A radial base neural network approach for emotion recognition in human speech

Lal Hussain[1*,2], Imran Shafi[3], Sharjil Saeed[2], Ali Abbas[2], Imtiaz Ahmed Awan[2], Sajjad Ahmed Nadeem[2], Syed Zaki Hassan Kazmi[2], Saeed Arif Shah[2], Saqib Iqbal[2], Bushra Rahman[4]

[1*]Quality Enhancement Cell, The University of Azad Jammu and Kashmir, City Campus, 13100, Muzaffarabad, Pakistan

[2]Department of Computer Science & IT, The University of Azad Jammu and Kashmir, City Campus, 13100, Muzaffarabad, Pakistan

[3]Faculty of Computing & Technology, Abasyn University, Islamabad, Pakistan

[4]North Suffolk Cardiology, Stony Brook Medicine, 101 Nicolls Road, Stony Brook, NY 11794, USA

Summary

Emotions play a vital role during verbal communication in our daily life as only the textual information cannot convey the complete information. Emotions in human speech is a complex phenomenon, which vary from person to person based on gender, anger, varying activities and spoken languages. In this work, a novel technique based on artificial neural networks (ANN) is proposed to recognize real-time emotions such as anger, disgust, fear, happiness, sadness and surprise. First, the noise and silence are filtered from recorded speech using adaptive filtering. Secondly, the acoustic and statistical features are extracted from the filtered speech. Set of uncorrelated features are obtained by using principal component analysis (PCA). The input and target features are used to train the feed forward neural network (FFNN), generalized regression neural network (GRNN), Elman network and radial basis feed forward neural network (RBFNN). Performance analysis based on test results indicates that the RBFNN gives better performance and recognition rate than FFNN, GRNN and Elman network.

Keywords:

Speech Signal Processing, Neural Network, Human Computer Interaction, Support Vector Machine, Receive Operating Curve, Principle Component Analysis

1. Introduction

Emotion recognition plays an important role in our daily lives; however, scientific knowledge is limited regarding the human emotions. There are many applications of HCI; the knowledge of which is gained through emotional experience in the human as well as it provides a relation between effective expressions and emotional experience. Recently many novel database have been created, which contains the emotional expressions. Most of these databases contain the visual, speech or audio visual information. The visual data usually contains the face and or body gestures. While, audio data contains the varying genuine or acted emotions of human speech in many languages. Some examples of such database include MMI is containing web based posed, facial expressions as well as audio video emotions of human [1].

There are many kinds of emotional robotics such as emotional robots, intelligent robots and cognitive robots. A stimulus in human brain produces changes in emotion. The information processed is then directed to the central nervous system (CNS), motor nervous system and autonomic nervous system (ANS). The basic emotion defined by Ekman viz. Anger, happiness, disgust, fear, sadness, surprise etc. are then determined by CNS based on the information received [2].

The cognitive process in human is supported by emotions that help us to work effectively and efficiently. Moreover, it helps us to mediate our social interactions. There are many automatic recognition models used widely in games and communication channels etc. [3].

During conversation, meanings and words as well as emotions are delivered. Human emotions are recognized by speech and facial expressions. When the people communicate two kinds of information is exchanged such as linguistic information (Verbal) and paralinguistic information (emotion state, tone, gestures etc.). There are multimodals HCI through which human interact with computer systems e.g. fatigue detector in intelligent automobile systems help the driver to monitor his/ her vigilance and he/she may apply appropriate actions to avoid from the accident. Moreover, during conversation auditory signal take the information of many kinds. When the people communicate, the verbal part does not convey the correct message without the pertinent utterance. Various kinds of emotions technically help us to perceive the information. Moreover, the auditory features of speech signals are estimated during input of the audio signals. There are many researches in psychology, which provides the acoustic features extracted from the emotional speech. The prosodic features include pitch, energy, speech rate while, spectral features involve MFCC and voice quality parameters [4].

In Southern California University, speech emotion research group has integrated the acoustic features and spoken words to differentiate between positive and

Manuscript received August 5, 2017
Manuscript revised August 20, 2017

negative emotions by using Probabilistic Neural Network (PNN) and Hidden Markov Model (HMM) with Bayesian, Support Vector Machine (SVM), nearest Neighbour Method (NNM) and C4.5 classifier. These classifiers and models help us to recognize basic emotions [5, 31].

The intelligent systems use automatic speech recognition systems like virtual agents, mobile phones, in-car interfaces and many other pattern recognition models. This system also helps and increases the acceptance among potential users. Some other applications of automatic speech recognition include the call centres, interest recognition database, spontaneous speech evaluation and other prototypical emotions. Besides, there are some other emotion recognition systems to recognize real-life emotions such as dialogue systems, surveillance, tasks and media retrieval. Likewise, for acoustic features, the linguistic information is derived from the Dynamic Bayesian Network (DBN). In addition, Sensitive Artificial Listener (SAL) database is used for emotion recognition which contains the natural colour speech and helps the recognition rate from high to low valence state [6].

The multimedia contents from the user are guessed through Bio inspired multimedia by interpreting response during media appreciation. For example, during sonification rules in brainwave music interface, EEG characteristics are mapped such as intensity, note and pitch. In order to model the human emotional responses such as facial expressions, speech and other body gestures, some conventional methods are used. Bio-signals are recorded from automatic nervous systems in periphery including skin conductance, respiration, electromyography etc, which provide more complex detail in comparison with audio visual information [7].

Face, pedestrian, objects and facial expressions can be recognized using Haar like features. While to detect speech, speaker recognition, gender classification and emotions are recognized using Mel-Frequency Cepstrum Coefficient (MFCC). In addition, standard deviation and average on each frame of the input signal are helpful to recognize the acceleration signal. Besides, SVM and decision tree classifiers are used to recognize human activities based on control parameters [8].

HCI is used efficiently to recognize different emotions based on varying behaviour of the user. Any kind of communication can be made efficient and effective through particular emotion. The most important and effective mode of communication is attained through speech emotion recognition. Human speech comprises of linguistic information and emotions. LPCs, MFCCs, voice energy and fundamental frequency formants are used to recognize speech and spears. There are diverse applications of automatic emotion recognition such as speech recognition systems, forensics, text to speech systems, humanoid, robotics and medical domains. Linguistics and psychological fields are closely related to emotional recognition systems [9].

Human feelings can be best described by speech and emotions of human. To detect these emotional states some methods such as physiological measurements (involving blood volume, heart rate, blood pressure, conductance level, skin resistance, papillary response, skin temperature, respiration rate and brain potential) are used. Besides it, human emotions play a vital role to make rational decisions, perception and learning as well as many other biased cognitive tasks [10].

Computer visions systems help us to monitor the people and play an important role in our lives. Human face detection, face tracking, person identifiers, action recognition, gender classification and age estimations helped us in HCI systems to control the passive surveillance in smart buildings. These systems restrict the access of people to certain areas based on stated features while some other demographic information make decision such as number of women authenticated to enter the store [11].

Many classifiers are used to recognize speech emotions which include Linear Discriminant Classifier (LDC), K-nearest Neighbourhood Classifier to recognize both positive and negative emotions of speech signals. Moreover, feature selection is made by SVM. Emotions can be recognized using Probabilistic Neural Network (PNN) and Gaussian Mixture Model (GMM), while basic acoustic features such as intensity, pitch, and MFCC extract the basic features of emotion from the speech signals [12].

Several feature selection methods are published in the literature such as random forest ensemble learning, decision tree and genetic algorithms. Besides, Canonical Correlation Analysis (CCA) can be used to extract the most appropriate and relevant features of the emotional states. Because, all extracted features are not important to recognize the particular emotions. Thus, feature selection methods help to remove irrelevant features and select the most appropriate features which maintain the classification accuracy and low computational complexity [13].

2. Artificial Neural Network- A review

Neural network processor is a parallel distributed processor. It can have the natural tendency to store the experiential knowledge, which is made useful for us. The neural network is like the human brain in the aspect that network acquired knowledge through the learning process. Moreover, in order to store that knowledge synaptic weights are used.

The emotion recognition is proposed in real time based on age, gender and varying activities. The data is collected

from diverse people and University students before training to reflect emotions in a real sense and in different sessions keeping in view the varying activities, situations and paralinguistic properties. All sounds recorded are in Urdu language. In case of noisy situation, FFT and Adaptive filtering techniques are applied to filter the sound from noise to extract exact features. To extract the correct information from speech signals, pre and post processing techniques are used. The features extraction results are passed to four different neural networks to check the performance. The Regression analysis and Mean Square Error has helped us to judge the performance.

Artificial neural intelligence is used in forecasting problems by the economists. Predictions are required for any inventory is to acquire or to build a factory that there would be no serious depression for the next few months. Change point detection theory has many developments. Moreover, there are some new directions which are emerged and these problems are widely used in dynamic systems and time series problems. ANN is a biological inspired computational model. It consists of processing elements known as neurons and connection between them. The weights are bounded to these connections. Neural network is used in many applications but has limited applications in remotely sensed imagery (RSI) [14]. At present 120,000 gigabytes of Land sat series data is collected over the last three to four decades. This data is stored in the EROS Data Centre. Besides, a large number of aerial photographs and other RSI data are also archived in various locations of the world. The scientists and engineers are facing big challenges to use the huge amount of data to help fanners. The confidentiality, integrity and authenticity of the information resources is protected by the cryptosystems [15]. In order to meet the encryption and decryption standard specifications, these systems are required to meet the stringent specifications regarding the information security. In the past few years, there is an increasing interest in the applications of cryptography through the neural networks. These cryptography applications include the visual cryptography, management, generation and exchange of protocol, pseudo random generator etc. Moreover, ANN is used to predict the diseases like cancer and cardiology illness [16]. It helps the physician with decision support in future. ANN is used in real time clinical prediction. The term diagnostic means to detect and isolate of faults or failure. It is a process to predict a future state based on the present and historic conditions.

3. Data Acquisition

The current research is conducted to recognize the real time emotions in human speech based on age, gender and varying activities. A questionnaire is developed on different activities on each basic emotion. The target groups are trained to speak the utterances accordingly. A total of 149 utterances are recorded in Urdu language from male and female. Furthermore, there are 28, 23, 23, 29, 23 and 23 utterances are recorded each for anger, disgust, fear, happy, sad and surprise emotions respectively. In addition, there are 15, 14, 10, 16, 10 and 10 utterances of male while 13, 09, 13, 13, 13 and 13 utterances from female are recorded corresponding to anger, disgust, fear, happy, sad and surprise respectively. The utterances are recorded of 10 to 15 seconds for each gender. Each frame of sound is comprised of 256 as frame size and there are 480 frames in each sound. A signal channel of 16-bit digitization is used.

4. Methods

The proposed approach can be best described by the block diagram given in Figure 1. The method to recognize emotions consists of three primary modules namely (i) Pre-processing, (ii) Neural network processing and (iii) Post-processing module. De-noising Filter and Feature Extraction modules are sub-modules of pre-processing.

Preprocessing

During this phase two approaches are adopted to extract the proper information namely i) Filtering and ii) features extraction. To remove the silence and noise from speech signals if exists, FFT and adaptive filtering from noise cancellation are used. These techniques are used because of the advantage to reduce noise without affecting the original features of speech signal [17]. In second phase of preprocessing, features are extracted from the speech emotion sounds. In this phase acoustic features such as pitch, intensity are calculated. Moreover, the prosodic features such as statistical features of MFCC and frequency contours are calculated. When people communicate, emotions greatly influence on human daily activities. These emotions in human varies from person to person [18] based on gender, age and nature of activities. Emotion is basically a complex structure of interactions among human, feelings of happiness, sadness and anxiety etc. Moreover, it is a cognitive process of appraising. The pitch of sound and intensity differ from person to person with respect to age and gender, entropy also affects the emotional states. A total of 149 utterances are recorded of male and female based on age and gender. Moreover, feature selection is helpful to remove the irrelevant information and data and extract the most appropriate and relevant data [19]. Likewise, to make the efficiency of training of the neural network, pre-processing is required at the inputs and targets. In this case inputs and targets are scaled before the networking training. The purpose of this scaling is to specify the range for inputs and targets. The

premnmx command is used to scale these inputs and targets.

$$[pn, minp, maxp, tn, mint, maxt] = premnmx(p, t);$$

$$net = train(net, pn, tn);$$

The matrices p and t are used to hold the original inputs and targets of the network. While the normalized inputs and targets are set in pn and tn. Moreover, the minimum and maximum values of original inputs and targets are contained in minp and maxp, while the original targets are contained in mint and maxt. Likewise, Principal Component Analysis (PCA) is used. It is helpful, when the input vector has larger dimensions. PCA is used to reduce the dimensions. Three effects are provided by this technique. Initially, it has uncorrelated the components of the input vector by orthogonalizing them. Secondly, the components vectors are ordered with respect to the larger variations. Finally, the data sets with least variations are eliminated.

4.1 Filtering

There are 149 utterances of speech signals recorded and normalized. A database is maintained after filtering from unwanted noise in order to extract exact features from speech signals. Adaptive filtering from noise cancellation is used. There are many ways and approaches to reduce noise from the speech signals, but adaptive filter gives more advantages over the other methods. It increases the performance and quality of speech signals. For instance, the parameters; computational complexity, rate of convergence, ability to track sudden change of parameters, residual error level are helpful metrics which determine that which method produce better result. This filter enhances the quality of speech signal and is helpful to correctly recognize the emotion in speech signals. Many factors affect the quality of speech perception such as a variety of sources, loudness, noisy environment etc. So, in order to achieve the quality sound as produced by the original person we must have such methods which remove unwanted noise. Adaptive filter function as follows:

The input signal = desired signal d(n) + interfering noise v(v)

Mathematically

$$x(n) = d(n) + v(n) \quad (1)$$

Variable filter comprises of finite impulse response equal to the filter coefficient.

$$w_n = [\omega_{n(0)}, \omega_{n(1)}, \omega_{n(2)}, \dots\dots\dots\omega_{n(n)}]^T \quad (2)$$

Error signal is computed as

$$e(n) = d(n) - \hat{d}(n) \quad (3)$$

Variable filter is calculated by the desired signal convolving it with the input signal with the impulse response.

$$\hat{d}(n) = w_n * x(n) \quad (4)$$

Where

$x(n) = [x(n), x(n-1), \dots\dots\dots, x(n-p)]^T$ is the input signal vector.

Updating rule for variable filter

$$w_{n+1} = w_n + \Delta w_n \quad (5)$$

Where Δw_n is the correlation factor for filter coefficient.

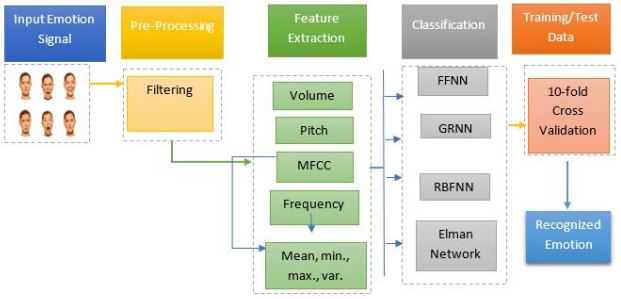

Fig. 1 Proposed System Design for emotion recognition based on Neural Networks

The Fig. 1 shows the schematic diagram to illustrate the complete picture for emotion recognition using neural networks. In the first step, the emotion signals are read and further preprocessed for filtering and features extraction. The set of extracted features are passed as input to set of neural network classifiers. In the next phase, the data is divided in to training and test using 10-fold cross validation and relevant emotion is recognized.

4.1.1 Features Extraction

In this paper acoustic features such as pitch, volume and prosodic features such as Frequency sum, maximum, minimum and MFCC mean, maximum, minimum and variance. Cepstral analysis is used to estimate the pitch. Feature extraction [20] is necessary in recognizing particular emotion future in human. Some MATLAB algorithms are programmed for features extraction. For different type of problems, researchers extract relevant features for classification purposes. Rathore et al. (2015) used [22] geometric features for automatic colon cancer detection, [23] employed ensemble methods based on hybrid features for colon cancer classification, [24] used Support vector machine kernels for load forecasting, [25-28] employed complexity based methods to quantification and analysis of physiological signals, [29] extracted complexity based features to classify the normal and pathological subjects for heart rate variability analysis and [30] extracted image texture features to detect and classify the human faces and non-faces. [33] employed Machine learning classifier for seed classification. Time- Frequency Wavelet based Coherence was computed by [32] to distinguish the EEG resting state from Eye closed (EC) to Eye open (EO). Recently, Hussain et al. [34] computed complexity base features to quantify the dynamics of EEG motor movements with EC and EO conditions.

4.1.1.1 Volume

The loudness of an audio signal is one of the most important features to obtain the acoustic perception in the speech signals. The term volume is interchangeably used as intensity, energy and volume. For simplicity, we have used it as volume. It represents the sampling amplitude within a frame. Mathematically, volume is calculated as:

$$\text{Volume} = \sum_{i=1}^{n} |s_i| \qquad (6)$$

Where [Si] is the ith sample within the frame and n is the size of the frame.

4.1.1.2 Pitch

Pitch is an important feature of audio signals, particularly for quasi-periodic signals. The voiced sounds are taken from human speech and monophonic music from most music instruments. The Pitch represents the vibration frequency of sound source of audio signals. Likewise, pitch is the fundamental frequency of audio signals, which is equal to the reciprocal of the fundamental period. Typical Pitch ranges according to the following ranges.
Male: ranging from 85 – 155 Hz
Female: Ranging from 165- 255 Hz
Singer: Ranging from 80-1100 Hz

4.1.1.3 MFCC

It is commonly used acoustic feature to recognize the emotions in human speech. It is helpful in recognizing the human spoken languages such as speech recognition. During the noisy signals, to improve the robustness log-mel amplitude is raised to suitable powers. It is done before applying the DCT. MFC is the short representation of power spectrum based on linear cosine transform of power spectrum of frequency based on mel scale. MFCC co-efficient collectively make up an MFC. They are derived from the audio clip from its cepstral representation. The difference between cepstrum and mel frequency cepstrum is that MFC frequency bands are distributed equally using mel scale, while in comparing it with normal cepstrum they provide more closely approximate response of human auditory system. So MFC is more powerful for the representation and recognition of sound.

4.2 Neural Network Processing

Neural networks are image processing and data mining tools. The basic idea is to implement human thoughts as an algorithm on computer for efficient results. As human brain consists of neurons, which send the activation signals to each other resulting to create intelligent thought. Similarly, Artificial Neural Network (ANN) also consists of number of neurons, which send activation signals to each other. ANN approximates multiple inputs and outputs. Thus, it is used in a variety of applications like data mining, classification problems, clustering, function approximation, prediction, image processing, speech and emotion recognitions. This thesis covers the real-time emotion recognition in human speech using ANNs. Emotion recognition is a supervised learning problem. Various classifiers may be used for its recognition including Bayesian Learning, Support Vector Machine, Linear Discriminant Analysis, Multilayer Neural Network, FFNN and Hidden Markov Model (HMM) to capture temporal transitions [21].
There are many types of neural networks to solve a variety of problems. The common type is Feed Forward Neural network (FFNN) while other types include Radial Base Function Neural network (RBFNN), Self-Organizing Maps (SOM) and Recurrent Networks (RN). The basic emotions have already been identified using FFNN, GRNN [20]. In this paper, RBFNN is used and the results are compared with GRNN, FFNN and Elman neural networks.

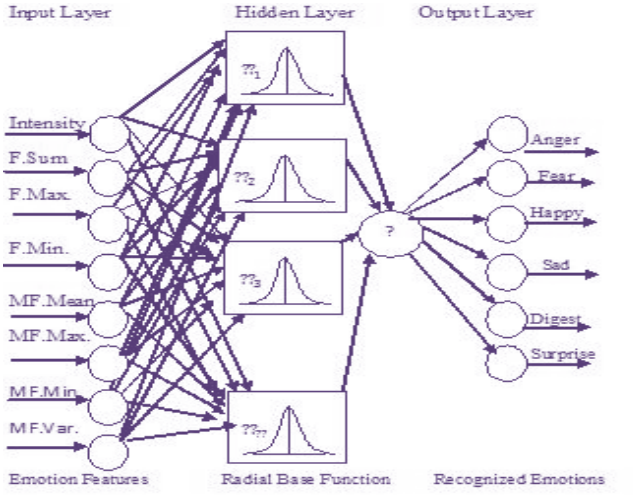

Fig. 2 RBFNN Architecture for emotion recognition, F (Frequency), MF (MFCC)

4.3 Post- Processing

Before the experiments, pre-processing is applied on data. In order to see the output after simulation, it is required to obtain the original scale. Thus, a subroutine poststd and postmnmx is required for corresponding prestd and premnmx. In this research, a set of training data is normalized with premnmx and the network trained using the normalized data. The network is then simulated, unnormalized the output of the network using postmnmx. Finally, the regression analysis is performed between the network outputs (unnormalized) and the targets to check the quality of the network training. The routine poststd post processes the network training set that is preprocessed by routine prestd. The purpose of this routine is to converts the data back into unnormalized units.

After network simulation that is trained in preprocessing, its output is converted back to the original units using following commands:

$an = sim(net, pn);$

$a = postmnmx(an, mint, maxt);$

Here an correspond to tn. The un-normalized network output a is in the same as original targets t.

Besides, if premnmx is used to preprocess data, then when the trained network is used with new inputs they should be preprocessed with the minimum and maximums that are computed for the training set. That can be accomplished with the routine tramnmx. Following commands are used;

$$pnewn = tramnmx(pnew, minp, maxp)$$
$$anewn = sim(anewn, mint, maxt)$$
$$anew = postmnmx(anewn, mint, maxt)$$

5. Results Analysis and Performance Evaluation

The network is trained with 149 emotions comprising of each basic emotion as anger, disgust, fear, happy, sad and surprise of both male and female. Out of 149 utterances, there are 74 utterances are from male and 75 from female. The performance of the network is judged using Mean Square Error (MSE), Regression Analysis (RA), multiple linear regression, Post training Analysis, and Principal Component Analysis (PCA). Initially, the PCA is applied on data set which significantly reduces the data set. The data set is divided into validation, testing and training subsets. The distribution on data set is made such as ¼, ¼ and ½ on validation set, test set and training set respectively. However, the original data is equally spaced at all points. As shown in the Figure 4 below using RBFNN that validation set errors and test set errors are similar in characteristics and significantly no over fitting has been appeared.

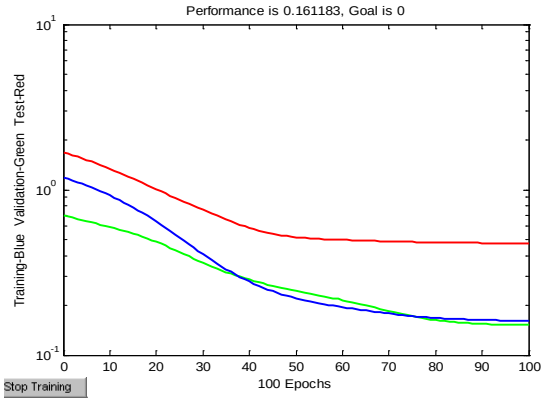

Fig. 3 Performance on Validation, test and Training set using Elman Network

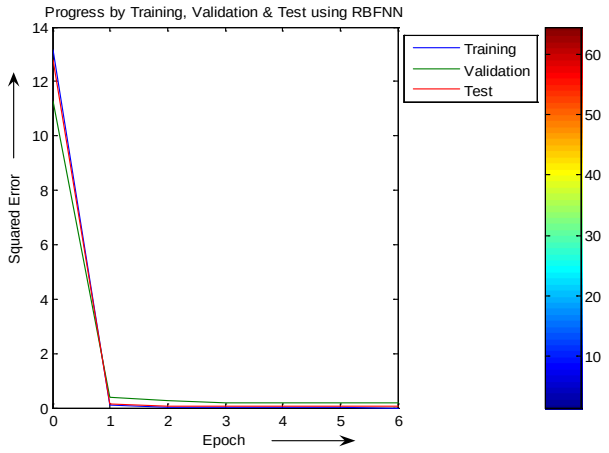

Fig. 4 Performance on Validation, test and Training set using RBFNN

Moreover, the Figure 4 is drawn using FFNN among validation, test and training set.

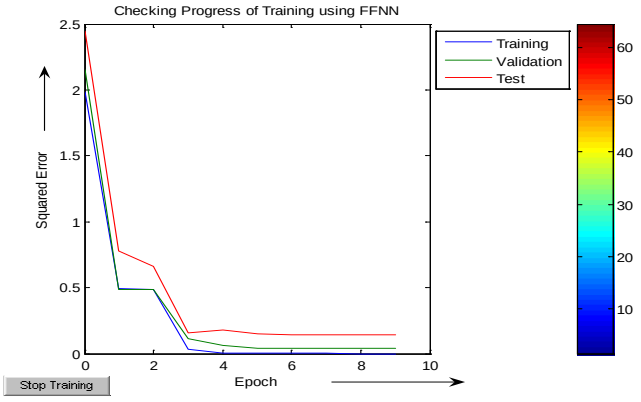

Fig. 5 Performance on Validation, test and Training set using FFNN

The Multiple Linear Regression is applied on network training and testing. The results obtained so far are *R*2 statistic, F Statistics P value and Error Variance. *R*2 statistic indicates that there are very low variations in the observations. F statistic (for the hypothesis test that all the regression coefficients are zero), and the p-value associated with this F statistic. The results of training (73 % as in X) and testing (27 % as in y) and the corresponding difference (d) shows that there are very least variations in training and testing data of each class of basic emotion. Besides, covariance is computed against each of the emotion and features gives the least result of variations.

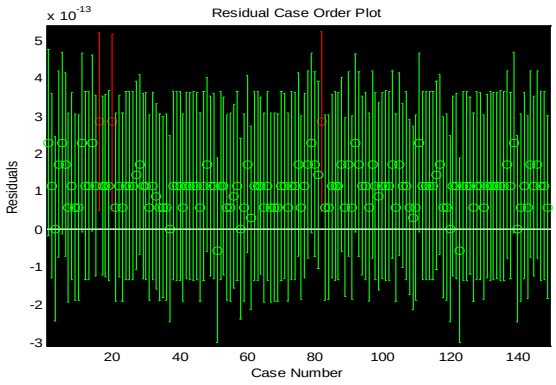

Fig. 6 Residual plot against case number

Multiple linear regressions is used to give response between predictor observations and observed responses. In this case, the plot in figure 6 above shows the residuals plotted in case order (149 emotions by row). The 95% confidence intervals about these residuals are plotted as error bars. The red lines indicate that these points are outlier not crossing the zero line i.e. these points are not lying in this range of desired data points.

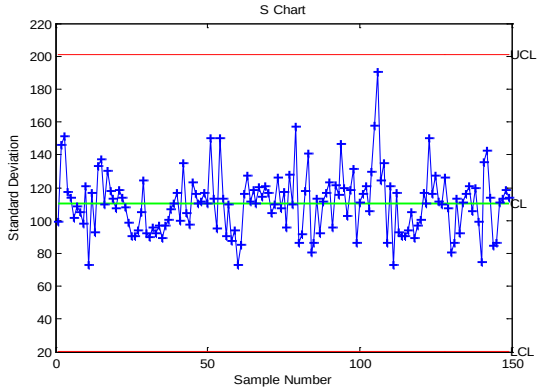

Fig. 7 Standard Deviation Chart of Group responses

The graph in Figure 7 contains the sample standard deviation s for each group, a center line at the average s value, and upper and lower control limits. From the figure 50, it is clear that all points are within the control limits, so the variability within subgroups is consistent with what would be expected by random chance

Least Mean Square (LMS) algorithm is used to calculate the Mean Square Error (MSE). It has minimized the sum of square of errors. The input is applied to the network and network output is compared to the target. The error is computed between the target output and network output. Moreover, the average of sum of the error is minimized.

$$\boldsymbol{mse} = \frac{\boldsymbol{1}}{\boldsymbol{Q}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{Q}} \boldsymbol{e(k)^2} = \frac{\boldsymbol{1}}{\boldsymbol{Q}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{Q}} (\boldsymbol{t(k) - a(k)})^2 \qquad (7)$$

The LMS algorithm adjusts the weights and biases of the linear network.

Table 1: Emotions vs Features

Motion	G	Max. Vol.	Frequency			MFCC			
			Sum	Max.	Min.	Mean	Max.	Min.	Var.
Anger	F	67.40	-3.80	0.66	-0.83	-19.10	2.15	-37.97	249.34
	M	76.51	-6.95	0.85	-0.93	-17.74	2.75	-50.72	280.40
Disgust	F	46.82	-6.43	0.38	-0.57	-16.69	1.83	-36.35	275.40
	M	53.55	-14.55	0.47	-0.65	-17.02	2.55	-37.97	416.10
Fear	F	101.97	-10.67	0.88	-1.00	-18.98	2.28	-50.72	294.31
	M	16.17	-3.89	0.26	-0.21	-18.11	3.14	-36.90	294.49
Happy	F	57.15	-6.37	0.72	-0.82	-16.36	2.29	-36.35	250.10
	M	37.03	-5.25	0.46	-0.40	-17.05	2.79	-37.93	290.45
Sad	F	40.25	-5.78	0.51	-0.59	-16.45	2.04	-38.16	264.68
	M	21.06	-4.75	0.22	-0.28	-17.81	3.93	-34.53	314.56
Surprise	F	70.91	-10.43	0.79	-0.73	-16.81	2.52	-37.97	303.02
	M	67.77	-7.54	0.88	-0.72	-18.43	3.54	-35.77	276.99

Table 2: Network performance using MSE and Regression Analysis

Network Model	MSE	Regression Analysis		
		M	B	R
FFNN	1.9975e+003	0.9440	-9.3734	0.9009
GRNN	0.0056	146.6049	386.5623	1.0000
RBFNN	4.1328e-025	146.6100	382.2900	1
Elman Network	1.9946e+004	-0.2724	518.4446	-0.5814

From the table above MSE is calculated against FFNN, GRNN, RBFNN and Elman Network. The values against each network in MSE column represents that RBFNN has least MSE in comparison with other networks. Very low error represents the better recognition rate.

Moreover, the network performance is also measured using Regression Analysis (RA). This is done on network response and corresponding targets. The routine postreg is used to perform this analysis. Mathematically,

$$[\boldsymbol{m, b, r}] = \boldsymbol{postreg}(\boldsymbol{t', s'}); \qquad (8)$$

Network output and corresponding targets are passed to postreg. Three parameters are returned by postreg routine. Here m and b corresponds to slope and y intercept, whereas r corresponds to the correlation coefficient. If the value of r equals to 1, it represents best fit, i.e. output is equal to the targets. After training the network, real series data comprising of features and emotions is processed. There is a 127 by 8 matrix by splitting this matrix into new data sets, first 109 as training data points and 18 as testing data points, two from each emotion.

Table3: Confusion Matrix of Emotions in Human Speech

		Predicted					
		Anger	Disgust	Fear	Happy	Sad	Surprise
Actual	Anger	17	0	0	3	0	0
	Disgust	0	16	0	0	1	0
	Fear	0	0	17	0	0	0
	Happy	0	0	0	21	0	0
	Sad	0	0	0	0	17	0
	Surprise	0	0	0	0	1	16

The data is normalized by pre and post processing and applied regression analysis on FFNN, GRNN, RBFNN and Elman Network. As shown in the figure 8 and table

value of r is 1 in GRNN & RBFNN correctly fit the data points of testing and training, however, MSE of RBFNN is small than GRNN. This concludes that RBFNN has more perfectly recognized particulars emotion than FFNN, GRNN and Elman Network.

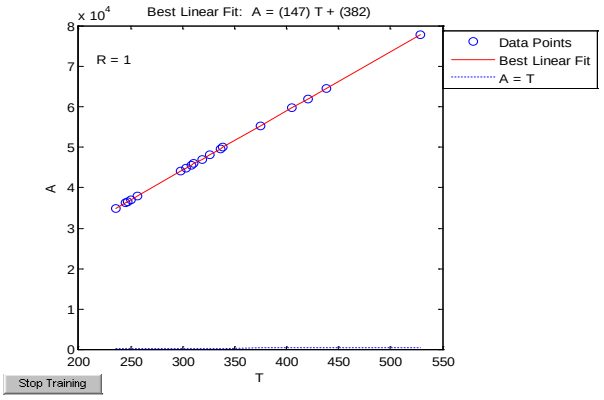

Fig. 8 Regression Analysis using RBFNN

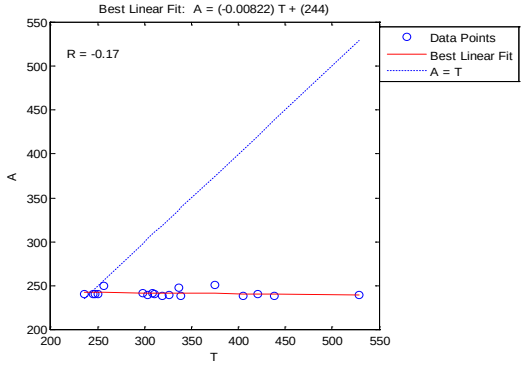

Fig. 9 Regression Analysis using Elman network

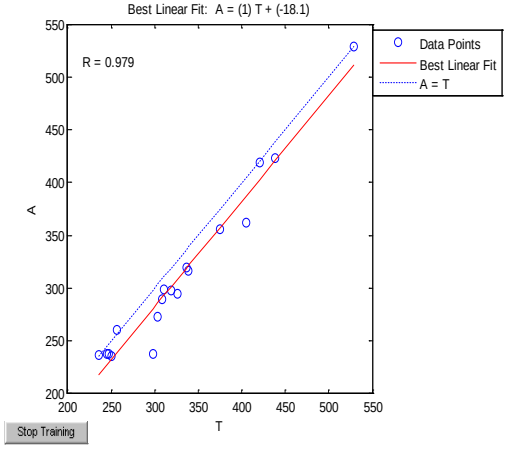

Fig. 10 Regression Analysis using FFNN

Real time emotions are recorded from male and female based on age, varying activities. The network is trained using FFNN, GRNN, Elman network and RBFNN. The acoustic and MFCC prosodic features are calculated against each basic emotion. The table 1 about emotion vs features shows that intensity of sound during anger in male is more than that of female. However, this value of intensity is very high in female during fear state. It is also observed that female's intensity is more than that of female during happy, surprise and sad states. However, the intensity of male is more than female in anger and disgust state. The other statistical and prosodic features of MFCC and frequency vary accordingly as that the intensity changes.

$$Accuracy = \frac{True\ Positive + True\ Negative}{All} \qquad (9)$$

$$= \frac{104}{109} = 95.41\ \%$$

The network using RBFNN is trained with 109 emotions containing 20, 17, 21, 17 and 17 emotions from Anger, Disgust, Fear, Happy, Sad and Surprise respectively. The confusion matrix shown in the figure depicts an accuracy of 95. 41 % between Actual and Predicted emotions.

The statistical and prosodic features are calculated from each sound after reading from the database. These features are passed to the neural network as input and networks are trained. Moreover, the scaling on inputs and targets is made using PCA and RA.

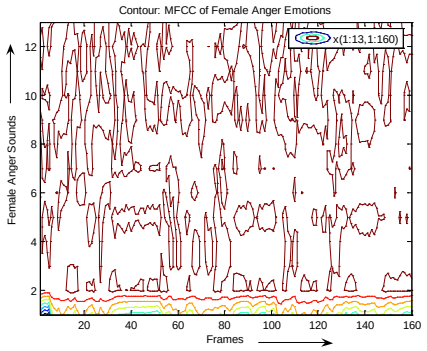

Fig. 11 MFCC Contour of Female anger utterance

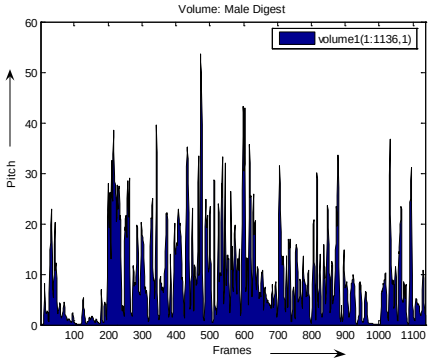

Fig. 12 Volume of male Disgust utterance

The Figures 11 to 14 are the graphical representations of male and female emotions. Here Figure 11 and 13 are MFCC contours during anger and disgust states of female and male respectively. From Table 1 and Figure 11 and 13, it is seen that variance during disgust state in male is near to twice than that of female during anger. Moreover, it is also observed that Figure 13 is much denser than Figure 11.

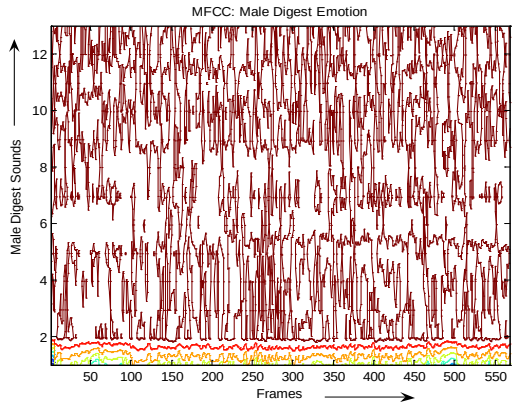

Fig. 13 MFCC contour of Male Disgust utterance

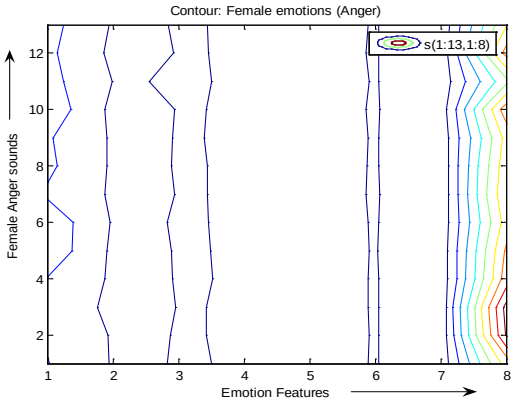

Fig. 14 Frequency contour of Female Anger utterance

In addition, MFCC features are calculated, it is seen that the MFCC variance of male are more than that of female. It changes almost in a constant ratio during each emotion. However, it is seen that during the fear state, it is almost the same value.

The frequency sum, maximum and minimum varies accordingly as that the intensity (Volume) as shown in Figure 12. In addition, MFCC features are calculated, it is seen that the MFCC variance of male are more than that of female. It changes almost in a constant ratio during each emotion. However, it is seen that during the fear state, it is almost the same value.

The features are calculated against each emotion of male and female. From the results it is observed that during anger intensity of sound of male is more than that of female. However, during the fear state female has very higher degree of intensity than that of male. It is also observed from the results that female have more intensity than male when they are in the state of fear, happiness, sadness or surprise conditions, while the male has greater intensity during anger, and disgust states.

6. Conclusions and future work

In the present work, human emotions such as anger, disgust, happiness, sadness and surprise are recognized based on age, gender and varying activities. Pre-processing is applied before training in order to extract the appropriate information. The inputs and targets are scaled before passing to the network using the PCA. The noise and silence are filtered from the recorded speech signals using adaptive filtering. In addition, the acoustic features like volume, intensity and statistical features are computed using MFCC. These features help to recognize any particular emotion. The data set is processed by the chosen ANN types. The scaled outputs of the ANN are de-normalized by post-processing. The experimental results obtained by the RBFNN demonstrate the minimum MSE, up to 98% recognition rate in human varying activities and a single correlation coefficient, which ensure that RBFNN best fit the data. In future, we aim including other spoken languages of Azad Jammu and Kashmir, Pakistan to be test for varying emotions.

References

[1] Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing, 3(1), 42-55.

[2] Kim, D. H., & Baranyi, P. (2011, January). Novel emotion dynamic express for robot. In Applied Machine Intelligence and Informatics (SAMI), 2011 IEEE 9th International Symposium on (pp. 243-245). IEEE.

[3] Kleinsmith, A., Bianchi-Berthouze, N., & Steed, A. (2011). Automatic recognition of non-acted affective postures. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41(4), 1027-1038.

[4] Tawari, A., & Trivedi, M. M. (2010). Speech emotion analysis: Exploring the role of context. IEEE Transactions on multimedia, 12(6), 502-509.

[5] Fu, L., Wang, C., & Zhang, Y. (2010, July). A study on influence of gender on speech emotion classification. In Signal Processing Systems (ICSPS), 2010 2nd International Conference on (Vol. 1, pp. V1-534). IEEE.

[6] Wollmer, M., Schuller, B., Eyben, F., & Rigoll, G. (2010). Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. IEEE Journal of Selected Topics in Signal Processing, 4(5), 867-881.

[7] Lin, Y. P., Wang, C. H., Jung, T. P., Wu, T. L., Jeng, S. K., Duann, J. R., & Chen, J. H. (2010). EEG-based emotion

recognition in music listening. IEEE Transactions on Biomedical Engineering, 57(7), 1798-1806.
[8] Nishimura, J., & Kuroda, T. (2010). Versatile recognition using Haar-like feature and cascaded classifier. IEEE Sensors Journal, 10(5), 942-951.
[9] Shah, F. (2009, December). Automatic emotion recognition from speech using artificial neural networks with gender-dependent databases. In Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT'09. International Conference on (pp. 162-164). IEEE.
[10] Siraj, F., Yusoff, N., & Kee, L. C. (2006, June). Emotion classification using neural network. In Computing & Informatics, 2006. ICOCI'06. International Conference on (pp. 1-7). IEEE.
[11] Rao, G. M., Babu, G. R., Kumari, G. V., & Chaitanya, N. K. (2009, March). Methodological Approach for Machine based Expression and Gender Classification. In Advance Computing Conference, 2009. IACC 2009. IEEE International (pp. 1369-1374). IEEE.
[12] Ser, W., Cen, L., & Yu, Z. L. (2008, December). A hybrid PNN-GMM classification scheme for speech emotion recognition. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on (pp. 1-4). IEEE.
[13] Cen, L., Ser, W., & Yu, Z. L. (2008, December). Speech emotion recognition using canonical correlation analysis and probabilistic neural network. In Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on (pp. 859-862). IEEE.
[14] Das, K., Ding, Q., & Perrizo, W. (2001). Artificial neural network applications on remotely sensed imagery. In Info-tech and Info-net, 2001. Proceedings. ICII 2001-Beijing. 2001 International Conferences on (Vol. 3, pp. 510-515). IEEE.
[15] Schmidt, T., Rahnama, H., & Sadeghian, A. (2008, September). A review of applications of artificial neural networks in cryptosystems. In Automation Congress, 2008. WAC 2008. World (pp. 1-6). IEEE.
[16] Ghavami, P., & Kapur, K. (2011, June). Prognostics & artificial neural network applications in patient healthcare. In Prognostics and Health Management (PHM), 2011 IEEE Conference on (pp. 1-7). IEEE
[17] Wang, Y., & Guan, L. (2004, October). An investigation of speech-based human emotion recognition. In Multimedia Signal Processing, 2004 IEEE 6th Workshop on (pp. 15-18). IEEE.
[18] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. IEEE Signal processing magazine, 18(1), 32-80.
[19] Rong, J., Li, G., & Chen, Y. P. P. (2009). Acoustic feature selection for automatic emotion recognition from speech. Information processing & management, 45(3), 315-328.
[20] Khanchandani, K. B., & Hussain, M. A. (2009). Emotion recognition using multilayer perceptron and generalized feed forward neural network.
[21] Yang, B., & Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. signal processing, 90(5), 1415-1423.
[22] Rathore, S., Hussain, M., & Khan, A. (2015). Automated colon cancer detection using hybrid of novel geometric features and some traditional features. Computers in biology and medicine, 65, 279-296.
[23] Rathore, S., Hussain, M., Iftikhar, M. A., & Jalil, A. (2014). Ensemble classification of colon biopsy images based on information rich hybrid features. Computers in Biology and Medicine, 47, 76-92.
[24] Hussain, L., Nadeem, M. S., & Shah, S. A. A. (2014). SHORT TERM LOAD FORECASTING SYSTEM BASED ON SUPPORT VECTOR KERNEL METHODS. International Journal of Computer Science & Information Technology, 6(3), 93.
[25] Hussain, L., Aziz, W., Alowibdi, J. S., Habib, N., Rafique, M., Saeed, S., & Kazmi, S. Z. H. (2017). Symbolic time series analysis of electroencephalographic (EEG) epileptic seizure and brain dynamics with eye-open and eye-closed subjects during resting states. Journal of physiological anthropology, 36(1), 21.
[26] Hussain, L., Aziz, W., Saeed, S., Shah, S. A., Nadeem, M. S. A., Awan, I. A., ... & Kazmi, S. Z. H. (2017). Quantifying the dynamics of electroencephalographic (EEG) signals to distinguish alcoholic and non-alcoholic subjects using an MSE based Kd tree algorithm. Biomedical Engineering/Biomedizinische Technik. DOI: https://doi.org/10.1515/bmt-2017-0041
[27] Qumar, A., Aziz, W., Saeed, S., Ahmed, I., & Hussain, L. (2013, December). Comparative study of multiscale entropy analysis and symbolic time series analysis when applied to human gait dynamics. In Open Source Systems and Technologies (ICOSST), 2013 International Conference on (pp. 126-132). IEEE.
[28] Hussain, L., Aziz, W., Nadeem, S. A., Shah, S. A., & Majid, A. Electroencephalography (EEG) Analysis of Alcoholic and Control Subjects Using Multiscale Permutation Entropy. Journal of Multidisciplinary Engineering Science and Technology, 1(5), 380-387.
[29] Hussain, L., Aziz, W., Nadeem, S. A., & Abbasi, A. Q. (2014). Classification of Normal and Pathological Heart Signal Variability Using Machine Learning Techniques. International Journal of Darshan Institute on Engineering Research and Emerging Technologies, 3(2), 13-18.
[30] Hussain, L., Aziz, W., Kazmi, Z.H. and Awan, I.A., 2014. Classification of Human Faces and Non-Faces Using Machine Learning Techniques. International Journal of Electronics and Electrical Engineering 2 (2), 116-123
[31] Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia, 16(8), 2203-2213.
[32] Hussain, L., & Aziz, W. (2016). Time-Frequency Spatial Wavelet Phase Coherence Analysis of EEG in EC and EO During Resting State. Procedia Computer Science, 95, 297-302.
[33] Ajaz, R. H., & Hussain, L. (2015). Seed classification using machine learning techniques. International Journal of Electronics and Electrical Engineering 2 (5), 1098-1102.
[34] Hussain, L., Aziz, W., Saeed, S., Shah, S.A., Nadem, M. S. A., Awan, I. A., Abbas, A., Majid, A., & Kazmi, S. Z. H. (2017) Complexity analysis of EEG motor movement with eye open and close subjects using multiscale permutation entropy (MPE) technique. Biomedical Research-India; 28 (16), 7104-7111.

Lal Hussain is a Programmer at Quality Enhancement Cell, University of Azad Jammu and Kashmir, Muzaffarabad Pakistan. He obtained his MS in Communication and Networks from Iqra University, Islamabad, Pakistan in 2012 with Gold medal. He received Ph.D. from Department of Computer Science & Information Technology, University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan in February 2016. He worked as visiting PhD researcher at Lancaster University UK for six months under HEC International Research Initiative Program. He is author of more than 10 publications of highly reputed peer reviewed and Impact Fact Journals. He presented various talks at Pakistan, UK and USA. His research interest includes Biomedical Signal Processing with concentration on complexity measures, Time-Frequency representation methods, Cross Frequency Coupling to analyze the dynamics of neurophysiological and physiological signals. Other research interest includes Neural Networks and Machine Learning classification and prediction problems etc.