RECURRENT MODEL FOR ACOUSTIC MODELLING IN SPEECH AND SPEAKER RECOGNITION

Santosh Kumar Pateriya[1], Prof. Rupesh Kumar[2]
[1]M.Tech Scholar, Dept of Electronics & Communication Engineering, AITR, Bhopal
[2]Assistant Professor, Dept of Electronics & Communication Engineering, AITR, Bhopal

ABSTRACT

speech is considered to be the most potent peerless & unrivaled method of articulating the information with one another in human beings than any other modes of communication available & as now electronics has evolved & flourished like anything the most powerful human-made innovation viz a personal computes are also perceived to cognize the speech & to make it a primary interface of interaction with PC's over extant available interface viz keyboard mouse etc. In order to achieve this adequacy, there comes an obligation of developing a system that can understand & interpret the speech. An Automatic Speech Recognition (A.S.R's) is one such effective tool enabling this aid of taking human speech as an input via a microphone or a receiving device & generating equivalent output as a response. It reflects its obligatory significance of becoming a versatile interface amid human & PC's & it can be obtained by ASR environment with ease. Having its various portent utilization, for example, commanding & controlling, speech-to-text, running installed application & stock programs of PC's, database querying, voice interpretation, dialect observation & other office relevant tasks. In current scenario due to a standard computer interface & its limited awareness has constrained its utility to more than half of the available population who are not proficient to operate & handle PCs. As humans find speech to be the most contended & straightforward mode of communication.similar mode is expected from a PC too to address with a user in its local dialects. & its realization to practicality will enable even a most common human being to efficiently avail oneself to extricate the true potential of modern PC's. our country India is a land of a diverse & wide class of various dialects & linguistics there is an essential exigency of a potential Automatic Speech Recognition (A.S.R.) environment into which local dialects can be housed. hence This paper condenses the ASR systems & methods of evaluating different phases of speech recognition. The application presentation deals with bringing out the importance of basic concepts of Signals and Systems in the field of speech communication and it's contribution to the technological advances.Speech processing typically involves a basic representation of a speech signal in a digital domain which requires limiting the band width of the signal, sampling it at a certain corresponding rate and storing each sample with an adequate resolution. But our focus in the field of speech processing is in communication. Speech can be represented in terms of a signal carrying some message content or information. Speech signals can be thought of signals in both continuous and discrete domain.

ABBRIEVIATION

SPECH RECOGNITION SYSTEM	SRS
DISCRETE COSINE TRANSFOR	(DCT)
SUPPORT VECTOR MACHINES	(SVM)
LINERA PREDICTIVE CODING	(LPC)
DISCRETE WAVE CHANGE	(DWT)
FAST FOURIER TRANSFORM	(FFT)
DIFFERENTIATE RELATIVE HIGHER ORDER AUTOCORRECTION SEQUENCES SPECTRUMS	(DRHOASS)
MEL FREQUENCY CEPSTRAL COEFFICIENT	(MFCC)
HIDDEN MARKOV DISPLAY	(HMM)
CONVLUTION NEURAL NETWORKS	(CNNS)
JOINT DENSITY GAUSSIAN MIXTURE MODEL	(JD-GMM)
DEEP NEURAL NETWORK	(DNN)

KeyWords: *Speech, Speaker Recognition, Automatic Speech Recognitions framework, neural.*

INTRODUCTION

Speech recognition is a process by which a computer takes an audio signal (recorded or input instantaneously) & converts them into words in real time it is achieved by utilizing systematic procedure of certain steps & software that helps to provide the output data which is a recognized speech interpreted by computer with speech recognition systems (SR) it is a process of decoding acoustic speech signal captured by microphone or telephone to a set of words in real time they are frequently deployed as a dictation software, intelligent personal assistant in personal computers, smartphones,web browsers, voice commanding/dialing, system controlling/navigation devices & gadgets designed specifically for physically challenged people & many more similar kind of applications while designing of a SR system many crucial factors are worth considering as it has to deal with numerous challenges like speakers voice accompanied by surrounding noise which affects the degree of accuracy in speech recognition.

Accent of speaking varies from person to person & it is a very big challenge A speaker may speak something very quickly & all of the words spoken have to be individually recognized accurately the speech An acoustic model is created by taking audio recordings of speech & their audio recordings of speech & their text transcriptions & using software to create statistical representations of the sounds that makes up each word It is used by a speech recognition engine to recognize speech the speech recognition are classified into two categories speaker dependent SR systems which work by learning the unique characteristics of a single person's voice depending on the speaker for training & speaker independent SR systems which are designed to recognize anyone's voice, so no training or learning g is required Accuracy is usually rated with WORD ERROR RATE (WER) whereas speed is measured with real-time factor other measures of accuracy include single WORD ERROR RATE (SWER) & COMMAND SUCCESS RATE (CSR). Factors affecting the accuracy of an SR system vocabulary size & confusability, speaker dependence vs independence, isolated, discontinued, or continuous speech, task & language constraints, Read vs spontaneous speech, adverse condition. Automatic Speech Recognition systems (ASR's) works on two approaches Templates based & statistics based which will be elaborated in further following sections in details.

LITERATURE REVIEW

RELATED WORK

Batista et al. [4] PSO is applied to locate the best information of each class (design) to be prepared by SVM and there is an analysis of the contrast between utilizing or not this streamlining. The digits of zero to nine in Brazilian Portuguese dialect are perceived consequently by SVM. Those digits are pre-prepared utilizing mel-cepstral coefficients and Discrete Cosine Transform (DCT) to create a two-dimensional framework utilized as contribution to the PSO calculation for producing the ideal information.

Zade et al. [5] associated with Support Vector Machines (SVM) to models of Speech Recognition Systems in outlook of MFCC and LPC highlights for Azerbaijani Datasets. This DataSet has been operated for speech recognition by Multi-layer Artificial Neural Network and accomplished a few outcomes. The basic intention of this work is applying SVM procedures to the Azerbaijan Speech Recognition Systems. The variety of consequences of SVM with various Kernel capacities is broke down in the preparation procedure. It is demonstrated that SVM with outspread premise and polynomial portions give better recognition comes about that Multi-layer Artificial Neural Network.

Kanisha and Ganeshan et al [6] In this work, from the info speech signal via perceiving the substance includes three phases, for example, the preprocessing, highlight extraction and Multi Support Vector Machine (SVM). The flag is prepared and clamor free flag is created by handling the flag and the highlights are extricated. For advance these highlights diverse enhancement calculations are used. From this calculation the ideal highlights, for example, top signal frequency, Tri-unearthly component, and discrete wave change (DWT) accomplish the APSO method. These ideal highlights are given as the contribution of the multi SVM and the flag in testing process, the flag accurately perceive the content. From the outcomes the enhancement algorithm (APSO) gets the 97.8% precision contrasted with the current system SVM direct portion work.

Dev et al. [7] profoundly wrangled on the issue of expanding the quality of speech front-closes and propelled an imaginative sequence of MFCC vector assessed by methods for three stages. In the main stage, the relative higher request auto-correlation coefficients were proficiently mined. From that point the extent array of the subsequent discourse flag was surveyed by methods for the quick Fourier change (FFT) and it was recognized as far as frequency. In the last stage, the recognized greatness range was transformed into MFCC-like coefficients, named as MFCCs mined from the Differentiate Relative Higher Order Autocorrelation Sequences Spectrums (DRHOASS).

Henawy et al [8] have recommended the purpose of speech recognitions deliver an instrument which will perceive precisely the ordinary human dialogue from any speaker. The credentials rate of 98% was acquired utilizing the proposed include extraction system. The features in light of the Cepstrum give exactness of 94% for speech recognitions while the features in light of the brief time series energy in time space give precision of 92%. The features in light of formants frequencies give exactness of 95.5%. Obviously the features in view of MFCCs with precision of 98% give the finest exactness rate. So the structures rely upon MFCCs with HMMs might be suggested for recognition of the communicated in Arabic digits.

Huang et al. [9] have proposed a powerful mechanism for Chinese speech recognitions on little vocabulary estimate was open speech recognitions of Chinese words in outlook of Hidden Markov Models. The qualities of speech words are fashioned by sub-syllables of Chinese characters. Add up to 640 speech tests are noted by 4 local guys and 4 females with as often as possible talking capacity. The preparatory consequences of inside and outsides testing accomplish 89.6% and 77.5%, individually. The last accuracy rates for inside and outsides test in normal accomplishes 92.7% and 83.8%. The outcomes revealed that the methodologies for Chinese speech recognitions on little vocabulary are viable.

Poonkuzhali et al [10] have proposed the Speech was a standout amongst the majority of encouraging models by which individuals can express their moods like outrage, pity, and joy. Acoustic parameters of a speech signal like energy, pitch, Mel Frequency Cepstral Coefficient (MFCC) was essential in discover the condition of a man. The features get decreased to 16.6% of every 300 emphases. Subterranean insect Colony Optimization can choose the more instructive highlights without losing the execution.

Selveraj and Ganeshan [11] a novel speech recognition method in light of vector quantization and enhanced molecule swarm improvement (IPSO) is recommended. The recommended philosophy contains four phases, to be specific, (I) de-noising, (ii) feature extracting (iii) vector quantizations and (iv) IPSO based Hidden Markov display (HMM) method (IP-HMM). At to start with, the speech or frequencies are de-noised utilizing middle channel. Next, attributes, for example, top, pitch range, Mel recurrence Cepstral coefficients (MFCC), standard deviation, mean and least and most extreme of the flag are blackmailed from the de-noised flag. Following that, to achieve the preparation procedure, the separated qualities are given to hereditary calculation based codebook age in vector quantization. The underlying populaces are made by choosing irregular code vectors from the preparation set for the codebooks for the hereditary calculation process and IP-HMM helps in doing the acknowledgment. Now the innovativeness will be done as far as one of the hereditary operation hybrids. The projected speech recognitions approach offers 97.14% precision.

Najkar et al. [12] proposed a dynamic writing computer program are supplanted by a hunt technique which depends on particle group optimization process. The real thought is centered on creating an underlying populace of division vectors in the pattern seek space and enhancing the area of sections by a refreshing calculation. A few strategies are presented and assessed for the interpretation of particles and their comparing development structures. What's more, two division methodologies are investigated. The main strategy is the standard division which tries to augment the probability work for each contending acoustic model independently. In the following strategy, a worldwide division tied between a few models and the framework tries to recover the probability utilizing a typical tied division. The outcomes demonstrated that the cause of these elements is observable in finding the worldwide ideal while keeping up the framework exactness. The thought was tried on a separated word acknowledgment and telephone characterization undertakings and demonstrates its critical execution in both precision and computational multifaceted nature perspectives.

G. Saon and M. Picheny [13] described a set of deep learning procedures that proved to be mainly successful in attaining performance grows in word error rate on an existing huge vocabulary familiar speech recognitions benchmark tasks (“Switchboard”). They found that the finest performance is achieved by merging features from both recurrent and convolutional neural networks. They compared two intermittent architectures: partly unfolded nets with max-out activations and bi-directional extended short-term memory nets. Additionally, inspired by the success of convolutional systems for image organization, they considered a convolution networks with many convolutional layers and miniature kernels that form an approachable field with further non-linearity and fewer parameters than ordinary patterns. As soon as combined, these neural networks accomplish a word error rate of 6.2% on this tough job; this was the best testified rate at the time of this writing and is even additional outstanding given that human recital itself is predictable to be 4% on this data.

Vydana and Vuppala [14], In this work, the assessment of enduring networks have been investigated for of speech recognitions. Along with the profundity of the remaining system, the criticality of width of the lingering system has likewise been examined. It has been watched that at higher profundity, width of the nets is likewise an indispensable parameter for achieving huge enhancements. A 14-hoursubset of WSJ corpus is utilized for educating the speech recognition plans, it has been watched that the lingering systems have demonstrated much straightforwardness in joining even with a profundity substantially higher contrast with profound neural system. In this work, utilizing remaining systems a flat out diminishment of 0.4 in WER blunder rates (8% decrease in the relative mistake) is come to than the best performing profound neural system.

Guiming et al. [15] utilized the Convolution Neural Networks (CNNs) to acknowledge speech recognition. It is another sort of neural system that can diminish ghostly variety and model phantom relationships which exist in signals. Beside the paper uses Back Propagation to teach the neural network. During the whole experiment, the paper uses a collection of speech that recorded by ourselves as training data, and it uses the others to test the neural network. Experimental outcomes demonstrated that CNNs can efficiently implement isolated word recognition.

Zheng et al. [16] presents a phonetically-aware joint density Gaussian mixture model (JD-GMM) framework for voice conversion that no longer requires parallel data from source speaker at the training stage. Considering that the phonetic level features contain text information which should be preserved in the conversion task, we suggest a method that only concatenates phonetic discriminates features and spectral features take out from the same target speaker's speech to train a JD-GMM. After the mapping relationship of these two features is trained, we can use phonetic discriminant features from source speaker to estimate target speaker's spectral features at conversion stage. The phonetic discriminant features are takeout using PCA from the output layer of a deep neural network (DNN) in automatic speaker recognition (ASR) system. It can be understood as a low dimensional illustration of the senone posteriors. They compared the proposed phonetically-aware method with conventional JD-GMM method on the Voice Conversion Challenge2016 training database. The experimental outcomes showed that their proposed phonetically-aware feature method can obtain similar performance compared to the conventional JD-GMM in the case of using only goal speech as training data.

Sandanalakshmi et al. [17] presented well-organized speech to text converter for phone application is offered in this work. The major intention is to make a framework which would give ideal execution as far as precision, multifaceted nature, postponement and memory prerequisites for portable condition. The speech to content converter contains of two phases that is front-end examination and example acknowledgment. The front end investigation entails preprocessing and aspect extraction. The conventional voice activity detection processes which track only energy cannot well classify potential speech from input because the undesirable part of the speech also has some energy and appears to be speech. In the suggested system, VAD that computes energy of high frequency part distinctly as zero crossing rates to discriminate noise from speech is used. Mel Frequency Cepstral Coefficient (MFCC) is utilized as highlight extraction plot and Generalized Regression Neural Network is utilized as recognizer. MFCC gives little word mistake rate and upgraded include extraction. Neural Network progresses the exactness. Along these lines a little database containing all conceivable syllable articulation of the client is sufficient to give acknowledgment accuracy more like 100%. Along these lines the proposed procedure interests acknowledgment of constant speaker free applications like cell phones, PDAs and so forth.

PROPOSED METHODOLOGY

This section of the research work briefly describe our propose methodology which uses k-mean, LPC, LPCC, Huffman, Gaussian filter and Neural network for the speech recognition and conversion.

1.1 K-MEANS CLUSTERING

In K-Means [41], data points can just have a place with a signal cluster. Let X = {x1,x2,x3,… ,xn} be data points set and K ={k1,k2,… … .,kk} the set of cluster center. Where n is the length of data points and k indicates the quantity of clusters. Recognizable proof of cluster is finished by utilization of a few parameters. Here Euclidian separation is used which determines the separation of data point from each cluster center.

The Euclidian distance D is characterized as:

$$\text{Dist}((x, y), (r, s)) = \sqrt{(x - r)^2 + (y - s)^2} \qquad (1)$$

Where (x, y) represents to the coordinates of data point and (r, s) are directions of the cluster center. Based on this parameter every datum indicates are assigned out the cluster having the base distance from it. Presently by utilizing the information point in the cluster new cluster center are resolved.

$$Ki = (1/C_i) \sum_{J=1}^{C_j} x_i \quad (2)$$

Where 'I' speaks to group number,
Ki is the new group focus;
Cj signifies number of aggregate data points having a place toward the i_{th}group. The procedure of count of Euclidian distance and new bunch cluster Centroid proceeds until the point that no data point is apportioned to another cluster i.e. no data point changes its position.

Stage 1: Randomly select c questions, each protest communicates an underlying cluster center.
Stage 2: Calculate the distance between each example in one gathering and each clustering center, and afterward put the specimen into the subset of the closest clustering center.
Stage 3: Calculate the normal of tests in each clustering subset as its new clustering center.
Stage 4: Repeats the procedure (Step 2 and Step 3) until the point that the target work unites.
In this paper, we chiefly think about the K-means clustering by utilizing a great deal of BP neural systems' weights and limits as unique data, these nets are prepared essentially before the outfit learning for enhancing the assorted variety between neural systems.

1.2. LINEAR PREDICTIVE CODING (LPC)

Linear Predictive coding (LPC) is a tool used generally in audio signal handling and speech preparing for addressing the phantom envelope of a driven signal of speech in stuffed casing, using the info of a straight prescient model. [27] It is a champ among the best speech examination, and a champ among the most significant procedures for encoding awesome quality speech at a low piece rate and gives incredibly correct evaluations of speech parameters. LPC researches the speech motion by assessing the arrangements, removing their effects from the speech signal, and surveying the power and repeat of whatever remains of the buzz. The path to ousting the organizations is called inverse filtering, and the remaining sign after the subtraction of the isolated showed hail is recognized as the development. The numbers which depict the power and recurrence of buzz, the organizations, and the buildup signal, can be secured or conveyed somewhere else. LPC consolidates the speech motion by pivoting the strategy: use the buzz parameters and the store to influence a source to signal, use the arrangements to make a channel (which addresses the tube), and run the source through the channel, achieving talk. Since talk signals change with time, this method is done on short bits of speech signal, which are called outlines; all things considered 30 to 50 outlines for each second give rational speech with extraordinary weight.

1.3 LINEAR PREDICTIVE CEPSTRAL COEFFICIENTS (LPCC)

This is similarly a remarkable system and for the most part used to remove the features from speech signal. For audio frame LPC parameters can enough delineate vitality and recurrence range. The base of clearing up audio signal range, showing an example acknowledgment is set by the eventual outcome of extending logarithm which restricts the snappy contrast in frequency range, more bound together and better for brief time character and it is significance of Cepstrum got from special range. One of the typical at this very moment otherworldly estimation at the present time used are LPC construed cepstral coefficients (LPCC) and their backslide coefficients LPCC shows the qualifications of the common structure of human verbal tract and is figured through accentuation from LPC Parameters to the LPC Cepstrum.

1.4 GAUSSIAN FILTER

This is a filter whose motivation reaction is a Gaussian capacity (or a guess to it). Gaussian filters have the characteristics of having no overpass to a stage work input while limiting the ascent and fall time. This conduct is firmly linked with the way that the Gaussian filter has the base conceivable gathering delay. It is viewed as the perfect time area filter, similarly as the sinc is the perfect frequency space filter. These properties are imperative in regions, for example, oscilloscopes and advanced media transmission system.
The one-dimensional Gaussian filter has drive reaction given by

$$g(x) = \sqrt{\frac{a}{\pi}} e^{-ax^2} \quad (3)$$

1.5 NEURAL NETWORK

Neural network are outstanding for catching complex nonlinear relations display in speech signal. Neural network can be ordered into three classes: FeedForward Neural Network (FFNN), FeedBack Neural Network (FBNN) and the blend of both; the last one is known as Competitive Learning Neural Network (CLNN). Feedforward neural network can be operated for order and capacity estimation; criticism neural network can be applied for design stockpiling and example affiliation [42].The number of hubs in shrouded layers, concealed layers preparing for the system and handling units have noteworthy effect on the model execution. At the point when the info and yield designs are the same the system plays out the autoassociation, else it runs heteroassociation. An autoassociative neural system (AANN) catches the dispersion of the information relying upon the imperatives of the structure. It ought to be specified that heteroassociative neural system (HANN) basically utilized for a type of mapping capacity between the info and yield designs sets of the system [42], [43].

As specified before, FFNN is required to catch the association among the info and yield include vectors of the given preparing information; it is realized that a neural system with two concealed layers can understand any nonstop vector esteemed capacity. The actuation work for the units is direct at the info layer is straight, and nonlinear at the shrouded layers. The three impact variables of the system are the extent of the preparation set, the design of the neural system, and the many-sided eminence of the concern [44].

IMPLEMENTATION RESULTS

In this section, the performance of our suggested method of Speech recognition and conversion using k-means, LPC, LPCC, Gaussian filter and neural network is deliberated. Experimental setting: computer frequency of 3.2 GHZ with 4 GB memory, Software Environment MATLAB 2012A. The simulation of projected method is applied on following speech namely: India, Bhopal, Amit Sahu, Vinay, Vinay Singh, Indore, Delhi, Asta and SIRT. In Gaussian filter, K-means algorithm is applied to acquire a cluster number explicit to each observation vector and sets the centroid of the examination vector. After gathering the model, it precedes one centroid for each of cluster K and raises to the cluster number closest to it. K-mean algorithm is illustrated as the squared distances in the middle of each observation vector and its centroids. In the training segment parameters of Gaussian filter model are formed iteratively by Huffman coding. Euclidean distance is realized among observation vector and its cluster centroids to equate the spoken term with the present database. The word which is identified is revealed as text in the output.

CONCLUSION

Speech recognition systems are an indispensable part of the ever-advancing field of human computer interaction it will revolutionize the way people interacts with electronic devices especially PCs because PCs are multipurpose powerful machines capable of performing complex tasks with ease with speech recognition common person would be able to use the actual performance of the computers because if the machine could successfully understand the human speech it will make its handling too easy AI (artificial intelligence) is the latest example of it improved accuracySpeech recognition has a big potential in becoming an importantfactor of interaction between human and machine in the near future.A speaker independent speech recognition system has been proposed tocombine the advantages of Artificial Neural Networks and Hidden MarkovModels. The parameters of the Artificial Neural Networks and HiddenMarkov Model subsystems can influence each other. Encouraged by the results of the above described experiment, it may be stated that global optimization of a hybrid Artificial Neural Networks-Hidden Markov Model system would give some significant performance benefits. We have seen how such a hybrid system could integrate multiple Artificial Neural Networks modules, which may be recurrent. A combination of Probabilistic neural network and Recurrent Neural Network recognizes 98% of the phonemes correctly, followed by a Hidden Markov Model which recognizes words at better accuracy for English Speech Corpus. The results show reasonably good success in recognizing continuous speech from various speakers, for a large vocabulary. The different modules were analyzed in their respective domains and were successfully verified for different speech input files. We designed the 4 stages of speech recognition - the general processing, Prepossessing, Phoneme Recognition and Text Recognition in software. The speaker independent speech recognition systems were successfully trained to recognize speech inputs that were recorded using a microphone as well as speech samples obtained from the database.Key research challenges for the future are acoustic robustness, useof multiple word pronunciations and efficient constraints for the access ofavery large lexicon and well-organized methods for extracting conceptual representations from word hypotheses. The recognition system presented in our proposed method performs stage mode speaker independent speech recognition. The obtained results can be improved by fine tuning the system with larger training databases. The next step would be to recognize live speech, which would require more resources including larger speech databases, acoustic models and exhaustive vocabularies to produce goodrecognition results

In human-computer interaction technology, the voice recognition technology is a complex but increasingly important work, attracted a large number of researchers to join in. And this presentation we have shown that various speech recognition technique which have few advantages and problems identified So, On the basis of previous literature survey we have proposed novel approach using Neural Networks to improve the Speech Recognition quality because of its ability to implicitly detect complex nonlinear relationships between dependent and independent variables. At the end of Simulation, proposed work is increased by 20% approx. it improves the sensitivity , minimizes the Computation time , improve peak signal to noise ratio , reduces the word error rate which enhances the rate of detection of words improves Speech Recognition Rate and enhances the

REFERENCES

[1] Mohsen Fallahnezhad, Mansour Vali, Mehdi Khalili , "Automatic Personality Recognition from Reading Text Speech", 25th Iranian Conference on Electrical Engineering (lCEE2017)

[2] Sri Harsha Dumpala, Sunil Kumar Kopparapu, "Improved Speaker Recognition System for Stressed Speech using Deep Neural Networks", International Joint Conference on Neural Networks (IJCNN2017)

[3] David Dov, Ronen Talmon," Kernel-Based Sensor Fusion With Application to Audio-Visual Voice Activity Detection", IEEE Transactions on Signal Processing (Dec2016)
[4] Gueorgui Pironkov, Stephane Dupont, Thierry Dutoit, "Speaker-Aware Multi-Task Learning for Automatic Speech Recognition", 23rd International Conference on Pattern Recognition (ICPR2016)
[5] Wang Fei, Xiaofeng Ye, Xing Zhang, "Research on speech emotion recognition based on deep auto-encoder", IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER2016)
[6] Johannes A Louw, Avashlin Moodley, "Speaker Specific Phrase Break Modeling with Conditional Random Fields for Text-to-Speech", Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech2016)
[7] Roman Jarina, Roman Jarinay, Peter Pocta, Michal Chmulik," Automatic speaker verification on narrowband and wideband lossy coded clean speech", International Workshop on Biometrics and Forensics (IWBF2016)
[8] A Rajeswari, P Sowmbika, P Kalaimagal, M Ramya, M Ranjitha, "Improved emotional speech recognition algorithms", International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET2016)
[9] Hilal H. Nuha, M Abido, "Firefly algorithm for log-likelihood optimization problem on speech recognition", 4th International Conference on Information and Communication Technology (ICoICT2016)
[10] Gracieth Cavalcanti Batista, Washington Luis Santos Silva, Angelo Garangau Meneze, "Automatic Speech Recognition Using Support Vector Machine and Particle Swarm Optimization", IEEE 2016.
[11] Kamil Aida-zade, Anar Xocayev, Samir Rustamov "Speech Recognition using Support Vector Machines", IEEE-2016.
[12] S S Poorna, C Y Jeevitha, Shyama Jayan Nair, Sini Santhosh, "Emotion recognition using multi-parameter speech feature classification", International Conference on Computers, Communications, and Systems (ICCCS2015)
[13] I. E. Henawy, W. Khedr, O. ELkomy and A. Z. Abdalla, "Recognition of phonetic Arabic figures via wavelet based Mel Frequency Cepstrum using HMMs", Journal of Housing and Building National Research Center, Vol. 10, pp. 49-54, 2014.
[14] Preeti Saini, Parneet Kaur, "Automatic Speech Recognition: A Review", International Journal of Engineering Trends and Technology-(IJETT2013)
[15] Anil Kumar Vuppala "Neural Network based Feature Transformation for Emotion Independent Speaker Identification" Springer International Journal of Speech Technology Report No: IIIT/TR/2012.