Evaluation of Phonetic System for Speech Recognition on Smartphone

Gulbakshee J Dharmale, Dipti D Patil

Abstract*: This paper presents detailed study and performance evaluation of phonetic system by comparing it with various classification techniques of automatic speech recognition such as Neural Network, Hidden Markov Model, Support Vector Machine and Gaussian Mixture Model. In the phonetic system, recognized speech is processed by using language processing i.e. matching phonemes and hence generates more correct output text. The accuracy of speech recognition of ASR classifier and phonetic system is evaluated on day to day human to machine communications, using high-quality recording equipment, while the results for enhancement of existing systems is done on everyday android phones, and evaluated for normal conversations in Hindi and English language. Classifier is used to classify the fragmented phonemes or words after the fragmentation of the speech signal. Different classification techniques are implemented and comparing accuracy of speech recognition of different classifier. It is seen that GMM is better at the classification of signal data, outcomes of performance evaluation shows that GMM outperforms the other three classifiers in terms of accuracy by more than 20%. This result is compared with implemented phonetic system which shows that ASR accuracy, using phonetic system is better than GMM. We observed 6% improvement in ASR accuracy with phonetic system.

Keywords: Automatic Speech Recognition (ASR), Mel Frequency Cepstral Coefficient (MFCC), Gaussian Mixture Model (GMM), Speech Enhancement, Hidden Markov Model (HMM)

I. INTRODUCTION

Automatic Speech Recognition is a technology which permits machine to take out oral contained from a speech signal and produce a text message by using feature extraction and classification techniques. The ASR technology uses artificial intelligence and computational methodology of signal processing. The speech signal is produced through different parts of the mouth with the help of changing air pressure outside the mouth. Then these changes can be sampled periodically and recorded in a digital waveform. This recorded wave form carries all the information about the spoken word. Then, features of the speech signal is extracted and segmentation done to achieve speech recognition. Classification technique is applied to classify the fragmented phonemes or words after the fragmentation of the speech signal and provide output text.

The accuracy of speech recognition depends upon the idea of a signal which tends to be exchanged off because of the environment in which it is recorded or talked. Phonetic system is implemented to increase accuracy of speech recognition. The phoneme is the lowest unit of phonetics; it's created by vowels as well consonants. Phonemes are the basic units of a word. Word fragmentation involves fragmentation of a specified input word into the respective phonemes so that the speech recognizer is only needed to recognize the particular set of different phonemes within the word rather than having to recognize all the different individual words.

ASR system is divided into three major blocks, first feature extraction of speech signal, second segmentation and last in classification phase, classify segmented word and produce output text. ASR system is explained as below;

A. Feature Extraction part using MFCC

Feature extraction is used to derive expressive features from the improved and windowed speech signal to enable the classification of sounds.

Mel Frequency Cepstral Coefficient (MFCC) is useful feature extraction techniques used in speech recognition system based on the frequency plot using the Mel scale. It concentrates on only certain frequency components. These filters are non-uniformly independent of frequency.

MFCC is linearly distributed within the Mels or first 1000 Hz and is then logarithmically distributed above 1000 Hz [1]. This method is based on Mel scale. Basically, the MFCC is computed by taking a linearly spaced frequency scale signal and then multiplying it with a set of triangular band pass filters. To better reflect the dynamic changes of the speech, its first and sometimes second derivative of the input feature vectors of the speech signal. The frequency component converting to their Mel scale equivalent is given by following equation;

$$Mel(k) = 1127\log\left(\frac{1+k}{700}\right) \quad (1)$$

where k is the frequency factor to be converted to its equivalent Mel value;

- Pre-emphasis Filter: It is a simple signal processing filter which decreases the amplitudes of lower frequency bands and increases the amplitude of high-frequency bands.
- Framing and Windowing: Speech signal is mobile in nature. Framing can use to construct an immobile speech signal. The speech signal is split up into smaller frames overlapped with each other. [2].

Revised Manuscript Received on August 05, 2019

Gulbakshee J. Dharmale, computer science and engineering department, Sant Gadge Baba Amravati University, Amravati, India
Dipti D. Patil, Information technology department, MKSSS's Cummins College of Engineering for Women, Pune, India.

- Windowing is the next step which is used to remove discontinuities at the boundaries of frames. Windowing method used is Hamming Window.
- FFT (Fast Fourier Transform): In this process, each frame in the time domain converts to a frequency plot [3].
- Mel Filter: A set of 20 triangular bandpass filter is multiplied by the output of the FFT to get log energy of each triangular bandpass filter [4]

B. Acoustic Model

It is the main part of the training. The acoustic model provides a mathematical function of the acoustic information and phonetics. It uses speech signals from a training database. Several examples are available for acoustic modeling. Hidden Markov Model (HMM) is widely practiced and accepted model for training and realization [5].

C. Language model

The language model contains the structural constraints available in the language to get the probabilities of occurrence of a word followed by the sequence of n-1 words. It plays an important role in training. The language models distinguish between a phrase and word with analogous sound [6]. The speech recognition system uses bi-gram, tri-gram, and n-gram language models.

D. Pattern Recognition

Pattern recognition is found to be the foremost common and wide adopted techniques of speech recognition. This technique uses pattern recognition and pattern comparison. The leading characteristic of this technique is that it utilizes a well structured and integrated mathematical framework [7]. This mathematical framework assists in formulating consistent representations of speech patterns; therefore, lead to the acquirement of a lot of correct answers. Pattern recognition is further divided into two approaches, i.e., random approach and model approach.

E. Acoustic Phonetic Approach

The foremost primitive approaches of speech recognition were supporting the method of locating sounds and speeches. One in all the key objectives of such activities was to supply adequate labels to the sample sounds, so as to acknowledge the patterns of the sound. It's vital to note that such ways are found to be the muse of the acoustic-phonetic approach. As per the notion of the acoustic-phonetic approach, there exist phonemes (phonetic units) and finite units among language. This unit of acoustic-phonetic approach is extensively categorized by the gathering of acoustic properties that are sometimes evident within the speech signal [8].

In this paper, Section 1 provides a brief description about ASR techniques including feature extraction technique MFCC, pattern matching approach. NN, HMM, SVM, and GMM classification techniques explained in section 2. Section 3 gives an idea about implementation of above classifiers and phonetic model. Result and analysis of performance evaluation of different ASR classifiers and comparison with phonetic model is described in section 4. Conclusion of this paper provides in section 5.

II. ASR CLASSIFICATION TECHNIQUES

Speech recognition is one of the most significant application areas of digital signal processing. Speech recognition systems have divided into two levels, the first level of ASR system is the feature extraction level using Linear Predictive Cepstral Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCC). The classification technique is the second level using Neural Networks (NN), Hidden Markov Models (HMM), Support Vector Machine (SVM) and Gaussian Mixture Model (GMM). Classification techniques are used to classify the respective fragmented words or phonemes after an effective fragmentation of the speech signal. The most common techniques used for this activity are NN, HMM, SVM and GMM. The more details of theses classification techniques are described as follows;

A. Neural Network

Neural Networks are utilized to pattern recognition because of the capacity to prepare them to classify patterns. NN system must have memory to get the patterns related to phoneme vectors. Memory is the property that output of the framework rely not just on the present input to the framework, but also relies upon past inputs into the framework. Recurrent network must be used to achieve this task. The hidden layer neurons of yield are reintroduced and weighted to the hidden layer neurons of input. This causes present hidden layer outputs to rely on past hidden layer outputs, hence providing the system memory [9].

B. Hidden Markov Model

Hidden Markov models utilizing likelihood guide to sort input discourse into content information. As illustrated in [10] the HMMs utilizing a chain of words connected with one another to create a training model, which is then assessed with the testing information to get the most ideal recognized outcome. The essential element of HMM is markov property or probabilistic model which expresses that present state, future state and past states are autonomous to one another it implies the states are determined independently. The HMM can be considered as black box, where the progression of states visited over time is hidden and the series of output symbols produced over time is visible, thus it is known as a hidden markov model.

C. Support Vector Machine

Support vector machine is most useful machine learning technique utilized for pattern acknowledgment. It is a supervised learning which takes a set of input records and expects the outcomes relies upon its preparation on standard accessible information. The basic idea of this technique is to nonlinearly record input information to some high dimensional space, where the information can be straightly isolated and subsequently give preferred regression or classification results [11].

D. Gaussian Mixture Model

It is an unsupervised learning algorithm and density estimator. GMM performs better as it requires less preparation and test information; consequently the memory necessity is less [12].

It is utilized as a classification method to analyze the features extorted from the MFCC with the reserved pattern. GMM is characterized by its Gaussian allocation and each Gaussian allocation is deliberate by its mean, variance and weight of the Gaussian allocation. A likelihood expansion calculation is utilized to assess GMM factors from preparing information [13]. Different ASR classification techniques are explained in table I.

Table I. Description of different ASR classifiers

Sr. No.	ASR Technique	Details of ASR Classification Techniques
1	Neural Network (NN)	• Suitable for pattern recognition • Simply adjust to the robust and new surroundings • More Training of data is needed • Self-learning and self-organizing
2	Hidden Markov model (HMM)	• Supports large vocabulary size • Training Complex • Continuous and isolated word recognition
3	Support vector Machine (SVM)	• Ability to deal with the robust and high dimensional input vector • Computational cost increases with gain in several classes • Needs fixed length input
4	Gaussian mixture Model (GMM)	• Training is composite • Supports large vocabulary size • Independent speaker and continuous word recognition (Ratnadeep, Deshmukh & Alasadi. 2018).

III. PHONETIC SYSTEM FOR SPEECH RECOGNITION

The phonetic system is developed to improve the accuracy of speech recognition. Initially, speech recognition has done by the using a MFCC based Hidden Markov Model (HMM) classifier. The classifier is trained and a state probability model is developed, this model consists of chains of frequently occurring words connected via the probability of each of the words occurring together. The training corpus consists of a large set of words taken from online sources like UCI textual dataset repository, twitter and other social media datasets [14].The word linkages are then stored in the database via forward indexing, and this index is then further used for comparison. Whenever a new speech sample is taken, it is first converted into text using existing speech to text technique; the resulting text is then given to the HMM-map. This contextual HMM-map matches the occurrence of the words recognized by the speech to text API with the probability map initially created. If the probability is lower than a given threshold, then other words matching the phonetic interpretation of the given word are selected, which have a higher probability of occurrence. The recognized word is then replaced by this new word, and the process is continued for the entire recognized text.

Steps of the phonetic model given below:

1. Take input speech from the user, mark it as S
2. Divide the speech into segments, say those segments are S_1, S_2, S_3, ... S_n
3. Each segment is a different spoken word
4. Find MFCC features of each segment, mark it as F_1, F_2, F_3, ... F_n
5. Compare these features with the stored database using HMM, and get the output words W_1, W_2, W_3, ... W_n.
6. Combine the words to form the sentence
7. Apply phonetic analysis on the sentence by checking the HMM probabilities
8. Correct the output based on the words with highest probability
9. Produce the result as the output text.

To evaluate efficiency and accuracy of speech recognition by phonetic system, back propagation Neural Network, Hidden Markov Model, Support Vector Machine and Gaussian Mixture Model-based techniques using android speech engine has been implemented. The speech engine allows for real-time processing of the input sounds and works by dividing the input speech signal into segments, where each segment is a word spoken by the user. Each of the words is then given to an MFCC extraction unit, which extracts the features of the sound samples. These features are then compared with the standard IIT Bombay's Hindi Word net of words. The IIT Hindi Word net contains a full list of Hindi words, which are used as a database for this system [15]. The feature matching process is done with the help of the respective classification technique which are implemented in their standard form.

Implemented classification techniques have compared with the phonetic system to find the comparative analysis of speech recognition accuracy of phonetic system and existing classifiers.

The flow diagram for evaluating accuracy of speech recognition is as shown in given figure 1.

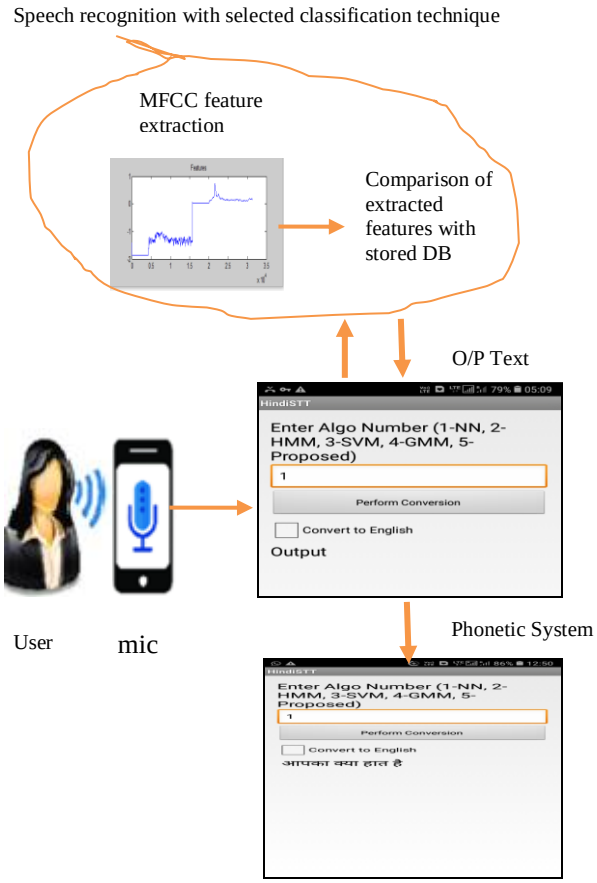

Fig. 1: Process for performan[illegible] of ASR classification techniques.

IV. ANALYSIS AND RESULT

The performance of ASR classifier is evaluated on Android phone, using high-quality recording equipment, while the results for the enhancement of existing systems is done on an everyday human to machine communication, and evaluated for normal conversations in the Hindi language.

Receiver operating characteristics (ROC) curve is plotted to explain the performance of ASR classifiers. ROC curve is used in signal detection speculation to represent a mapping between true positive rate and false positive rate of ASR classifiers [15].

The recognition rate of four ASR classifiers such as NN, HMM, SVM and GMM describes by ROC curve as shown in figure 2. ROC curve is plotted between true positive rate and false positive rate. There are four possible parameters are true positive means correctly recognized words (TP), Incorrectly recognized words are categorized as true negative (TN), a word which is not spoken, but recognized is labelled as a false positive (FP) and spoken words but not recognized are counted as false negative (FN). These four instances are used to calculate tp rate, fp rate, accuracy, sensitivity, and specificity. Different points in ROC space show positive and negative recognition. One upper left point (0, 0.9) represents perfect speech recognition, which indicates GMM performs better than other three classifiers. Tp rate and fp rate are calculated by given formula;

$$tp\ rate = \frac{Positive\ Correctly\ Recognized\ word\ (TP)}{Total\ Positive\ Words\ (P)} \quad (2)$$

$$fp\ rate = \frac{Negative\ Incorrectly\ Recognized\ words\ (FN)}{Total\ Negative\ words\ (N)} \quad (3)$$

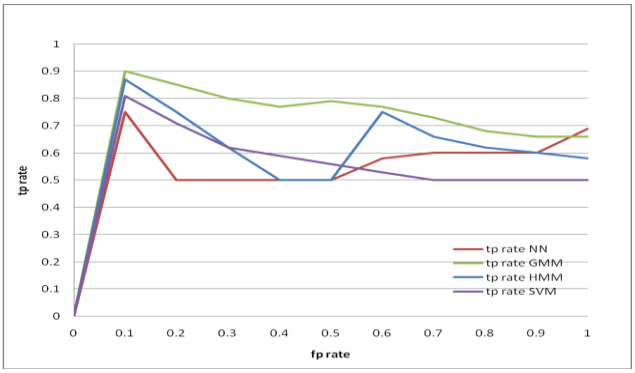

Fig. 2. ROC Curve For Four Different ASR Classifiers

An accuracy of each classifier is evaluated for short and long sentences used in day to day communication and delay in speech recognition is calculated. Speech recognition accuracy and delay in recognition can analyze on different android mobile device for each of the ASR classification algorithms. Accuracy of speech recognition can be calculated using following equations;

$$Accuracy = \frac{TP + TN}{P + N} \quad (4)$$

From results it observed that accuracy of GMM is better than other classifiers also delay in recognition is minimised as shown in table II. Accuracy and delay in speech recognition for each ASR classifier are described in figure 3 and figure 4.

Table II. Comparison of delay and accuracy of speech recognition using different ASR classifiers.

Sr. No.	ASR classifiers	Delay in recognition (ms)	Accuracy (%)
1	MFCC + NN	0.024	64
2	MFCC + SVM	0.024	71
3	MFCC + HMM	0.023	80
4	MFCC + GMM	0.023	83

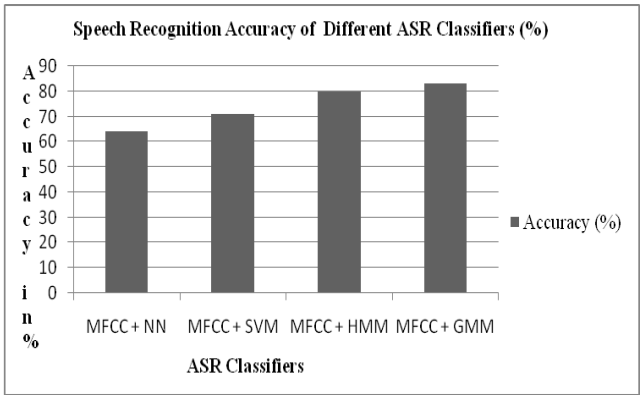

Fig. 3. Comparison of speech recognition accuracy of different ASR classifiers

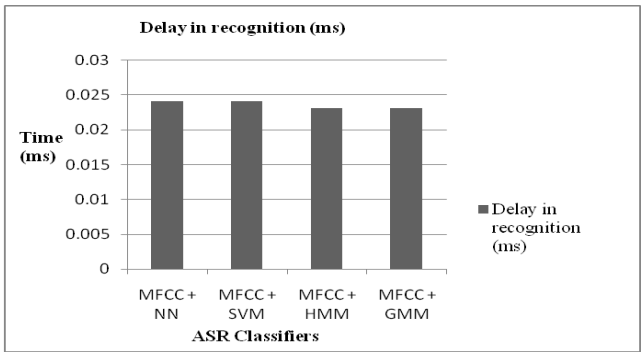

Fig. 4. Delay in speech recognition with different ASR classifiers

Figure 5. shows the comparison of speech recognition accuracy of existing ASR classifiers with proposed phonetic model for both English and Hindi language.

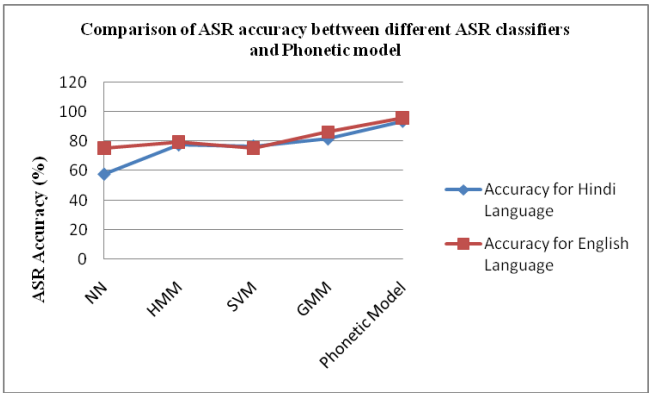

Fig 5. Comparison Of Speech Recognition Accuracy Of Phonetic System And Existing ASR Classifiers For Hindi And English Language

Accuracy of speech recognition using a phonetic system for Hindi language is 93.6 and English language 96 % respectively as shown in figure 6.

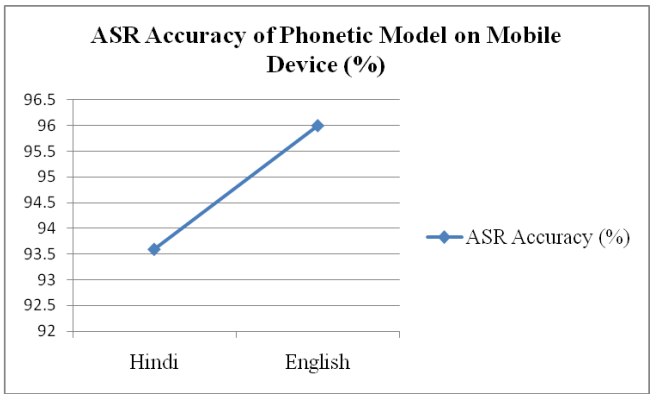

Fig. 6. Speech Recognition Accuracy Of Phonetic System On Mobile Device For Hindi And English Language

Long speech input can be recognized by of the implemented system for selected classifier on the Android device for Hindi language which can be seen from figure 7.

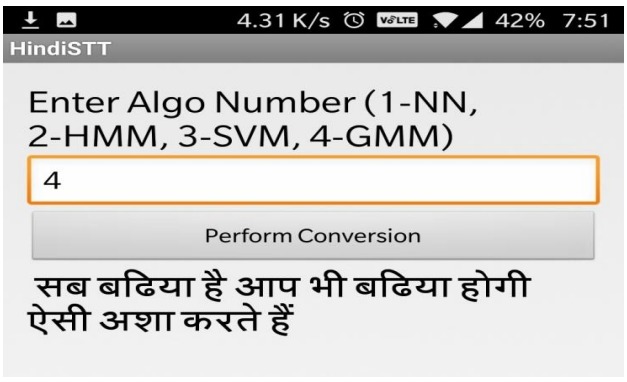

Fig 7. Long Text Conversion With The Selected (GMM) Classification Technique

An output of speech to text conversion using the phonetic model for English and Hindi Language is represented in figure 8.

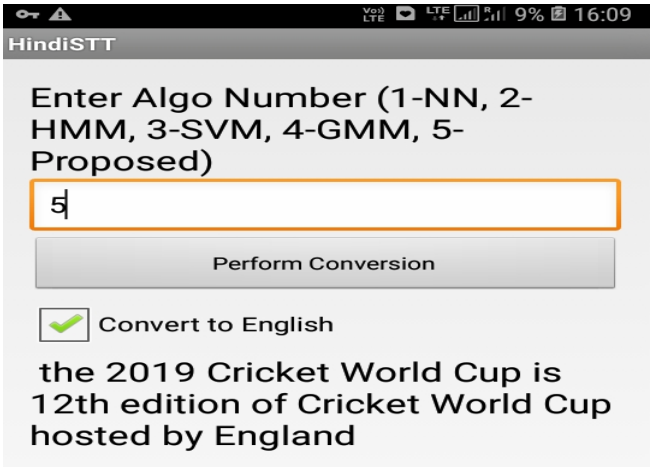

Fig. 8. Speech Recognition With Phonetic System For English Language

V. CONCLUSION

The performance of different ASR classifiers has evaluated and it is observed that Mel frequency components along with HMM-based methods provide better classification accuracy than other methods hence this technique is used in phonetic system. In real time implementation, GMM outperforms other algorithms and provides very high classification accuracy. To appraise performance of the phonetic system, it is compared with different ASR classifiers. From the results it is stated that ASR accuracy with the system improved by 6% since in phonetic system, recognized text is further mapped to contextual HMM-map to produce more accurate output text. Accuracy of speech recognition with phonetic system is 96 % for English language and 93.6 % for Hindi language.

REFERENCES

1. T. Kamm, H. Hermansky, and A. G. Andrea, " Learning the Mel-scale and optimal VTN mapping. Center for Language and Speech Processing", Workshop, 1997.
2. Jyoti B. Ramgire, Prof. Sumati M.Jagdale, "A Survey on Speaker Recognition with Various Feature Extraction and Classification Techniques", International Research Journal of Engineering and Technology (IRJET). Vol. 3, No. 4, 2016, pp. 709-712.
3. Pratik K. Kurzekar, Ratnadeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P. Shrishrimal, "A Comparative Study of Feature Extraction Techniques for Speech Recognition System", International Journal of Innovative Research in Science, Engineering and Technology. Vol. 3, No. 12, 2014, PP. 18006-180016.
4. Ms. R.D. Bodke, Prof. Dr. M. P. Satone,"A Review on Speech Feature Techniques and Classification Techniques", International Journal of Trend in Scientific Research and Development. Vol. 2, No. 4, 2018, pp. 1465-1469.
5. Lahdesmaki. H., and Shumleuch, A., "Learning the Structure of Dynamic Bayesian Networks from Time Series and Steady state Measurements", Machine Learning (ML). Vol. 71, No. 2, 2008, PP. 185-217.
6. Ms. Jasleen Kaur, Prof. Puneet Mittal., "On Developing an Automatic Speech Recognition System for Commonly used English Words in Indian English", International Journal on Recent and Innovation Trends in Computing and Communication. Vol. 5, No. 7, 2017, pp. 87-92.
7. Essa. E., Tolba, A., Elmougy. S., "Combined Classifier Based Arabic Speech Recognition", International Journal in Speech Recognition and Computer-Human Interaction. Vol. 4, No. 2, 2008, PP. 11-15.
8. Garg.A Rehg. V., "Audio-visual Speaker Detection Using Dynamic Bayesian Network", Vol. 1, No. 1, 2011, pp. 19-27.

9. Bhuvaneshwari Jolad, Dr. Rajashri Khanai., "Different feature extraction techniques for automatic speech recognition: a review", International journal of engineering sciences & research technology. Vol. 3, 2018, pp. 181-188.
10. R. Thiruvengatanadhan, "Speech Recognition using SVM", International Research Journal of Engineering and Technology (IRJET). Vol. 5, No. 9, 2018, pp. 918-921.
11. Sumita Nainan, Vaishali Kulkarni, "A Comparison of Performance Evaluation of ASR for Noisy and Enhanced signal using GMM", International Conference on Computing, Analytics and Security Trends (CAST) College of Engineering Pune, India. IEEE. 2016, pp. 486-494.
12. Ratnadeep R. Deshmukh, Abdulmalik Alasadi, "Automatic Speech Recognition: A Review", Signal processing and Cmputer Vision, 2014, pp. 464-470.
13. www.cfilt.iitb.ac.in>hindi_version
14. Gulbakshee Dharmale, Dipti D. Patil, V. M. Thakare, "Implementation of Efficient Speech Recognition System on Mobile Device for Hindi and English Language", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 10, No. 2, Feb. 2019
15. Tom Fawcett, "An Introduction to ROC analysis. pattern recognition letters", Vol. 27, 2006, pp. 861-874.

AUTHORS PROFILE

Ms. Gulbakshee J. Dharmale pursued B.E. Computer Science & Engineering from SGB Amravati University, Amravati in 2006 and M.tech. Computer Engineering from Dr. Babasaheb Ambedkar Technical University, Lonere in year 2011 and pursuing Ph.D. Computer Engg. From SGB Amravati University, Amravati. She is life member ISTE since 2011. She has published 5 research papers including in IEEE and its also available online. Her main research work focuses on Artificial Intellegence and Machine learing.

Dr. Dipti D. Patil pursued M.E. Computer Engineering from TSEC Mumbai University, Mumbai in 2007 and Ph.D. Computer Engineering from SGB Amravati University, Amravati in 2014. She is currently working as Associate Professor in MKSSS's Cummins College of Engineering for Women, Pune since 2014. She is member of BoS-Information Technology, SPPU, LMISTE, LMCSI. She has published more than 40 research papers in reputed journals including in Scopus and conferences including in IEEE and it's also available online. Her research work focuse on Machine Learning , Pattern Recognition, Classification, Neural Networks and Artificial Intelligence.