The New Technique Enhancing of Automatic Speech Recognition System for ODIA Language using HTK Based On Hidden Markov Model (HMM)

Priyabrata Sahu, Sunil Ku Panigrahy, Umakant Bhaskar Gohatre

Abstract: ***The purpose of this paper is to address the application to an Indian Regional Language, ODIA of a single word Automatically Speech Recognition System (ASRS). The toolkit is based on Hidden Markov Model (HMM). The details was obtained from 8 ODIA Language speakers. The program is then qualified for 205 different terms in ODIA. Samples from six separate speakers have again been obtained. This is then evaluated in real time. A GUI has been created to enhance the system's interactivity. We used and introduced the test framework for development of the GUI JAVA application. A comprehensive model of an ASR framework was developed to explain each HTK resource using HTK library modules and software. The findings of the experiment indicate that the overall machine efficiency is 93.45%.***

Keywords: **HMM, HTK, P-ASR, Automatic Speech Recognition system.**

I. INTRODUCTION

Thinking is the vocalization of human speech. They use devices such as the keyboard, mouse and screen respectively to connect with a computer, use apps. An ASR framework is an alternative to the hardware interface. [1] This is used as an input for automated speech recognition and is captured on the microphone. The transmitted signal is then translated as similar as possible to the spoken data into a series of text data [2]. The key challenges with applying an ASR program are the various modes of expression of people. The environmental problems are still impacting it.

Our ASR system's key goal is to turn a voices signal recorded by a computer into a text message set. This text message should be precise and effective, regardless of the user, speaker or setting. Automatic speech recognition (ASR) system's main application fields are dictation, program control, automated telephone call processing and query-basic management. Idealization. The main applications of this are program control, automatic telephone call processing and access to query based information systems, such as travel information. Automatic speech recognition systems. Taking into account all the difficulties and broad implementations, the paper aims to create a GUI-driven word recognizer for HMM related vocabulary that is separate and isolated [3, 4].

A. *Motivation.*

Automatic speech work in recent decades a machine's understanding has received significant interest. The checked literature shows how many development projects in this area over the last 50 years have been sponsored by organizations such as AT&T Bell Labs, IBM, DARPA and Microsoft. There is still a great deal of work in this field. Nevertheless, the academic organizations have concentrated on developing an ASR framework for currently designed applications. “Microsoft SAPI, Dragon-Speech and IBM by speech are several examples”. However, the popularity with speech recognition has created diverse individuals involved in creating solutions for the citizens with their area, whether they be native or local. In India, there are no such effective speech recognition systems for ODIA language that have been established for different languages like Hindi [5], Bengali, Panjabi, etc. Regional significance has also facilitated the exploration of speech recognition in the unexplored field of ODIA.

II. RELATED WORK.

Every portion of the paper provides a critical analysis of plays related to this work. The experimental, real-time and isolated Punjabi word recognition was implemented by R. Kumar [6]. The present research focuses on a functional reconnoiter system using a linear predictive application with vector quantification and dynamic programming. The performance of a specific word recognition system based on the speaker for a limited vocabulary can also be evaluated using Concealed Markov (HMM) or Dynamic Time Warp (DTW) approaches. The Hidden Markov Program based in particular on recognizer word recognition practices. R. Now, again. Kumar [7] has developed an autonomous, real-time and isolated Punjabi word framework [7].

The mechanism is understood through the methods of Vector Quantization and Dynamic Time Warping (DTW). The device pattern (check token features) is observable by using dynamic time warm synchronization in accordance with each reference pattern and the system data store for the task (LPC coefficients or LPC-dependent coefficients) is for the process testing.

Revised Manuscript Received on May 21, 2020.

***** Correspondence Author

Priyabrata Sahu *, Assistant Professor in the Department of Computer Science Engineering and Applications, Indira Gandhi Institute of Technology, Odisha, India

Sunil Ku Panigrahi, Research Scholar, Department of computer Engineering, KIIT University, Odisha**, India**

Umakant Bhaskar Gohatre, Assistant Professor, Department of Engineering, University of Mumbai, Mumbai, India 400 701

The program developed for this independent term for limited vocabulary. A related word recognition program for Hindi was developed by R. Kumar [5].
The framework was built with a secret toolkit of the Markov System (HTK). The machine was qualified to recognize every series of words from 15 participants, men and women alike. The test results from five speakers was again used to assess the recognizer's efficiency. The aim of this paper is to discuss the architecture and the implementation of an isolated ODIA word recognizer consisting of 205 terms in both speaker and independent real time environments.

III. STATISTICAL FRAMEWORK OF AN ASR.

An ASR approach as shown in fig.1 involves five key elements: acoustic feature extraction research, acoustic model based on HMM arithmetic, language models, voice dictionary and decoder. / Photo. Photo. 1: AAD Block Diagram. The sound waves captured from the front end of the microphone are generated by the acoustic processing unit. The data is first converted into a vector sequence function. The decoder is instead delivered.
Then we obtain the effect by using auditory, language and pronunciation templates by decoding machine. The following four phases can be separated primarily into 1.Signals parameterization, using an extraction method like MFCC or PLP2. The following four phases can be split. Gaussian mixture (GMMs) Acoustic rating, 3. Chain simulation with secret Markov (HMMs) and 4.Generating the strategic hypothesis with a data ranking. (Acoustic and language models and models of production). The optimal output is chosen using a decoder as the final output [2].

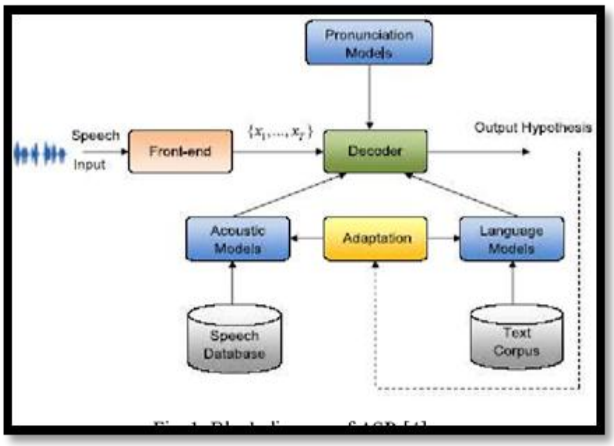


Figure 1: Block Diagram of ASR

A. *ODIA Language Phonology.*

In Indo-Aryan languages in the Indian state of Odisha and areas of West Bengal, ODIA is spoken at around 40 million people and is common in Chhattisgarh, and Andhra Pradesh. ODIA is a different official language of India. It is the first language of Odisha and the second official language of Jharkhand. It has a long tradition in literature and is often known as a Classical Tongue. Bengali and Assamese are loosely related to ODIA. The ODIA language is one of the several descends of the Brahmi language in Ancient India, the Kalinga type.

$$\mathrm{P}(Y_n \in A | X_1 = x_1, \dots, X_n = x_n) = \mathrm{P}(Y_n \in A | X_n = x_n). \quad (1)$$

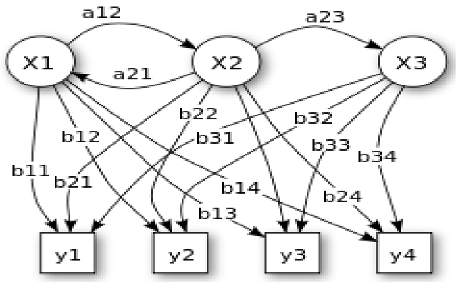


Figure 2: Network sections

A secret Markov cycle can in a discrete form be interpreted as a generalization of the replacement urn problem (where each item is transferred from the urn to the original urn before the next step). [5]

$$\varphi(x) = \frac{e^{-x^2}}{\sqrt{\pi}} \quad (2)$$

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) \quad (3)$$

$$f(x) = \sqrt{\frac{\tau}{2\pi}} e^{-\tau(x-\mu)^2/2}. \quad (4)$$

$$f(x) = \frac{\tau'}{\sqrt{2\pi}} e^{-(\tau')^2(x-\mu)^2/2}. \quad (5)$$

Such components cannot be represented in simple functions, and therefore are defined as special functions. However, other estimates are known; see below. Both roles have strong relation

$$\Phi(x) = \tfrac{1}{2}\left[1 + \operatorname{erf}\left(\tfrac{x}{\sqrt{2}}\right)\right] \quad (6)$$

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \quad (7)$$

$$\int \Phi(x)\,dx = x\Phi(x) + \varphi(x) + C.$$

$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}\left[x + \frac{x^3}{3} + \frac{x^5}{3\cdot 5} + \cdots + \frac{x^{2n+1}}{(2n+1)!!} + \cdots\right] \quad (8)$$

Consider this example: there is a genius in a region not apparent to an observer. The space comprises X1, X2, X3, etc. The balls, each numbered y1, y2, y3, include a well-known combination of balls ... In that space the genius picks an urn and draws the urn randomly. They placed instead the ball on a transport line, which enables the spectator to track the ball's series, but not the urn series from which it was taken. The genie has a certain system for selecting urns; the urn choice for the n-th ball just relies on a random number and the ballot preference for the (n-1)-th ball. The choice of urn does not depend directly on the urns selected before the last urn; this is therefore referred to as the markov procedure.

$$Y = y(0), y(1), \dots, y(L-1) \quad (9)$$

$$P(Y) = \sum_X P(Y \mid X) P(X), \quad (10)$$

$$X = x(0), x(1), \dots, x(L-1). \quad (11)$$

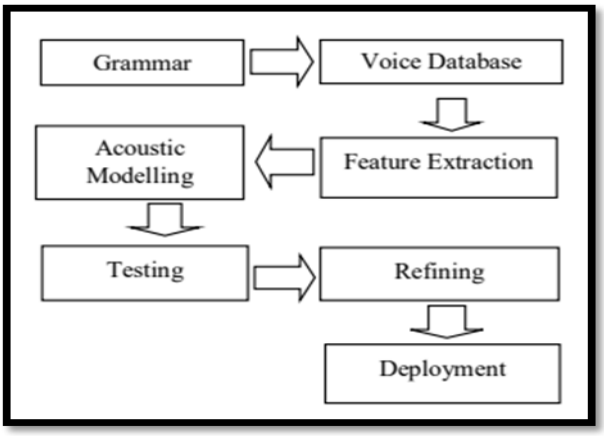


Figure 3: Process steps system diagram

The goal is to convert human speech computer signals into text or written types that are important to other grammars. Automated voice recognition system, or auto voice recognition system (ASR). The microphone recorded and interpreted the human voice signal to produce a standard document. Therefore, ASR may also be used for other uses, for example by merely providing a code order to execute a series of tasks. For e.g. someone at home asks the machine to shut the bathroom lights off and shut on without the electric switch needing to press sufficiently to send a signal. Another instance of ASR is when someone drives a vehicle and needs to notify their smartphones about the current position or the destination route details or even about the electric switching situation in the building.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad (12)$$

The ASR is divided into tiny ASR words (SVASR) as well as broad ASR vocabulary (LVASR) depending on the amount of terms endorsed. Compared to LVASR, SVASR recognizes only a few terms. An indication of SVASR is when electric switches are turned on at home. Another indication is a computer recognition program or guidance to dial automated calls from a telephone book touch.

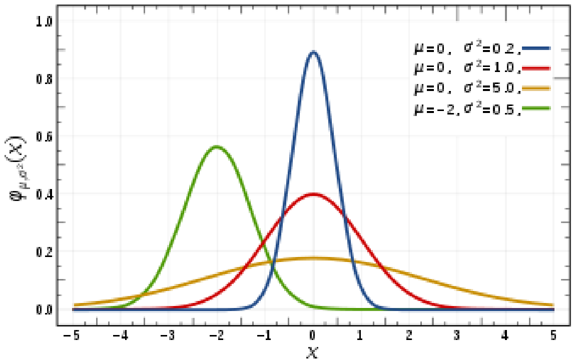


Figure 4: Graphical Representation 1

Vowels and vowel diacritics with k

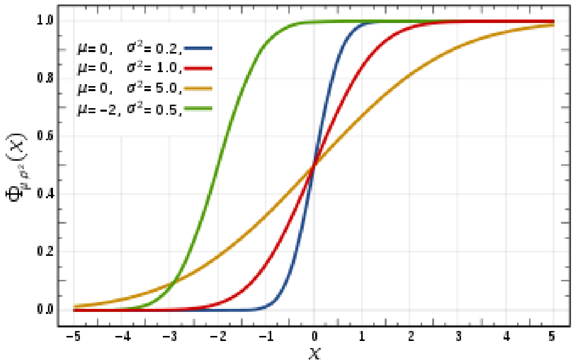
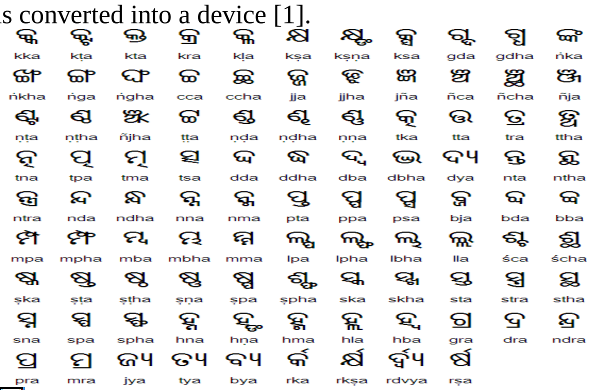


Figure 5: Consonats rate

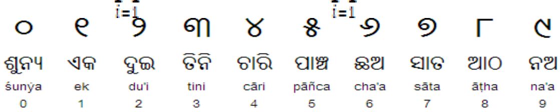
LVASR is used in comparison as a keyboard alternative for typing search engines with subjects and infinite terms normally. The automated translation system between two separate languages, for example between English and Indonesian and vice versa, may also be used by LVASR. There is no need to learn the language of the interlocutor in this situation for the speaker with an LVASR (smartphone) as it is converted into a device [1].



$$\sum_{\text{term in doc}} P(\text{term}) = 1 \qquad (13)$$

$$P(\text{query}) = \prod_{\text{term in query}} P(\text{term}) \qquad (14)$$

$$P(w_1,\dots,w_m) = \prod_{i=1}^{m} P(w_i \mid w_1,\dots,w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)},\dots,w_{i-1})$$



The language ODIA is formed from the alphabet Kalong, one of the several descendants of the ancient India's Brahmin language. The first documented event ... ODIA consonant letters have an intrinsic vowel, as in many abide scripts. ODIA script includes 35 consonants and 11 vowels.

IV. ODIA-ASR (O-ASR)

The microphone recorded and interpreted the human voice signal to produce a standard document. Therefore, ASR may also be used for other uses, for example by merely providing a code order to execute a series of tasks. For e.g. someone at home asks the machine to shut the bathroom lights off and shut on without the electric switch needing to press sufficiently to send a signal.

A. *System Description.*

The O-ASR Architecture is implemented using Hidden Markov Model Toolkit (HTK) 3.4[8]. For O-ASR, the Ubuntu 11.10 version Linux has been used. A graphical user interface is structured to render the software more user-oriented, intuitive and responsive relative to the Java Environment. In 205 different ODIA words, the system is educated and the word pattern is used for recognition.

$$H(x_1,x_2,\ldots) = \sum_{s} V(s) \quad (15)$$

$$\langle f \rangle = \sum_{x_i} f(x_1,x_2,\ldots)P(x_1,x_2,\ldots)$$

$$= \frac{1}{Z(\beta)}\sum_{x_i} f(x_1,x_2,\ldots)\exp\left(-\beta H(x_1,x_2,\ldots)\right) \quad (16)$$

B. *System Architecture & Implementation.*

The O-ASR device design consists primarily of four parts, i.e. data preparation for testing, sound processing, Acoustic model creation and Interface decoder.

$$\frac{\partial}{\partial \beta_k}(-\log Z) = \langle H_k \rangle = \mathrm{E}[H_k] \quad (17)$$

C. *Training Data Preparation.*

In this step the speech signal is registered and labelled. 205 distinct terms in ODIA are educated in the applied framework. The data is captured using a unidirectional microphone with the aid of a .wav recording device. The recorded.wav files are stored in HTK format.

The sample rate of transmission is 16 kHz. The data has been documented by eight speakers, and any phrase is said in a data file three times, and three samples are taken by each voice. Thus samples (205 * 3) of eight different speaker files, totaling 4920 (205 * 3 * 8), were given for the speaker archives. The waveforms of speech are identified with a Wave Suffer labeling device. The composition of the labeling includes seven consecutive regions as the phrase is spoken in a script three times: beginning from silence, documenting term, recording of silence and finishing from silence. The called .lab file is a simple text file used in acoustic models software development.

D. *Acoustic Analysis*

The devices for speech recognition cannot manage waveforms directly. They will be more streamlined and effective in their representation. This phase is referred to as acoustic research. In a variety of acoustic dimensions, the original waveform can be translated. The method Met Frequency Campestral Coefficient (MFCC) was used for extracting functions. Framing: the signal is separated into successive frames (usually about 20ms and 40ms) with overlap about them, which are used as a computational phase for the MFCC.

$$Z(\beta) = \sum_{x_i} \exp\left(-\sum_k \beta_k H_k(x_i)\right) \quad (18)$$

$$Z(\beta) = \mathrm{tr}\left[\exp\left(-\sum_k \beta_k H_k\right)\right] \quad (19)$$

Windowing: A windowing (for example hamming) method multiplies each frame by extracting: each window is fitted with an acoustic coefficient vector (which provides a compact representation of the frame spectral properties). Settings file (.cone) is a text file defining different parameters such as speech file format (STF), function extraction technique (MFCC), time frame period (25ms), frequency interval (10ms), amount of MFCCs (12), etc. Throughout both preparation and the decoding process of the program, Acoustic Vector (.mock) files are used. For this reason, the H copy HTK device is used.

V. ACOUSTIC MODEL GENERATION.

An acoustic model is characterized as a model reference to which unknown expressions can be described by comparisons. Two acoustic versions are usable, respectively. Create term and pattern phoneme. Term model was used because it is appropriate for the usage of limited vocabulary and the predictive method to device testing Secret Markov Modeling (HMM).

The first HMM initialization is carried out using a sample during this step of deployment. For each term in the dictionary, this prototype must be produced. The same topology applies to all HMM's, and the established topology is comprised of 4 active (OB) and two non-emitting (Initial) states with no OB). Single Gaussian diagonal matrix distributions are used as observer functions and are defined in a text description file called the prototype as vector and variance. The HTK Hint method is used for initialization of this predefined prototype along with acoustic vector (.mock files) and testing labels (.lab files).

$$P\left(w_i \mid w_{i-(n-1)},\ldots,w_{i-1}\right) = \frac{\mathrm{count}\left(w_{i-(n-1)},\ldots,w_{i-1},w_i\right)}{\mathrm{count}\left(w_{i-(n-1)},\ldots,w_{i-1}\right)} \quad (20)$$

$$P(w_m|w_1,\ldots,w_{m-1}) = \frac{1}{Z(w_1,\ldots,w_{m-1})}\exp\left(a^T f(w_1,\ldots,w_m)\right) \quad (21)$$

The HTK method H Rest is used for the calculation of optimum HMM parameters (transitional likelihood, mean and variance vectors for each measurement function) in the second stage of its implementation. Such iterative step is called a re-evaluation and replicated with each HMM. It's done a lot. This re-estimates convergence of displays by change (convergence factor).

$$P(w_t|w_{t-k},\ldots,w_{t-1}) \quad (22)$$

$$P(w_t|w_{t-k},\ldots,w_{t-1},w_{t+1},\ldots,w_{t+k}) \quad (23)$$

The final step, known as a convergence test, of the creation of the acoustic model is repeated until the absolute convergence factor value does not decline from one

H rest iteration to another. Throughout our system implementation the re-estimation method is repeated five times. Therefore, five HMMs in the vocabulary are generated by word.

A. *Task Definition.*

The basic design of the recognizer, i.e. the vocabulary model, and ODIA word dictionary, before reaching the final stage of the testing of the built framework. The model for prediction is to be described (task dictionary). The Job grammar is entered in a text file, which is defined in an expanded Backus near (EBNF). The job grammar is assembled with HTK HP ars tool to create the.slf job network.

A text file task language establishes the connection between the name of the HMM and the name of the grammatical element. The names of the signs are also incorporated in the above correspondence, since such names signify the signals the receiver of acknowledgement will make. The grammar variables name is dealt with automatically, if not provided, by default for performance purposes.

$$\frac{1}{T}\sum_{t=1}^{T} \sum_{-k\leq j\leq k, j\neq 0} \log P(w_{t+j}|w_t) \tag{24}$$

Planting takes place on the same noise conditions during the evaluation and training levels on machines with Windows operating systems. The planting app is from HTK. HVite. The speaker frequently and arbitrarily pronounces the numbers via the Bluetooth microphone. The speech based on the qualified acoustic HMM-GMM model is registered and recognized by HVite. The documents kept will be utilized where the fault happens to optimize the program.

$$Z(\beta) = \sum_{x_i} \exp(-\beta H(x_1, x_2, \ldots)) \tag{25}$$

$$Z(\beta) = \exp \int (-\beta H(x_1, x_2, \ldots))\, dx_1 dx_2 \cdots \tag{26}$$

A criterion for validating research findings is used when referring to a real-life program as the operator of the electric switch for domestic electrical appliances. HTK combines this function with a frame likelihood parameter. A order is performed in the upper layer if only the frame likelihood meets the threshold. The instruction is ignored if the value is below the threshold.

B. *System Testing*

This stage is used by HTK device H Viet for the production of outpost [9] to produce a transcript of the Unknown Utterance. The same is accurate for a testing corpus ready and is also used to translate the test signal in sequence of acoustic vectors (.mock) with the usage of HTK device H Copy. The method H Viet uses the token transmitting algorithm Viterbi Algorithm to process the signal that coincides with the Markov model of the recognizer. A filtering module is then used to process the transcription file, which removes the known term from the file and displays it in text form.

$$Z(\beta) = \mathrm{tr}(\exp(-\beta H)) \tag{27}$$

$$Z = \mathcal{D}\varphi \exp \int (-\beta H[\varphi]) \tag{28}$$

C. *Performance Analysis*

To make the software easier and more user friendly it has a graphical user interface with two buttons of recording and display. The user only clicks on the capture and takes up the microphone sound signal. Only after clicking on the view icon can the identified data be shown. As demonstrated. This can be done through two Java system shell files. The HTK H Copy and H Viet tools in the first shell script are disabled, and its output is then filtered by the 2nd shell script instruction. The added approach is therefore easier and more abstract

VI. CONCLUSION

The generated transcription files in the HVite method are linked to the corresponding initial reference transcription files. The HTK method review of machine results is carried out using HTK. The equations below demonstrate that N is the number of words in the analysis, D is the number of removals, s is the number of substitutions and I is the number of insertions.

Right percentage (PC) = (N-D-S)/N to 100 Where the PC defines a term error number in above-equation.

Percentage Precision (PA) = (N-D-S-1)/N to100 Where PA is set at the term accuracy point in the above equation.

Word Error Rate (WER) = 100% accuracy ratio.

Where the term error rate (WER) in the above equation is among the parameters for deciding unit output.

REFERENCES

1. R. Rabiner,and B.H. Juang, Fundamentals of Speech Recognition, PHI,New Jersey,1993.
2. R K Aggarwal, and M.Dave "Acoustic modeling problem for automatic speech recognition system: conventional methods (Part I) International journal speech Technology, Springer, Vol.14, issue 2, 2011.
3. R. Rainer, "A Tutorial on Hidden Markov Models and Selected Applications in Speech recognition", Proc. Of the IEEE Vol.77, Issue 2, pp.257-286, 1989.
4. R. Rabiner, and B.H. Huang, "An introduction to hidden markov models," IEEE Acoust. Speech Signal Processing Mag., pp.4-16, 1986.
5. K. Kumar, R K. Agarwal, and A.Jain "A Hindi speech recognition system for connected words using HTK" International Journal Computational Systems Engineering, Vol. 1, 2012.
6. R. Kumar "Comparison of HMM and DTW for Isolated Word Recognition of Punjabi Language" In Proceedings of Progress in pattern Recognition, Image Analysis, Computer Vision, and Applications, Sao Paulo, Brazil. Vol. 6419 of Lecture Notes in Computer Science (LNCS), pp.244-252, Springer Verilog, November 8-11, 2010.
7. R. Kumar and M. Singh. "Spoken isolated word Recognition of Punjabi Language using dynamic time Warping Technique" Demo in Proceedings of Information System for Indian Languages, Punjabi University, Patiala, India, March 9-11,2011.Vol. 139 of communication in computer and Information Science (CCIS),Page 301,Springer Verilog.
8. HTK "Hidden Markov Model Toolkit', available at http://htk.eng.cam.ac.uk,2012.
9. K. John and A. W. Black, "The CMU ARCTIC Speech Databases," in 5th ICSA Speech Synthesis Workshop - Pittsburg, 2004, pp. 223–224.
10. S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. Acoust., vol. 28, no. 4, pp. 357–366, 1980.
11. K. Kumar, R.K. Aggarwal "Hindi Speech Recognition System using HTK" International journal of Computing and Business Research ISSN Vol.2 issue 2 May 2011
12. Kacur, J., Rozinaj, G Practical Issues of Building Robust HMM Models Using HTK and SPHINX Systems. In: Mihelic, F., Zibert, J. (eds.) Speech Recognition, Technologies and Applications. I-Tech, pp. 171–192 (2008) Google Scholar

13. Sovilj-Nikić, S., Delić, V., Sovilj-Nikić, I., Marković, M Tree-based Phone Duration Modeling of the Serbian Language. Electronics and Electrical Engineering (Elektronika ir Elektrotechnika) 20(3), 77–82 (2014)

AUTHORS PROFILE

Priyabrata Sahu received the Master of Technology degree in Computer Science from Fakir Mohan University, Odisha, India. He is currently working as an Assistant Professor in the Department of Computer Science Engineering and Applications, Indira Gandhi Institute of Technology, Odisha, and India. He has more than 11 years of experience in teaching and learning. His research interest includes Artificial Intelligence, Natural Language Processing, and Data Science. He is a professional member of ISTE, IE.

Sunil Ku Panigrahi received the Master of Technology degree in Computer Science from Biju Pattnaik University of Technology, Odisha, India. He is currently doing his research work at KIIT University, Odisha, and India. He has more than 14 years of Teaching and learning experience. His research interest includes Natural Language Processing, Big Data Analytics, and Artificial Intelligence

Umakant Bhaskar Gohatre is graduated in BE in Electronics and Communication Engineering, and Master in VLSI and Embedded System Design and is working as faculty of Engineering and Technology since 7 years under Engineering college of University of Mumbai affiliated colleges. His area of Research is Mathematical Computational Models, Machine Learning, Image Processing, and Computer Vision. Embedded system, Big Data. He has published more than 30 Research papers in high indexing journals, and Present Research work more than 26 National and International Conferences.