ISSN: 0976-3104

ARTICLE

ISOLATED WORD RECOGNITION SYSTEM FOR SPEECH TO TEXT CONVERSION USING ANN

Sunanda Mendiratta[1], Neelam Turk[2], and Dipali Bansal[3]

*[*1]YMCA University of Science and Technology, Faridabad, INDIA*

[2]Department of Electronics, YMCA University of Science and Technology, Faridabad, INDIA

[3]Electronics Engineering Department Faculty of Engineering and Technology, Manav Rachna International University, Faridabad, INDIA

ABSTRACT

The capacity of a device or a program to listen, identify various sounds is referred as Speech recognition and recognize some known languages from the spoken words and for human machine interface the Automatic Speech Recognition (ASR) system is helpful. In recent periods, for the ASR system, lot of research works has been developed but the concerns in that system arevast, because of the improper techniques used for the feature selection. Proper features are selected in this paper in order to develop a superior ASR system and convert the spoken word into corresponding text. Three phases are comprised in the proposed system; preprocessing, feature extraction and classification. Initially, from the source the spoken word is detected by the preprocessing phase and the noise level is reduced. Then, totally eight features are extracted containing five statistical features and three common in the feature extraction phase. Then, the classifier is trained by using these features referred as artificial neural network (ANN) with back propagation (BP). The eight features are use as the training dataset for BP algorithm and based on this feature the corresponding text displayed at the output of the proposed ASR system. The proposed system is implemented in the working platform of MATLAB. The implementation result prove that the system provide superior classification and converts the text properly. The classification accuracy examined based on performance metrics. Ultimately, the overall performance of the proposed ASR system is superior and well suitable for human machine interface.

KEY WORDS
Automatic Speech Recognition (ASR);speech signal preprocessing ;feature extraction; isolated word recognition; speech to text conversion; statistics features; back propagation (BP) algorithm.

Published: 20 October 2016

***Corresponding Author**
Email:
sunandamendiratta712@gmail.com

Guest editor: Dr. Yuvaraja T

INTRODUCTION

Speech recognition (SR) is the process by which a computer (or other machine) identifies spoken words [1]. Basically, it refers to talking to your computer and it correctly recognizes what you are saying. The SR techniques are also referring as computer Speech Recognition (CSR), or Automatic speech recognition (ASR) [2]. Usually, the process of converting speech signals to a sequence of words occur through an algorithm implemented using a computer program. About the mechanisms for mechanical realization of human speech capabilities for reasons ranging from technological curiosity, to desire to automate simple a task which necessitates human machine interactions and research in automatic speech recognition by machines has fascinated a high deal of attention for sixty years [3].

The most part of research has been motivated by People's desire in speech processing to construct mechanical models for imitating human verbal communication capabilities [4]. The fundamental objective of Speech recognition area is to create techniques and methods for speech input to machine [5]. Today, the automatic speech recognition methods find broad application in tasks based on significant advances in statistical modeling of speech that need human machine interface containsquery based information methods and automatic call processing in telephone networks provide updated travel data, stock price quotations, climatic reports etc., [6]. In common, environmental noise can easily influence the free space communication signals like the speech signal [7].

In ASR method, numerous conceivable types of environmental distortion remains a challenging issue in noisy environments and compensate these distortions precisely is difficult[8].Using some essential featuresthe speech signal is trained and then based on the trained feature the speech signal tested or recognized. Subsequently, training and testing are the two major process of basic speech signal recognition method [9]. For poor recognition performance in noise the mismatch between test condition and training was the important purpose [10].Numerous strategies are created in order to improve the performance and to lower this mismatch. For grouping these methods,the two fundamental categories are model adaptation method and Feature enhancement method [11]. Using Model adaptation techniques the probability distributions of the recognizer is remunerated straightly and in presence of noise the feature extraction methods are helpful in speech recognition [12-14]. Short-term energy, formants, Pitch, MFCCs, cross-section areas and Teager energy operator-based features are the most fascinating features for the significant speech recognition [15].

Denoising feature vectors received at the test time can work as the feature recognition strategies and these are normally trained from clean speech, so that they match the recognizer's acoustic models [17]. As the computation of these methods are simpler than model domain techniques they are more attractive and the recognizer independently implemented. For instance, at the spectrum level methods such as spectral subtraction [18] work, spectral-MMSE [19] work on the features specifically. Generally, a prior speech model is used by a class of techniques in Gaussian mixture model (GMM) to assist the enhancement process. For instance, stereo recordings of noisy andclean speech is used by SPLICE and using a GMM from noisy to clean speech a piecewise linear mapping was studied[20]. Whereas, enhanced performance have been indicated by front-end techniques on numerous tasks make point-estimates of the clean speech features. Additionally mismatch between the acoustic model and the features are because of errors in these estimates, resulting in performance degradation. These methods can enhance recognition accuracy in noisy conditions [21], [22] and obtain superior execution.

The remaining paper is summarized as follows. The recent research works related to the ASR system is examined in Part 2. Part 3 explains the proposed ASR system for speech to text conversion. Part 4 demonstrates the experimental implementation and performance comparison. Then, finally resulting sections gives the conclusion and references.

RELATED WORK

Based on the speech intelligibility some of the recent work is listed below:-

Yixiong Pan *et al.*[23] have recognized three emotional states: happy, sad and neutral. The actual researched characteristics consist of: vitality, message, linear predictive range html coding (LPCC), Mel-frequency range coefficients (MFCC) and Mel-energy range powerful coefficients (MEDC). The German Corpus (Berlin Database of Emotional Speech) and self-built Chinese emotional databases are used by Support Vector Machine (SVM) classifier for training. In unlike databases, the various combinations of features are compared in the resultant section. The complete investigational consequences exposed that the feature combination of MFCC+MEDC+ Energy had the highest accuracy rate on both Chinese emotional database and Berlin emotional database.

Navdeep Kaur*et al.*[24] has proposed a system to identify the speech dependent on the best possible element as well as in addition to the dialect. Speaker recognition is used to recognize a person from a spoken phrase. The SR system developed based on the neural network and they utilized twenty three phrases to the SR when using the taken out characteristics LPCC, REMOTE CONTROL andLPC by talk transmission as well as element vectors are created. For identification processes of various speakers and languages and for training, Neural Network back propagation learning algorithm is used. 25 speakers are comprised in the database used for that system .575 neurons are contained in the ANN model input layer and 25 neurons in the output layer and through the experimental verification the average recognition score was 93.38%.

Ant Colony Optimization method used in feature selection algorithm has been proposed by C. Poonkuzhali *et al.* [25]for automatic speech recognition. Consider the input as the appeared actual speech signal and through MFCC feature extraction approach, extracts 39 coefficients using MFCC. The unwanted features are withdrawn by the Ant Colony Optimization (ACO) technique. The fresh confirmation with their methodology exhibited that the last number connected with features removed significantly.

Petko N. Petkov *et al.*[26] has proposed a speech pre-enhancement method based on matching the recognized text to the text of the original message. This qualifying measure was appropriately estimated due to the probability with the appropriate transcription provided a good estimate with the noisy speech features. Along with drop in the actual signal-to-noise relation, in the profile of atmospheric noise speech intelligibility diminishes. They have carried out speech pre-enhancement system that optimizes the actual proposed qualifying measure to the variables of a couple different speech modification strategies within a good energy-preservation restriction. Inearlier knowledge of transcription the process proposed was necessary with the transported message and acoustic speech models from a computerized speech recognition system. This performance demonstrated a major development exceeding healthy speech and a reference point system that optimizes perceptual-distortion-based aim intelligibility measure.

An automatic speech recognition (ARS) technique have been presented by Siddhant C. Joshi *et al.* [27] using back propagation neural network. The strategies designed in that paper may expand to some common applications including sonar target recognition, missile tracking and grouping of submerged acoustic signals. Back-propagation neural network algorithm used the proposal training trials and their ideal result convictions to perceive specific patterns by changing the real service beliefs linked with the nodes and weight several links relating its nodes. A real trained network appeared later uses feature recognition inside ASR program

Deniz Baskent*et al.*[28] has presented a study on speech enhancement system for the hearing impaired persons based on phonemic restoration (PR).To measure the speech signal interrupted intermittently two conditions are used and its recognition are interruptions stay quiet and with high noise break interruptions are loaded. They used linear amplification whereasseveralcontemporary hearing aids (HAs) gave compressive amplification. Behind this choice the reason was that the study was the chief to give baseline PR data with hearing-impaired (HI) listeners, and compressive solutions could have had sudden impacts because of their nonlinear nature.

Cross adaptation of language model was investigated by X. Liu*et al.*[29], either carry out as a stand-alone system combination method or used together with acoustic model cross adjustment to improve large vocabulary continuous speech recognition (LVCSR) method. Three kinds are in language models that includes a multi-level LM that both syllable as well as word arrangements are modeled, a word level neural network LM, and the linear combination of the two were cross adjusted. The investigational results on a state-of-the-art speech recognition task suggested complimentary features exist on multiple layers of the progressive hierarchy among highly diverse sub-frameworks.

Wooil Kim*et al.*[30] have presented three approaches acquired using missing-feature techniques to enhance the speech recognition accuracy. The first innovation of that paper was Frequency-dependent classification, which engage independent classification. The second innovation was Colored-noise generation using multi-band partitioning, which involved the use of masking noises with artificially-introduced spectral and temporal variation in training the Bayesian classifier. The third innovation was an adaptive method to assess the priori values of the mask classifier , which determined if a specific time-frequency section of the test data was solid or not. It demonstrates that these advancements provide improved speech recognition accuracy on a small vocabulary test.

PROPOSED ASR SYSTEM

The common objective of the automatic speech recognition system (ASR) is to enhance performance in man machine interaction. Motivates the proposed ASR system to classify the signal and convert the classified signals the corresponding text signal. Four stages are comprised in the proposed system they are preprocessing, feature extraction, optimal feature selection and recognition. The architecture of the proposed ASR system is shown in [Fig. 1].

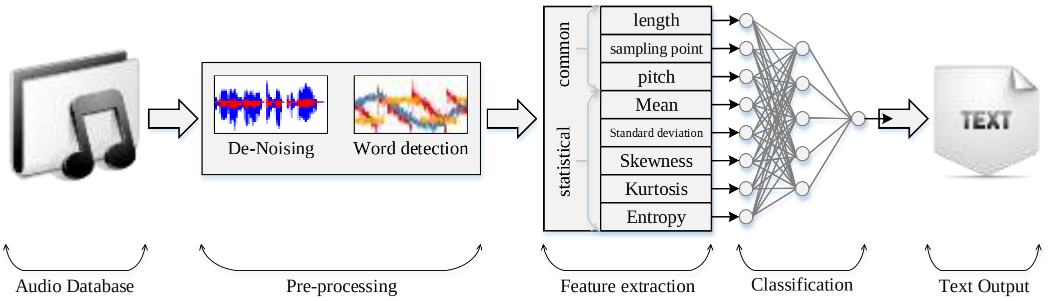


Fig. 1: Architecture of Proposed ASR system

Preprocessing the input recoded audio signal for removing noise and to detect the word. Then, the features are extracted such as sampling point, word length, pitch and the five statistical features like Mean, Variance, Skewness, Kurtosis, and Entropy. Subsequently, select the optimal level of features for the classifier training. Then, recognize the spoken word and displays the corresponding text. The detaileddescriptions of the proposed ASR system is given in following sections

Input Database

The input database contains dissimilar persons spoken various speech signals. For the proposed ASR system, consider these stored speech signals as the input signal and equation (1) represents the input speech signal given beneath as,

$$S_i = x_i(t) \big| i = 1,2,\ldots,N \tag{1}$$

Where, 'S_i' is i^{th} input signal in database and 'N' represents the total no of speech signals in database.

Preprocessing

In ASR method, preprocessingis the first phase where analog speech signal is managed at the recording time, which varies with time.The signal is processed by digital means, as it is important to sample the continuous-time signal into a discrete-time discrete valued (digital) signal. The speech is partitioned into frames or a sequence of uncorrelated segments as the properties of a signal change gradually with time and the sequence is processed as if properties of each frame has fixed. Under this postulation,the features of each frame is removed based on the sample inside the frame. Moreover, the original signal will be replacedby the feature vector in the additional processing, which means the speech signal converted from a time varying analog signal into a sequence of feature vectors. The method of varying speech samples sequencesinto feature vectors characterizing events in the probability space known as Signal Modeling [31].

The function of preprocessing is to derive a set of parameters presented from the transmission medium in a form to represent speech signals advantageous for subsequent processing and the sampled speech signal is processed and representation is produced independent on amplitude variations, speaker stress and noise. Both the frequency and time domain methods are used in the preprocessing step, amid this Time domain approaches are usually easy to implement and directly dealing with the speech signal'swaveform with parameters of zero crossing ratesand energy. Some form of spectral analysis are involved by Frequency domain approaches that are not straightlyobvious in the time domain. In speech recognition the most widely used are the latter methods.

Background Noise Removal

In the genuine situations, the background noise is normally created through fans, air conditioning, fluorescent lamps, type writers, PC's, back conversation, footsteps, traffic, opening and shutting the entryways etc., the speech recognition method'sdesigners frequently have less control over these things. Except for impulse noise sources like type writers,certain kind of noise in nature is additive and they are usually steady state [32]. Based on the environment, about 60 dB to 90 dB the levels of noise will vary. The most commonly used technique is Head mounted close speaking microphone to reduce the impact of the background noise. At normal conversational level when a speaker generating speech and the microphone filtering the speech signal then the average speech level increment by around 3dB each time. To remove the background noise the filter is given by Equation (2).

$$E_s = 10 \times \log_{10}\left\{ \in + \frac{1}{N}\sum_{n=1}^{N} S^2(n) \right\} \qquad (2)$$

Where, the 'E_s' is log energy of a block of 'N' samples and '$\in$' is a small positive constant added to prevent the computing of log zero. '$S(n)$' be the n^{th} speech sample in the block N samples.

Speech Word Detection

The speech recognition want to process the utterance consisting of silence, speech and other background noise is need to process in. The presence of speech detection entrenched in the background noise and dissimilar non-speech events known as an end point detection or speech activity detection or speech detection. For numerous reasons, a fine end point detection algorithm influences the performance of framework for accuracy and speech. First, before recognition the silence frame isremoved and for both speech and noisethe accumulated utterance likelihood score focus more on the speech portion of an utterance [33]. Second, it is difficult to precisely model silence and noise inaltering environments [34] and limit this effect using background noise frame in advance. Third, non-speech frames are removed, whenthe computation time is considerably lowered the number of non-speech frames is large[35-39].

Steps for speech word detection are as follows:

Step 1: Measurements for endpoint detection
Zero crossing count, NZ is the number of zero crossing the block. Equation (2) denotes the log energy Es of a block of length N samples. Equation (3) characterizes the normalization auto-correlation coefficient at unit sample delay C1.

$$C_1 = \frac{\sum_{n=1}^{N} S(n)s(n-1)}{\sqrt{\left[\sum_{n=1}^{N} S^2(n)\right] \times \left[\sum_{n=0}^{N-1} S^2(n)\right]}} \qquad (3)$$

Step 2: Filter for end point detection

Assume one utterance may have possible pauses separate several speech segments. To determine each segment detect a pair of endpoints named segment beginning and ending points. There is reliablya starting point is followed by a raising edge and a descending edge before an ending point on the energy contours of utterance. In the methodology at first, the edge is detected and after that the equivalent endpoints are detected. For accurate and robust endpoint detection a detector is required to detect all possible endpoints from energy feature. Equation (4) postulates the endpoint detection has the feature for one dimensional short term energy in the data sample as given beneath

$$E(l) = 10 \times \log 10 \sum_{j=n(l)}^{n(l)+l-1} o(j) \qquad (4)$$

Where, '$o(j)$' is data sample, 'l' is window length, '$E(l)$' is frame energy in decibel, '$n(l)$' is number of first data sample in the window.

Step 3: Energy normalization

The purpose of normalization of energy is to normalize the utterance energy E_l . Equation (5) specifies the normalization of energy for finding the maximum energy value $E_{\max}$ over the words as,

$$E_{\max} = \max(E_l), 1 \le l \le L \tag{5}$$

Subtracting $E_{\max}$ from E_l to give $\hat{E}_l = E_l - E$. In this way the peak energy value of each word is zero decibels and the recognition system is relatively insensitive to the difference in gain between different recordings. There is constrictions in performing the above calculations that word energy contour normalization cannot take place until locating the end of the word.

Feature Extraction

In the ASR method Feature Extraction is a crucial process, based on the feature of the system provides the precise recognition of words. The clarity of the recognized speech is enhanced based on the feature extraction and selection. Feature extraction is exceptionally focused in this proposed ASR system. The common features such assampling point, word length, pitch and the statistical features such as Standard deviation, Mean, Kurtosis,Skewness and Entropy of the speech signal are considered.

Statistical Feature Extraction

Mean: Mean is the statistician's jargon for the average estimation of a signal represented as μ (a lower case Greek mu) and found generally as you would expect as all samples are collectively added, and divide by N. These are represented as a mathematical form in Equation (6).

$$\mu = \frac{1}{N}\sum_{i=0}^{N-1} x_i \tag{6}$$

Standard deviation:Similar to the average deviation is the standard deviation, but instead of amplitude the averaging occurs with power. This isachieved by squaring each deviation before taking the average (bear in mind, power ∝ voltage^2).To compensate for the beginning square, take square root in order to complete the process. As anequation form it is represented in Equation (7).

$$\sigma = \frac{1}{N-1}\sum_{i=0}^{N-1}(x_i - \mu)^2 \tag{7}$$

Skewness:Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean in probability theory and statistics. The value ofskewness can be negative or positive, or even undefined. For n sample values, a natural method for moment's estimator of the population skewness is shown in eqn. (8) as given beneath

$$\gamma = \frac{\frac{1}{N-1}\sum_{i=0}^{N-1}(x_i - \mu)^3}{\sigma^3} \tag{8}$$

Kurtosis: Kurtosis is any measure of the "peakedness" of the probability distribution of a real-valued random variable in probability theory and statistics. Kurtosis is a descriptor with the shape of a probability distribution, with the same idea of skewness and distinctive methods for evaluating it for a theoretical distribution and relating methods for estimating it from a populationsample,same asskewness. Primarily peakedness arethe various interpretations of Kurtosis (width of peak), lack of shoulders and tail weight (distribution primarily peak and tails, not in between) the precise measures are interpreted. The Kurtosis is calculatedusing equation (9) as,

$$Kurt = \frac{\frac{1}{N-1}\sum_{i=0}^{N-1}(x_i - \mu)^4}{\sigma^4} \tag{9}$$

Entropy of the speech signal:Entropy of the speech signal is valuable to estimate the differential entropy of a method or process with some observations in dissimilar engineering fields includes image investigation, independent component investigation, speech recognition, genetic investigation, time delay estimation and manifold learning. Histogram-based estimation is the simplest and the most commonly used approach in this paper. The formula used for calculating the histogram-based entropy estimation is given in Equation (10).

$$Entropy = -\sum_{i=0}^{N-1} x_i \log\left(\frac{x_i}{w(x_i)}\right) \tag{10}$$

Where, w is the width of the i^{th} bin.

Word Identification

In this phase, the identified spoken words is converted into text. The identificationaccuracy greatly depends on the classification technique and the accuracy is identified highly. To tackle this problem an artificial intelligence (AI) is penetrated. The AI algorithm used in this paper for classification is artificial neural network (ANN).

ANN Based Word Recognition

Duplicating the neural structure and working of the human brain is the principal aim of the programmed computational model, ANN. It consist of an interrelated structure of artificially created neurons that for data exchange it function as pathways. Artificial neural networks are adaptiveand flexible, adjusting and learning every different external or internal stimulus. In sequence artificial neural networks are used, and data processing, pattern recognition systems, modeling and robotics. The ANN comprises of a single output layer and a singleinput layer and to one or more hidden layers. Except the input layer all nodes are composed with neurons. Depending on the issue the number of nodes in each layer differs. The architecture complexity of the network is dependent upon the number of nodes and hidden layers. ANNtraining is to find a set of weights that would give desired values at the output when at its input presented a dissimilar pattern. Training and testing is the two main process of an ANN. An example of a simple artificial neural network is shown in [Fig.2.

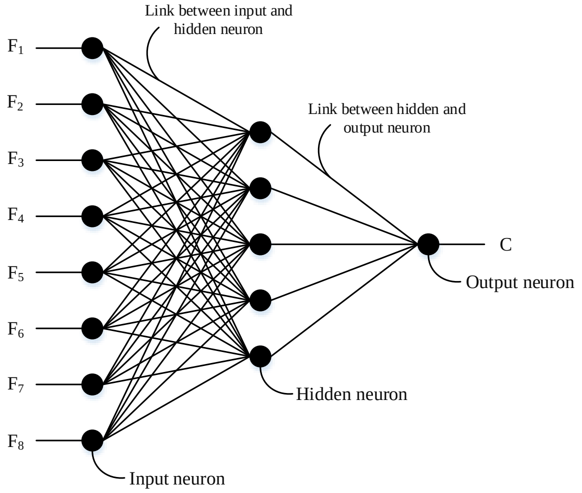

Link between input and hidden neuron
F1
F2
F3
F4
F5
F6
F7
F8
Link between hidden and output neuron
C
Output neuron
Hidden neuron
Input neuron

Fig. 2: Structure of ANN

The input for the ANN gets total of eight features and as the output of the ANN corresponding text can be obtained. Three common features and five Statistical features are included in the eight features. Statistical features such as standard deviation, mean, kurtosis, skewness and the common features such as sampling point, word length, and pitch. Thus, eight inputs (eight features) and corresponding word output is contained in the proposed ANN architecture. The two main process of a classification algorithm is training and testing.

Training and Testing of ANN

In this phase, using the extracted features the ANN is trained. In training, the input and the output is defined and then fix the appropriate weight so that in the testing phase the classifier (ANN) can able to predict the apt object (word). In a common classifier algorithm the major part is the training phase. In the proposed system the back propagation (BP) algorithm is used for training. In the training the process involved is given as follows and [Fig. 3] shows the architecture of back propagation neural network (BPNN).

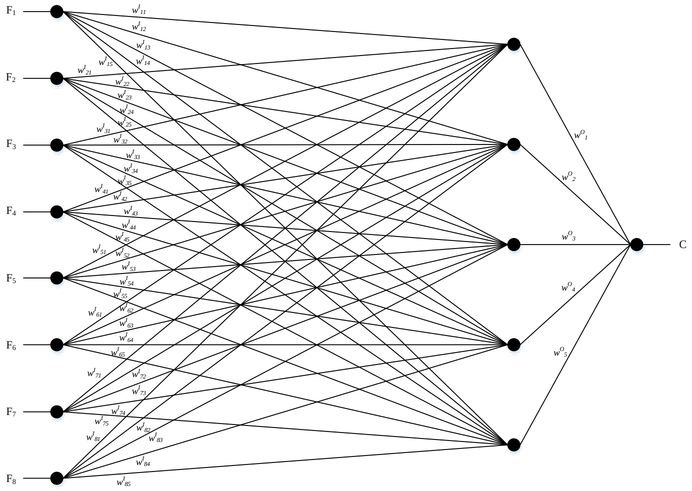


Fig. 3 : Proposed back propagation neural network

The proposed ANN consist of one output units, eight input units and M hidden units (M=5). In the back propagation algorithm's forward pass, first to the hidden layer the input data is transmitted andthen to the output layer. In the hidden layer each node gets input from the input layer and with appropriate weights multiplexed and summed. Using the equation (11) the output of the neural network is achieved which isgiven below.

$$C = \sum_{j=1}^{M} \frac{w_j^O}{1 + \exp(-\sum_{i=1}^{N} F_i w_{ij}^I)} \tag{11}$$

In eqn. (11), 'F_i' is the i^{th} input value and 'w_j^O' is the weights assigned between hidden and output layer, 'w_{ij}^I' is the weight assigned between input and hidden layer and M is the number of hidden neurons. The non-linear transformation of the resulting sum is the output of the hidden node. The same processis followed by the output layer. From the output layer the output values are compared with target values and for the neural network calculatethe learning error rate, which is given in eqn (12) as

$$\partial_k = \frac{1}{2}(Y - C)^2 \tag{12}$$

In eqn. (12), ∂_k is the k^{th} learning error of the ANN, Y is the desired output and C is the actual output. The error between the nodes is transmitted back towards the hidden layer by the backward pass of the back propagation algorithm. Then, for some other training dataset the training is repeated by changing the weights of the neural network.

The following steps describe the minimization of error by back propagation algorithm.

i. First, to hidden layer neurons the weights are assigned. The input layer has a constant weight, for the output layer neurons the weights are randomly chosen. Then, using Eqn. (11) the output is calculated.

ii. Then, compute the back propagation error using the Eqn. (13).

$$BP_{error} = \sum_{k=1}^{Z} \partial_k \tag{13}$$

Where, 'BP_{error}' is Back Propagation error, '∂_k' learning error rate of k^{th}training data set. Then the weight deviation in the hidden neuron is finding by using the equation (14).

$$\Delta w = BP_{error} \cdot \gamma \cdot \delta \tag{14}$$

Where, 'Δw' is the weight deviation, 'γ' is the learning rate, which usually ranges from 0.2 to 0.5, 'δ' is the average of hidden neurons output.

$$\delta = \frac{1}{T} \cdot \sum_{n=1}^{T} H_n \qquad (15)$$

Here, $H_n = \dfrac{1}{1 + \exp\left(-\sum_{i=1}^{N} F_i w_{in}^{I} \right)}$

Where, ‘ δ ’ is the average of hidden neurons output. ‘ N ’ is the total no of input neurons, ‘T’ is the total no of training and ‘ H_h ’ is the h^{th} output at hidden neuron or activation function at input side. Then find the new weights by using the equation (16) given below.

$$w^{new} = w + \Delta w \qquad (16)$$

Where, ‘ w^{new} ’ the new weight or updated weight and ‘ w ’ is the current weight.
Then repeat the process until the BP error gets minimized and it satisfies $BP_{error} < 0.1$. If the BP error reaches a minimum value, then the ANN is ready for classification.
Subsequently, process the testing, in testing the ANN undertaken for the verification of classification accuracy. In this phase, other than the data used in the training a new set of data is verified by the classifier. Once the classification accuracy is satisfied by ANN it can be used for the live application for the intended classification.

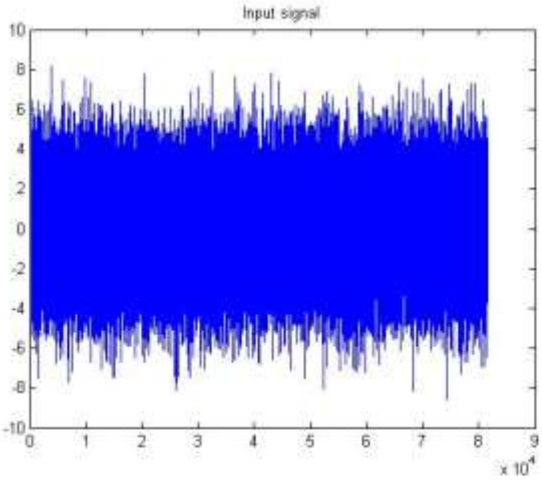
RESULTS

The proposed system for the automatic speech recognition is implemented in the working platform of MATLAB with the following system specification.

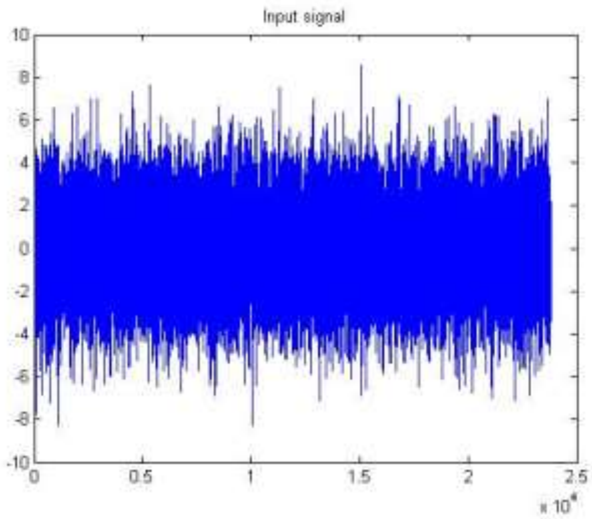
Processor	: Intel i5 @ 3GHz
RAM	: 8GB
Operating system	: windows 7
Matlab version	: R2013a

The ASR system is a peculiar process to made interface between man and machine .The normal machines don’t have any sense to identify the attitude of a common man. One can make a system to interact with human with the help of artificial intelligence. Everyone could interact with or control the machine but for the interaction among various attitudes speech is one of the special and easiest way. A speech recognition system is important in this case in order to identify the speech or spoken word precisely. Consequently, by converting these recognized words or speech into the corresponding electrical signal the machine could realize the speech of an ordinary man. An ASR system is developed in this paper to recognize the spoken word and convert it to corresponding text. For implementing the proposed system the research tool in this section is Matlab. Using some recorded speech signals the recognition performance is analyzed.

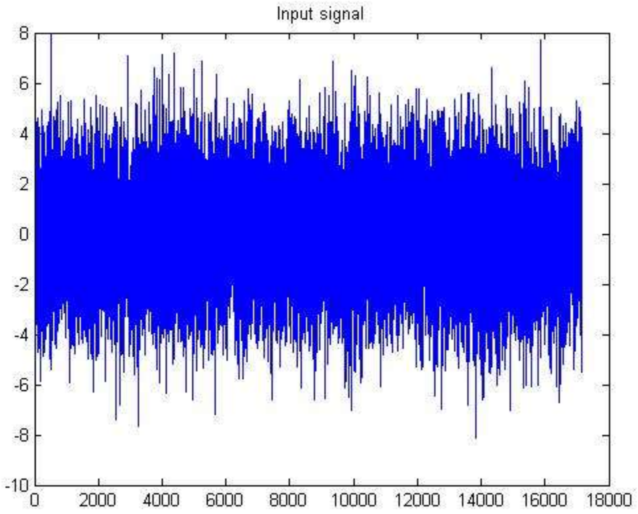
In the input audio database we have stored 68 recorded speech signals. Amongst 68, for the performance analysis three signals such as Apple, Badam and Cow are used. The input speech signal apple, Badam and Cow are shown in [Fig. 4].



(a) Apple

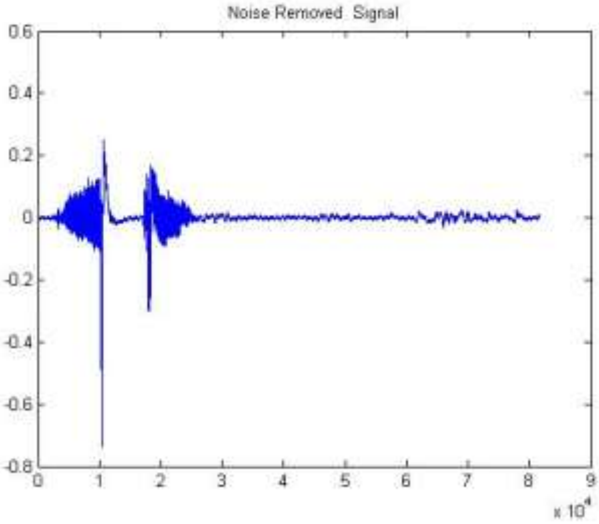


(b) Badam

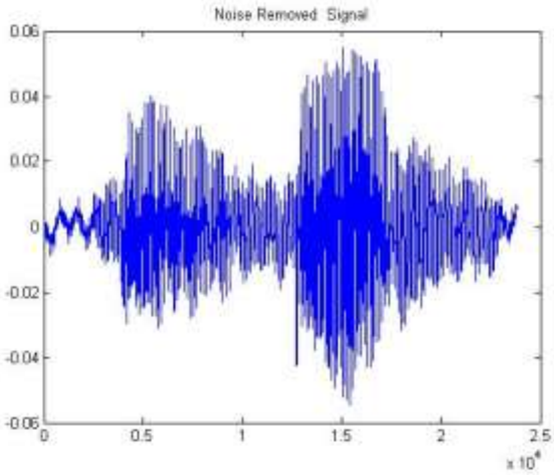


(c) Cow

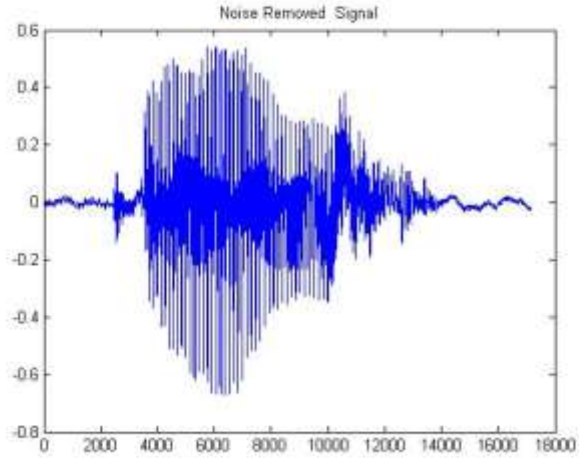
Fig. 4: Input speech signal :(a) Apple, (b) Badam, (c) Cow



(a)

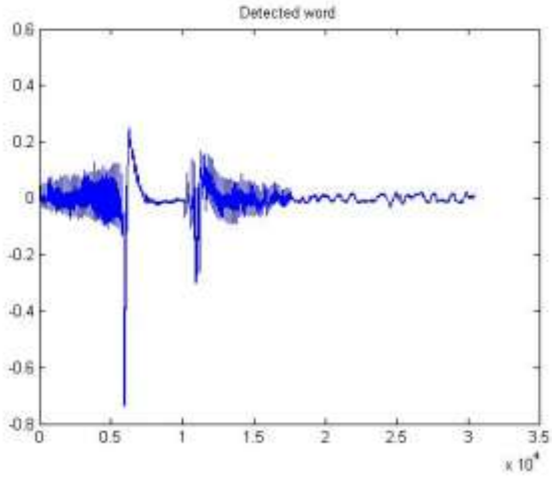


(b)

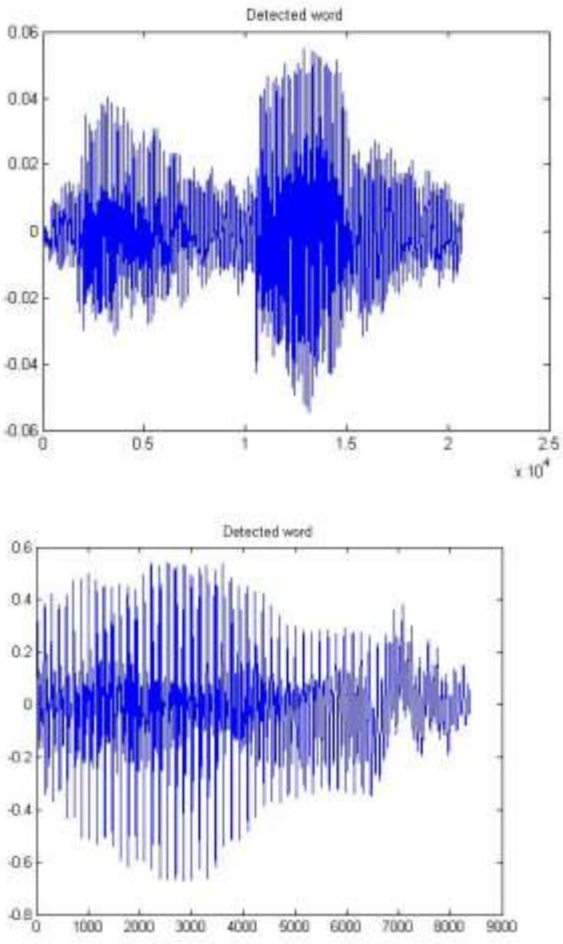


(c)

(i):(a) Apple, (b) Badam, (c) Cow



(a)



(b)

(c)

(ii): (a) Apple, (b) Badam, (c) Cow

Fig. 5: Preprocessed signal; (i) De-noised and (ii) Detected word

……………………………………………………………………………………………………

In [Fig. 5] the pre proposed signal shown for three different speech signals Apple, Badam and Cow. To detect the isolated word and for noise removal the preprocessing step is used. The statistical and the common features of the spoken signal are followed after the preprocessing stage. After preprocessing the common features are obtained easily .Using the eqn. (6) to eqn. (10) the statistical features are obtained and **[Table 1]** gives the obtained statistical values.

Table 1: Values of Statistical features

Features	Numerical value of Signal Samples		
	Apple	Badam	Cow
Mean	-0.0000077	0.0022	-0.00033
Variance	0.0011	0.0549	0.0173
Skewness	-0.00000023	-0.0000087	-0.0000057
Kurtosis	0.00000000017	0.000028	0.00000071
Entropy	5683.3	15510	5242.4

The eight features including common and statistical are given to the ANN to display the corresponding text and for the classification of signal. The proposed ASR system's obtained output was shown in [Fig. 6]. The corresponding text of spoken word is the outcome of ASR system.

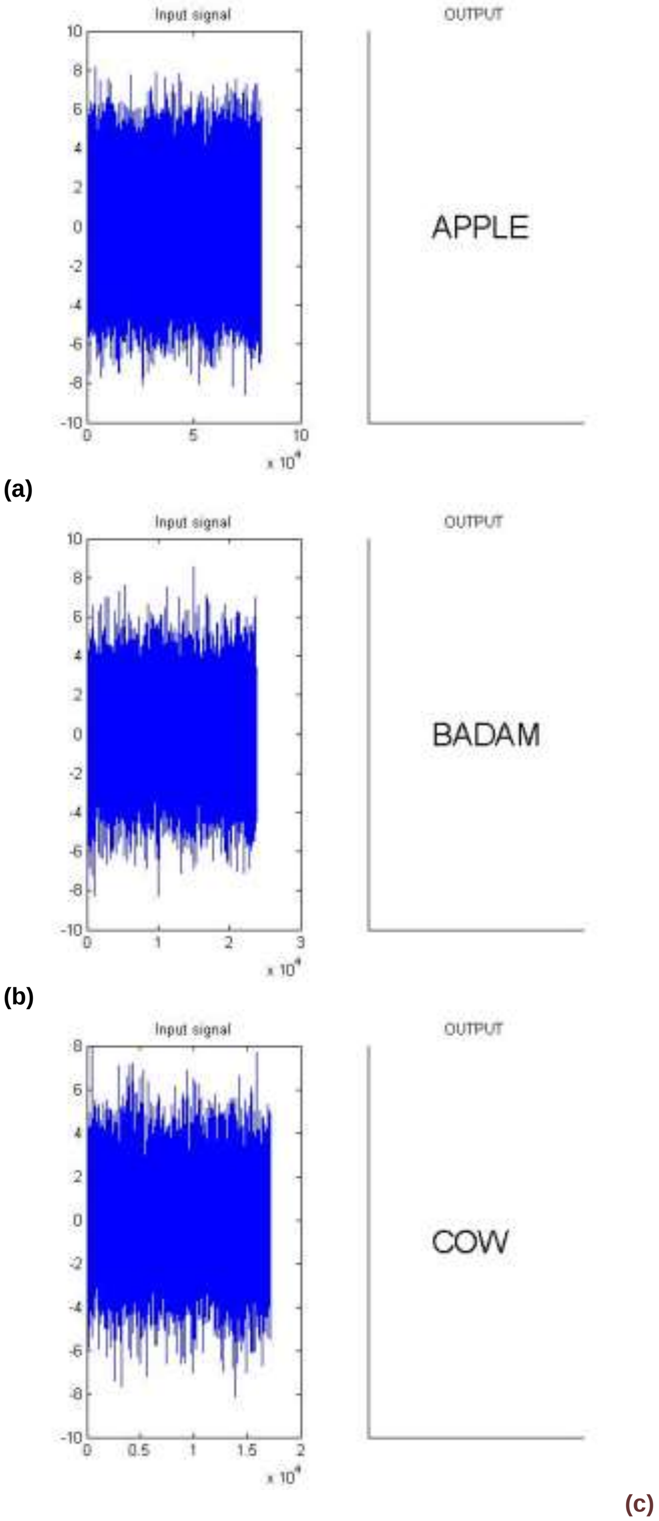


Fig. 6: Text display; (a) Apple, (b) Badam, (c) Cow

Then, analyze the performance of ANN classifier based on the Sensitivity, classification Accuracy, Specificity, Negative Predictive Value (NPV), Positive Predictive Value (PPV), False Discovery Rate (FDR), Matthews Correlation Coefficient (MCC) and False Positive Rate (FPR).For these performance measures the graphical representation is shown in [Fig. 7].

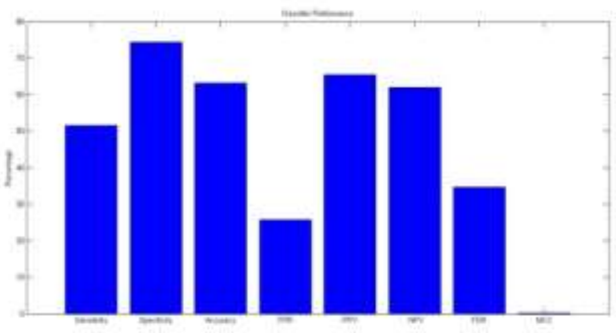


Fig.7: Classification Performance of proposed ASR

The classifier performance of the proposed ASR system is shown in [Fig. 7]. The performance obtained for Specificity, Sensitivity, FPR, Accuracy, NPV, PPV, MCC and FDR are 74%, 50%, 26%, 62%, 60%, 65%, 0.19% and 35% respectively. These results displays that the proposed ASR system can satisfy the required performance like accuracy sensitivity and specificity for identifying the spoken word. We can show that the proposed system is a suitable approach for automatic speech recognition from these performance analyses and result.

CONCLUSION

The spoken speech signal is converted in to corresponding text by the proposed automatic speech recognition system. Three stages are comprised in the proposed ASR system; at first stage preprocessing the spoken speech signal to detect the isolated word and to remove noise. The sum of eight features such as sampling point, word length, Mean, pitch, Skewness, Variance, Entropy and Kurtosis are extracted in the second stage. In the final level, the word spoken gets recognized with the help of features extracted in the previous stage and the corresponding text is displayed. With back propagation training using ANN classifier the identification of word occurs. The proposed system's performance was scrutinized for three signal samples like Badam, Apple and Cow explained the corresponding waveform. Then based on the accuracy, classification accuracy and specificity the classifier's performance was analyzed. The complete performance shows that for the man machine interaction the proposed system for the ASR is the best options. The ASR system we proposed provide enhanced performance so it is suitable for a real time human and machine interaction over speech. The ASR system's performance will further improved by the influence of a novel classifier approach in future.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

The authors are thankful to Hon'ble Dean and Management

FINANCIAL DISCLOSURE

None.

REFERENCES

[1] Bedros, Renee, Charles PB, Arch WB, Francis MC, Brian Al, George HQ, David Spoor, Stephen RS, and David JW. [1993] Multimedia interface and method for computer system. U.S. Patent 5,208,745, issued May 4, 1993.

[2] Trentin, Edmondo, and Marco Gori. [2001] A survey of hybrid ANN/HMM models for automatic speech recognition .Neurocomputing, 37(1):91-126.

[3] Furui and Sadaoki. [2005] 50 years of progress in speech and speaker recognition. SPECOM 2005, Patras, 1-9.

[4] Childers DR, Cox V, DeMori R, Furui S, Juang BH, Mariani JJ, Price P, Sagayama S, Sondhi MM, and Weischedel R. [1998] The past, present, and future of speech processing.IEEE signal processing magazine, 15(3):24-48.

[5] Gaikwad, Santosh K, Bharti WG, and PravinYannawar. [2010] A review on speech recognition technique.International Journal of Computer Applications, 10(3):16-24.

[6] JuangBH, and Rabiner LR.[2005]Automatic speech recognition–a brief history of the technology development. Georgia Institute of Technology.Atlanta Rutgers University and the University of California, Santa Barbara, 1.

[7] Krim Hamid, and Mats Viberg. [1996] Two decades of array signal processing research: the parametric approach.Signal Processing Magazine, 13(4):67-94.

[8] Li Jinyu, Li Deng, Dong Yu, Yifan Gong, and Alex Acero. [2009] A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions.Computer Speech & Language, 23(3):389-405.

[9] Anusuya MA, and S.K. Katti SK. [2009] Speech Recognition by Machine: A Review.International Journal of Computer Science and Information Security, 6(3):181-205.

[10] Viikki, Olli, and Kari Laurila. [1998] Cepstral domain segmental feature vector normalization for noise robust

speech recognition.Speech Communication,25(1):133-147.

[11] Kalinli, Ozlem, Michael LS, and Alex Acero. [2009] Noise adaptive training using a vector Taylor series approach for noise robust automatic speech recognition.In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 3825-3828.

[12] Leggetter CJ, and Woodland PC.[1995] Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models.Computer Speech Languages, 9(2):171-185.

[13] GauvainJ-L, and Lee CH.[1994] Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains.IEEE Transactions on Speech Audio Process, 2:291–298.

[14] Dupont, Stéphane, and JuergenLuettin. [2000] Audio-visual speech modeling for continuous speech recognition.IEEE Transactions on Multimedia,2(3):141-151.

[15] Ververidis, Dimitrios, and Constantine Kotropoulos. [2006] Emotional speech recognition: Resources, features, and methods. Speech communication, 48(9):1162-1181.

[16] Kyung Hak Hyun, Eun Ho Kim and Yoon-KeunKwak. [2007] Emotional Feature Extraction Based On Phoneme Information for Speech Emotion Recognition. In Proceedings of 16th IEEE International Symposium on Robot and Human interactive Communication, 802-806.

[17] Kalinli, Ozlem, Michael LS, JashaDroppo, and Alex Acero. [2010] Noise adaptive training for robust automatic speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 18(8):1889-1901.

[18] Boll S.[1979] Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustic Speech Signal Process, 27(2):113-120.

[19] Yu D. Deng L, J Droppo J, Wu J,Gong Y, and Acero A. [2008] A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition.In Proceedings of ICASSP, Las Vegas, NV, 4041-4044.

[20] Deng L, Acero A, Plumpe M, and Huang X. [2000] Large-vocabulary speech recognition under adverse acoustic environments. In Proceedings of ICSLP, Beijing, China, 806–809.

[21] Saon G, Huerta JM, and Jan EE. [2001] Robust digit recognition in noisy environments: The IBM Aurora 2 system.In Proc. Interspeech, Aalborg, Denmark, 629–632.

[22] Cui X and Alwan A. [2005] Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR.IEEE Transaction on Speech Audio Processing, 13(6):1161–1172.

[23] Yixiong Pan, PeipeiShen and LipingShen. [2012] Speech Emotion Recognition Using Support Vector Machine. International Journal of Smart Home, 6(2):101-108.

[24] NavdeepKaur and Sanjay Kumar Singh. [2012] Data Optimization in Speech Recognition using Data Mining Concepts and ANN.International Journal of Computer Science and Information Technologies, 3(3):4283-4286.

[25] Poonkuzhali C, Karthiprakash R, Valarmathy S, and Kalamani M. [2013] An Approach To Feature Selection Algorithm Based on Ant Colony Optimization For Automatic Speech Recognition. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2(11):5671-5678.

[26] Petko NP, Gustav EjeHenter, BastiaanKleijn W.[2013] Maximizing Phoneme Recognition Accuracy for Enhanced Speed Intelligibility in Noise.IEEE Transaction on Audio, Speech and Language Processing, 21(5):1035-1045.

[27] Siddhant CJ, Cheeran AN. [2014] MATLAB Based Back-Propagation Neural Network for Automatic Speech Recognition. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 3(7):10498-10504.

[28] DenizBaskent, Cheryl L, Eiler and Brent Edwards. [2010] Phonemic restoration by hearing-impaired listeners with mild to moderate sensorineural hearing loss.Hearing Research, 260(1–2):54–62.

[29] Liu X, Gales MJF, and Woodland PC. [2013] Language model cross adaptation for LVCSR system combination. Computer Speech and Language, 27(4):928–942.

[30] Wooil Kim and Richard MS. [2011] Mask classification for missing-feature reconstruction for robust speech recognition in unknown background noise. Speech Communication, 53(1):1–11.

[31] Picone L.[1993] Signal modeling technique in Speech Recognition. IEEE ASSP Magazine, 81(9):1215-1247.

[32] Hwang T, and Chang S. Energy Contour enhancement for noisy speech recognition. International Symposium on Chinese Spoken Language Processing, 1:249-252.

[33] Rabiner L, and Sambur M.[1976] Some Preliminary experiments in the recognition of connected digits.IEEE Transactions on Acoustics, Speech and Signal Processing, 24(2):170-182.

[34] Abdulla W. [2002] HMM–based techniques for speech segment extraction.Scientific programming, IOS Press, Amesterdam, The Netherlands, 10(3):221–239.

[35] Becchetti C, and Ricotti L. [2004] Speech Recognition Theory and C++ Implementation.John Wiley & Sons, Wiley Student Edition, Singapure, 121-188.

[36] Ney H.[2003] An optimization algorithm for determining the end points of isolated utterances.In Proceedings of IEEE International Conference on Acoustics, Speech, and Siganl Processing (ICASSP), 7(3):26-41.

[37] T Yuvaraja, M Gopinath.[2014] Fuzzy Based Analysis of Inverter Fed Micro Grid inIslanding Operation International Journal of Applied Engineering Research ISSN 0973-4562 9()16909-16916.

[38] Yuvaraja Teekaraman, Gopinath Mani. [2015]Fuzzy Based Analysis of Inverter Fed Micro Grid in Islanding Operation-Experimental Analysis International Journal of Power Electronics and Drive System (IJPEDS) 5(4): 464~469

[39] T Yuvaraja, K Ramya.[2016] Implementation of Control Variables to Exploit Output Power for Switched Reluctance Generators in Single Pulse Mode Operation IJE TRANSACTIONS A: Basics 29(4): 505-513.