International Journal of Pure and Applied Mathematics
Volume 118 No. 20 2018, 2969-2977
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Enhancement of the Speech with Emotion Recognition using Artificial Intelligence

MR.S. LOKESH.M.E..,
Assistant professor, Department of Electronics and Communication Engineering,
Vel Tech, Avadi – Alamathi Road,
Avadi, Tamil Nadu, India.
lokesh@veltechengg.com

Banupriya.V, Jaya pretheena.P
UG Scholar Department of Electronics and Communication Engineering,
Vel Tech, Avadi – Alamathi Road,
Avadi, Tamil Nadu, India.
Banupriyavenkatesan.v@gmail.com , jayapretheena0304@gmail.com

***Abstract*-Automatic speech emotion recognition (SER) plays an important role in HCI systems for measuring people's emotions has dominated psychology by linking expressions to group of basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise). In this work, we introduce a split based enhancement to signal which we are loading. This paper describes about speech enhancement and speech emotion recognition from speech signal based on features analysis and NN-classifier. Speech enhancement is done based on the phase and frequency value of the particular signal.The recognition system involves speech emotion detection, features extraction and selection and finally classification. These features are useful to distinguish the maximum number of samples accurately and the NN classifier based on discriminant analysis is used to classify the six different expressions. The simulated results will be shown that the filter based feature extraction with used classifier gives much better accuracy with lesser algorithmic complexity than other speech emotion expression recognition approaches.**

Index terms: NN-Neural Network, HCI-Human Computer Interface, GUI-Graphical User Interface, MFCC-Mel Frequency Cepstrum Coefficient.

1. INTRODUCTION

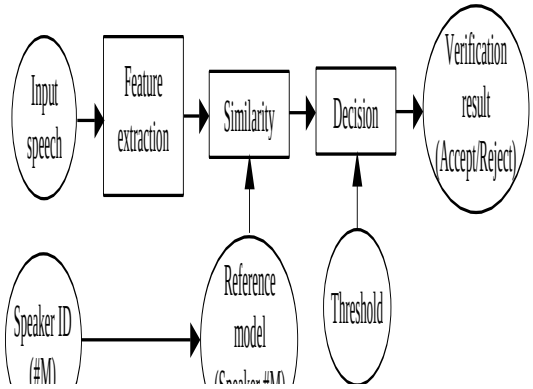
Generally the sounds of spoken language were studied at specific stages (1) phonetic additives of spoken words, e.g., vowel and consonant sounds, and (2) acoustic wave patterns. A language maybe broken down into a totally small variety of fundamental sounds, called phonemes (English has approximately forty). [7]An acoustic wave is a series of changing vibration patterns styles (usually in air), but we are more accustom to "seeing" acoustic waves as their electrical analog on an oscilloscope (time presentation) or spectrum analyzer (frequency presentation). Also visible in sound analysis are dimensional patterns called spectrograms, which show frequency (vertical axis) vs. time (horizontal axis) [10].

2. SPEAKER RECOGNITION PRINCIPLES

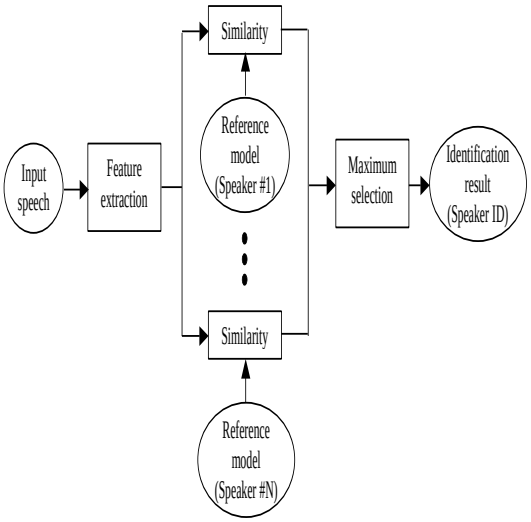
Speaker recognition is divided into 2 types,identification and verification,Speaker identification demonstrates for determining what registered speaker gives utterance. Speaker verification, same time to do process of selecting or eliminating the unique claim of a speaker. Figure 1 shows the fundamentals of speaker identification and verification systems. The system shows detailed text-independent speaker identification system therefore its aim is to findthe person who speaks nearer to speaker of what is saying.

At top level, all recognition systems undergoes two important characteristics (refer to Figure 1). Feature

extraction is the included for extractaction of little amount of information from the voice signal that will be used later to identify each speaker. Feature matching introduces for real steps of finding the unknown speaker by differentiate with extracted features from others voice to compare givens speakers.Figure shows the speaker verification and speaker identification.



Speaker Verification



Speaker Identification

Figure 1. Speaker Recognition Systems

2.1 Speech Feature Extraction

The idea is to convert the speech waveform,into a set of features (at a considerably lower information rate) by (DSP) for further analysis. This is always said to be signal-processing session.
The speech signal is normally a time varying signal (it is called quasi-stationary). An example is shown in Figure 2. When determine bya sufficiently a small time period (between 5 and 100 msec), mainly they are fairly stationary.

A large density of possibility is parametrically finding the speech signal for the speaker recognition content, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others[20].

MFCC's are depend upon the known variation between the human ear's critical bandwidths of frequency, filters are employed linearly at low frequencies to high frequencies have been used to determine the phonetically main area of speech. This is said mel-frequencyscale,that is linear frequency spacing in the range of 1000 Hz and less than that and a logarithmic spacing more than 1000 Hz.

2.2. Mel-Frequency CepstrumCoefficients Processor

A below diagram is the structure of an MFCC processor is shown in Figure 2. The speech input is gradually recorded at a sampling rate more than 10000 Hz. This sampling frequency was decided to minimize the factors of aliasing in the ADC

These sampled signals may determine all frequencies till 5 kHz, which occurs more energy of sounds that are given by humans. As seen earlier the importantance of the MFCC processor is to describe the behavior of the human ears. Better than speech waveforms MFFC's are less susceptible to mentioned variations.

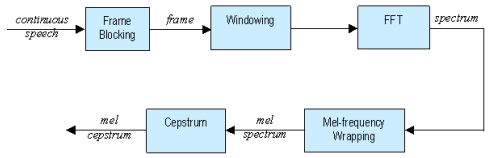


Figure 2.Block diagram of the MFCC processor

2.3. Frame Blocking

Here the continuous speech signal is framed into the frames ofN samples, the adjacent frames are separated by M(M<N).It continuous for until all the speech is

accounted for themselves one or more.Typical values for N and M are N=256.

2.4.Windowing

The next process is to processing the window by each unique frame to minimize the signal discontinuities at the starting and last session of each frame. The technique is to minimize the spectral distortion by usage of window to taper the signal to zero at the starting and last of each frame. wediscribes the window as $w(n), 0 \le n \le N-1$, here N is the number of samples in individual frame, then the result of windowing is the signal[8].

$$y_l(n) = x_l(n)w(n), \quad 0 \le n \le N-1$$

Typically the Hamming window is used here

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

2.5 Fast Fourier Transform (FFT)

Now Fast Fourier Transform, that converts every frame of N samples the time domain into the frequency domain.[6] It is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples $\{x_n\}$, as follow:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,..., N-1$$

normallyX_k's are complex numbers since only absolute values (frequency magnitudes) are consisdered. The output sequence $\{X_k\}$ is interpreted as follow: positive frequencies $0 \le f < F_s/2$ correspond to values $0 \le n \le N/2-1$, where negative frequencies $-F_s/2 < f < 0$ correspond to $N/2+1 \le n \le N-1$. Here, $F_s$denotes the sampling frequency.

2.6 Cepstrum

This is last step, here convertingmel spectrum back to time. The output said to be the mel frequencycepstrum coefficients (MFCC). [6]The cepstralof the speech spectrum is a good determinationof the local spectral characteristics of the signal for the determined frame analysis.Since the mel spectrum coefficients (and so their logarithm) are real numbers, we may themthem to the time domain using the Discrete Cosine Transform (DCT). Therefore denoting themel power spectrum coefficients is the results of last process $\tilde{S}_0, k = 0,2,..., K-1$, we can find the MFCC's, $\tilde{c}_n$, as

$$c_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad n=0,1,...,K-1$$

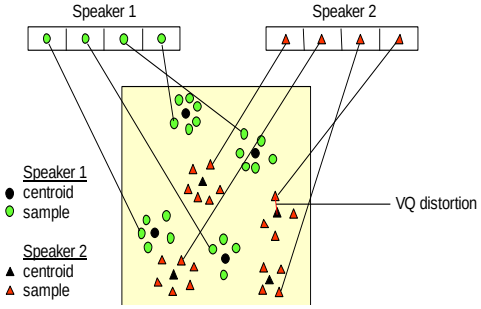
3.FEATURE MATCHING

The demerits of speaker recognition depend on a topic in scientific and engineering hence said pattern recognition. The aim of pattern recognition is to divide objects of interest into categories or classes[9]. The objects of interest are normally called patterns and since are series of acoustic vectors that are derived by input speech. The classes here says speakers. Since the classification modules here is applied on extracted features, they may also referred to as feature matching[13-20].

The remaining patterns are used for testing the classification algorithm; these patterns are collectively said be test set.

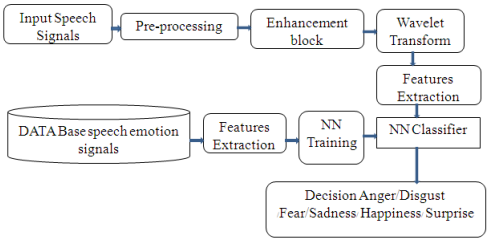
Figure 3 shows a conceptual diagram of recognition process.Itrefer to the acoustic vectors of speaker 1 while the triangles of the speaker 2[17]. While training phase, using the clustering algorithm.The distance from a vector to closest code word in codebook is called a VQ-distortion.

Figure 3.Conceptual diagram illustrating vector quantization codebook formation.



3.1Proposed Method Analysis:-

Figure 4.Classifier Block



3.2 Input Voice Signal

A input signal is one of the frequencies within part of the audio range,that used in the transmission range. In telephony,the using range of voice frequency band range from approximately 300Hz to 3400 Hz.

3.3 Preprocessing

3.3.1 Sampling

Sampling is done for function varying in space , time or any other dimension and similar outputs are obtained in two or more dimension.The sampling frequency or sampling rate fs is defined as the number of samples obtained in one second (samples per second , thus fs=1/Ts).

3.3.2QuantizationQuantization, in mathematics and DSP, step of mapping a big set of input signals to a light set – such as roundingvalues.Aalgorithm that undergoes quantization is called**Quantizer**. The error made by quantization is known as quantization error or round-off error. Quantization is depends to some degree in around all digital signal processing, as the step of define a signal in digital form normally says as rounding. Quantization forms for the core of frequently all lossy compression algorithms.The first type, commonly called rounding quantization, is the mainly usedfor many applications, to show the use of approximate definition for few quantity that is to be noted and used in further steps[13]. This module says the easy method of approximation that uses in everyday arithmetic. This may also includes analog-to-digital conversion of a signal. Hence the purpose is mandatory to retain as signal fidelity as possible. In these the design always centersnearly managing the approximation error

3.4.WaveletDecomposion

Awave is an modulating function of time or space that is periodic. It an infinite length continuous function in time or space[14]. In attractive wavelets are localized waves. A waveletis a waveform of an efficiently small duration that has an average value of zero [11].Basically the wavelets are obtained from a single prototype wavelet Called the motherwavelets[2].

3.5.FeatureExtraction Speech signal contains huge number of parameter which shows emotion content .If any change in these parameter indicate change in the emotions. The proper choice of feature vector is most important one. So many approaches toward automatic recognition of emotion [17]. Feature vector may be divided into long-time and short-time feature vectors. The long-time is estimated once. Almost all length of the utterance,hence the short-time ones are determined over window of usually less than 100ms. The long-time approach identifies emotions more efficiently. Short time features uses interrogative phrases which has large pitch contour and a larger pitch standard deviation.**3..6.NNClassifier**Neural network are predictive models lossy depends on the action of biological neurons[16]. The selection of the “neural network” was of the great PR successes of the Twentieth description such as “A network of weighted, addictive values with nonlinear transfer function”. Hence, despite the name, neural networks are far from “thinking machines” or “artificial brains”. A typical artificial neural network might have a hundred neurons. In comparison, the human nervous system is believed to have about $3x10^{10}$neurons[19].

4.ALGORITHMUSED The algorithm is divided into four types:i)Feed-forwardcomputationii)Back propagation to the output layeriii)Back propagation to the hidden layeriv) Weight updates

4.1 DWT Algorithm representation The discrete wavelet transform is gained by applying complementary low-pass and high-pass filters. And subsequent decimation (H and L). Both H and L are applied to data vector x_1, x_2, ..., x_8. The result of H is the four wavelet coefficients for the first resolution; the output of L is the four coefficients of the scaling function[12]. The wavelet coefficients of the other resolution levels are obtained by iterating the low- and high-pass filtering steps on the coefficients of the scaling function [3].

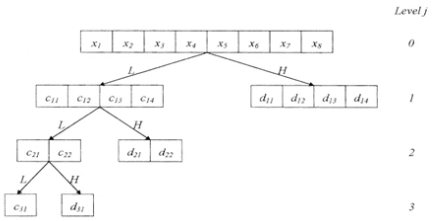


Figure 5. Dwt Algorithm Representation

5 .EXPERIMENTAL RESULT

5.1 Features of Angry Signal

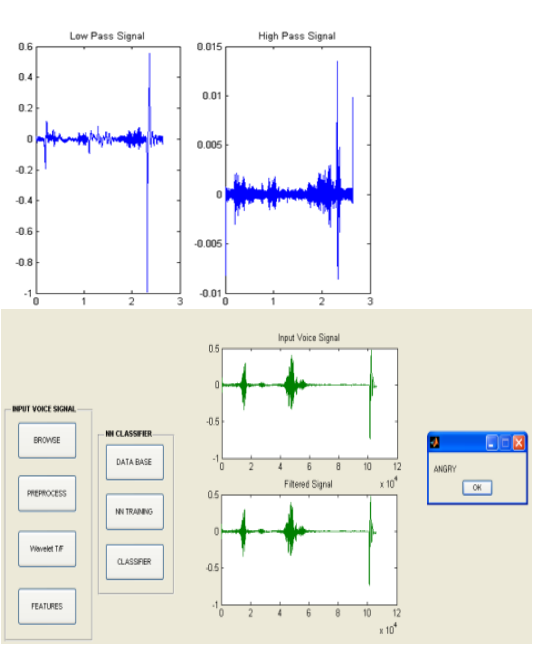


Figure 6.Classification response of Angry Signal

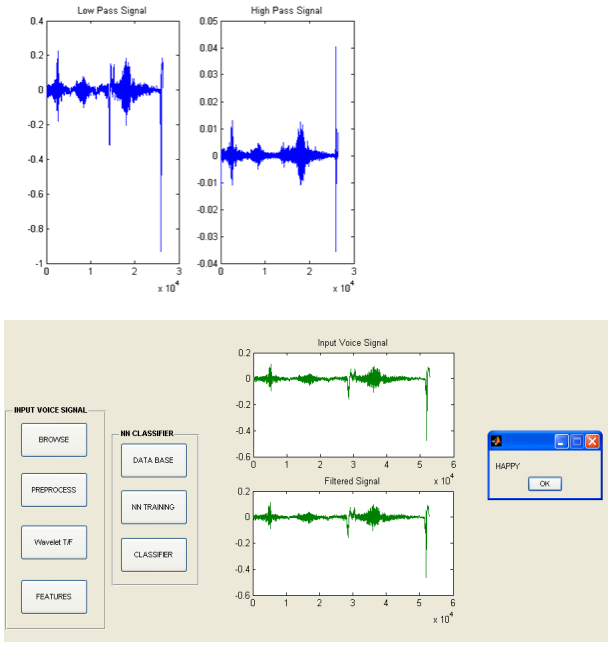


Figure 7.Classification response of Happy Signal

5.1.1 Preprocessing

Fig 6 shows the classification response of angry signal. In that browser can take the recorded voice (input) signal and preprocessing is used to remove the unwanted noise signal (i..e)the signal other than the input signal and filtering process is done here.

5.1.2 Wavelet Transform

The wavelet transform is used to divide a continuous time function into wavelets.

5.1.3 Feature Extraction

The feature extraction explains the value of entropy and zerocrossing. Such that feature extraction of peak point is 0.5543, mean signal is -0.9941 and minimum level signal is 1.9137e-005.

5.1.4 NN Trainer and Classifier

Database that compare the input signal with the other loaded signal. NN trainer that trains the nn classifier to give the accurate emotion of the loaded input signal. Finally the emotion of the input signal is successfully shown in the figure 6.(i.e) the experimental result of the Angry signal.

5.2 . Features of Happy Signal

5.2.1 Preprocessing

Fig 7 shows the classification response of happy signal. In that browser can take the recorded voice (input) signal and preprocessing is used to remove the unwanted noise signal (i.e) the signal other than the input signal and filtering process is done here.

5.2.2 Wavelet Transform

The wavelet transform is used to divide a continuous time function into wavelets.

5.2.3 Feature Extraction

The feature extraction explains the value of entropy and zero crossing. Such that feature extraction of peak point is 0.2268, mean signal is -0.9305 and minimum level signal is -9.1273e-004

5.2.4 NN Trainer and Classifier

Database that compare the input signal with the other loaded signal. NN trainer that trains the nn classifier to give the accurate emotion of the loaded input signal. Finally the emotion of the input signal is successfully shown in the figure 7. (i.e) the experimental result of the Happy signal.

5.3Features of Sad Signal

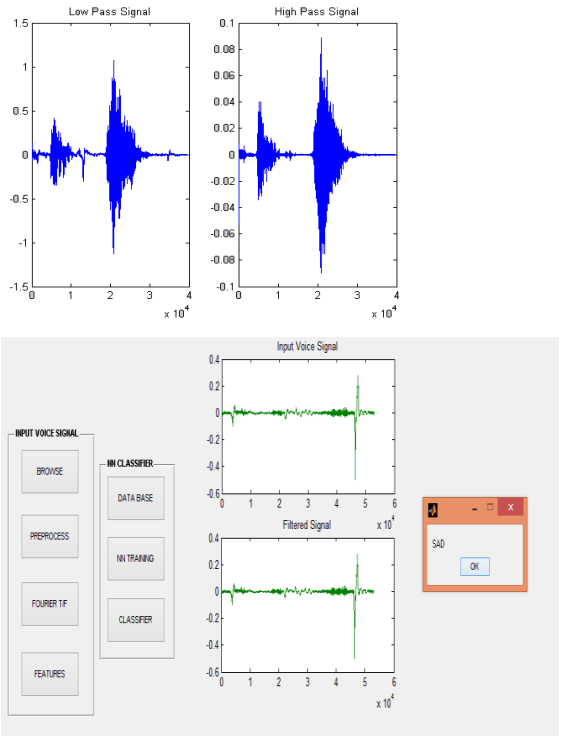


Figure 8.Classification response of Sad Signal

5.2.1 Preprocessing

Fig 8 shows the classification response of happy signal. In that browser can take the recorded voice (input) signal and preprocessing is used to remove the unwanted noise signal (i.e) the signal other than the input signal and filtering process is done here.

5.2.2 Wavelet Transform

The wavelet transform is used to divide a continuous time function into wavelets.

5.2.3 Feature Extraction

The feature extraction explains the value of entropy and zero crossing. Such that feature extraction of peak point is 1.0781, mean signal is -1.1281 and minimum level signal is -1.1351e-005.

5.2.4 NN Trainer and Classifier

Database that compare the input signal with the other loaded signal. NN trainer that trains the nn classifier to give the accurate emotion of the loaded input signal.

Finally the emotion of the input signal is successfully shown in the figure 8. (i.e) the experimental result of the Sad signal.

	Person1	Person2	Person3
Happy Signal	-0.23489	-0.04037	-0.01666
Angry Signal	-0.17551	-0.31076	-0.273
Normal Signal	-0.16705	-0.06045	-0.11124
Sad Signal	-0.14658	-0.00109	-0.00618

Figure 9.Different Emotional Signal Response

In nn classifier training a database has been created for the above mentioned four emotions for 3 persons shown in fig 9. Its peak, minimum, average values are calculates and a overall mean value for each emotion of the individual is stored in a table. Based on the above table the neural network classifier is trained for each of emotion. Thus when a input signal is given the neural network classifier compares its mean value with those values in the table and estimate the correct emotion of the speaker.

6. CONCLUSION

In this work, most recent work done in the field of Speech Emotion Recognition is discussed. Frequently used method used methods of feature extraction and several classifier performances are reviewed. Success of emotion recognition is dependent on appropriate feature extraction as well as proper classifier selection from the sample emotional speech. It can be seen that Integration of various features can give the better recognition rate.

Classifier performance is need to be increased for recognition of speaker independent systems. The application area of emotion recognition from speech is expanding as it opens the new means of communication between human and machine. It is needed to model effective method of speech feature extraction so that it can even provide emotion recognition of real time speech. This speech emotion recognition has various application in robotics and medical field. The sole purpose of this process is to help in man machine interaction. In robotics, Robot can recognize emotional information as contained in the human speech signals for friendly interaction with human beings, and eventually satisfactory performance and effect are realized. In call-center application.the experiments results indicate that the proposed method provides very stable and successful emotional classification performance and it

promises the feasibility of the agent for mobile communication services.

7. REFERENCE

[1] Y. Li and Y. Zhao, "Recognizing Emotions in Speech Using Short- Term and Long-Term Features," Proc. Int'l Conf. Spoken Language Processing, pp. 2255-2258, 1998.

[2] N. Virag, "Single channel speech enhancement based on maskingproperties of the human auditory system," *IEEE Trans. Speech AudioProcess.*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[3] P. Loizou, "Speech enhancement based on perceptually motivatedbayesian estimators of the magnitude spectrum," *IEEE Trans. SpeechAudio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.

[4] T. Vogt, E. Andre, "Improving Automatic Emotion Recognition from Speech via Gender Differentiation", inProc. of Int'l Conf. of Language Resources and Evaluation,2006.

[5] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll,"Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech," Proc. Int'l Conf. Spoken Language Processing, pp. 2225-2228, 2007.

[6] H. Hu and C. Yu, "Adaptive noise spectral estimation for spectral subtractionspeech enhancement," *Signal Process., IET*, vol. 1, no. 3, pp.156–163, Sep. 2007.

[7] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction,"*Speech Commun.*, vol. 50, no. 6, pp. 453–466, 2008.

[8] C. S. Seelamantula and T. V. Sreenivas, "Blocking artifacts in speech/audio: Dynamic auditory model-based characterization andoptimal time-frequency smoothing," *Signal Process.*, vol. 89, no. 4,pp. 523–531, 2009.

[9] A. P. Stark and K. K. Paliwal, "Group-delay-deviation basedspectral analysis of speech," in *INTERSPEECH*, 2009, pp.1083–1086.

[10] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*. SanRafael, CA, USA: Morgan and Claypool, 2009, Synthesis Lectures onSpeech and Audio Processing

[11] J. I. Marin-Hurtado and D. V. Anderson, "Distortions in speech enhancementdue to block processing," in *Proc. ICASSP*, Mar. 2010, vol.1, pp. 4774–4777..

[12] K. Paliwal, K. Wojcicki, and B. Shannon, "The importanceof phase in speechenhancement,"*SpeechCommunication*,vol. 53, no. 4, pp. 465 – 494, 2011.

[13] T. Gerkmann and M. Krawczyk, "Mmse-optimal spectralamplitude estimation given the stft-phase," *Signal ProcessingLetters, IEEE*, vol. 20, no. 2, pp. 129 –132, feb. 2013.7466.

[14] Dr. AntoBennet, M, SankarBabu G, Natarajan S, "Reverse Room Techniques for Irreversible Data Hiding", Journal of Chemical and Pharmaceutical Sciences 08(03): 469-475, September 2015.

[15] Dr. AntoBennet, M ,Sankaranarayanan S, SankarBabu G, " Performance & Analysis of Effective Iris Recognition System Using Independent Component Analysis", Journal of Chemical and Pharmaceutical Sciences 08(03): 571-576, August 2015.

[16] Dr. AntoBennet, M, Suresh R, Mohamed Sulaiman S, "Performance &analysis of automated removal of head movement artifacts in EEG using brain computer interface", Journal of Chemical and Pharmaceutical Research 07(08): 291-299, August 2015.

[17] Dr. AntoBennet, M "A Novel Effective Refined Histogram For Supervised Texure Classification", International Journal of Computer & Modern Technology , Issue 01 ,Volume02 ,pp 67-73, June 2015.

[18] Dr. AntoBennet, M, SrinathR,RaishaBanuA,"Development of Deblocking Architectures for block artifact reduction in videos", International Journal of Applied Engineering Research,Volume 10, Number 09 (2015) pp. 6985-6991, April 2015.

[19] AntoBennet, M &JacobRaglend, "Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images", Journal of Computer Science, vol. 8, no. 9, pp. 1447-1454, 2012.

[20] AntoBennet, M &JacobRaglend, "Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering", European Journal of Scientific Research, vol. 66 no.1, pp.120-129, 2011.