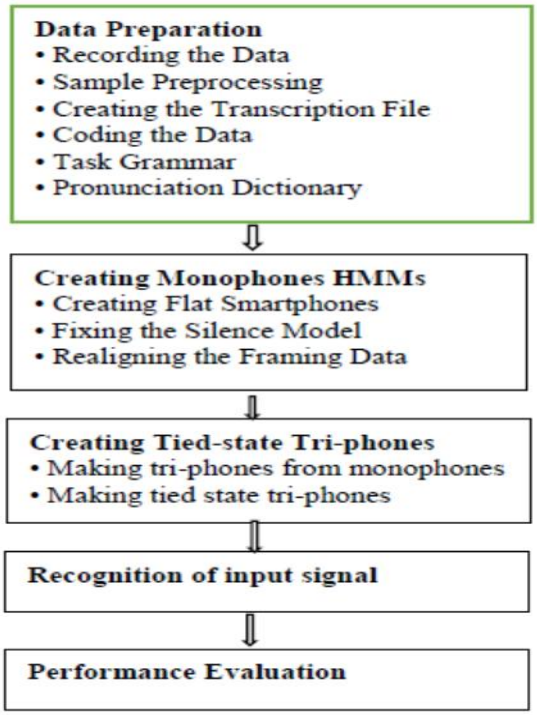


Fig.3 Flow diagram of proposed continuous voiced Odia digit recognition system

4. Coding the (Audio) Data: In this step, sequences of feature vectors are extracted from the raw speech signals. This is required because, HTK can't process the raw *.wav* files directly. So, the raw (*.wav)* files are always need to be converted into to a feature vector format called as: MFCC format. The ***HCopy*** tool of HTK is used to convert the *.wav* files to MFCC format. There are two options to do it. The ***HCopy*** command can be executed for each audio file created earlier to obtain its equivalent MFCC file. It will take more time as we need to consider all files individually for converting to its corresponding MFCC files. Second options is, we can create a text file containing a list of audio files and the corresponding MFCC files. Then, the text file can be used as a parameter to the ***HCopy*** command for obtaining all the MFCC files at once [31]. The second approach is executed for our purpose.

5. Task Grammar: A recognition Grammar basically defines the limitations on what the Speech Recognition Engine (SRE) can anticipate as input. The SRE listens for a set of words and/or set of phrases. Once these processed words or phrases are heard by the system, the SRE returns the corresponding words or phrases. It is a bit concern to know that, grammar should contain the same words for which the acoustic models are trained. The grammar used for this continuous digit recognition process is based upon the modified BNF (Backus-Naur Form) format.

Sample Grammar

S: S_S DIGIT E_S

The production rule defined in the grammar contains "S" is the starting sentence symbol. S_S and E_S correspond to the "start silence" that occurs just before the start of the utterance and the "end silence" that comes after the utterance. With this grammar, one substitution per sentence is allowed, taking colon ":" as the extractor. The *.voca* file holds word definitions for each word defined in the *.grammar* file. A simple grammar with one word category is used for our purpose. DIGIT can be substituted with any one of the ten digits described in *.voca* file.

Sample Voca

%S_S
<s>sil
%E_S
</s> sil
% DIGIT
ଆଠ ଆଠ
ଏକ ଏକ
ଚାରି ଚାରି
ଛଅ ଛଅ
ତିନି ତିନି
ଦୁଇ ଦୁଇ

ନଅ ନଅ
ପାଞ୍ଚ ପାଞ୍ଚ
ଶୂନ ଶୂନ
ସାତ ସାତ

6. Pronunciation Dictionary: The initial step in creating the pronunciation dictionary is taking the words in the grammar and generating a list of the sorted words. The sorted words are considered as one per line, with actual pronunciations (set of phonemes that build a word).It can be constructed by the expert linguist of that language. In our proposed work, it is done easily as there are only ten number of Odia digits/words to be considered [32].

To build a pronunciation dictionary using HTK, the following steps are used:

- Considered a *trans.txt file:* This is the set of words/digits we have recorded in the previous step.
- Created a *wlist* file from the *trans.txt* file: This is a sorted list of unique words that present in the *trans.txt* file.
- Created the pronunciation dictionary: This is computed by adding pronunciation information to the words in *wlis*t file.

B. The Monophones HMMS

The initial step for HMM training is to express a prototype model called "proto". The flat start monophones, macro files are created and re-estimated monophones are done in ordered to use in the next step of the proposed system. In realignment of the training data word-to-phone mapping operation is performed. In this case the ***HVite*** command will consider all pronunciations for each word (in the case, multiple pronunciations are there for a single word), and then output the best matched pronunciation for the acoustic data.

C. Creating tied-state Triphones

The pronunciation of a word/digit present in the dictionary file, contains a list of phonemes (also called monophones).To generate a triphone (set of 3 phones) which can be obtained from monophones, the "L" phone (the left-hand phone) precedes "M" phone(middle phone) and the "R" phone (right-hand phone) follows it. The triphone is acknowledged in the formula of "L-M+R". We have marked over that, HMMs are essentially the statistical signs of the phones that build up a word. An HMM is buildup with many 'states', and these states can be shared. The way the “sp” and “sil” phones can be shared with their center 'state', states for phonemes in the triphone are also shared. These tied of states are called as “senones”. Decision tree clustering was used here which allows previously unseen di-phones and tri-phones to be synthesized.

4. Performance Evaluation

The equations represented below provides the method for evaluating the performance of our proposed system.

Digit Correct rate (DCR) = (NW − DW − SW)/ NW × 100
Digit Accuracy rate (DAR) =
(NW − DW − SW − IW)/ NW × 100
Digit Error Rate (DER) = 100% − DAR

Here, we considered NW=Number of words/digits in the test corpus, DW=Number of deleted words/digits, SW=Number of substituted words/digits and IW=Number of inserted words/digits required for matching. DER is considered as one of the major parameter for the performance measurement [19].

Table-1 represents the Odia digits with its equivalent English word, the symbols used in Odia language and corresponding Unicode.Table-2 depicts the pronunciation dictionary for Odia digits.This dictionary is created considering the phonemes for each of the digits. It may be noticed that some of the digits can have more than one pronunciation. Hence we can have more than one phonemes represented for those digits.

Table 1. Odia digit with corresponding words, words in Roman and English language, symbol and Unicode

Odia Digit (word)	Odia Digit (symbol)	pronunciation
ଏକ	[୧]	ଏକ,ଏ ◌କ ଏ
ଦୁଇ	[୨]	ଦୁଇ
ତିନି	[୩]	ତିନି,
ଚାରି	[୪]	ଚାରି
ପାଞ୍ଚ	[୫]	ପାଞ୍ଚ,ପାଞ୍ଚ୍ ଅ
ଛଅ	[୬]	ଛଅ
ସାତ	[୭]	ସାତ,ସା ◌ତ ଏ
ଆଠ	[୮]	ଆଠ,ଆ ◌ଠ ଏ
ନଅ	[୯]	ନଅ
ଶୂନ	[୦]	ଶୂନ,ଶୂ ଉ ନ୍
SENT-START SENT- END	[ ]	sil
	[ ]	sil

Table.2 Pronunciation dictionary for Odia digits

Odia Digit (word)	Odia Digit (Roman) (English word)	Odia Digit (Symbol)	Unicode
ଏକ	eka (One)	୧	0B67
ଦୁଇ	dui(Two)	୨	0B68
ତିନି	tini(Three)	୩	0B69
ଚାରି	chāri(Four)	୪	0B6A
ପାଞ୍ଚ	pāñcha(Five)	୫	0B6B
ଛଅ	chhaa(Six)	୬	0B6C
ସାତ	sāta(Seven)	୭	0B6D
ଆଠ	āṭha(Eight)	୮	0B6E
ନଅ	naa(Nine)	୯	0B6F
ଶୂନ	suna(Zero)	୦	0B66

IV. RESULT ANALYSIS AND DISCUSSIONS

A. Comparison concerning word/digit level and phoneme level approach

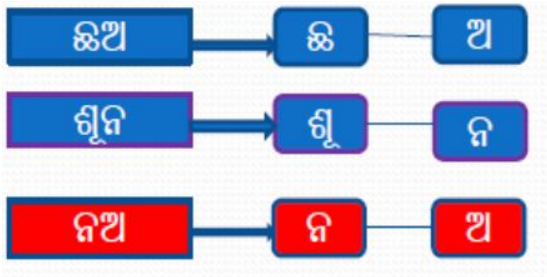
In the word/digit level recognition, we required a training set which contains a set of voice recordings of the similar words by many speakers. Using that training set, an HMM corresponding to each of the vocabulary words/digits is trained. It is then tested to recognize the words/digits using the testing set. The testing set may be considered for a new set of recordings by multiple speakers. So all together we need ten set of HHMs which can recognize ten different digits for the Odia language. The experiment was conducted using the isolated Odia digit recognition approach [19]. Fig. 4 depicts the recognition result of one sample containing ten digits using isolated approach.

Each sample of testing set contains the voices of ten digits corresponding to a single mobile number. In our case the percentage of sentence recognition rate is 70%, which is displayed in the line Sent. Since the grammar holds one digit/word in one sentence, so the correct rate of recognition for words/digits is also 70%.This is represented in Word line. Accurately recognized testing word/digit is denoted as HW=7.Similarly, SW=3 reflects the errors obtained by substituting 3 words/digits/.NW=10 represents the total number words/digits present in a sample, which is considered for testing. Similarly, the test is evaluated for other nine samples that is for other nine mobile numbers. In our experiment, it has been found that, the average correct word/digit recognition is 70.7%.

```
==============HTK Results Analysis=========================
-------------Overall Results ---------------------------------
Sent: %Correct=70.00  [HW=7, SW=3, NW=10]
Word: %Corr=70.00,  Acc=70.00   [HW=7, DW=0, SW=3,IW=0, NW=10]
===========================================================
```

Fig. 4 Result of running one sample (Isolated approach)

In the phoneme level approach, each word/digit can be segmented into two or more number of phonemes. The number of HMMs required for each digit depends upon the number of phonemes required for that digit. So we can have network of HMMs for recognition of the digits. For example, two phonemes are required for representing the Odia digits ଶୂନ (suna) and ଛଅ (chhaa).The corresponding HMMs for the phonemes are trained to recognize those digits. Recognition in phoneme level approach has more advantage than the isolated approach. The reason is, even if we don't have voice database that is not trained for the Odia digit ନଅ (naa), still it



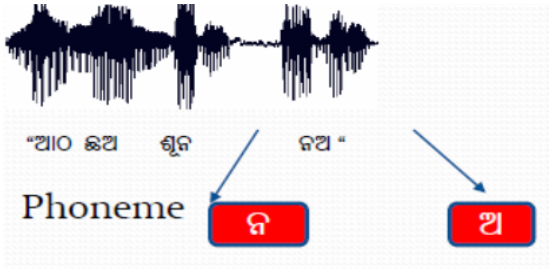


Fig. 5 Recognition of unknown digit ନଅ from the phonemes of the digits ଛଅ and ଶୂନ

can be obtained from the phonemes of ଶୂନ (suna) and ଛଅ (chhaa).The HMMs for ନ (na) and ଅ (aa) can be obtained from ଶୂନ (suna) and ଛଅ (chhaa) respectively. Fig. 5 depicts the phonemes corresponding to the Odia digits ଛଅ (chhaa), ଶୂନ (suna) and ନଅ (naa) and also shows how ନଅ (naa) can be obtained even it is not trained earlier. The experiment was conducted with phoneme level approach for continuous Odia digit recognition by considering the mobile numbers. The result shown in Fig.6 is quiet encouraging compared to the previous approach.

```
==============HTK Results Analysis=========================
-------------Overall Results ---------------------------------
Sent: %Correct=80.00  [HW=8, SW=2, NW=10]
Word: %Corr=80.00,  Acc=80.00   [HW=8, DW=0, SW=2,IW=0, NW=10]
===========================================================
```

Fig. 6 Result of running one sample (phoneme level approach)

B. Comparison between Isolated and N-gram Language Model

Speech recognition systems are generally categorized as either isolated system or continuous system. In isolated word/digit recognition, a tiny pause is to be considered between each spoken word, whereas in continuous speech recognition it doesn't require any such pauses. Isolated word/digit recognition involves two major stages: a training stage and testing stage. In the training stage, the system is updated for building an acoustic model for each word/digit. In our case, we have built the HMM's for each of the Odia digits and trained these to recognize the Odia digits from ଶୂନ (suna), ଏକ (eka) to ନଅ (naa). In the testing stage, isolated digits are recognized by considering the acoustic models of these digits. The sentences of our *trans.txt* file (transcription file obtained after recording) contains the mobile number represented in the words of Odia digits. Each line of the corresponding file is augmented with a distinct symbol <s> at the starting of the line, which will give the bigram count for the first word and a special symbol </s> which will mark as the end of the line. We considered the unigram count for each of the Odia digits. Bigram probability for ten Odia digits (corresponding to a mobile number) in our database of mobile numbers is considered for the evaluation.Table-3 shows the bigram counts for a mobile number "ଆଠ ଦୁଇ ଚାରି ନଅ ଆଠ ପାଞ୍ଚ ପାଞ୍ଚ ଛଅ ସାତ ଦୁଇ (āṭha dui chāri naa āṭha pāñcha pāñcha chhaa sāta dui)" (୮୨୪୯୮୫୫୬୭୨(in Odia), (8249855672(in

English))". After normalization (dividing each cell by the correct unigram for its row taken from the unigram count table represented in Table-5), the bigram probabilities are represented in Table-4. Note that some of the entries are zeros, that means some binary counts (conjugative digits) are not there in the corpus. Some binary count probability is high suggest that the digit pair is appearing more in the corpus. We can compute the probability of correct sequence of digits of a mobile number containing ten digits by multiplying the bigram probabilities as follows:

Pr(<s>ଆଠଦୁଇଚାରିନଅଆଠପାଞ୍ଚପାଞ୍ଚଛଅସାତଦୁଇ</s>)=Pr(ଆଠ|<s>)Pr(ଦୁଇ|ଆଠ)Pr(ଚାରି|ଦୁଇ)Pr(ନଅ|ଚାରି)Pr(ଆଠ |ନଅ)Pr(ପାଞ୍ଚ|ଆଠ)Pr(ପାଞ୍ଚ|ପାଞ୍ଚ)Pr(ଛଅ|ପାଞ୍ଚ)Pr(ସାତ|ଛଅ)Pr(ଦୁଇ|ସାତ)Pr(</s>|ଦୁଇ)= 0.2 X 0.078 X 0.83X 0.073 X 0.197 X 0.496 X 0.158 X 0.132 X 0.047 X 0.078 X 0.14=9.860992906785E-10

Multiplying ample n-grams altogether will always result a numerical underflow because the probabilities are lie between 0 to 1. So to avoid numerical underflow, log probabilities can be used for further computation. Multiplying raw probabilities can be obtained by adding the corresponding log probabilities. Then just taking the exponential (exp) of the log probability (logprob) the original probability can be obtained. We also computed the trigram probability for the same mobile number. The proposed system is executed for thrice, once for each language model (unigram, bigram and trigram).It has been found that the trigram model gives more accurate transcription compared to other two models. The results in table-6 shows, the average sentence accuracy rate in different approaches considering 10 untrained mobile numbers. The performance of recognition using tri-gram approach is found to be better than both isolated, unigram and bi-gram approach.

Table-3 Bigram counts for a mobile number consisting of 10 digits

	ଆଠ	ଦୁଇ	ଚାରି	ନଅ	ଆଠ	ପାଞ୍ଚ	ପାଞ୍ଚ	ଛଅ	ସାତ	ଦୁଇ
ଆଠ	6	10	10	8	6	6	6	20	16	10
ଦୁଇ	0	0	6	2	0	4	4	4	16	0
ଚାରି	6	6	12	6	6	12	12	2	4	6
ନଅ	30	12	12	8	30	8	8	10	14	12
ଆଠ	6	10	10	8	6	6	6	20	16	10
ପାଞ୍ଚ	12	0	4	6	12	12	12	10	2	0
ପାଞ୍ଚ	12	0	4	6	12	12	12	10	2	0
ଛଅ	8	6	10	2	8	12	12	12	4	6
ସାତ	16	10	6	18	16	6	6	14	8	10
ଦୁଇ	0	0	6	2	0	4	4	4	16	0

Table-4 Bigram probabilities for 10 digits of a mobile number (Approximation up to two digits after decimal point)

	ଆଠ	ଦୁଇ	ଚାରି	ନଅ	ଆଠ	ପାଞ୍ଚ	ପାଞ୍ଚ	ଛଅ	ସାତ	ଦୁଇ
ଆଠ	.47	.08	.06	.06	.47	.47	.47	.16	.13	.08
ଦୁଇ	0	0	.83	.03	0	.06	.06	.06	.22	0
ଚାରି	.07	.07	.15	.07	.07	.15	.15	.02	.05	.07
ନଅ	.2	.08	.08	.05	.2	.05	.05	.07	.10	.08
ଆଠ	.47	.08	.06	.06	.47	.47	.47	.16	.13	.08
ପାଞ୍ଚ	.16	0	.05	.08	.16	.16	.16	.01	.02	0
ପାଞ୍ଚ	.16	0	.05	.08	.16	.16	.16	.01	.03	0
ଛଅ	.1	.07	.12	.02	.1	.14	.14	.14	.05	.07
ସାତ	.16	.08	.05	.14	.16	.05	.05	.11	.06	.08
ଦୁଇ	0	0	.83	.02	0	.06	.06	.06	.22	0

Table-5 Unigram count for the 10 digits

ଏକ	ଦୁଇ	ତିନି	ଚାରି	ପାଞ୍ଚ	ଛଅ	ସାତ	ଆଠ	ନଅ	ଶୂନ
80	72	84	82	76	86	128	128	152	112

Fig.7 shows overall average sentence accuracy rate (%) of the continuous voiced Odia digit recognition system using different approaches. The results obtained using the combination of HMM and MFCC techniques for the proposed system are tolerable, but can be enhanced further to attain better recognition rates. Table-7 depicts the better recognition rates (%) acquired from previous research works in juxtaposition to our proposed CVODR system.

Table.6 Average sentence accuracy rate

Approach	Average sentence accuracy rate
Isolated	70.70%
Unigram	80.50%
Bi-gram	82.80%
Tri-gram	86.60%

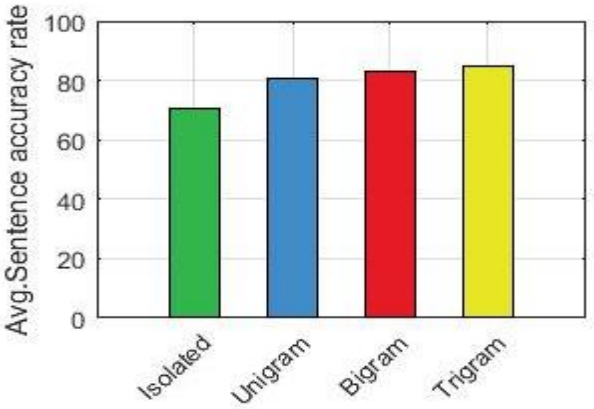


Fig. 7 Average sentence accuracy rate for Odia voiced digit recognition with different approaches

Table 7. Comparison of recognition rates (%) with our continuous voiced Odia digit recognition system

Reference	Feature Extraction	Feature classification Technique	Recognition rate (%)
R.Thangarajan et.al [33]	MFCC	HMM	Speaker Dependent: 88.82 Speaker Independent: 92.06
M. Z Bhotto et.al [34]	MFCC	VQ	70-85
S. Mohammad et.al [35]	MFCC	VQ	88.88
S.M. Ahadi et.al[36]	MFCC (noisy) MFCC (noiseless)	HMM (noisy) HMM (noiseless)	28 - 78 86
Proposed system (CVODR)	MFCC	HMM	Isolated: 70.7 Unigram:80.5 Bigram:82.5 Trigram:86.6

V. CONCLUSION AND FUTURE WORK

In this work we implemented a continuous voiced Odia digit recognition system (CVODR) for recognizing the mobile numbers for the Odia language using HTK Toolkit. The results shown above quantify that tri-gram language model with phoneme level approach gives the better performance corresponding to bi-gram, unigram and isolated approach. The research effort may further be extended by considering large volume of training corpus with different speakers of varying ascents. The recognition system is always sensitive to varying pronounced techniques and altering scenarios used while recording the voices, so the correctness of the system is an inspiring region to work upon. Hence, various speech improvements and noise reduction procedures may be applied for building the system further accurate, fast and efficient.

In future, the system may be further improved using machine learning approach which can automate the system for more accuracy and robust.

REFERENCES

1. Pranab Das, Prerana Das, Kakal Acharjee and Vijay Prasad “Voice recognition system: Speech-to-text.” Journal of Applied and Fundamental Sciences, pp-191- 195, Vol-1, 2015.
2. R. Errattahi, Asmaa El Hannani, and Hassan Ouahmane. "Automatic speech recognition errors detection and correction: A review." *Procedia Computer Science* 128 (2018): 32-37.
3. Karpagavalli, S., R. Deepika, P. Kokila, K. Usha Rani, and E. Chandra. "Automatic Speech Recognition: Architecture, Methodologies and Challenges-A Review." *International Journal of Advanced Research in Computer Science* 2, Vol.No. 6, 2011.
4. Saurabh Chatterjee, Project guides: Harish Karnick, Srinivasan Umesh, “Speech Recognition in Indian Languages” Btp term1 report.
5. R. Rabiner, and B. H. Juang, “Fundamentals of Speech Recognition” *Prentice-Hall International*, New Jersey, 1993.
6. Megha Agrawal and Tina Raikwar. "Speech recognition using signal processing techniques." *International Journal of Engineering and Innovative Technology* 5.8: pp.65-68, 2016.
7. S.Pannirselvam, G. Balakrishnan. Article: Comparative study on preprocessing techniques on automatic speech recognition for Tamil language. *IJCA Proceedings on National Conference on Research Issues in Image Analysis and Mining Intelligence* (NCRIIAMI 2015), 2:25–28, 2015.
8. Abolfazl Rezaei, Nasser Salehi. An introduction to speech sciences (acoustic analysis of speech). *Iranian Rehabilitation Journal*, 4(1):5–14, 2006.
9. M.Dua, R.K.Aggarwal , V. Kadyan ,S. Dua “Punjabi Automatic Speech Recognition Using HTK”, International Journal of Computer Science Issues (IJCSI), pp.359-364,Vol.-9,Issue-4,2012.
10. Daniel Jurafsky, James H. Martin “Speech and Language processing” chapter 3 N-gram language Models, Draft of September 23, 2018.
11. P. Mohanty and A. K. Nayak, "Design of an Odia Voice Dialer System," 2019 5th National language (5th NLC-2019) IOSR at Ravenshaw University, Cuttack, Odisha, 04-06 February 2019.
12. B.Medhi,P. H. Talukdar“ Isolated Assamese Speech Recognition using Artificial Neural Network”, International Symposium on Advanced Computing and Communication (ISACC), Silchar,(IEEE) pp. 141-148,2015.
13. S.Mohanty, B. K Swain “Markov Model Based Oriya Isolated Speech Recognizer-An Emerging Solution for Visually Impaired Students in School and Public Examination”, International Journal of Computer & Communication Technology-2010. (IJCCT) Vol. 2 Issue-2, 2010.
14. A .Mohmed, K.Ramchandran.“ HMM/ANN hybrid model for continuous Malayalam speech recognition” International Conference on Communication Technology and System design (ELSIVIER), pp-616-622, 2012.
15. S.Mohanty, B. K Swain, “Speaker Identification using SVM during Oriya Speech Recognition”, International Journal Image, Graphics and Signal Processing,Vol-7,Issue No-10,pp-2836,2015.
16. A. Firoze, M. S. Arifin, and R. M. Rahman, “Bangla user adaptive word speech recognition: Approaches and comparisons,” International Journal of Fuzzy System Applications (IJFSA), vol. 3, no. 3, pp. 1–36, 2013.
17. Babita Saxena, Charu Wahi, "Hindi Digits Recognition System on Speech Data Collected In Different Natural Noise Environments", International Conference on Computer Science, Engineering and Information Technology (CSITY 2015) February 14~15, 2015.
18. Cini Kurian, Kannan Balakrishnan. “Connected digit speech recognition system for Malayalam language”. Sadhana, Vol-38(6), pp1.339– 1346, 2013.
19. P. Mohanty and A. K. Nayak, "Isolated Odia Digit Recognition Using HTK: An Implementation View," 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), Changsha, China, pp. 30-35, 2018.
20. S.Renjith, A. Joseph,A. Babu KK , “Isolated Digit Recognition for Malayalam an Application perspective”, International conference on Control Computing(ICCC), pp.190-193,2013.
21. D.S.Kulkarni, R.R. Deshmukh, V.J.L Patil, P.P Shrishrimal, S.D Waghmare, A.M Oirere “Marathi Isolated Digit Recognition using HTK” IJCA Proceedings on International Conference on Cognitive Knowledge Engineering ICKE 2016(2):42:45, January 2018.
22. G. Muhammad, Y. A. Alotaibi and M. N. Huda, "Automatic speech recognition for Bangla digits," 2009 12th International Conference on Computers and Information Technology, Dhaka, pp. 379-383, 2009.
23. M. D. Rudresh, A. S. Latha, J. Suganya and C. G. Nayana, "Performance analysis of speech digit recognition using cepstrum and vector quantization," 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Mysore, pp. 1-6, 2017.
24. Lakshmi, A., and Hema A. Murthy. "A new approach to continuous speech recognition in Indian languages." Proceedings national conference communication. 2008.
25. D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, pp. 4945-4949,2016.

26. M. Wageesha, D. Karunathilake, T. Madhushani, N. Galagedara, and D. Dias. "Sinhala speech recognition for interactive voice response systems accessed through mobile phones." In 2018 Moratuwa Engineering Research Conference (MERCon), pp. 241-246. IEEE, 2018.
27. Ceaparu M, Toma SA, Segărceanu S, Gavăt I. "Voice-Based User Interaction System for Call- Centers, Using a Small Vocabulary for Romanian." International Conference on Communications (COMM) (IEEE), Jun 14, pp. 91-94, 2018.
28. P. Mittal and N. Singh "Development and analysis of Punjabi ASR system for mobile phones under different acoustic models" International Journal of Speech Technology, 22(1), pp.219-230, 2019.
29. Palaz, Dimitri, Mathew Magimai-Doss, and Ronan Collobert. "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition." Speech Communication, 108, pp 15-32, 2019.
30. H.Muralikrishna, T.Ananthakrishna and K.Shama, "HMM based isolated Kannada digit recognition system using MFCC," International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, pp. 730-733, 2013.
31. http://htk.eng.cam.ac.uk/ftp/software/htkbook-3.5.alpha-1.pdf
32. http://www.voxforge.org/home/dev/acousticmodels/linux/create/htkjulius/tutorial/data-prep
33. R. Thangarajan, M. Natarajan, M. Selvam, "Word and Triphone Based Approaches in continuous Speech Recognition for Tamil Language," WSEAS Transactions on Signal Processing, ISSN: 1790-5022, Issue 3, Volume 4, March, 2008.
34. M.Z., Bhotto and M.R., Amin, "Bangali Text Dependent Speaker Identification Using Mel Frequency Cepstrum Coefficient and Vector Quantization," 3rd International Conference on Electrical and Computer Engineering, Dhaka, Bangladesh, pp. 569-572, 2004.
35. Shariah, Mohammad AM Abu, et al. "Human computer interaction using isolated-words speech recognition technology." 2007 International Conference on Intelligent and Advanced Systems. IEEE, 2007.
36. S.M., Ahadi, H., Sheikhzadeh, R.L., Brennan, and G.H., Freeman, "An Efficient Front-End for Automatic Speech Recognition," IEEE International Conference on Electronics, Circuits and Systems (ICECS2003), Sharjah, United Arab Emirates, 2003.

AUTHORS PROFILE

Prithviraj Mohanty is currently pursuing his Ph.D. degree in Computer Science at Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha. He received his B.Tech. degree in Computer Science and Engineering from Biju Patnaik University of Technology (BPUT), Odisha and master (Computer Science and Engineering) degree from KIIT University, Bhubaneswar. His research interest includes speech recognition, machine learning, artificial intelligence, grid computing and distributed computing.

Dr. Ajit Kumar Nayak received his master degree and Ph.D. (Computer Science and Engineering) from Utkal University, Bhubaneswar, Odisha. He is currently working as a professor and head of the department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha. He has rich teaching and research experience and he has nearly 40 reputed journal/conference publications. His research interests include wireless sensor network, speech recognition, optical character recognition, text summarization, image processing and machine learning.