An Dual Tree Complex Wavelet Transform Based Approach for Automatic Emotion Speech Recognition along with Artificial Neural Network

R. Benickson Jeyaraj[1], A. Joel Livin[2], A. Kowski Rajan[3]

[1] M.E Student, Applied Electronics, LITES Thovalai ,India, [2] A.Joel Livin, HOD, Assistant Professor, ECE Dept, LITES Thovalai, India, [3] A.Kowski Rajan , Assistant Professor, ECE , LITES Thovalai,
[1]benicksonjeyaraj@gmail.com[2]joellivin.ece@lites.edu.in[3] kowski234762@gmail.com

Abstract

Automatic Speech Emotion Recognition from speech finds greater significance in better man machine interfaces and robotics. Speech emotion based studies closely related to the databases used for the analysis. The Automatic Speech emotion recognition is a standard diagnostic tool to distinguish the different types of emotions in speech signals. The detection of emotions is a challenging task since the small variations in speech signals cannot be distinguished. In this paper, dual tree complex wavelet transform (DTCWT) based feature extraction technique for automatic classification of emotions in speech signals is proposed. The feature set comprises of complex wavelet coefficients extracted from the DTCWT decomposition of a speech signal in conjunction with five other features (AC power, mean, standard deviation, kurtosis, and skewness) extracted from the signal. This feature set is classified using artificial neural network. The performance of the proposed feature set is compared with statistical features extracted from the sub-bands obtained after decomposition of the signal using discrete wavelet transform (DWT) and with five other features (AC power, mean, standard deviation, kurtosis, and skewness) extracted from the signal. The experimental results indicate that the DWT and DTCWT based feature extraction technique classifies speech emotion signals with an overall accuracy of 94.21% and 96.72%, respectively when tested over four types of emotions from eNTERFACE 2005 database.

Keywords*: Emotion Recognition, Discrete Wavelet Transform, Dual Tree Complex Wavelet Transform, artificial neural network, feature extraction, classification.*

I. INTRODUCTION

The interface between man and machine will become more meaningful if the machines can recognize the emotional contents. Emotions are the backbone of human interactions and are closely related to rational thinking perception, cognition and decision makings [1]. Emotional cues can be analyzed from speech, facial expressions and gestures. In this work we are focusing on recognizing emotions from speech [2]. Theories for emotional standards are mainly classified to two. One deals with discrete approach and the other deals with dimensional approach. The discrete approach related to universal basic emotions whereas dimensional approach characterizes and distinguishes different emotions [3]. Since speech is the primary medium for interaction, speech based emotional studies are more significant. Emotions in speech do not alter the linguistic contents of speech but changes its effectiveness. Automatic emotion recognition systems finds applications in Human-Computer Interfaces (HCIs), humanoid robotics, text-to speech synthesis systems, forensics, lie detection, interactive voice response systems etc [4].

Speech recognition is a pattern classification issue and speech recognition frameworks utilize detached word recognition [5]. Current speech recognition frameworks utilize a pattern matching methodology. The classifier, which is ordinarily in light of hidden Markov models (HMM) [6]. Hidden Markov model (HMM)-based speech recognition advances have grown extensively and can now get a high recognition execution. Voice dictation frameworks, spoken dialogue frameworks, and speech data interfaces are illustrative speech applications that utilization these modern innovations. These advancements lead us to anticipate that speech input interfaces will be inserted in practical applications. The improvement of speech input interfaces implanted in versatile terminals obliges recognition precision, scaling down, and low-power consumption [7]. Past examination on custom equipment portrayed the execution of the HMM calculation utilizing application specific integrated circuits (ASICs) [8] and field-programmable gate arrays (FPGAs) [9, 10].

There are a several sorts of parametric representations for the acoustic signals. Among them the Mel-Frequency Cepstrum Coefficients (MFCC) is the most broadly utilized. There are numerous reported works with MFCC, particularly on the improvement of the recognition accuracy [11]. The utilization of Mel Frequency Cepstral Coefficients can be considered as one of the standard technique for feature extraction [12]. The utilization of around 20 MFCC coefficients is basic in ASR, in spite of the fact that 10-12 coefficients are frequently thought to be adequate for coding speech [13]. The most prominent drawback of utilizing MFCC is its sensitivity to noise because of its dependence on the spectral structure. Strategies that use data in the periodicity of speech signals could be utilized to overcome this issue, despite the fact that speech likewise contains aperiodic substance [14]. It has been observed that feature extraction algorithms and classification method performs important role in the area of stuttered events recognition [15].

II. RELATED WORKS

Thomas M et al. [16] have proposed dual tree complex wavelet transform (DTCWT) based feature extraction technique for automatic classification of cardiac arrhythmias. The feature set comprises of complex wavelet coefficients extracted from the fourth and fifth scale DTCWT decomposition of a QRS complex signal in conjunction with four other features (AC power, kurtosis, skewness and timing information) extracted from the QRS complex signal. This feature set is classified using multi-layer back propagation neural network. The performance of the proposed feature set is compared with statistical features extracted from the sub-bands obtained after decomposition of the QRS complex signal using discrete wavelet transform (DWT) and with four other features (AC power, kurtosis, skewness and timing information) extracted from the QRS complex signal. The experimental results indicate that the DWT and DTCWT based feature extraction technique classifies ECG beats with an overall sensitivity of 91.23% and 94.64%, respectively when tested over five types of ECG beats of MIT-BIH Arrhythmia database.

El Ayadi M et al. [17] have provided survey of speech emotion classification addressing three important aspects of the design of a speech emotion recognition system. The first one is the choice of suitable features for speech representation. The second issue is the design of an appropriate classification scheme and the third issue is the proper preparation of an emotional speech database for evaluating system performance. Conclusions about the performance and limitations of current speech emotion recognition systems are discussed in the last section of this survey. This section also suggests possible ways of improving speech emotion recognition systems.

Ramakrishnan S, and El Emary I. M [18] presented a wide range of features employed for speech emotion recognition and the acoustic characteristics of those features. Also in that paper, they analyzed the performance in terms of some important parameters such as: precision, recall, F-measure and recognition rate of the features using two of the commonly used emotional speech databases namely Berlin emotional database and Danish emotional database. Emotional speech recognition was being applied in modern human-computer interfaces and the overview of 10 interesting applications is also presented in that paper to illustrate the importance of that technique.

Mencattini A et al. [19] proposed the use of a PLS regression model, optimized according to specific features selection procedures and trained on the Italian speech corpus EMOVO, suggesting a way to automatically label the corpus in terms of arousal and valence. New speech features related to the speech amplitude modulation, caused by the slowly-varying articulatory motion, and standard features extracted from the pitch contour, have been included in the regression model. An average value for the coefficient of determination R2 of 0:72 (maximum value of 0:95 for fear and minimum of 0:60 for sadness) is obtained for the female model and a value for R2 of 0:81 (maximum value of 0:89 for anger and minimum value of 0:71 for joy) is obtained for the male model, over the seven primary emotions (including the neutral state).

Pérez-Espinosa H et al. [20] reported the results obtained from experiments with a database of emotional speech in English in order to find the most important acoustic features to estimate Emotion Primitives which determine the emotional content on speech. They are interested in exploiting the potential benefits of continuous emotion models, so in that paper they demonstrate the feasibility of applying this approach to annotation of emotional speech and they explore ways to take advantage of this kind of annotation to improve the automatic classification of basic emotions.

III. PROPOSED METHODOLOGY FOR AUTOMATIC SPEECH EMOTION RECOGNITION

The objective of this paper is to present efficient recognizer using emotional speech recognition techniques in order to produce emotional states of human. This proposed technique consists of few steps including pre-processing, feature extraction, classification for recognition. The pre-processing is done in order to increase the effectiveness of feature extraction process. Feature extraction is to extract the features of the given speech signal. Then, based on the extracted features the given to classification for the recognition of emotions. The process flow of the proposed automatic speech emotion recognition system is shown in figure 3.1.

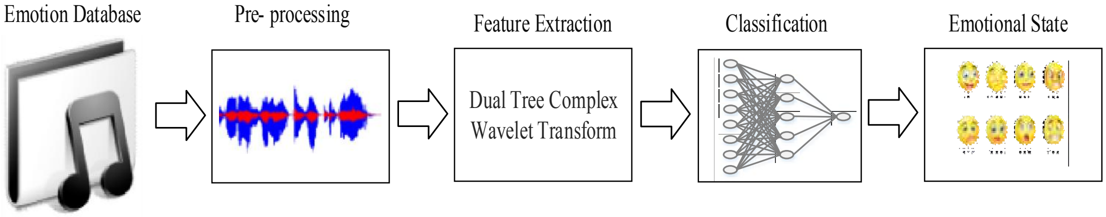

Figure 3.1: Architecture of Proposed Method

3.1 Input Database

The input database contains various speech signals which are spoken by different persons. These stored speech signals are considered as the input signal for the proposed ASR system. The input speech signal is represented as in equation 1.

$$S_i = x_i(t) \,\middle|\, i = 1,2,\ldots,N \rightarrow (1)$$

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

3.2Pre-processing of Speech Signals

The input speech signals have noise in it which will influence the recognition process. Consequently, the first imperative step in speech recognition is the pre-processing of the speech signals or acoustic signals which is executed to remove avoidable waveform of signal and to shorten the task of recognition. In this pre-processing, Wiener filter is applied to evacuate the noise which deals with the standard of spectral subtraction. This diminishes the noise by assessing the noiseless signal and afterward comparing with original signals. It take up that noise is stationary in examination with non-stationary signal in this way subtracting it from the original signal. After this process, signals are pre-emphasized to normalize the word counters by diminishing the high spectral dynamic range. The signal went through high pass FIR with distinctive component value. This procedure is finished by detecting the end points of the signal and evacuating the silence. Finally, a noise removed is obtained at the pre-processing step without the loss of original data which is given to feature extraction process.

3.3 Dual Tree Complex Wavelet Transform Based Feature Extraction

The Self-Organizing Map (SOM) was first introduced by TeuvoKohonen. It can organize multidimensional data in such a way that similar samples are closer to each other and less similar samples are further away from each other. Hence, it can be used to generate a codebook by setting the size of the SOM to be the size of the codebook.

This paper proposes the feature extraction technique based on a DTCWT technique which consists of complex wavelet coefficients extracted from the fourth and fifth scale of detail coefficients (D4 and D5). The obtained features are appended by four other features (i.e. power, mean, standard deviation, kurtosis, and skewness) extracted from the each speech signal.

- Discrete Wavelet Transform

The wavelet transform of a signal allows the representation of a signal in multiple scales and provides simultaneous localization of time and frequency. This is achieved by the decomposition of the signal over dilated (scale) and translated (time) versions of a prototype wavelet. An input signal is decomposed by using a low pass filter and high pass filter followed by down sampling in each stage.

The output of the first stage high pass filter gives the detail coefficient D1, whereas the low pass filter gives the approximation coefficient A1. The decomposition of a signal up to sample three scales is shown in Figure 3.2.

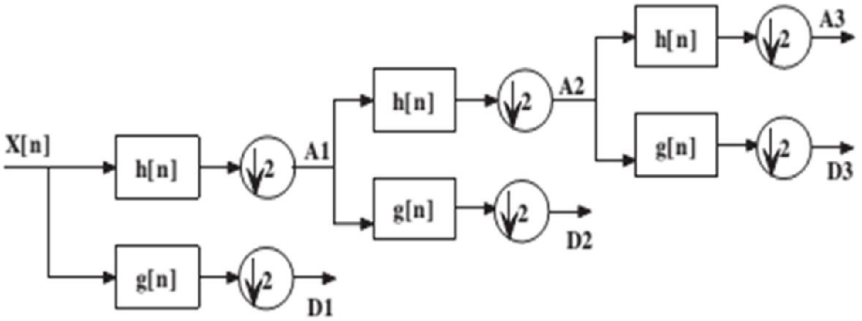

Figure 3.2: Three-level sub-band decomposition using DWT technique

The low pass and high pass filters used in each stage are quadrature mirror filters (QMF) since they satisfy the QMF condition given by

$$H(z)H(z^{-1})+H(-z)H(-z^{-1})=1 \rightarrow (2)$$

$$G(z)=zH(-z^{-1}) \rightarrow (3)$$

where H(z) and G(z) are the z transform of low pass and high pass filters respectively. They are related in time domain as

$$g[L-1-n]=(-1)^{n}.h[n] \rightarrow (4)$$

where L is the length of the filter. The output of low pass and high pass filters for an input signal x[n] is given by

$$y_{lpf}(k)=\sum_{n} x(n).h[-n+2k] \rightarrow (5)$$

$$y_{hpf}(k)=\sum_{n} x(n).g[-n+2k] \rightarrow (6)$$

The prototype wavelet used in our experimental study is Daubechies wavelet of order 4 (Db4) due to its morphological similarity with the speech signal. The feature extraction technique using DWT can be summarized as:

- Decompose the speech signal to eight levels using 1D DWT technique.
- The decomposition levels correspond to the detail coefficients D1-D8 and approximation coefficient A8. From each level five statistical parameters are computed i.e. Power, Mean, standard deviation, skewness and kurtosis of the wavelet coefficient.

- Dual Tree Complex Wavelet Transform

Though the conventional DWT technique is a powerful tool for analyzing ECG signals, it lacks the property of shift invariance. The DWT technique has been used for feature extraction. The conventional DWT technique is a powerful tool for analyzing 1D signals but it suffers from problems like oscillation, shift variance and aliasing. The detection and modeling of signal with singularities become complicated because the wavelet coefficient oscillates with positive and negative values around the singularities. The DWT technique lacks the property of shift invariance due to the down sampling operation at each stage of DWT implementation. Hence, the energy of the wavelet coefficient changes significantly for a small time shift in the input pattern. Due to this undesirable property the DWT coefficients fail to distinguish input signal shifts. A well-known way of providing shift invariance is to use the undecimated form of the dyadic filter tree but this suffers from increased computation requirements and high redundancy in the output. Wavelet coefficients are calculated using iterative time discrete operations with the non-ideal high and low pass filters. Therefore, aliasing can appear. Inverse DWT cancels aliasing, but only if the wavelet coefficients are not processed. The problem of aliasing will result in artifacts in the reconstructed signal. The DTCWT is a simple technique which overcomes the DWT shortcomings.

In the reported literature [21], the dual tree complex wavelet transform (DTCWT) features are used for automatic classification of cardiac arrhythmias. So here use DTCWT for automatic speech emotion recognition. The dual tree approach uses two real wavelet filters: one for acquiring the real part of the transform, and the other for the imaginary part. The combination of two such filters is termed as an analytic filter. The analytic filters give a new structure equivalent to two standard DWT filter bank structures operating in parallel as shown in Figure 3.3. The sub-band signals of the upper DWT (Tree A) can be interpreted as the real part of a complex wavelet transform and sub-band signals of the lower DWT (Tree B) can be interpreted as the imaginary part. Each tree uses different sets of filters that satisfy perfect reconstruction conditions. Let $h_0(n)$ and $h_1(n)$ be the low-pass and high-pass filters of the upper filter bank and $g_0(n)$ and $g_1(n)$ be the low pass and high pass filters of the lower filter bank. The filters $h_0(n)$ and $g_0(n)$ should be designed such that the wavelet functions are approximate Hilbert transform pair and they are related as

$$g_0(n) = h_0(n-0.5) \rightarrow (7)$$

The structure of 1D Q-shift dual-tree proposed by Kingsbury [22] is shown in Figure 3.3. In the implementation of DTCWT, there are two sets of filters used, one set of filters at level 1 and the other set of filters at all higher levels. The filters beyond level 1 have even length but are no longer strictly linear phase and have a group delay of approximately 1/4. The required delay difference of 1/2 sample is achieved by using the time reverse of the tree and filters in tree b [23]. Thus dual-tree complex wavelet transform employs two real DWTs, and complex coefficients only appear when the two trees are combined. The combination of two such trees helps in optimal representation for signals containing singularities (jumps and spikes).

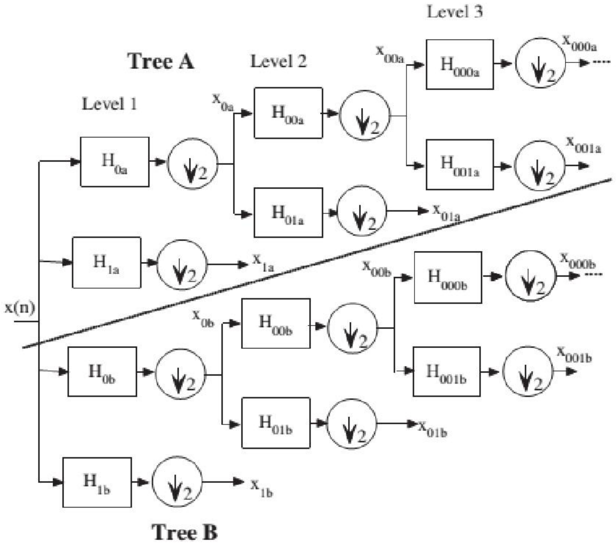

Figure 3.3: Dual Tree CWT

The DTCWT provides approximate shift invariance [24] with a limited redundancy factor of 21 for 1D signals, which is substantially lower than the undecimated DWT. The approximate shift invariance property of DTCWT is determined experimentally by determining the energy of fourth scale detail coefficients obtained from the decomposition of different levels. The variation in the energy of the complex wavelet coefficient is very small when compared to the energy of DWT coefficients. Thus the application of complex wavelet transform turns out to be superior to the DWT technique in pattern recognition problems.

The filters used in each stage are of length 10 [25]. The feature extraction technique using DTCWT is summarized as follows.

- Extract the speech signal by selecting a window of 256 samples around the signal
- Decompose the signal to eight resolution scales by using 1D DTCWT.
- Choose the features of DTCWT from 7th and 8th scale and compute the absolute value of the real and imaginary coefficients (detail coefficients) from each scale.
- Perform 1D FFT on the selected features and take the logarithm of the Fourier spectrum. The shift invariant property of DTCWT and FFT helps in classifying the signals efficiently.

In addition to the complex wavelet based features, five other features are extracted from the input speech signal.

a. AC Power of signal: The use of some sort of power measures in speech recognition is fairly standard today. Power is rather simple to compute. It is computed on frame by frame basis as

$$P(n) = (1/N_s)\sum_{m=0}^{N_s-1}\left(w(m)s(n-N_s/2+m)\right) \rightarrow (7)$$

Where Ns is the number of samples used to compute the power, s(n) denotes the signal, w(m) denotes the window function, and the ,and n denotes the sample index of center of the window.

b. Mean: The mean is represented by which provides the average value of a signal. It is produced by adding all samples together, and divide by N which is given by

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} C_{ij} \rightarrow (8)$$

c. Standard Deviation: The standard deviation is alike to the average deviation, excluding the averaging is done with power in its place of amplitude. This is attained by squaring each of the deviations before taking the average. The mathematical form is given by

$$\sigma_i = \sqrt{\left(\frac{1}{N}\sum_{j=1}^{N}(C_{ij}-\mu_i)^2\right)} \rightarrow (9)$$

d. Skewness: Skewness is a measure of the degree of asymmetry of a distribution. Skewness of the sample signal is calculated by using

$$\gamma = \frac{\frac{1}{n}\sum_{i=0}^{n}(y_i-\mu)^3}{\sigma^3} \rightarrow (10)$$

e. Kurtosis:Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case. The mathematical form is given by

$$K_i = \sqrt{\frac{N}{24}}\left(\frac{1}{N}\sum_{j=1}^{N}\left(\frac{C_{ij}-\mu_i}{\sigma_i}\right)^4 - 3\right) \rightarrow (11)$$

where i=1, 2… l is the decomposition level and N denotes number of coefficients of detail at each decomposition level.

3.3 Classification using Artificial Neural Network

The ANN consists of a single input layer, and a single output layer in addition to one or more hidden layers. All nodes are composed of neurons except the input layer. The number of nodes in each layer varies depending on the problem. The complexity of the architecture of the network is dependent upon the number of hidden layers and nodes. Training an ANN is to find a set of weights that would give desired values at the output when presented with different patterns at its input. The two main process of an ANN is training and testing. Figure 3.4 shows the structure of artificial neural network.

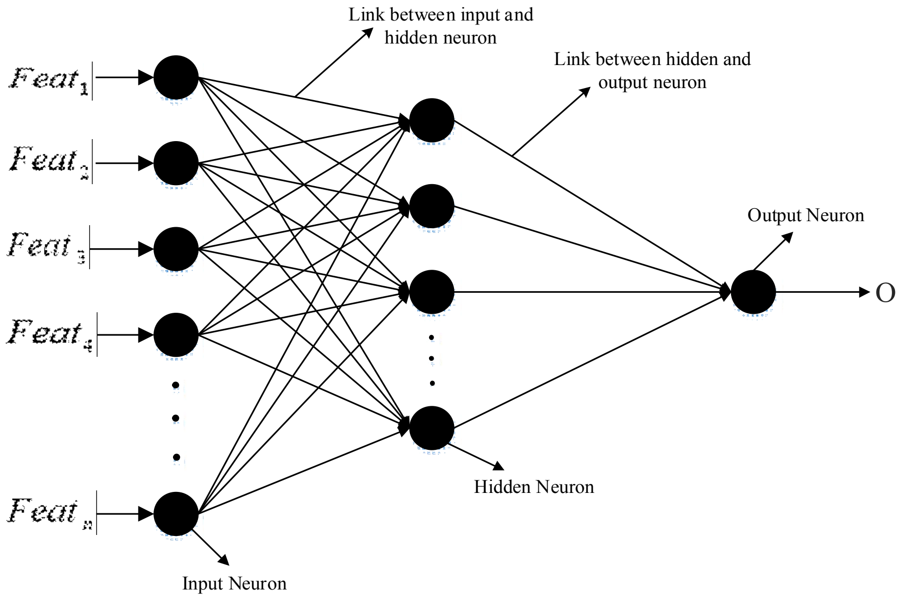

Figure 3.4: Structure of ANN

The proposed neural network consists of 24 input neurons they are feature vector obtained at the feature extraction process, one output neuron and M hidden units (M=4). First, the input data is transmitted to the hidden layer and then, to the output layer. This is called the forward pass of the back propagation algorithm. Each node in the hidden layer gets input from the input layer, which are multiplexed with appropriate weights and summed. The output of the neural network is obtained by the equation given below.

$$O = \sum_{m=1}^{M} \frac{w_m^O}{\left(1 + \exp\left[-\sum_{n=1}^{N} Feat_n w_{nm}^I\right]\right)} \rightarrow (12)$$

where C represents final output, m represents hidden neurons, n represents input neurons,w_m^o represents weights assigned between hidden neuron and output neuron, w_m^I represents weights assigned between input neuron and hidden neuron and F_n is the nth input value. The output of the hidden node is the non-linear transformation of the resulting sum. Same process is followed in the output layer. The output values from the output layer are compared with target values and the learning error rate for the neural network is calculated, which is given in equation.

$$err_l = \frac{1}{2}\left(\mathrm{Re}\,al - Obt\right)^2 \rightarrow (13)$$

Where represents lth learning error, Real represents actual output and Obt represents obtained output. The error between the neurons is transmitted back to the hidden layer. This is called the backward pass of the back propagation algorithm. Then the training is repeated for some other training dataset by changing the weights of the neural network. The error can be minimized using back propagation algorithm. Initially the weights are assigned to hidden layer neurons. The input layer has a constant weight, whereas the weights for output layer neurons are chosen randomly. Then, the output is calculated using equation output equation. The back propagation error $BacPr_{error}$ is calculated by using equation.

$$Bac\,\mathrm{Pr}_{error} = \sum_{l=1}^{k} err_l \rightarrow (14)$$

The weight deviation in the hidden neurons is calculated by using below equation

$$\Delta w = Bac\,\mathrm{Pr}_{error} \cdot \gamma \cdot \delta \rightarrow (15)$$

Where Δw is the weight deviation, δ is the learning rate, which can be chosen between 0.2 to 0.5, and γ represents the average of hidden neurons output which is given by

$$\gamma = \frac{1}{T}\sum_{i=1}^{T} H_i = \frac{1}{T}\sum_{i=1}^{T}\left(\frac{1}{1+\exp\left(-\sum_{n=1}^{N} Feat_n w_{nm}^{I}\right)}\right) \rightarrow (16)$$

where, γ is the average of hidden neurons output, N represents total number of input neurons, T represents total number of training samples and is the ith activation output at input side. The new weights is calculated by using equation which is given below.

$$w_{new} = w + \Delta w \rightarrow (17)$$

Where, w_{new} is the new weight and w is the current weight. This process is repeated until theBackPr$_{error}$<0.1 . If the Back propagation error reaches a minimum value, then artificial neural network is ready for classification. The flowchart of the proposed method is shown in figure 3.5.

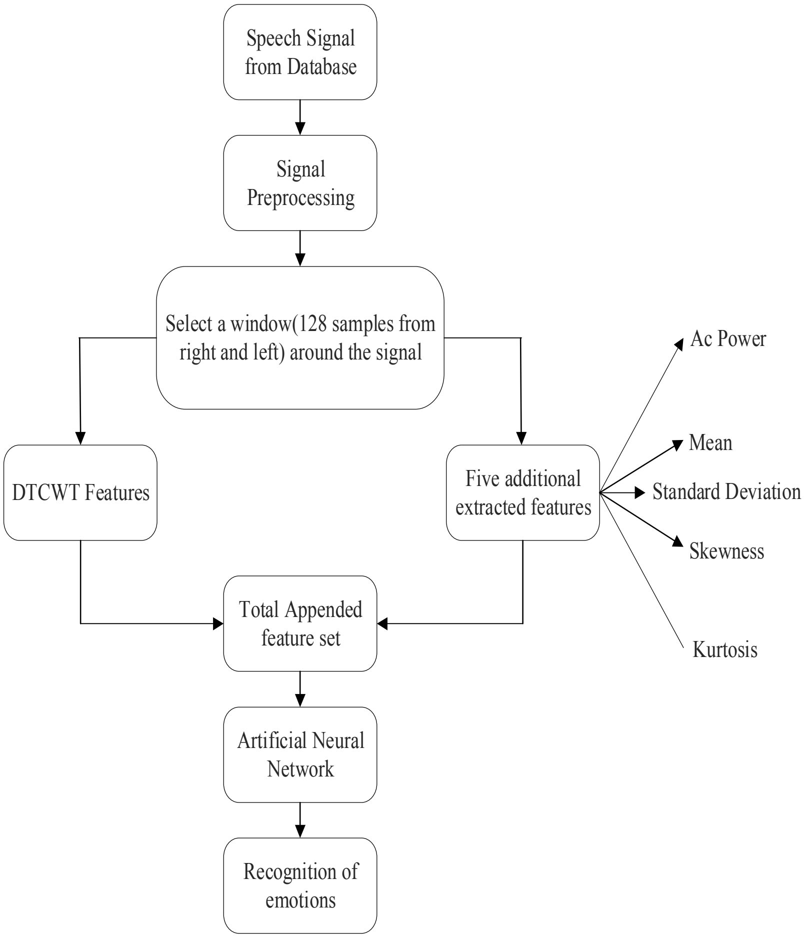
Speech Signal from Database
Signal Preprocessing
Select a window(128 samples from right and left) around the signal
DTCWT Features
Five additional extracted features
Ac Power
Mean
Standard Deviation
Skewness
Kurtosis
Total Appended feature set
Artificial Neural Network
Recognition of emotions

Figure 3.5: Flowchart of Proposed Method

IV. RESULT & DISCUSSION

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 8
Matlab version	: R2013a

In this paper, “eNTERFACE 2005 database” are used for evaluating the performance of the proposed method. The network is trained using 2000 signals (500 from Happy, 600 from Sad, 600 from Anger, and 300 from Neutral) and tested over the remaining data which corresponds to 22112 2500 beats (11464 900 from Happy, 3222 710 from Sad, 4311 300 from Anger, and 3115 590 from Neutral). The sample speech signal taken from the database is shown in figure 4.1.

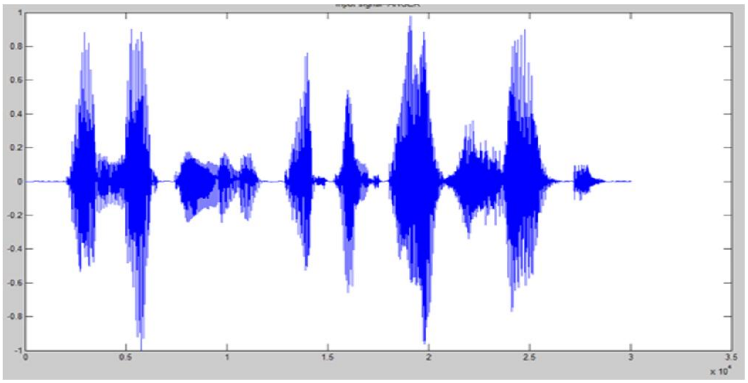

Figure 4.1: Sample Input Speech Signal

By applying pre-processing technique to this noisy signal we will get noise removed signal which is shown in figure 4.2.

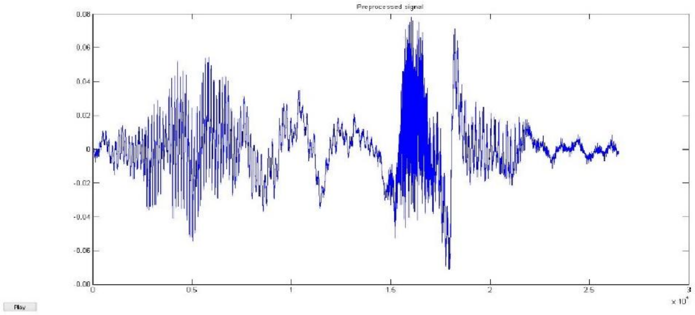

Figure 4.2: Pre-processed Speech Signal

Then by applying 8-level DWT decomposition of the speech emotion signal we get the decomposed signal which is shown in figure 4.3.

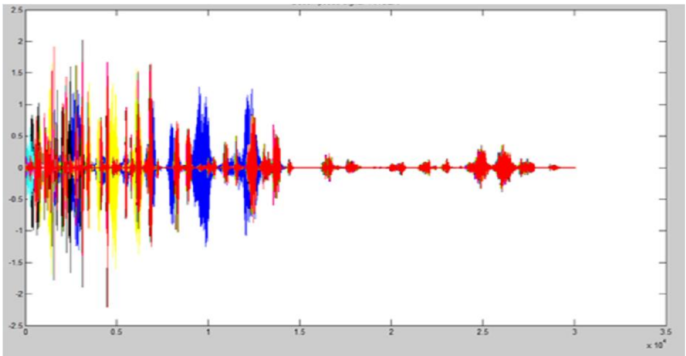

Figure 4.3: Decomposed Signal

The Filters used in each stage of DTCWT are of length 10. The sets of coefficient used in first level are shown in Table 4.1 and the remaining levels are shown in Table 4.2.

Table 4.1: Filter Coefficients used in first level

Tree A		**Tree B**	
H0a	**H1a**	**H0b**	**H1b**
0.00000000	0.00000000	0.01122679	0.00000000
-0.08838834	-0.01122679	0.01122679	0.00000000
0.08838834	0.01122679	-0.08838834	-0.08838834
0.69587998	0.08838834	0.08838834	-0.08838834
0.69587998	0.08838834	0.69587998	0.69587998
0.08838834	-0.69587998	0.69587998	-0.69587998
-0.08838834	0.69587998	0.08838834	0.08838834
0.01122679	-0.08838834	-0.08838834	0.08838834
0.01122679	-0.08838834	0.00000000	0.01122679
0.00000000	0.00000000	0.00000000	-0.01122679

Table 4.2: Filter Coefficients used in remaining level

Tree A		**Tree B**	
H00a	**H01a**	**H00b**	**H01b**
0.03516384	0.000000000	0.00000000	0.035163840
0.00000000	0.000000000	0.00000000	0.000000000
-0.08832942	-0.114301840	0.11430184	0.088388340
0.23389032	0.000000000	0.00000000	0.233890320
0.76027237	0.587518300	0.5875183	-0.760272370
0.5875183	-0.760272370	0.76027237	0.587518300
0.00000000	0.223890320	0.23389032	0.000000000
-0.11430184	-0.088388340	0.08832942	-0.114301840
0.00000000	0.000000000	0.00000000	0.000000000
0.00000000	-0.035163840	0.03516384	0.000000000

The classification results of ANN classifier using feature sets are shown in Table 4.3. The diagonal elements indicate the correctly classified beats corresponding to their respective classes. From the table, it is clear that 20 Happy, 30 Sad, 29 Anger, and 24 Neutral are misclassified using DTCWT technique. Similarly for DWT technique, 40 happy, 45 sad, 46 anger and 44 neutral signals are misclassified. From these results we can see that the utilization of DTCWT in placement of DWT shows improved results in classification of signals

Table 4.3: Classification Results obtained from the eNTERFACE 2005 database

Class	**Confusion Matrix (DWT)**				**Confusion Matrix (DTCWT)**			
	Happy	**Sad**	**Anger**	**Neutral**	**Happy**	**Sad**	**Anger**	**Neutral**
Happy	860	15	10	15	880	3	8	9
Sad	11	665	28	6	9	680	20	1
Anger	21	14	254	11	7	15	271	7
Neutral	14	5	25	546	5	0	19	556

It is important to classify speech emotion signals accurately. Classification performance is evaluated using four common metrics as given below.

$$Accuracy(Acc) = \frac{TP+TN}{TP+TN+FP+FN} \rightarrow (18)$$

$$Sensitivity(Sen) = \frac{TP}{TP+FN} \rightarrow (19)$$

$$Specificity(Spe) = \frac{TN}{TN+FP} \rightarrow (20)$$

$$Positive\Pr edicitivty(Ppe) = \frac{TP}{TP+FP} \rightarrow (21)$$

Where, TP, TN, FP and FN denotes true positive, true negative, false positive and false negative respectively. Accuracy is the measure of overall system performance over all the available classes, Sensitivity is the fraction of real events that are correctly detected among all real events, Specificity is the fraction of non-events that has been correctly rejected and Positive predictivity is the fraction of real events in all detected events. The overall detection sensitivity of DTCWT based feature is high due to the efficient classification of Happy, Sad, Anger and Neutral signals. Classification performance of the feature sets are shown in Table 4.4. The proposed method gives the highest classification sensitivity with average accuracy of 96.72%, sensitivity of 94.93%, specificity of 93.32% and positive predictivity of 96.32%.

Table 4.4: Classification Performance of DTCWT Based Technique

Method	**Class**	**Performance Matrix (%)**			
		Accuracy	**Sensitivity**	**Specificity**	**Positive Predictivity**
Discrete Wavelet Transform	**Happy**	93.23	95.56	91.32	90.78
	Sad	93.80	93.66	93.81	87.01
	Anger	90.49	84.67	88.34	99.41
	Neutral	99.32	92.54	99.95	96.49
Dual Tree Complex Wavelet Transform	**Happy**	97.77	97.78	97.66	97.63
	Sad	95.53	95.77	95.66	94.07
	Anger	97.05	90.33	83.09	97.53
	Neutral	96.53	95.86	96.87	96.08

V. CONCLUSION

In this paper, a novel technique is proposed for classifying emotions in speech signals using DTCWT based feature set. Five features (AC power, Mean, Standard Deviation, kurtosis, and skewness) extracted from each speech signal concatenated with the features extracted from the seventh and eighth decomposition levels of DTCWT, are used as total feature set. The performance of this method is compared with DWT based statistical features. In this paper, the ANN is used as a classifier because it has ability to learn and generalize, smaller training set requirements, fast operation, and ease of implementation. The major advantage of this network is that it finds the nonlinear surfaces separating the underlying patterns which is generally considered as an improvement on conventional methods and the complex class distributed features can be easily mapped by neural network. The proposed method has shown a promising sensitivity of 94.93% which indicates that this technique is an excellent model for automatic speech emotion. The performance of the proposed method is compared with DWT based statistical features and it is seen that the DTCWT feature set achieves higher recognition accuracy than DWT based feature.

REFERENCES

[1] [1] Nwe T. L, Fo S. W, and De Silva L. C, “Speech emotion recognition using hidden Markov models”, Speech communication, Vol. 41, No. 4, pp. 603-623, 2003.

[2] Lalitha S. L, Madhavan A, Bhushan B, and Saketh S, “Speech emotion recognition”, In Proceedings of IEEE International Conference on Advances in Electronics, Computers and Communications (ICAECC), pp. 1-4, 2014.

[3] Nicholson J, Takahashi K, and Nakatsu, “Emotion recognition in speech using neural networks”, Neural computing & applications, Vol. 9, No. 4, pp. 290-296, 2000.

[4] Koolagudi S. G, and Rao K. S, “Emotion recognition from speech: a review”, International journal of speech technology, Vol. 15, No. 2, pp. 99-117, 2012.

[5] Manikandan J, Venkataramani B, Girish K, Karthic H and Siddharth V, “Hardware implementation of real-time speech recognition system using TMS320C6713 DSP”, In Proceedings of IEEE International Conference on VLSI Design, pp. 250-255, 2011.

[6] Nadeu C, Macho D andHernando J, “Time and frequency filtering of filter-bank energies for robust HMM speech recognition”, Speech Communication, vol. 34, no. 1, pp. 93-114, 2001.

[7] Yoshizawa S, Wada N, Hayasaka N and Miyanaga Y, “Scalable architecture for word HMM-based speech recognition and VLSI implementation in complete system”, IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 53, no. 1, pp. 70-77, 2006.

[8] Han W, Hon K W, Chan C F, Lee T, Choy C S, Pun K P and Ching P C, “An HMM-based speech recognition IC”, In Proceedings of IEEE International Symposium on Circuits and Systems, ISCAS'03, vol. 2, pp. 744–747, 2003.

[9] Melnikoff S J, Quigley S F and Russell M J, “Implementing a simple continuous speech recognition system on an FPGA”, In Proceedings of IEEE Annual Symposium on Field-Programmable Custom Computing Machines, pp. 275-276, 2002.

[10] Rodríguez-Andina J J, Fagundes R D R and Júnior D B, “A FPGA-based Viterbi algorithm implementation for speech recognition systems” In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'01), vol. 2, pp. 1217-1220, 2001.

[11] Han W, Chan C. F, Choy C. S, and Pun K. P, “An efficient MFCC extraction method in speech recognition”, In Proceedings of IEEE International Symposium on Circuits and Systems, pp. 145-148, 2006.

[12] Ahmad K. S, Thosar A. S, Nirmal J. H, and Pande V. S, “A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network”, In Proceedings of IEEE International Conference on Advances in Pattern Recognition (ICAPR), pp. 1-6, 2015.

[13] Hagen A, Connors D. A, and Pellom B. L, “The analysis and design of architecture systems for speech recognition on modern handheld-computing devices, “In Proceedings of the IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis, pp. 65-70, 2003.

[14] Ishizuka K, and Nakatani T, “A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition”, Speech communication, vol. 48, no. 11, pp. 1447-1457, 2006.

[15] Ai O. C, Hariharan M, Yaacob S, and Chee L. S, “Classification of speech dysfluencies with MFCC and LPCC features”, Expert Systems With Applications, vol. 39, no. 2, pp. 2157-2165, 2012.

[16] Thomas M, Das M. K, and Ari S, “Automatic ECG arrhythmia classification using dual tree complex wavelet based features”, AEU-International Journal of Electronics and Communications, Vol. 69, No. 4, pp. 715-721. 2015.

[17] El Ayadi M, Kamel M. S, and Karray F, “Survey on speech emotion recognition: Features, classification schemes, and databases”, Pattern Recognition, Vol. 44, No. 3, pp. 572-587, 2011.

[18] Ramakrishnan S, and El Emary I. M, “Speech emotion recognition approaches in human computer interaction”, Telecommunication Systems, Vol. 52, No. 3, pp. 1467-1478, 2013.

[19] Mencattini A, Martinelli E, Costantini G, Todisco M, Basile B, Bozzali M, and Di Natale C, “Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure”, Knowledge-Based Systems, Vol. 63, pp. 68-81, 2014.

[20] Pérez-Espinosa H, Reyes-García C. A, and Villaseñor-Pineda L, “Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model”, Biomedical Signal Processing and Control, Vol. 7, No. 1, pp. 79-87, 2012.

[21] Thomas M, Das M. K, and Ari S,“Automatic ECG arrhythmia classification using dual tree complex wavelet based features”, AEU-International Journal of Electronics and Communications, Vol. 69, No. 4, pp. 715-721, 2015.

[22] Kingsbury N,“Complex wavelets for shift invariant analysis and filtering of signals”, Applied and computational harmonic analysis, Vol. 10, No. 3, pp. 234-253, 2001.

[23] Afsar F. A, and Arif M,“Robust electrocardiogram beat classification using discrete wavelet transform”, In Proceedings of IEEE International Conference on Bioinformatics and Biomedical Engineering, pp. 1867-1870, 2008.

[24] Selesnick I. W, Baraniuk R. G, and Kingsbury N. G,“The dual-tree complex wavelet transform”,IEEE Signal Processing Magazine,Vol. 22, No. 6, pp. 123-151, 2005.

[25] Manoharan S.“A dual tree complex wavelet transform construction and its application to image denoising”, International Journal of Image Processing (IJIP), Vol. 3, No. 6, pp. 293.