IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

An Efficient Implementation of Speech Recognition based on Curvelet Transform and Artificial Neural Network

Nidamanuru Srinivasa Rao[1*], Chinta Anuradha[2], SV Naga Sreenivasu[3]

[1]Dept. of CSE, Acharya Nagarjuna University, Guntur, India
[2]Dept. of CSE, VR Siddhartha Engineering College, Vijayawada, India
[3]Dept. of CS, Narasaraopeta, Guntur, India

[*]*Corresponding Author: rao_75nidamanuru@rediffmail.com, Tel.: 919290081612*

Available online at: www.ijcseonline.org

Received: 20/Mar/2018, Revised: 28/Mar/2018, Accepted: 19/Apr/2018, Published: 30/Apr/2018

***Abstract* -** Speech Recognition is ability to translate a dictation or spoken words to text format. In the field of electronics and computers, speech has not been used much more due to the complexity and different types of sounds and speech signals. However, with traditional methods, processes and algorithms, we can simply process the speech signals and identify the text. This paper presents an efficient speech recognition system based on discrete curvelet transform (DCT) and Artificial Neural Network (ANN) methods to enhance the identification rate. This research article comprised in two distinct phases, a feature extractor and a recognizer is presented. In Feature Extraction phase, Curvelet transform extract the features called curvelets from the given input speech signal and elements of these signals which support in gaining higher recognition rates. For feature matching, Artificial Neural Networks is used as classifiers. The performance evaluation has been demonstrated in terms of accurate recognition rate, maximum noise power of interfering sounds, miss rates, hit rates, and false alarm rate. The accurate classification rate was 98.3 % for the sample speech signals. Performance comparisons with similar studies found in the related literature indicated that our proposed ANN structures yield satisfactory results and improve the recognition rates.

***Keywords*—**Speech Recognitiov, Curvelet Transform, Feature Extraction, Artificial Neural Network.

I. INTRODUCTION

Speech is the most efficient mode of communication between peoples. This, being the best way of communication, could also be a useful interface to communicate with machines. Past few decades, developers have analysis speech for an extensive of applications ranging from mobile transportation to automatic evaluation machines. Speech recognition reduces the transparency caused by exchange of communication techniques. Speech signal communication not used a large amount in the field of computers and electronics because of the complexity. However, with current methods and algorithms can process the speech signals simply and identify the text. There are dissimilar ways to speech recognition in Artificial Neural Networks, Hidden Markov Model (HMM), Vector Quantization (VQ), support vector machine etc. This paper presents an authority speech recognition system based on discrete curvelet Transform (DCT) and ANN methods to enhance the identification rate.

The Research article is organized as follow: Section II provide some essential background and a summary of related work in speech identification and neural networks [1]. Section III states the Mathematical Formulation of speech recognition. Section IV presents our research with classification network. In Section V, compares the performance of our optimized systems against many other systems. Finally, Section VI describes the conclusions of this paper and suggestions for future work.

II. RELATED THEORY

Speech recognition is the process of extracting and determining information conveyed by a speech signal using computers. Speech recognition is currently used in many real time applications, such as mobile phones, smart devices, computers, and security systems. However, these systems are far from perfect in correctly classifying human speech into words. Speech recognizers consist of a feature extraction stage and a classification stage. The feature extraction can be considered as a dimensionality reduction process that attempts to capture the essential characteristics of the speech with less memory requirements for signal representation.

Speech recognition is a charming appliance of digital signal processing in the real-world applications. It is still a rising area and carries tough potential in the near future as computing power enhance. Speech recognition process needs deep processing due to huge samples per window. The growth of techniques to signal processing in the lack of

models lead to queries for the processes of making signals using ANN.These techniques recognize stationary signals within a given time and lack of capability to process localized proceedings correctly. Curvelet analysis has been confirmed as efficient signal computing techniques for a different signal processing issues.

Many different methods, algorithms, and mathematical models developed to help speech analysis and speech recognition. This section points out advances and techniques that have been and are being applied to the speech recognition process. Thiang, et al. offered speech identification using Linear Predictive Coding and Artificial Neural Network for domineering movement of mobile robot [6]. Akkas Ali et al. described automatic speech recognition method for Bangla words. He can be feature extraction was done by LPC and Gaussian Mixture Model. Hindered words recorded in thousand times which gave 84% accuracy [8]. Ms.Vimala et al. proposed speaker independent isolated speech detection system for Tamil language. Multiple designs like Feature extraction, acoustic model and pronunciation dictionary were applied using HMM which produced 88% of accuracy in 2500 words [9].

Cini Kurian et al. developed diverse acoustic models for Malayalam continuous speech recognition. HMM is used to compare and assess the Context Dependent, Context Independent models and Context Dependent models from this Context Independent model gain 21%. Hemdal & Hughes et al. took the basis of finding speech sounds and providing labels their exist a flat number of distinctive phonetic units in spoken language which are generally characterized by a set of acoustics characteristics differ with respect to time in speech signal. Suma Swamy et al. proposed an efficient speech recognition mechanism experimented with Mel Frequency Cepstrum Coefficients (MFCC), Vector Quantization (VQ) and HMM, she recognize 98% recognize accuracy. Annu Choudhary et al. stated that automatic speech recognition mechanism for connected and isolated words of Hindi language by with the support of Hidden Markov Model Toolkit (HTK). Hindi words are used for extracted dataset by MFCC; the recognition system achieved 95% accuracy in remote words and 90% in associated words.

Preeti Saini et al. presented Hindi automatic speech recognition based on HTK. Isolated words are recognizing the speech with ten conditions in HMM architecture, it generated 96.61%. MdMaya Moneykumar et al. developed word identification for speech recognition scheme in Malayalam language. This work was done with syllable based fragmentation with the support of HMM on MFCC for feature extraction of speech signal. Jitendra Singh Pokhariya and Dr. Sanjay Mathur introduced Sanskrit speech recognition using HTK and MFCC for feature extraction, both generates 95.2% to 97.2% accuracy respectively. Geeta Nijhawan et al. stated a real time speech recognition mechanism for Hindi words. MFCC using Quantization Linde, Buzo and Gray (VQLBG) algorithm for feature extraction. Hauptman et al. proposed Voice Activity Detector (VAC) system to detect speech signal from closed-captioned television. The television information would be used in recognition system for training of speech recognition. Speech recognizing typically involves multiple models for pronunciations, acoustics and language. The acoustic model uses neural networks (NN) and / or Hidden Markov Models. These approaches require accurate training data, generated by the laborious process of humans listening to speech and typing the words.

III. MATHEMATICAL FORMULATION OF SPEECH RECOGNITION

Speech recognition is a multi level pattern recognition process; acoustical signals are tested and prepared into an order of sub word phrases, units, sentences and words. The continuous speech waveform is initially separated into frames with stable span [2]. Each frame gives features represented by discrete parameter vectors; assume that duration of a single vector, i.e. one frame of speech curve form can be considered as stationary. This is not severely accurate but it is approximation of speech recognition. For each spoken word, let O = o1o2……oτ be a sequence of parameters vectors, where ot is at time and t∈ {1, … … . . , τ}. Given a dictionary D with words yi ∈ D, the recognition problem is summarized by

$$\widetilde{W} = \arg_{w}^{\max} P(Y|J) \qquad (1)$$

Where $\widetilde{Y}$ is the recognized word, P is the probability measure and Y = y1 ……yk a word sequence. Bayes' Rule permits to transform (1) in a suitable calculable form:

$$P(Y|J) = \frac{P(J|Y)P(Y)}{P(O)} \qquad (2)$$

Where P(O|W) represents the acoustic model and P(Y) the language model; P(J) can be unheeded. The set of prior probabilities P(Y), the majority of spoken word probably based on the likelihood P(J| yi). The combination of the acoustic probability and language probability model is weighted. So the language model is scaled by an empirically represented constant s, called language model scale factor (LMSF). LMSF is represented empirically to get best performance on recognition. This weighting has a side effect as a penalty for inserting new words. We add a scaling factor p word insertions called word insertion penalty (WIP) also calculated empirically. Thus equation one becomes

$$\tilde{Y} = \arg_{w}^{max} P(Y|J)\, P(Y) s|Y|^{p} \quad (3)$$

In the log domain, the total likelihood is calculated as

$$\log_{|y|}\ \tilde{Y} = \log_{|y|} P\ (J|Y) + s \log_{|Y|} P(Y) + p \quad (4)$$

Where |Y| the length of the word sequence Y, s is the language model scale factor (LMSF) and p, word insertion penalty. Global posterior probability is P(N|J, θ) such that N is the model given the acoustic data J and the parameters θ. Express the possibility of global posterior probability in terms of local posteriors $P(p_l^n | q_k^{n-1}, j_n, \theta)$ (where q_k^n denotes the specific state qk of N at time n) and language model priors. We have

$$P(N|O) = \sum_{l_1}^{L} \sum_{l_N}^{L} P(q_{l_1}^1, \dots\dots, q_{l_N}^N, N|J) \quad (5)$$

Where the posterior probability of the condition in sequence and modal can be fragmented into the multiplication of an acoustic model over language models:

$$P(q_{l_1}^1, \dots\dots\dots, q_{l_N}^N, N|O) = P(q_{l_1}^1, \dots\dots\dots, q_{l_N}^N|J)$$

$$P(N|J, q_{l_1}^1, \dots\dots\dots, q_{l_N}^N) \quad (6)$$

$$\simeq P(q_{l_1}^1, \dots\dots, q_{l_N}^N|J)$$

$$P(N|\, q_{l_1}^1, \dots\dots, q_{l_N}^N) \quad (7)$$

After Rewriting the above equation (7) becomes

$$P(N|J) \simeq \sum_{l_1,\dots,l_N} [\prod_{n=1}^{N} P(q_{l_n}^n |\, J_{n-c}^{n+d}) \frac{P(q_{l_1}^1|N)}{q_{l_n}^n}]\ P(N) \quad (8)$$

The posterior probability, where O_{n-c}^{n+d} is limited to local Context. With the Bayes rule, we can show that:

$$\frac{P(J_{n-c}^{n+d}|q_{l_n}^n)}{P(\ J_{n-c}^{n+d}\ )} = \frac{P(q_{l_n}^n|J_{n-c}^{n+d})}{P(q_{l_n}^n)} \quad (9)$$

Then we also have:

$$P(N|J) \simeq \sum_{l_1,\dots\dots l_N} [\prod_{n=1}^{N} P(J_{n-c}^{n+d}|q_{l_n}^n) \frac{P(q_{l_1}^1|N)}{P(\ O_{n-c}^{n+d}\ )}\ ]\ P(N) \quad (10)$$

The distinction between the hybrid and the likelihood approaches lies at the local level. The hybrid system estimates local posteriors and is then discriminate at the frame level. The likelihood system estimates local probability density functions. Both systems can give us an estimate of the global posterior.

IV. RESEARCH FRAMEWORK

4.1 Methodology

The curvelet transform is a multi scale directional transform that accept a best possible non-adaptive sparse demonstration of object, edges and curves. The time-frequency and multi-resolution property of curvelet transform for input speech signal, it is decomposed into different channels in frequency [3]. The breakdown procedure can be iterated with successive manner, so that single signal is decomposed into a lot of lower resolution components. The choice of the best decomposition phase hierarchy based on the nature way of signal analysis like low-pass filter range [4]. It is efficient technique for extracting non-stationary signals data. The extracted curvelet coefficients offer a solid version of power distribution in time and frequency of the signal. The following curvelet based speech recognition diagram represented in figure:

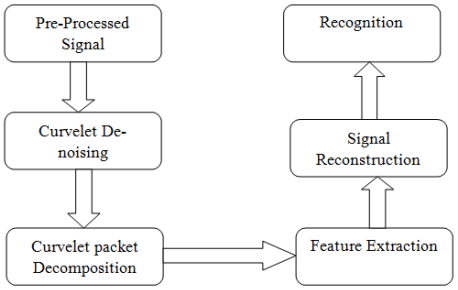

Figure 1. Curvelet Feature Extraction phases

4.2 Pre-processing

Speech recognition systems in this phase are used as subsequent feature extraction with increase efficiency and classification. Finally enhance the performance of speech signal recognition. At the end of this process, filtered and compressed frames of speech are forwarded to the feature extraction phase [5]. The following figure 2 represents the pre processing pipeline of speech signal.

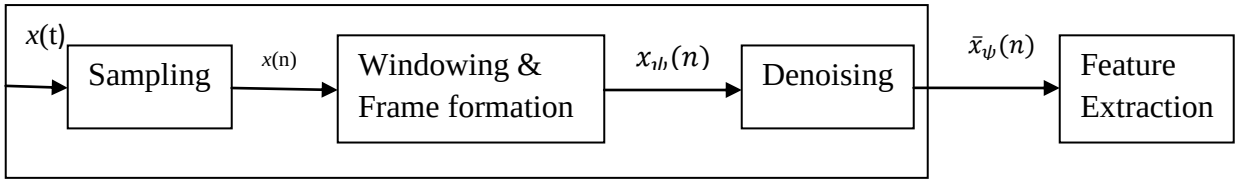

Figure 2. pre processing pipeline of speech signal

4.3 Curvelet Function

Construction of curvelet functions; initially require defining special window functions that fulfil certain acceptability conditions. Let us regard as the scaled Meyer windows:

$$U(t) = \begin{cases} 1 & |t| \le 1/3 \\ \cos[\frac{\pi}{2} v(3|t|-1)] & 1/3 \le |t| \le 2/3 \\ 0 & \text{else,} \end{cases}$$

$$Y(r) = \begin{cases} \cos[\frac{\pi}{2} v(5-6r)] & 2/3 \le r \le 5/6, \\ 1 & 5/6 \le r \le 4/3, \\ \cos[\frac{\pi}{2} v(3r-4)] & 4/3 \le r \le 5/3, \\ 0 & \text{else} \end{cases}$$

Where *v* is a smooth function satisfying

$$U(x) = \begin{cases} 0 & x \le 0, \\ v(x) + v(1-x) = 1, & x \in \mathbb{R} \\ 1 & x \ge 1, \end{cases}$$

An arbitrarily smooth window v is given by

$$U(x) = \begin{cases} 0 & x \le 0 \\ \frac{s(x-1)}{s(x-1)+s(x)} & 0 < x < 1 \\ 1 & x \ge 1 \end{cases}$$

The above two functions *V* (*t*) and *W*(*r*) satisfy the conditions

$$\sum_{l=-\infty}^{\infty} U^2(t-l) = 1, \quad t \in \mathbb{R} \quad (11)$$

$$\sum_{j=-\infty}^{\infty} Y^2(2^j r) = 1, \quad r \in 0 \quad (12)$$

4.4 Feature Classifier

Artificial Neural Networks are processing models stimulated by the brain. Machine learning and pattern recognition are capable of these models. It was working to recognize and classify vowel signals into their respective class. An artificial neuron with k given inputs converting a set $X \subset \mathbb{R}^k$ of input signals (a k-neuron on X) is a function.

$$E: \mathbb{R}^k \times X \ni (\vec{y}, \vec{x}) \to F(\vec{y}, \vec{x}) = f(\vec{y}, \vec{x}) \in \mathbb{R},$$

Where $\vec{w}$ is a weights vector, $\langle .,. \rangle$ is a actual scalar result, and f: $\mathbb{R} \to \mathbb{R}$ is called an activation function of the neuron. If f is a linear operator, then the neuron is called linear. A function

$$E^*:= E(\vec{y}, .) : X \ni \vec{x} \to E^*(\vec{x}) \in \mathbb{R},$$ is said to be a trained k-neuron on X.

The least-squares method is used for analysis of the learning process of a linear ANN. The properties of Gram matrices will be used to analyse linear training processes of ANN analysis. Consider an ANN consisting of a single linear M-neuron. It is sufficient to accept a neuron and identity the activation function. In this case, the square deviation function is given by:

With $s(x) = e^{-\frac{1}{(1+x)^2} + \frac{1}{(1+x)^2}}$

$$F(Y_1, \dots\dots\dots\dots, 1_N) = \sum_{n=1}^{n} [y(y_1, \dots\dots\dots\dots y_N)]^{(n)} - Z^n)^2$$

Where y ($w_1, \dots \dots \dots, w_M$) $^{(n)} = \sum_{m=1}^{M} x_m^{(n)} w_m$. Right side of the equation value calculates for the describing the learning process:

$$\frac{\partial F(w_1,\dots\dots 1_M)}{\partial w_{m'}} = \frac{\partial}{\partial w_{m'}} \sum_{n=1}^{N} [\sum_{m=1}^{N} x_m^{(n)} y_m - z^{(n)}]^2$$

Setting $H^{(n)} := \sum_{m=1}^{N} x_m^{(n)} y_m - z^{(n)}$,We obtain:

$$\frac{\partial E(w_1,\dots\dots 1_M)}{\partial w_{m'}} = \sum_{n=1}^{N} 2.H^{(n)} \cdot \frac{\partial (y^{(n)} - z^{(n)})}{\partial w_{m'}} = 2. \sum_{n=1}^{N} H^{(n)}. \frac{\partial y^{(n)}}{\partial w_{m'}} = 2. \sum_{n=1}^{N} H^{(n)}. \frac{\partial (\sum_{m=1}^{M} x_m^{(n)} m}{\partial w_{m'}} = 2 \sum_{n=1}^{N} H^{(n)} \cdot x_{m'}^{n} = 2. \sum_{n=1}^{N} x_{m'}^{n} \cdot [(\sum_{m=1}^{M} x_m^{(n)} w_m) - Z^{(n)}]$$

A linear one-layer ANN learning process is described by the system of T differential equations which are independent of each other. Each of them models the learning process of a single neuron. Indeed:

$$\frac{\partial E}{\partial w_{t'm'}} = \frac{\partial}{\partial w_{t'm'}} \sum_{n=1}^{N} \sum_{t=1}^{T} [(\sum_{m=1}^{M} x_m^{(n)} w_{t,m}) - z_t^{(n)}]^2 = \frac{\partial}{\partial w_{t'm'}} \{ \sum_{n=1}^{N} \{ \sum_{t=1,t\neq t'}^{T} [(\sum_{m=1}^{M} x_m^{(n)} w_{t,m}) - z_t^{(n)}]^2 \} + \frac{\partial}{\partial w_{t'm'}} \sum_{n=1}^{N} [(\sum_{m=1}^{M} x_m^{(n)} w_{t,m} - z_t^{(n)}]^2$$

Since the first component does not depend on $w_{t',m'}$,it is equal to zero. Thus:

$$\frac{\partial F}{\partial w_{t'm'}} = \frac{\partial}{\partial w_{t'm'}} \sum_{n=1}^{N} [(\sum_{m=1}^{N} x_m^{(n)} w_{t,m}) - z_t^{(n)}]^2 = 2. \sum_{n=1}^{M} x_{m'}^{n} [(\sum_{m=1}^{N} x_m^{(n)} w_{t',m}) - z_{t'}^{(n)}]$$

The unique signal can be synthesized using the inverse discrete curvelet transform (IDCT).

V. RESULTS AND DISCUSSION

Speech is typically sampled at a high frequency, this yield a sequence of amplitude values over time. Raw speech should be initially transformed and compressed, in order to simplify subsequent processing. Many signal analysis techniques are available which can extract useful features and compress the data by a factor of ten without losing any important information. Among the most popular is discrete curvelet transform. The result of signal analysis is a sequence of speech frames, these frames may be augmented by their own first and/or second derivatives, providing explicit information about speech dynamics; this typically leads to improved performance. The speech frames are used for acoustic analysis. In order to analyze the speech frames for their acoustic content, we need a set of acoustic models. There are many kinds of acoustic models, varying in their representation, granularity, context dependence, and other properties. Acoustic analysis is performed by applying each acoustic model over each frame of speech, yielding a matrix of frame scores. Frame scores are converted to a word sequence by identifying a sequence of acoustic models, representing a valid word sequence, which gives the best total score along an alignment path through the matrix. The end result of time alignment is a word sequence.

In speech processing, feature extraction is one of the major level to enhance speech processing applications in real world. A bulky set of feature extraction techniques is existing to apply on speech processing paths; however the division through curvelets is one of the popular techniques in these days for robustness. This research paper presents the progress and execution of curvelet transform technique using samples speech signals. The training and testing data set levels are recorded by using a PC-based audio input device. For resolution purposes the recoded samples are stored in matrices whose rows or columns represent as sample. The speech signal processing and analysis is done according to Section III, and the outcome weighted cepstral coefficients are the input to the pattern classifier. The following Table one shows the results of training and testing datasets using standard performance measures and defined as follows:

Table 1. Evaluated standard performance measures

S. No	Parameter	Training	Testing
1	Sensitivity	100%	100%
2	Specificity	100%	100%
3	Accuracy	100%	97.3%

Sensitivity = $\frac{Tp}{Tp+Fn}$, and

Accuracy = $\frac{Tp+Tn}{Tp+Tn+Fn+Fp}$

Specificity = $\frac{Tn}{Tn+Fp}$ and

Sensitivity = $\frac{Tp}{Tp+Fp}$

The following Figure 3 Performance comparison of training and test datasets are evaluated using standard performance measures:

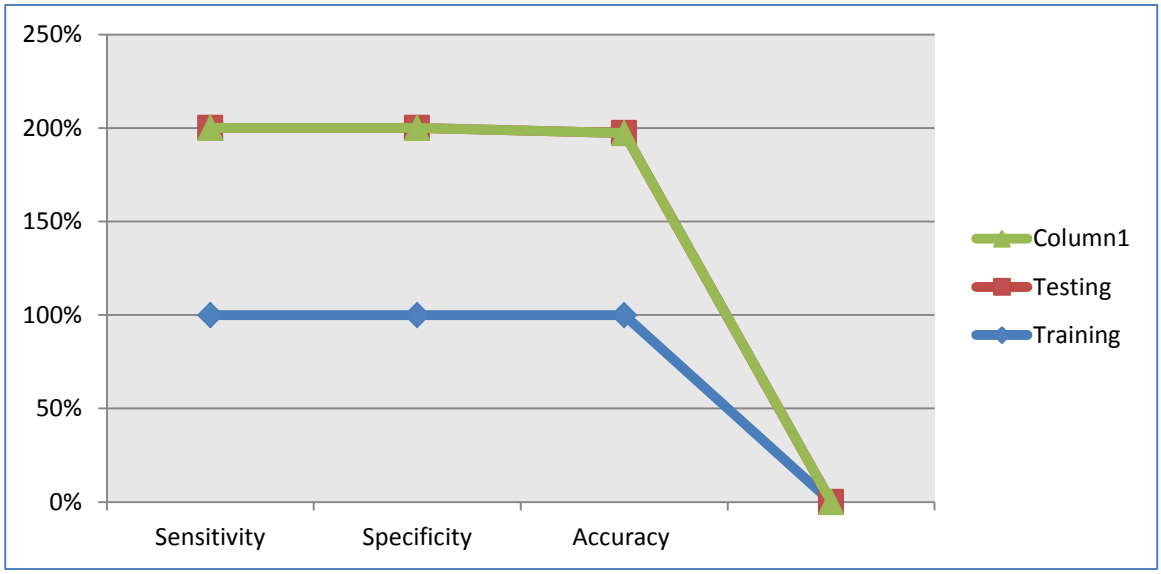

Figure 3. Performance comparison of training and test datasets

The results show a classification above 97.3%, which demonstrates the suitability of the method for recognition.

VI. CONCLUSION & FUTURE SCOPE

A speech recognition scheme requires solutions to the problems of both acoustic modelling and temporal modelling. However, modern algorithms and methods can process the speech signals simply and identify the text. This research concludes an expert speech recognition mechanism for isolated words based DCT and ANN methods was proposed. Artificial neural network play important role for classification of speech signal. They control similarly like human brain than conventional computer logic. This research article comprised in two distinct phases, they are feature extractor and recognizer. In Feature Extraction phase, Curvelet transform extract the features from the given input speech signal and elements of these signals which help to achieving higher recognition rate. Artificial neural networks are used as feature matching classifiers. The throughput evaluation has been confirmed in terms of accurate recognition rate, utmost noise power of interfering sounds, miss rates, hit rates, and false alarm rate. The rate of correct classification was about 98.3 % for the sample speech signals. The graphical results demonstrate that the proposed method can construct an accurate and robust classifier.The future of this technology is very promising and the whole key lies in hardware development as ANN need faster hardware.

REFERENCES

[1] B. Suksiri and M. Fukumoto," computer and information science", springer, Kochi University of Technology (KUT), Kami City, Japan, pp 15-26, 2016.

[2] http://recognize-speech.com/preprocessing.

[3] Thiang and S. Wijoyo," Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot", International Conference on Information and Electronics Engineering IPCSIT, Singapore, PP 121-131 , 2011.

[4] C. Kuriana, K. Balakrishnan, "Development & evaluation of different acoustic models for Malayalam continuous speech recognition", International Conference on Communication Technology and System Design,cochin , pp.1081-1088, 2011.

[5] Md. A. Ali , M. Hossain and Md. N. Bhuiyan," Automatic Speech Recognition Technique for Bangla Words", International Journal of Advanced Science and Technology, Vol. 50, 2013.

[6] V. Ca and V. Radhab," Speaker Independent Isolated Speech Recognition System for Tamil Language using HMM", International Conference on Communication Technology and System Design, Coimbatore, PP.1097 – 1102, 2012.

[7] P. Mishra and P. K. Mishra," A Study of various speech features and classifiers used in speaker identification", IJERT, vol. 5, issue 2, 2016.

[8] A. Choudhary, R.S. Chauhan and Gautam Gupta, "Automatic Speech Recognition System for Isolated & Connected Words of Hindi Language By Using Hidden Markov Model Toolkit (HTK)", in Proceedings of International Conference on Emerging Trends in Engineering and Technology,vol. 4, issue 6, pp.244– 252,2012.

[9] P. Saini, P. Kaur and M. Dua," Hindi Automatic Speech Recognition Using HTK", International Journal of Engineering Trends and Technology (IJETT) – Vol. 4, Issue **6**- June 2013.

[10] M. Moneykumar and E. Sherly," malayalam word identification for speech recognition system", An International Journal of Engineering Sciences, Special Issue iDravadian , Vol. 15, 2014.

[11] J. S. Pokhariya and S. Mathur, "Sanskrit Speech Recognition using Hidden Markov Model Toolkit", International Journal of Engineering Research & Technology (IJERT),Vol.3, Issue 10, pp.93-98, 2014.

[12] G. Nijhawan and Dr. M.K. Soni, "Real Time Speaker Recognition System for Hindi Words", International Journal of Information Engineering and Electronic Business, Vol. 6, pp. 35-40, 2014,.

Author Profiles

Mr. N srinivasaro received his B.Tech degree from JNTU Hyderabad, received his M.Tech degree from Acharya Nagarjuna University, Guntur, AP. His area of interests are Image processing, Computer Networks and Data Mining.

Ms. Ch. Anuradha received her B.Tech degree from Acharya Nagarjuna University, Guntur and received her M.Tech degree from JNTUK, Kakinada. She has 4 years of Industry experience and 5 years of techning experinece. She published 8 papers in National and Inter national journals. Her area of Interests are Image processing, Data Mining, Network Security.

Dr. SVN Naga Srinivasu received his Phd from Acharya Nagarjuna University, Guntur, Andhara Pradesh. Present he is working as a research supervisor at Nagarjuna University, Guntur. He published more number of papers in National and Inter National Journals. His area of Interests are Image Processing, Computer Networks, Software Engineering and Data Mining.

