IIOAB JOURNAL
ISSN: 0976-3104
|SPECIAL ISSUE| ENERGY, ENVIRONMENT, AND ENGINEERING
SECTION: E3

ARTICLE

FUZZY BASED SELECTION OF DWT FEATURES FOR AUTOMATIC SPEECH RECOGNITION SYSTEM FOR MAN MACHINE INTERACTION WITH CS-ANN CLASSIFIER

Sunanda Mendiratta[1*], Neelam Turk[2], 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]*Department Faculty of Engineering and Technology, Manav Rachna International University, Faridabad, INDIA*

ABSTRACT

Man machine interaction is the most commonly used technology in Computer Aided Systems and comes under various fields. Among them man machine interaction with automatic speech recognition is the advanced technology in this era. There are several technologies are put forward by several researchers but each of them has its own backlogs. In this paper we have proposed a methodology for man machine interaction system with automatic speech recognition by Means of fuzzy based Discrete Wavelet Transform (DWT) feature extraction and Artificial neural network with Cuckoo search optimization (CS-ANN) classifier. Here the input is first preprocessed using a sequence of preprocessing steps and then features are extracted from that by Means of DWT. After this the optimum number of features are selected by Fuzzy logic and finally the speech signals are recognized by training the ANN where the optimization is done through CS algorithm. The proposed methodology is implemented on MATLAB platform and the experimental results are presented and validated under different conditions.

KEY WORDS
Man-Machine interaction, Automatic speech recognition, Sampling, Hamming window, Harmonic decomposition, Discrete Wavelet Transform, Fuzzy model, Artificial Neural Network, Cuckoo search

INTRODUCTION

Speech signals contain a gigantic measure of data and can be depicted as having various levels of data. Voice based interfaces hold the way to understand this accomplishment. In this connection, Automatic speech recognition (ASR) systems in distinctive languages pick up significance [1]. ASR is an imperative undertaking in digital signal processing related applications. It is the procedure of automatically changing over the spoken words into written text by the PC framework [2]. In the course of recent decades speech recognition has made across the broad innovative advances in numerous fields, for example, call steering, automatic translations, data looking, data entry and so on [3]. Speech recognition has been expert by consolidating different algorithms drawn from diverse disciplines, for example, statistical pattern recognition, signal processing and semantics and so forth [4]. The percentage of the imperative uses of speech recognition are security gadgets, household apparatuses, PDAs, ATM machines and PCs [5]. Speech processing is one of the energizing zones of signal processing. The objective of speech recognition field is to create method and framework to produce for speech data to machine [6]. In view of significant development in statically displaying of speech, automatic speech recognition today find far reaching application in task that require human machine interface, for example, automatic call processing [7].

Speech recognition can be generally partitioned into two stages: feature extraction and classification [8]. Even though noteworthy advances have been made in speech recognition innovation, it is still a troublesome issue to outline a speech recognition framework for speaker-free, nonstop speech [9]. One of the crucial inquiries is whether the majority of the data important to recognize words is preserved while feature extraction stage. In the event that imperative data is lost during this stage, the execution of the accompanying classification stage is naturally handicapped and can never measure up to human ability [10]. Feature extraction can be comprehended as a stage to decrease the dimensionality of the input information, a diminishment which definitely prompts some data loss. Normally, in speech recognition, we partition speech signals into frames and extract features from every frame [11]. During feature extraction, speech signals are changed into an arrangement of feature vectors. At that point these vectors are exchanged to the classification stage. For instance, for the case of dynamic time warping (DTW), this succession of feature vectors is contrasted with reference information set. For the instance of hidden Markov models (HMM), vector quantization may be connected to the feature vectors [12-13] , which can be seen as a further stride of feature extraction. In either case, data loss during the move from speech signals to a succession of feature vectors must be kept to a minimum [14].

In different many-sided application, for example, speech recognition where the frameworks are created in light of genuine information, handling an extensive number of features is successive. Be that as it may, a significant number of the features are not pertinent to the issue of interest. Moreover, various features demand a lot of calculations, which back off the general procedure. Under these circumstances, naturally releasing insignificant features is important to accomplish a model that is exact and dependable in taking care of the issue at hand [15-16]. Also, utilizing the feature selection method lessens the computational expense, which improves the reaction time of the procedure. Feature selection (FS) is a standout amongst the most profitable research territories and has pulled in a lot of consideration in the course of recent decades. For the FS undertaking, two surely understood

Received: 7 September 2016
Accepted: 29 September 2016
Published: 2 Dec, 2016

***Corresponding Author**
Email: sunandamendiratta712@gmail.com
Tel.: +91-9677002684

Guest editor: Dr. Yuvaraja T

strategies for feature assessment are utilized. The first uses distance metrics to quantify the cover between distinctive classes [17]. Under this system, probability thickness functions of the example appropriation can likewise be considered. Thus, the subset for which the normal overlap is insignificant is considered as an answer. In the Meantime, intra-and in addition between class distances can be measured by considering the fuzziness and Entropy of the features [18]. During that time system, classification errors in light of the feature subset candidates are assessed. Therefore, the subset with insignificant misclassification is chosen as an answer. A few techniques in [19] have been thought about and has been actualized for the variable selection. In the Meantime, utilizing flexibilities to characterize the optimization issue is useful. For this reason, fuzzy set theory is utilized to classify the adaptabilities for the objective functions. This strategy prompts accomplishing additional exchange off to tackle this issue. Fuzzy optimization systems have been every now and again utilized as a part of the optimization of clashing objectives [20].

In this paper, we have proposed an automatic speech recognition system for the man machine interaction with DWT based feature extraction and ANN classifier optimized with Cuckoo search (CS) algorithm. The remaining of the paper is structured as follows: Section 2 gives the recent works related to the speech recognition system proposed by several researches recently. Section 3 gives in detail the proposed methodology and the processes associated with the proposed work. Section 4 discusses about the experimental results and discussion of our proposed methodology and finally section 5 concludes our paper.

RELATED WORKS

Kaya H et al. [21] have extended a recent discriminative projection based feature selection method using the power of stochasticity to overcome local minima and to reduce the computational complexity. The approach assigned weights both to groups and to features individually in many randomly selected contexts and then combined them for a final ranking. The efficacy of the method was shown in a recent paralinguistic challenge corpus to detect level of conflict in dyadic and group conversations. They advanced the state-of-the-art in this corpus using the INTERSPEECH 2013 Challenge protocol.

Cumani S, and Laface P et al. [22] have validated that a very small subset of the training pairs was essential to train the original PSVM model, and proposed two methods that permitted removing most of the training pairs that were not necessary, without damaging the precision of the model. This permitted intensely decreasing the computational resources and memory necessary for training, which became possible with enormous datasets comprising many speakers. They had evaluated these methods on the extended core situations of the NIST 2012 Speaker Recognition Estimation. Their consequences displayed that the precision of the PSVM trained with an appropriate number of speakers was 10%-30% better associated to the one attained by a PLDA model, reliant on the testing conditions. Since the PSVM precision expanded with the training set size, but for large numbers of speaker PSVM training did not scale well, their selection methods became applicable for accurate training discriminative classifiers.

To optimize a feature vector, Chatterjee S and Kleijn W. B et al. [23] had developed a new framework such that it imitates the human auditory system behavior. In an offline manner the optimization was conceded out on the basis of assumption that the local geometries of the feature vector domain and the perceptual auditory domain must be analogous. Along with a static spectral auditory model, using this principle they optimized and modified the static spectral Mel frequency cepstral coefficients (MFCCs) without regarding any feedback from the speech recognition system. Then the work was prolonged to comprise spectro-temporal auditory properties into manipulating a new dynamic spectro-temporal feature vector. Making use of a spectro-temporal auditory model, the dynamic feature vector was designed and optimized to integrate the human auditory response behavior across frequency and time. They exhibited that an essential development in automatic speech recognition (ASR) performance was achieved for any environmental condition, clean in addition to noisy.

Dikici E et al. [24] have presented some directions to enhance discriminative language modeling performance for Turkish broadcast news transcription. They used and compared three algorithms, namely, perceptron, MIRA and SVM, both for classification and ranking. They applied thresholding as a dimensionality reduction technique on the sparse feature set, and some sample selection strategies to decrease the complexity of training. In this paper, they also extended the earlier results by including the SVM classifier and classification and ranking versions of the MIRA algorithm for completeness. They also increased the feature space by incorporating higher order -grams, presented the relationship between ranking versions of perceptron and SVM, and gave a thorough statistical analysis and comparison of the results.

Pan S. T, and Li X. Y [25] have focused on field-programmable gate array (FPGA)-based robust speech measurement and recognition system, and the environmental noise problem was its main concern. To accelerate the recognition speed of the FPGA-based speech recognition system, the discrete hidden Markov model was used here to lessen the computation burden inherent in speech recognition. Furthermore, the empirical mode decomposition was used to decompose the measured speech signal contaminated by noise into several intrinsic mode functions (IMFs). The IMFs were then weighted and summed to reconstruct the original clean speech signal. Unlike previous research, in which IMFs were selected by trial and error for specific applications, the weights for

each IMF were designed by the genetic algorithm to obtain an optimal solution. The experimental results in this paper revealed that this method achieved a better speech recognition rate for speech subject to various environmental noises.

SPEECH RECOGNITION SYSTEM WITH DWT BASED FEATURE EXTRACTION AND CS-ANN CLASSIFIER

In the context of speech recognition, the recognition of the speech signal is done by extracting more relevant features that clearly characterizes the signal. Though there are several methods are employed in feature extraction, the incorporation of artificial intelligence in this field will produce better results. In this paper, we proposed a methodology for man machine interaction using Discrete wavelet transform (DWT) based features along with novel fuzzy based feature selection method and Hybrid Cuckoo search – Artificial neural network (CS-ANN) classifier for automatic speech recognition system. The proposed method consists of three stages. They are Signal Preprocessing, Fuzzy based DWT and Classification. Initially, the input recorded speech signal is preprocessed for removing noise and to detect the word. Then, the preprocessed signal is subjected to feature extraction process by using discrete wavelet transform applied to the speech signal where the preprocessed speech signal is decomposed into various frequency channels using the properties of wavelet transform. Here, in our proposed method, we are employing 8-level multi resolution wavelet transform so that at each decomposed level eight types of features like Mean, Standard Deviation, Skewness, Kurtosis, Entropy, Shannon Entropy, Log energy Entropy and Renyi's Entropy of the speech signal are extracted. Since, we are using 8-level DWT so the number of features extracted is high. So, a Fuzzy model is used to select the optimal features from speech signals which are extracted by DWT. Finally, the selected optimal set of features is used for training of Artificial neural network (ANN) classifier and then based on these features the spoken work is recognized and the corresponding text will be displayed. In order to improve the classification accuracy of the ANN the weights of the neural networks is optimized by using the Cuckoo search (CS) algorithm. The proposed method is implemented in the Matlab working platform and the results are compared with the previous methods under different conditions.

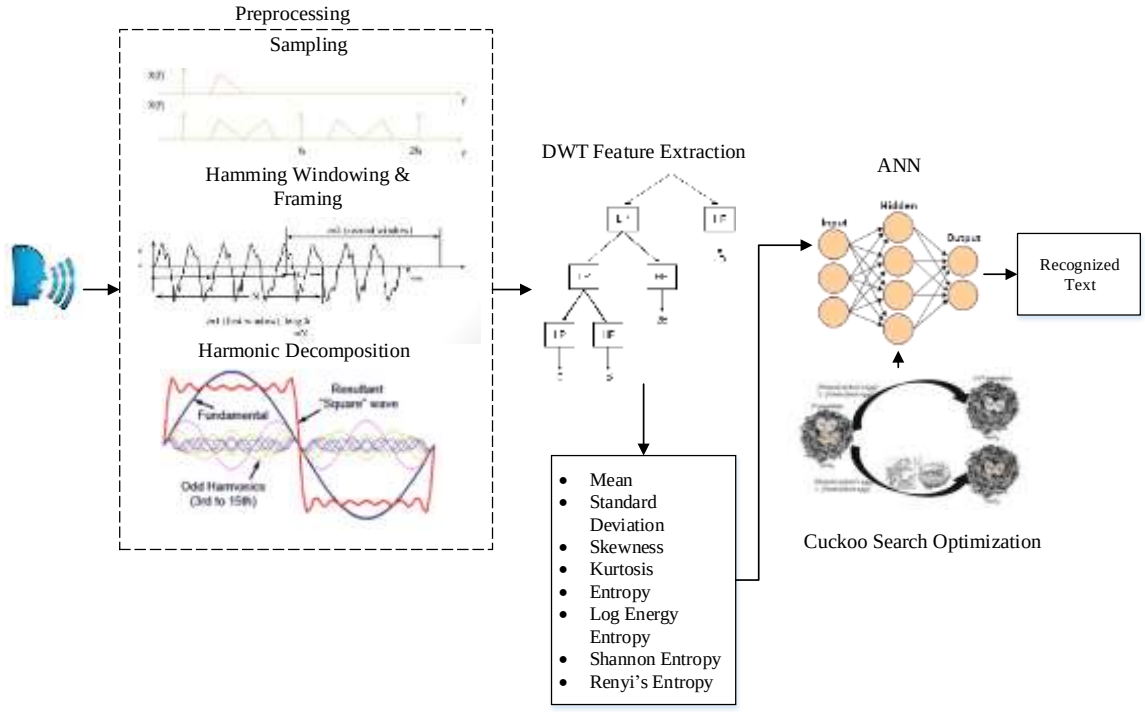

Fig.1: Block diagram of the proposed speech recognition system.

As shown in the above [Fig. 1], the input speech signal is initially undergoing some sequence of preprocessing steps as sampling the signal, producing frames by Hamming windowing process and denoising of the signal through the Harmonic decomposition process with Harmonic decomposition method. After preprocessing, the specified features from the signal is extracted through DWT and from the extracted features only the optimum number of features is chosen by Fuzzy Inference System (FIS). The optimally extracted features then employed in the training procedure of the ANN in which the optimization of the neural network is done through the CS optimization algorithm. Each of these processes are explained in detail in the following sections. Consider that the database D, contains 'N' number of speech signals as mathematically represented as $D = \{s_1, s_2, ...s_N\}$. From this the signal $s_i, i = 1,2, ...N$ is taken out and further operations are applied as discussed in the following sections.

Preprocessing of input speech

The input speech signal produced by the human talk, normally processed using some preprocessing techniques in order to make the signal to be suitable for further operations being applied to the signal. The purpose preprocessing the speech signal is to facilitate the signal processing, eliminate the silent frequencies and to remove the noises. Each of these preprocessing is performed through the following steps.

Sampling of the speech signal

The sampling of the speech signal s_i is performed to facilitate further operations on the speech signal by converting it into a digital format. Several types of sampling methods or available, but for processing the speech signal we have employed a Band pass Sampling method as the input signal in our methodology is of band pass type. For the band pass signals the spectrum of the signal $S(\omega)$ will be zero for the range of frequencies except $f_1 \le f \le f_2$. In general the frequency f_1 of the band pass signal is non-negative and will be greater than zero also the aliasing effect is zero when $f_s < 2f_2$, where f_s is the sampling frequency which is calculated using the equation (1).

$$f_s = \frac{1}{T} \tag{1}$$

where, $T = \frac{m}{2f_2}$ is the sampling interval and hence the equation (1) is modified as in equation (2).

$$f_s = \frac{2f_2}{m},\ m < \frac{f_2}{B} \text{ and } f_s = \frac{2KB}{m},\ f_2 = KB \tag{2}$$

where, B = Bandwidth of the signal

m =Number of Replications used in the sampling process (Can be any integer till $f_s \ge 2B$)

The sampled spectrum of the speech signal s_i of bandwidth B and the minimum sampling rate f_s is given by equation (3).

$$S_s(\omega) = \frac{1}{T}\sum_{i=1}^{N}\sum_{n=-\infty}^{\infty} s_i(\omega - 2nB) \tag{3}$$

where, n = time instant in discrete level, after sampled the points from the input speech signal, the samples are converted into frames by using Hamming window and this is explained in the following process.

Framing and Windowing

The input signal produced may contain some silent sounds which are not necessary in the speech recognition system to recognize that signal. Hence the signal is converted into frames with the which smooth the progress of signal analysis. In the process of framing the signal is converted into frames with the period of 20-30 ms applied at 10 ms intervals with an overlap of 50% between adjacent frames and this is referred to as short time spectral analysis. Here, we are employing Hamming window in the process of windowing which results in number of frames of the speech signal. Hamming window is the famous windowing technique usually preferred in speech signal processing as it has its own benefits. The process of windowing with Hamming window on speech signal is explained as follows;

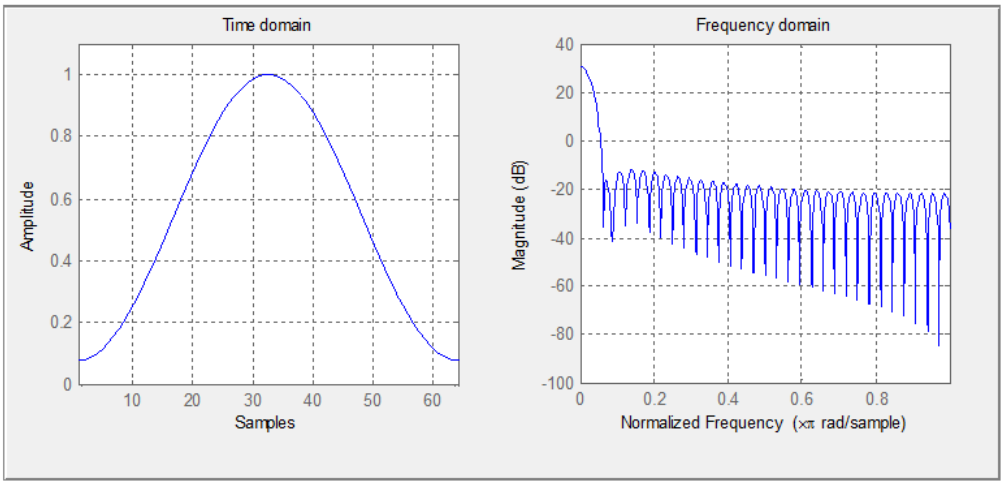

Fig.2: L point Hamming Window in Time and Frequency Domain (L=64).

In the above [Fig. 2], the L point Hamming window in time and frequency domain is displayed where L is the length of the window and it is considered here as 64. The Hamming window is generally represented as given in the equation (4) [26].

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N}\right) \tag{4}$$

where, N = Number of samples in each frame which is equal to L, L-1 or L+1.

In the process of windowing and framing, the sampled signal is initially divided into frames with the specified time intervals and then the hamming window is applied to each frame to produce the signal which is free from discontinuities. Consider that the input speech sample is represented as $s_i(n)$ for the *i*th sample, then the resulting frames after processed by the hamming window is given by the equation (5).

$$y_i(n) = \sum_{i=1}^{S} w(n)\, s_i(n) \tag{5}$$

where, S = Total number of frames produced
$s_i(n)$ = nth sample from the spectrum $S_s(\omega)$

The resulting S number of frames produced by the equation (5) is the speech frequencies which are free from the discontinuities. These frames further enhanced by the reduction of noise present and this is performed by the Harmonic decomposition process as explained in the following section.

De-noising frames by Harmonic decomposition

After converting the speech samples into frames without discontinuities the noise present in these samples is removed using denoising method and the type of noise present in the speech signal is the harmonic noise. Hence, the Harmonic decomposition [27] method is employed here as the noise reduction technique in which the input speech sample is decomposed into its fundamental as well as harmonic components. After, decomposing the speech sample into the components the order of the components which are greater than three could be ignored. The speech sample obtained from the windowing function is then denoised $y_i(n)$ by correlating it with the fundamental and harmonic components up to the order of three which is given mathematically in equation (6).

$$y_{di}(n) = \sum_{i=1}^{S}\sum_{j=1}^{3}(y_i(n) * h_j) \tag{6}$$

where, $y_{di}(n)$ = Denoised sample

h_j = Harmonic Levels of the input speech sample

Thus, the correlated signal produced are free from harmonic noise by filtering out the harmonic component after reconstruction. The preprocessed speech samples are then given to the feature extraction phase, in which the features for

the speech signal to aid the classification in recognition phase. The feature extraction phase is given in the following section.

Fuzzy Based DWT feature extraction

After, producing the de-noising signal of the speech input the features are extracted from the preprocessed signal for the purpose of recognizing the speech by training the classifier in the classification stage. There are several techniques are used to extract the features from the speech signal but the incorporation of artificial intelligence could produce most suitable features for the recognition of the speech signal. Hence in our proposed methodology, a novel fuzzy based DWT feature extraction is employed where eight different types of features are extracted such as Mean, Standard Deviation, Skewness, Kurtosis, Entropy, Shannon Entropy, Log energy Entropy and Renyi's Entropy from the signal using DWT feature extraction with eight level decompositions. After the features are extracted most of them not having necessary details to recognize the signal so that the fuzzy model is used to select the most suitable features. The process of DWT of extracting the features and the feature selection through the fuzzy model is detailed as follows.

DWT in feature extraction

A linear transformation that operates on a data vector is discrete wavelet transform (DWT) whose length is an integer power of two, converting it into a statistically diverse vector of identical length. In image compression and signal processing the wavelet transform has obtained an extensive acceptance. A signal's multi-resolution depiction is offered by the DWT which is very beneficial in examining "real-world" signals and therefore this transformation is accepted to excerpt the speech signal features in our proposed approach. Basically, by a DWT a distinct multi-resolution depiction of a continuous-time signal is acquired. It converts a series $a_0, a_1, \ldots a_n$ into one low pass coefficient series known as "approximation" and one high pass coefficient series known as "detail". $n/2$ is the Length of each series .In actual life conditions, such transformation is implemented recursively on the low-passseries till the preferred number of repetitions is attained.

The function is not incessant and henceforth not differentiable. Daubechies wavelets are the families of wavelets whose inverse wavelet transforms are adjoint of the wavelet transform i.e. they are orthogonal. Using Daubechies wavelets the wavelet transform result in gradually finer discrete samplings making use of recurrence relations. Each resolution scale is double that of the earlier scale. From data compression range to signal coding the discrete wavelet transform has wide range of applications. Hence, it is a device that divides data into varied frequency modules, and then each component with a resolution accorded to its scale is considered. DWT is calculated with a cascade of filters ensued by a factor of two sub-sampling. In [Fig. 3] the fundamental DWT architecture is shown.

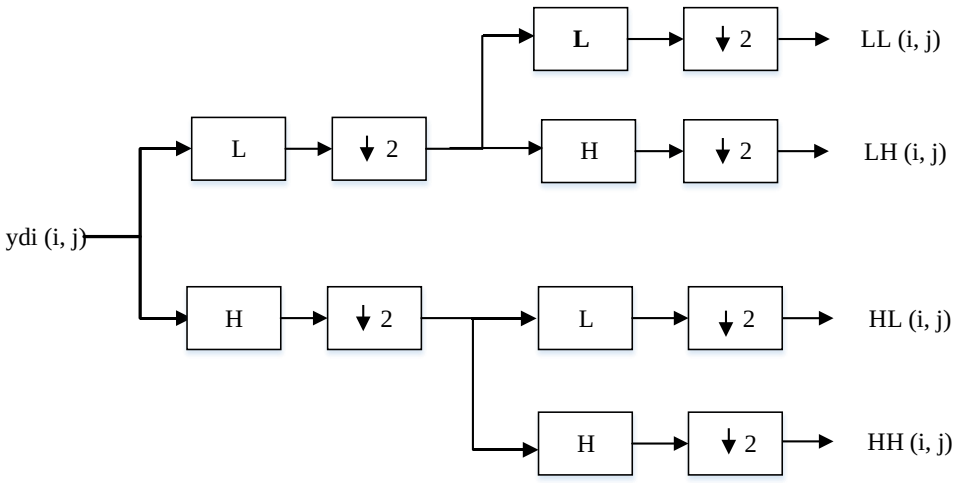

Fig. 3: General architecture of the 1-level DWT.

In the above [Fig. 3] the one level DWT is shown in which the speech signal is first split into low and high frequency bands with the Low pass and High Pass filters as given by the impulse responses denoted by $g(m)$ and $h(m)$ respectively and this is given in equations (7) and (8) and they are divided further to produce two low frequency bands and two high frequency bands and this is the complete one level decomposition by DWT. The ↓ symbol denotes the down sampling of the filtered element.

$$l_{i+1}(m) = \sum_{n=-\infty}^{\infty} y_{di}(n-2m)\, g(m) \quad (7)$$

$$h_{i+1}(m) = \sum_{n=-\infty}^{\infty} y_{di}(n-2m) h(m) \quad (8)$$

The process given in equations (7) and (8) can be continued until the required level of decomposition is reached by extending the architecture as given in the[Fig 3].The DWT's major feature is a function of multi scale representation. At different levels of resolution given function can be examined using the wavelets. The DWT is also orthogonal and could be invertible .In our proposed methodology we have employed 8-level DWT and hence the filtered coefficients produced for each signal is 32 from these coefficients we will extract the features mentioned above as in the following manner.

Mean, μ

The Mean value of the filtered coefficients is calculated at each level of the decomposition. If the low and high frequency components of level *k* is represented as l_{lk} , l_{hk} , h_{lk} and h_{hk} respectively, then the Mean value is calculated as given in equation (9).

$$Mean, \mu = \frac{1}{n}\sum (l_{ik} + h_{ik}), \qquad i = l \,\&\, h;\ \ n = 4 \quad (9)$$

Standard Deviation, σ

The Standard Deviation of the calculated filtered coefficients can be measured from the variance of the coefficients and it is the measure how the sample in the signal is far from the Mean value. The variance and the Standard Deviation from that is calculated as given in equations (10) and (11).

$$Variance, \sigma^2 = \frac{1}{n}\sum (x_i - \mu) \quad (10)$$

$$Standard\ deviation, \sigma = \sqrt{\frac{1}{n}\sum (x_i - \mu)} \quad (11)$$

Skewness, s

A significant distribution parameter is symmetry. Based on the definition, a skewed variable Mean is not situated at the distribution center. It might be observed that two signal segments can have the same Mean and Standard Deviation but diverse values for Skewness. Skewness is conveyed as in equation (12).

$$Skewness, s = \frac{E(x_i - \mu)^3}{\sigma^3} \quad (12)$$

s

k

$$Kurtosis, k = \frac{E(x_i - \mu)^4}{\sigma^4}$$

$H(s_n)$

$\{s_1, s_2, ...s_n\}$ $\qquad P(s_n) = P_n, 0 \le P_n \le 1, \sum P_n = 1$

$$I(s_n) = -\log P_n$$

$I(s_n)$

$H(s_n)$

$$H(s_n) = E[I(s_n)] = -\sum P_n \log P_n$$

$H_{Sh}(s_n)$

$$H_{Sh}(s_n) = \frac{H(s_n)}{\log m}$$

m

$H_{\log}(s_n)$

$$H_{\log}(s_n) = -\sum(\log(P_n))^2$$

$$H_{Ren}(s_n)$$

$$H_{Ren}(s_n) = \frac{1}{1-\alpha}\sum(\log(P_n)^{\alpha})$$

Fuzzy based feature selection

The wavelet features extracted as above are larger in number and such large quantity is not necessary to recognize the input signal. For that purpose the feature selection process is usually conducted after the feature extraction phase and the most common feature reduction techniques used are Principal Component Analysis (PCA), Isomap techniques etc. But these techniques consumes more time and results in considerable information loss. Hence including the human intelligence can produce apt features that are used better for the recognization of the speech signal and in our proposed methodology we select the required features by means of fuzzy logic. Here the method we have engaged is the fuzzy feature estimation index for a set of features which is described in terms of membership values indicating the degree of similarity between two features.
In fuzzy based feature selection, the parameter called evaluation index is calculated between a set of features based on the calculated membership functions. Initially the degree of membership μ_{uv} is measured between the feature sets *u* and *v*. With that membership function the evaluation index is measured and the feature sets which are having a minimum value of index is selected in training the neural network. The membership function μ_{uv} for the features is calculated as given in equation (19).

$$\mu_{uv} = \begin{cases} \left(1 - \frac{d_{uv}}{D}\right), & if\ d_{uv} \le D \\ 0, & otherwise \end{cases} \quad (20)$$

In equation (20), d_{uv} is the Euclidean distance which measures the similarity between the two features and *D* is the minimum distance between any given two features. The distance, d_{uv} is calculated as in equation (21).

$$d_{uv} = \sqrt{(u-v)^2} \quad (21)$$

And finally the evaluation index of the membership function is measured for all the features in the feature space as in equation (22).

$$e = \frac{1}{s(s-1)}\sum_u\sum_v \mu_{uv}(1-\mu_{uv}) \quad (22)$$

Where, s is the number of feature samples present in the respective feature space. This fuzzy evaluation index is determined for all of the eight features and then finally the respective features which are having minimum value of the index is considered in the further recognition process. After selecting the optimum features from the feature space the training and the testing of the ANN is carried out in the next stage in which the optimization of the network is done by Means of Cuckoo search algorithm and the process involved in this stage is detailed in the following section.

Recognition of speech with CS-ANN classifier

The optimally selected features from the feature extraction is then used for training the ANN for the purpose of recognizing the speech signal,but in normal back propagation algorithm is used for the optimization of weight function between the links. But here we have employed CSO algorithm to produce fair optimization of the network and the steps of this training procedure is given as follows after giving the basic introduction about ANN.

Artificial neural network (ANN)

A distinctive artificial neural network (ANN) has two types of basic constituents. They are neurons which are processed elements and links which are interconnections between neurons. Each link has a weighting parameter. From other neurons each neuron gets stimulus, progress the information, and an output is produced. Neurons are classified to input, output and hidden neurons. The input and output layers, are called the first and the last layers correspondingly, and the enduring layers are called hidden layers. Consider in the k^{th} layer the number of neurons as n_k . Let, the weight of the link is represented as w_{ij}^{k} between the j^{th} neuron of the $(k-1)^{th}$ layer

and the i^{th} neuron of the k^{th} layer. If for each neuron $w_{i0}^{\ k}$ is an additional weighting parameter, signifying the bias in the i^{th} neuron of the k^{th} layer. Before the neural network training the weighting parameters are initialized. In a systematic manner they are iteratively updated during training. Once the neural network is finished, the weighting parameters remain stable throughout the neural network usage as a model.

The training process of an ANN is to modify the biases and weights. The most prevalent technique to train, feed forward ANNs in numerous domains is the back-propagation (BP) learning. Though, one drawback of this method, which is a gradient-descent method, is that it necessitates a differentiable neuron transfer function. Also, as neural networks create intricate error surfaces with multiple local minima, instead of a global minimum the BP inclines to converge into local minima. In modern years, many enhanced learning algorithms have been proposed to overwhelm the handicaps of gradient-based methods. Because of conventional numerical methods' computational shortcomings in resolving complex optimization problems, researchers may depend on meta-heuristic algorithms. Over the last eras, numerous meta-heuristic algorithms have been applied successfully to several engineering optimization problems. For many complex real-world optimization problems, better solutions have been delivered by them in comparison with conventional numerical methods. Here we are employed Cuckoo Search optimization as a Meta heuristic algorithm for optimizing the ANN, the reason for this is explained in the next consecutive section.

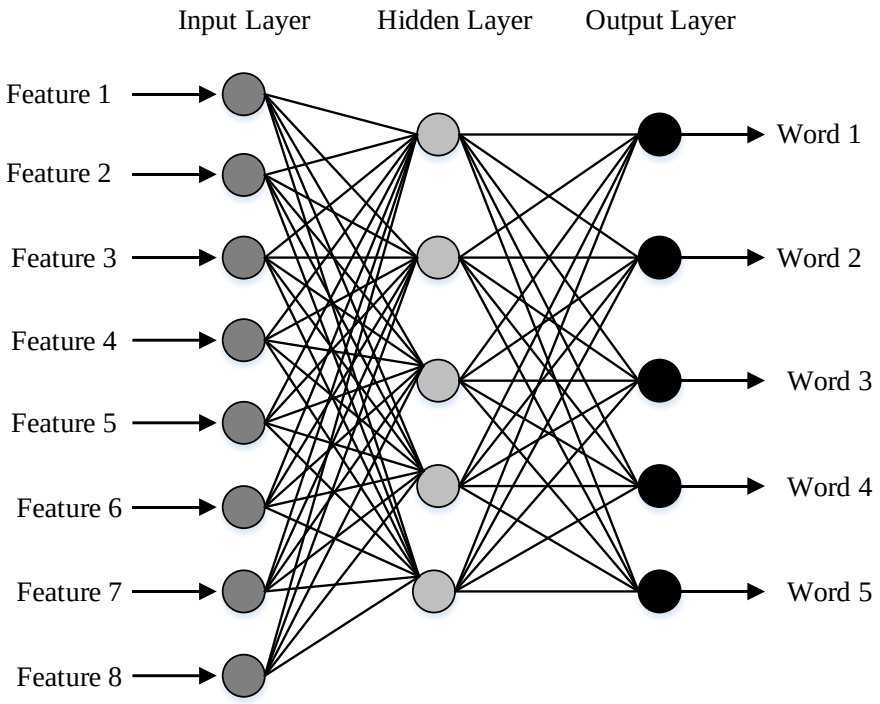

Fig. 4: Basic architecture of an ANN.

As given in the [Fig. 4] above, the input features are fed with the input layer at each input neuron. With this input the basis function is calculated at the hidden neurons present in the hidden layer with a weighting function assigned to the each link between the input layer and the hidden layer. Then the basis function is calculated at each output neuron of the output layer separately. These steps are repeated with another sample of the same words present in the signal and if any error produced Means the weighting function is optimized by Means of the CSO algorithm. The training and testing procedure of the ANN in recognizing the spoken words is explained as follows and the type of NN used here is the Feed Forward Neural Network (FFNN).

Training of ANN

Step 1.Define the selected features from the fuzzy evaluation index as the input neurons. We have employed eight types of features in recognizing the word and the fuzzy based feature selection technique will produce two best feature values for each type. Therefore, totally 16 feature values will be produced for training the NN as the input neurons.

Step 2.Then the basis function at each input neuron is to be calculated for each of the hidden neurons with the weighting function. This basis function calculation is given by the equation (23).

$$I_b = \sum_{j=1}^{J} u_j w_{jk}^{\ i} \quad (23)$$

where, J = Number of input neurons

u_j = Value of feature at each input neuron

$w_{jk}^{\ i}$ = weight value of the link between input neuron *j* and hidden neuron *k*

Step 3. After calculating the basis functions at the input neurons the activation function is calculated from that at each hidden neuron as given by the equation (24).

$$A_k = \left(1 + \exp\left(-I_b\right)\right)^{-1} \quad (24)$$

Step 4.Once the activation function is measured by each hidden neuron Means, the basis function of each of the output neurons is determined. This basis function calculation at the output side is given in equation (25).

$$O_b = \sum_{k=1}^{K} A_k w_k^{\ o} \quad (25)$$

where, $w_k^{\ o}$ = Weight value between the links of hidden neuron and output neuron.

The basis function produced at output side is the expected output to be produced at the recognizer. Sometimes, the value at the output neuron may deviate from the actual output and this is called learning error and denoted by the equation (26).

$$Learning\ Error, E_L = \frac{1}{2}\left(O_{act} - O_{obt}\right)^2 \quad (26)$$

where, O_{act} = Actual Output

O_{obt} = Obtained Output

Step 5.Optimization of Weight coefficients

The Cuckoo search optimization is employed in this part to select the corresponding weight coefficients so that the produced learning error is zero. Hence the fitness function used here is the learning error constrained to the weighting functions. The steps employed in the CSO algorithm are given in the following part.

Cuckoo search Optimization in Output, Weight Updating

Yang and Deb in 2009, developed a Cuckoo search algorithm which is one of the modern optimization algorithms that replicates some cuckoo species' breeding performance. Modern studies have exposed that CS is possibly far more effective than PSO and GA[28]. So that the CS algorithm is employed here for the optimization of the neural network to produce the zero learning error. The nature of the cuckoo search algorithm is detailed as follows.

Cuckoo proliferation approach

In communal nests some cuckoo species lay their eggs, though moderately a number of species employ in the obligate brood parasitism by laying their eggs in the host birds' nests (frequently other species). The brood parasitism principally falls into three classes, namely intra specific brood parasitism, cooperative breeding and the nest take over. After the eggs are laid, if the host birds could discover that the eggs are not their owns, they would either destroy the alien eggs or abandon their nests and new nests are built somewhere else; while some female cuckoo species can lay their eggs very specified in mimicry in pattern of the host bird's eggs. This lowers the probability of their eggs being discovered.

Many researches have presented that many insects and animals flight behavior might follow some distinctive characteristics of lévy flights. A common concern of lévy flights and random walk so as to attain new solution is offered in (27) and (28).

$$x^{i+1} = x^i + \beta \oplus Levy(\lambda), \qquad \beta > 0 \quad (27)$$

$$Levy(\lambda) = t^{-\lambda}, \qquad 1 \le \lambda \le 3 \quad (28)$$

where, x^{i+1} = New solutions (Here, learning rate)

By random levy walk ,some of the new solutions must be produced around the finest solution. Though, a significant fraction of new solutions must be formed by far field randomization. This would assure the algorithm not to be trapped in local optimums. So as to model the standard cuckoo search algorithm, the ensuing three idealized rules are developed.

- ✓ Just one egg at a time is laid by each cuckoo, and in a arbitrarily chosen nest it is dumped.
- ✓ The finest nests with high quality of eggs (solutions) would carry over to the next generation (algorithm iteration).
- ✓ The available number of host nests is constant, and each cuckoo's egg can be discovered by the host bird with the probability of $P_a \in [0,1]$.

Conferring to these three rules, the fundamental stages of CS can be précised as the pseudo code signified in[Fig. 5].

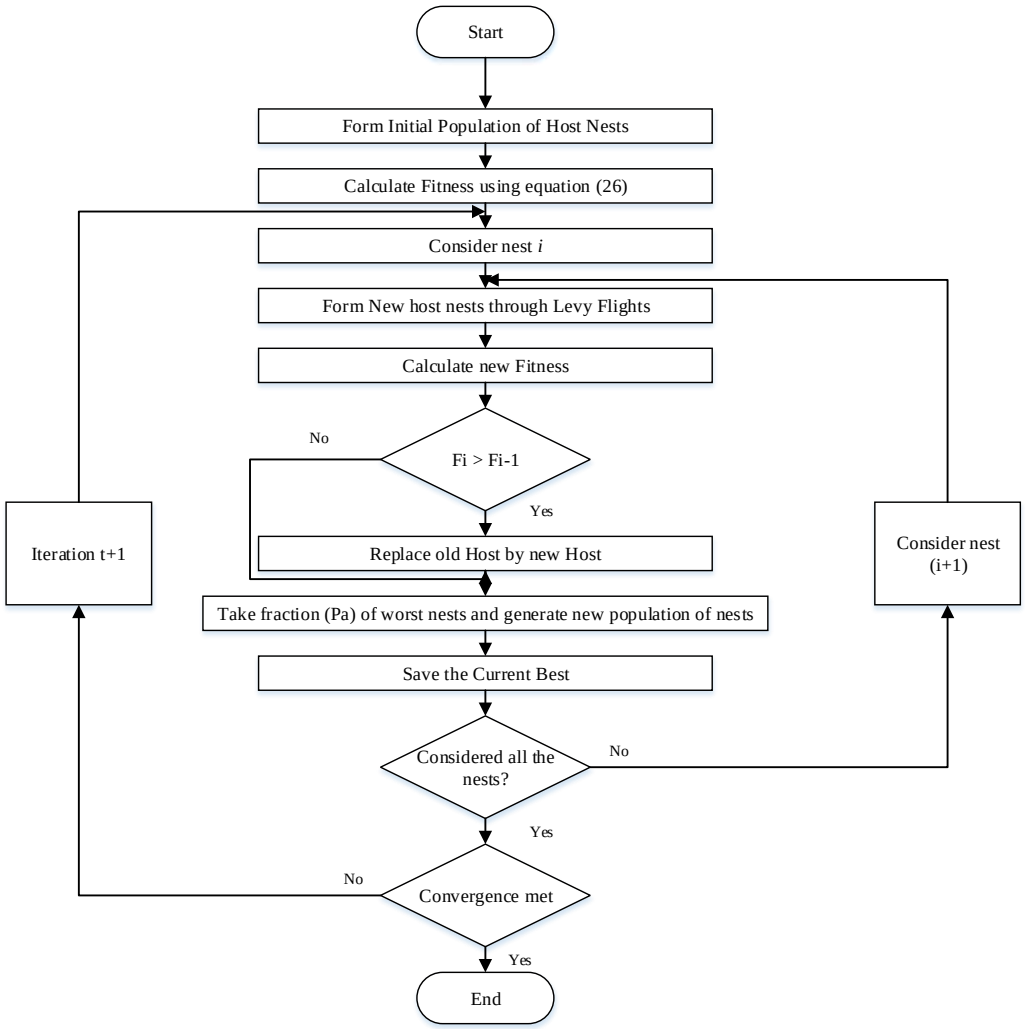

Fig. 5: Cuckoo search Algorithm.

As given in the flowchart depicted in [Fig. 5] above, initially the population (host nests) is defined with the weight coefficients as given in the section 3.3.1.1. After that the fitness function is calculated here the fitness function is the learning error of the neural network and the objective is the minimization of the error with best weighting coefficients. Once the fitness is calculated means the new host nests are formed through the levy flights by considering nest '*i*' and the fitness is calculated again. Then the calculated fitness value is compared with the one which is obtained in the previous stage and if the current one is better means the old host nest will be replaced with the newest one. Else the fraction of worst nests are taken to form the new population of nests. The best solution obtained till the process will be stored and if all the nests are considered or the maximum number of iterations are reached means the algorithm will be stopped and the best solution will be returned. Otherwise it will continues until are the nests are considered or else the maximum number of iterations reached. The weight coefficients are optimized in this way and the network is trained to produce the corresponding output values. After training the network with the maximum number of training samples the testing of the network is carried out and the testing procedure involved here is explained in the next section.

Testing of ANN

The testing procedure is also similar to that of the training of the neural network except that the learning error and hence the optimization is not done here. In the training stage the NN is trained to produce the words as recognized in the speech signal and in testing, the recognition performance of the network is validated. The experimental set up employed in our proposed work, the results, comparison with other works and the corresponding discussions are given in the experimental part.

RESULT AND DISCUSSION

In this paper, we have proposed a methodology of the ASR system for the man machine interaction with fuzzy based DWT feature extraction and the ANN optimized with the CSO algorithm. The proposed work is implemented in the working platform of MATLAB of version R2013a with the system configurations of Intel core i3 processor, 4GB RAM and Windows 8 Operating system. In this section, the dataset used in our proposed work, the results of the proposed method, the comparison results as well the discussions about the improvement of work is presented.

Dataset

In this paper, we use the Grid corpus [29] as a small-vocabulary ASR task to evaluate all approaches. The Grid database consists of 34,000 sentences that were uttered by 34 speakers, i.e., 1000 sentences per speaker. The task of the Grid corpus is to recognize sentences from a small vocabulary (51 words) with a fixed grammar of the form: command-color-preposition letter-digit-adverb. We have taken 100 speech data from the database among them 80% is used for training the recognizer and 20% is for testing the recognizer. The structure of the sentences in the grid database is given in the following [Table 1].

Table 1: Sentence Structure in Grid database

Command	Color	Preposition	Letter	Digit	Adverb
Bin	Blue	At	a-z (Except 'w')	0-9	Again
Lay	Green	By			Now
Place	Red	On			Please
Set	White	With			Soon

Preprocessing results of the speech signal

From this signal the preprocessing steps are applied to produce the enhanced version and to be suitable for the application of further operations. Initially, sampling of the signal is done to produce the sampled data at a sampling frequency of f_s = 10 kHz. After sampling, frames are produced with Hamming windowing at the specified intervals in section 3.1.2 and then noise present in the signals are removed with Harmonic Level decomposition. The preprocessing result of the sample signal from the database is given in the [Fig. 6] below.

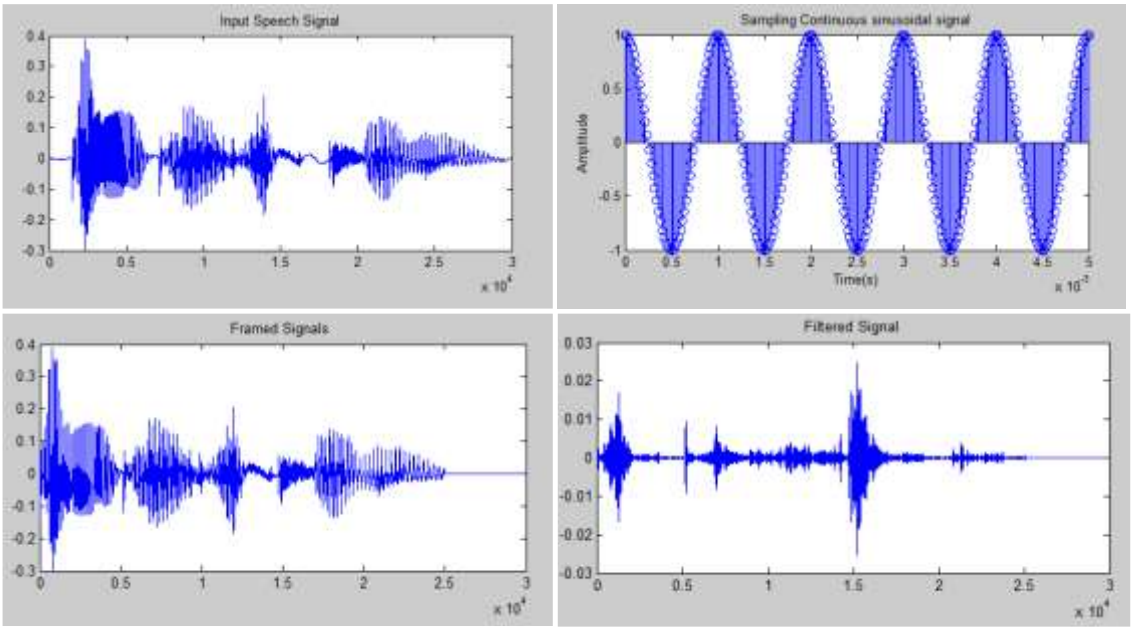

Fig. 6: Preprocessing result of the sample speech signal.

In [Fig. 6], the top left corner shows the original speech signal, the sampled signal produced is given in the top right corner, the frames of the signal is given in bottom left corner and the filtered signal is shown in the bottom right corner. After that eight different feature coefficients are extracted through 8-level DWT, which are totally 64 coefficients for each signal. From this two feature coefficients are selected optimally per each type of feature using the fuzzy model. The decomposed signal obtained with DWT feature extraction is shown in the following [Fig. 7].

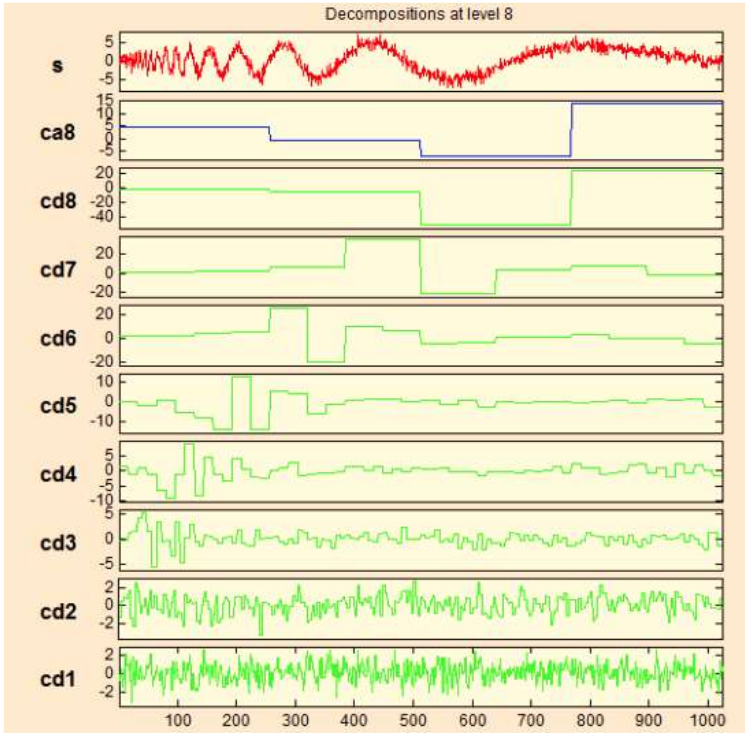

Fig.7: Eight level decomposition of the original speech signal by DWT.

The eight type of features calculated from these decomposition levels for the speech signal ' S_i ' is tabulated in the following [Table 2].

Table 2: Feature values of the signal S_i by DWT

Type of Feature	Feature Value
Mean	65.6667
Standard Deviation	58.0108
Skewness	-2.7095
Kurtosis	6
Entropy	0.5441
Shannon Entropy	-4.5223e^+5
Log energy Entropy	99.9938
Renyi's Entropy	-6.9389e^-16

Then the selected features from [Table 2] are used to train the NN in which the learning error is corrected by optimizing the weights at output neuron through CSO algorithm. The results of the speech recognition are given in the next section.

Results of recognition

In this section the results of our proposed speech recognition system for man machine interaction are presented. The performance of our proposed methodology is evaluated here based on the performance metrics such as recognition accuracy, word error rate, sensitivity, specificity, true positive rate and false positive rate.

Recognition accuracy

Recognition accuracy is defined as the performance metric which is used to measure the performance of a recognition system. The recognition accuracy R_a, is simply calculated using the following equation (29).

$$R_a = \frac{Number\,of\ recognized words}{Total\,number\,of\ words} \tag{29}$$

The greater the recognition accuracy of the system, the greater its recognition performance.

Word Error Rate (WER)

Word error rate is also the performance measure used to measure the recognition system performance. This is calculated directly from the equation (29) and given in equation (30).

$$WER = 1 - R_a \tag{30}$$

The lower the WER the better the performance of the speech recognition system.

Sensitivity

A measure of the capability of a method to properly recognize positive samplesis sensitivity. It could be computed using the subsequent equation.

$$Sensitivity = \frac{TP}{TP + FN} \tag{31}$$

The sensitivity value ranges between 0 and 1, where 1and 0mean best and worst recognition of positive samples, correspondingly.

Specificity

A measure of the ability of a method to recognize properly negative samples is specificity. It might be computed using the following equation.

$$Specificity = \frac{TN}{TN + FP} \tag{32}$$

The specificity value ranges between 0 and 1, where 0and 1 refers to worst and best appreciation of negative samples, correspondingly.

False positive rate (FPR)

It is the existence rate of positive test results in matters do not known to have the behavior for which an individual is being tested. The FPR is computed as in the following equation (33).

$$FPR = \frac{FP}{FP + TN} \tag{33}$$

False Negative rate (FNR)

It is the occurrence rate of negative test results in subjects referred to have the performance for which an individual is being verified. The FNR is computed as in the following equation (34).

$$FNR = 1 - TPR \tag{34}$$

ROC Curve

ROC curve clearly depicts the characteristic of any of the recognition system and it is the curve drawn between the TPR and FPR of the system. The ROC of our proposed methodology is shown in [Fig. 9]. The results of our proposed methodology in terms of these performance metrics are given in [Table 3] as well as in [Fig. 8].

Table 3: Results of Proposed Methodology

Performance Metric	Result
Recognition Accuracy (%)	95%
Word Error Rate (%)	5%
Sensitivity	0.95
Specificity	0.5
False Positive Rate (FPR)	0.5
False Negative Rate (FNR)	0.05

The results of our proposed methodology in terms of recognition accuracy and word error rate is represented graphically in the following [Fig. 8].

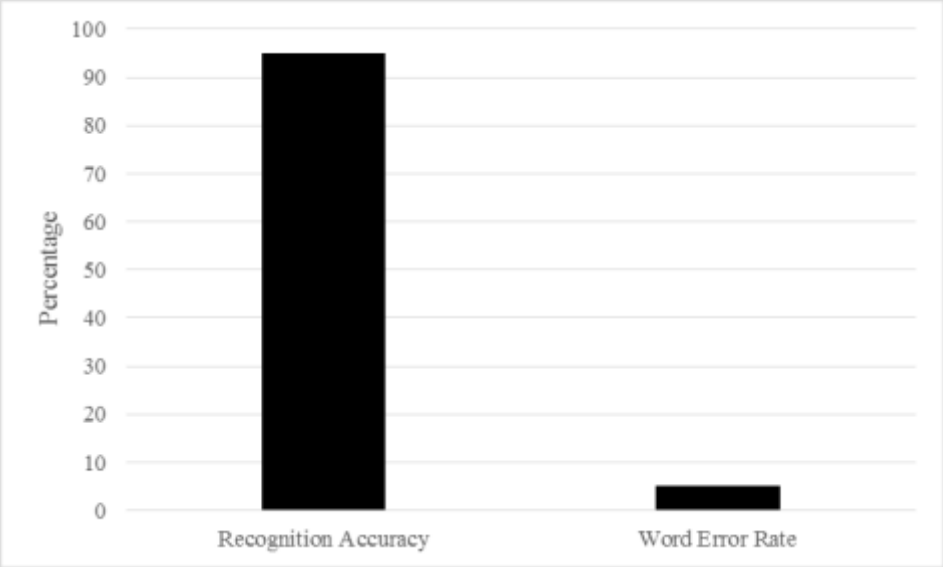

Fig.8: Performance of proposed methodology in terms of recognition accuracy and word error rate.

The performance of the recognizer is more understood by the performance metrics such as sensitivity, specificity, FPR and FNR for that the results obtained with our methodology is 0.95, 0.5, 0.5 and 0.05 respectively. This shows that most of the words are correctly recognized by our proposed methodology each signal. The results of our proposed method in terms of these parameters is presented in the following [Fig. 9].

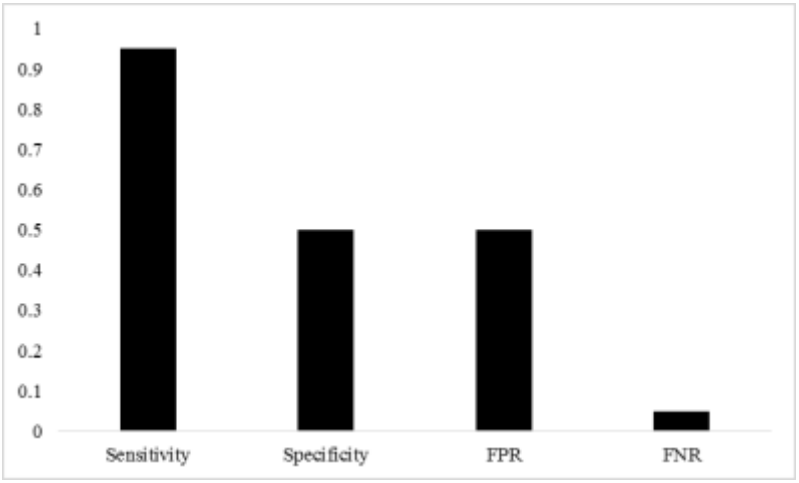

Fig. 9: Performance of proposed methodology in terms of Sensitivity, Specificity, False Positive Rate (FPR), False Negative Rate (FNR).

The corresponding ROC graph is presented in the following [Fig. 10]. It is the graph drawn between Sensitivity and FPR.

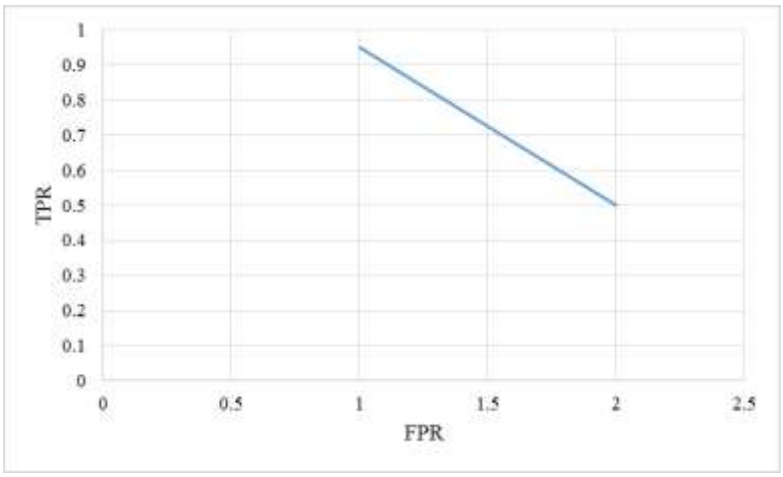

Fig. 10: ROC curve of our proposed methodology.

In order to prove the efficiency of our proposed methodology, it is compared with different techniques in terms of the performance metrics and this is presented in the following section.

Performance Comparison

The performance of our proposed man machine interaction system is validated with speech recognition with conventional back propagation and uncertainty-decoding based ASR [30] under the presence of babble noise at different levels with minimum and maximum number of features. The performance comparison results are given in [Table 4] and in [Fig.11].

Table 4: Performance Comparison of Proposed Methodology with other techniques

Performance Metric	Method														
	Proposed ASR with CS-ANN						**ASR without CS-ANN**						**Uncertainty-decoding based ASR (UD-ASR) [30]**		
	With 3 features (SNR in DB)			With 8 features (SNR in DB)			With 3 features (SNR in DB)			With 8 features (SNR in DB)			UD-GHF Decoding (SNR in DB)		
	0	5	10	0	5	10	0	5	10	0	5	10	0	5	10
Recognition Accuracy (%)	95	89.83	89.83	95	89.83	89.83	33.16	33.16	32.5	32.67	33.16	32.5	57.14	79.20	90.40
Word Error Rate (%)	5	10.17	10.17	5	10.17	10.17	66.84	66.84	67.5	67.33	66.84	67.5	42.86	20.8	9.6

The performance comparison results as given in the [Table 3] can be best understood by the result as given in the following [Fig. 11].

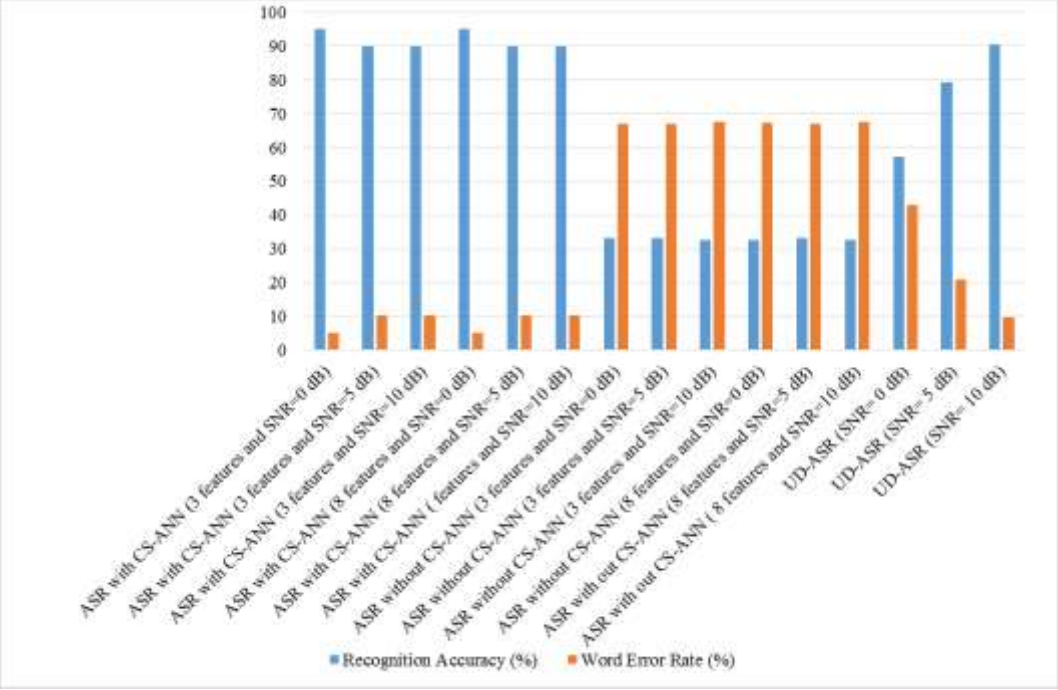

Fig. 11: Performance Comparison of Proposed Methodology (ASR with CS-ANN), ASR without CS-ANN and Uncertainty decoding based ASR with other techniques.

From the results of our proposed methodology as well as its comparison with the existing techniques the recognition rate achieved by our system is better than other methods such that the recognition accuracy is 95% with zero level of Babble noise and even after increasing the noise levels to 5 and 10 dB the accuracy of recognition is better than the method without optimization. Similarly UD-ASR method has the recognition accuracy

of 57.14%, 79.20% and 90.40% with the noise levels of 0 dB, 5 dB and 10 dB respectively and the discussion about these results is given in the following section.

DISCUSSION

The results shown in the [Tables 2-3] and in [Fig. 7-10] clearly depicts that better performance results are obtained with our proposed man machine interaction system for ASR compared with other techniques. The recognition rate achieved with our system is 95% which is better than the method proposed in [30] and without CS-ANN. This is clearly depicted in table 3 as well as in [Fig. 8], such that the system without CS algorithm in optimizing the ANN yields poor recognition results than the system with CS as well as the existing method [30]. However the accuracy of our proposed method is better than the method which we are taken into consideration even in the presence of babble noise at different levels, but the method in [30] achieved lower results compared to ours in lower SNR levels. The recognition results also shows that most of the words in the testing phase is identified clearly and hence the lower word error rate. Hence our proposed Fuzzy based DWT feature extraction produce better recognition and accuracy of the ASR system at low SNR levels and the accuracy level is maintained even the SNR level is increased which shows the efficiency of proposed method in achieving better accuracy results. The comparison result given in [Table 3] also shows that with the use of our proposed feature selection method the accuracy levels remains a stable one and with the CS-ANN classification the accuracy level is increased. In addition to that the optimization algorithm we have employed and the extracted features also have the major impact in recognizing the words as given by the comparison results.

CONCLUSION

In this paper we have proposed a methodology for man machine interaction system in ASR systems through fuzzy based feature selection and ANN optimized with CS optimization algorithm. The speech signals are initially converted into samples, frames using Hamming window and the noise levels are suppressed by harmonic decomposition. Next eight different features are extracted from the speech signal through DWT and the most relevant features selected by Means of fuzzy logic. The selected features are employed in training the NN in which the optimization of the network is performed by CSO algorithm. The experimental results are presented and compared with conventional technologies under various conditions and the results shows the efficiency of our proposed methodology. The experimental results given in section 4.3 reveals that the proposed ASR can achieve the recognition accuracy of 95% with lower word error rate of 5% which shows the betterment of our proposed work. The recognition accuracy achieved by our method is 95% with SNR level of babble noise at 0 dB, 89.83% at both the SNR levels of 5 dB and 10 dB respectively and it is a better result compared to the uncertainty decoding based method.

CONFLICT OF INTEREST
There is no conflict of interest between the authors regarding the manuscript.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
No financial contribution for my manuscript.

REFERENCES

[1] Harrag A. [2015] Nature-inspired feature subset selection application to arabic speaker recognition system, International Journal of Speech Technology 18(2): 245-255.
[2] El Ayadi M, Kamel MS, Karray F.[2011] Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44(3): 572-587.
[3] Furui S. [2010] History and development of speech recognition, In Speech Technology, Springer US, 1-18,.
[4] Yang B, Lugger M.[2010] Emotion recognition from speech signals using new harmony features, Signal processing, 90(5): 1415-1423.
[5] Patel I, Rao YS. [2010] A frequency spectral feature modeling for hidden markov model based automated speech recognition, In Recent Trends in Networks and Communications, Springer Berlin Heidelberg, 134-143.
[6] Barker J, Vincent E, Ma N, Christensen H and Green P. [2013] The PASCAL CHiME speech separation and recognition challenge, Computer Speech & Language 27(3): 621-633.
[7] Goh C ,Leon K. [2009] Robust computer voice recognition using improved MFCC algorithm, In Proceedings of International Conference on New Trends in Information and Service Science, 835-840.
[8] Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, Kingsbury B. [2012] Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Processing Magazine 29(6): 82-97.
[9] Siniscalchi S. M, Yu D, Deng L, Lee CH [2013] Exploiting deep neural networks for detection-based speech recognition, Neuro computing, 106:148-157.
[10] Dahl GE, Yu D, Deng L, and Acero A. [2012] Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, IEEE Transactions on Audio, Speech, and Language Processing 20(1): 30-42.
[11] Lee C, Hyun D, Choi E, Go J, Lee C.[2003] Optimizing feature extraction for speech recognition, IEEE Transactions on Speech and Audio Processing 11(1): 80-87.
[12] Al-Alaoui M A, Al-Kanj L, Azar J, Yaacoub E. [2008] Speech recognition using artificial neural networks and hidden

Markov models", IEEE Technology and Engineering Education (ITEE) 3(3): 77-86.
[13] Xi X, Lin K, Zhou C, and Cai J. [2005] A new hybrid HMM/ANN model for speech recognition, In Artificial Intelligence Applications and Innovations, Springer US, 223-230.
[14] Besson P, Popovici V, Vesin JM, Thiran JP, Kunt M. [2008] Extraction of audio features specific to speech production for multimodal speaker detection, IEEE Transactions on Multimedia 10(1): 63-73.
[15] Jensen R and Shen Q. [2007] Fuzzy-rough sets assisted attribute selection, IEEE Transactions on Fuzzy Systems 15(1): 73-89.
[16] Alcalá R, Gacto MJ, and Herrera F. [2011] A fast and scalable multi-objective genetic fuzzy system for linguistic fuzzy modeling in high-dimensional regression problems, IEEE Transactions on Fuzzy Systems, 19(4): 666-681.
[17] Weinland D, Ronfard R and Boyer E. [2011] A survey of vision-based methods for action representation, segmentation and recognition , Computer Vision and Image Understanding 115(2): 224-241.
[18] Zamani B, Akbari A, Nasersharif B and Jalalvand A [2011] Optimized discriminative transformations for speech features based on minimum classification error, Pattern Recognition Letters 32(7): 948-955.
[19] Chandrashekar G, and Sahin F. [2014] A survey on feature selection methods, Computers & Electrical Engineering, 40(1): 16-28.
[20] Mirhassani SM, Ting HN. [2014] Fuzzy-based discriminative feature representation for children's speech recognition, Digital Signal Processing 31:102-114.
[21] Kaya H, Ozkaptan T, Salah AA, Gurgen F. [2015] Random discriminative projection based feature selection with application to conflict recognition, IEEE Signal Processing Letters 22(6): 671-675.
[22] Cumani S, and Laface P. [2014] Large-scale training of pairwise support vector machines for speaker recognition, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 22(11):1590-1600.
[23] Chatterjee S, Kleijn WB.[2011]Auditory model-based design and optimization of feature vectors for automatic speech recognition, IEEE Transactions on Audio, Speech, and Language Processing 19(6): 1813-1825.
[24] Dikici E, Semerci M, Saraçlar M, Alpaydın E.[2013] Classification and ranking approaches to discriminative language modeling for ASR", IEEE Transactions on Audio, Speech, and Language Processing 21(2): 291-300.
[25] Pan ST and Li XY. An FPGA-based embedded robust speech recognition system designed by combining empirical mode decomposition and a genetic algorithm, IEEE Transactions on Instrumentation and Measurement 61(9): 2560-2572.
[26] Meseguer NA. [2009] Speech analysis for automatic speech recognition.
[27] T Yuvaraja, M Gopinath.[2014] Fuzzy Based Analysis of Inverter Fed Micro Grid inIslanding Operation International Journal of Applied Engineering Research ISSN 0973-4562 Volume 9, Number 22 (2014) pp. 16909-16916.
[28] 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
[29] T Yuvaraja, K Ramya. 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.