Emotional States Recognition by Speech Features, Its Need and Impact of Various ANN Model on Recognition Rate

Manish Kumar Pandey[1] Dr. Mohan Awasthy[2]
[1]Ph D Scholar [2]Principal
[1]CVRU, Bilaspur, India [2]RCET, Bhilai, India

Abstract— This paper focuses on enhancing requirements of the current voice command and speaking technologies with consideration of emotional states, so that the interaction with computers and smart devices may become more human. It indicates the study of recognition of human emotional states and shows the basic idea that how signal processing can apply for the same. Experiment is divided into two major and important parts; one is the analysis of speech samples where second is associated with modeling of neural network and its training method. Further we extend our experiment to study the analysis procedure. Here we actually show how the speech features can extracted from speech samples and how feed forward back propagation algorithm used to carry the training procedure. An emotion recognition rate of approximately 70% was obtained for five basic emotional states. The final phase of experiment illustrates that repetitive training of ANN significantly affect the recognition rate as concluding decision and suggests the boundary for optimum ANN training.

Keywords: Emotion state recognition, Artificial Neural Network, Back Propagation, Recognition Rate, Optimum ANN training

I. INTRODUCTION

Emotional state recognition is now become an important research area with the aim of civilizing the feature of human-computer interaction. Research in this domain emphasizes solving the technical difficulties concerned, from side to side design of ever more complex recognition algorithms. But elementary questions related to this field are remaining. How it is best implemented, to improve human-computer interaction? It is the most significant question, which is not being addressed because of the current state of the technology: the most common systems are simply not so good enough to solve such realistic applications.

Nonverbal communication plays a very important role in today's communication. With 2.71 billion smart phone users across the world the spread of smart phones are still increasing rapidly and use of variety of apps have become an inseparable part of our daily life, especially among the young age group. Talking to cell phones through voice commands in our very own human's natural language has also accelerated this movement which common now a days. However, current voice command technologies are mostly based on natural language processing and fantasizing about smart devices which can think, understand and talk with feelings like we do is still far-fetched. So, it is clear that the exchange of nonverbal information such as mood and emotions involved in speech will become vital in all forms of communication in near future.

The interaction between human and computer / smart devices has becomes more and more common. As such, the potential of communicating with human beings using both verbal and nonverbal communication channels will be vital. This will definitely make interactions between computers and humans more close and natural.

Although the importance of nonverbal aspects of communication has been recognized, until now most research has involved nonverbal information for images. In contrast the recognition of emotional states concerned in speech has been less treated. For all these reasons, we have studied the recognition of emotional states concerned in speech.

Making of a system to perform the emotion recognition task is a bit similar to speaker recognition problem but it has a quite different approach in the field of speech processing. The difference begins on both approaches from the database collection; it depends upon our aim, which had to be determined early. If we go for the speaker recognition task, it needed emotionless or neutral speech samples, but if we specifically choose to carry emotion recognition task, it is necessary to have a database of speech samples which contains emotions. This way emotion recognition is speaker independent mission, so we follow this strategy during the collection of database.

II. SETTING UP EXPERIMENT DESIGN AND PARAMETERS

A bit related experiment had been done before [3]. In there experiment they had created a database of 100 people (all males), with eight emotional states. They took a set of 100 words and record that with the professional radio artist. The other references had also experimented with large database. But, we follow quite opposite approach. For experiment purpose we carried with only 20 speaker samples with five basic emotional states. Speaker samples involved with this experiment are both male and female of various age group. We used appropriate sentences for collecting the samples for all emotional states. No professional voice artists took part in our experiment.

Human emotions are subject to feel, but here we have to count and categorize them. Exposure of emotion are wide, they may pointed out in combinations like disgusting sadness. We carried five basic emotions, which are accepted universally in human beings, as_

1) Disgust
2) Sad
3) Happy
4) Anger
5) Fear

Collected a total of 100 speech samples. (20 speaker samples * 5 emotion = 100 samples)

Hindi language was preferred for recording and collecting speech samples in our experiment. Samples were further carried for experiment were of equal length.

The variations in recorded voice of different speakers, according to mood / feelings are the key to identify the associated emotional state. These samples were analyzed

All rights reserved by www.ijsrd.com

and synthesized through AutoSignal™. The most of the features, related to analysis of the signals. The Fast Fourier Transform technique were applied for the synthesis of speech samples. The analyzed data outputs under following features_

1) Frequency
2) Magnitude
3) Real
4) Imaginary
5) Amplitude
6) Wavelength
7) Phase
8) dB Norm
9) PSD MSA (Power Spectral Density, Mean Squared Amplitude)

The training methodology implemented using the feed forward backpropagation neural network. Gradient decent rule used as training algorithm and output transformed in to linear function.

So, finally with the enough & sufficient difference in comparison to previous work in all aspects, extending from speech sample database to the unlike analysis & modeling process we obtained better level of recognition rate which was up to 70% as compared with 30% from previous work. Also, that the entire experiment suggests some interesting analytical and conceptual findings.

III. EXTRACTION OF SPEECH FEATURES AND SPEECH SYSTHESIS

In this stage we analyzed the collected samples, which was the first important phase of our experiment. This stage was responsible to address following five basic questions, necessary for analysis of any such database, for this kind of experiment.

1) Which data should be analyzed?
2) How data can be analyzed?
3) Which and how many features are significant and useful for this experiment?
4) Which technique should be suitable for synthesis of these samples?
5) Only one technique is enough or sufficient for synthesis, or it is, better to have another or combination of more then, one technique.

Auto Signal was used to analyze the database features of speech samples. Initially it imports signal in time domains. 'Fast Fourier Transform' method found suitable and used as well for speech synthesis, because it results in parameters which are relative and significant in terms of emotional state recognition.

The Fast Fourier Transform (FFT) decomposes a time-domain signal (which can be a function of time, spatial coordinates, or any time series abscissa) into complex exponentials (sines and cosines). A Fourier transform offers a complete picture of frequency space, but retains no information as to when in time a signal occurs. Thus a signal should either be wide-sense stationary, with a constant mean and variance across broad time segments, or we must care only qualitatively about whether a certain frequency content signal occurs somewhere in the time range being sampled. The FFT is a fast algorithmic route for producing the Discrete Fourier Transform (DFT). The forward and reverse discrete transforms are defined as follows:

$$X_n = \sum_{k=0}^{N-1} x_k e^{\frac{i2\pi kn}{N}}$$

$$x_k = \frac{1}{N} \sum_{n=0}^{N-1} X_n e^{\frac{-i2\pi kn}{N}}$$

While the DFT is very simple, it is an order n^2 procedure whereas the FFT is an n*log2(n) operation. The difference in processing times is vivid with large data sets. AutoSignal offers four different fast FFT algorithms. The Fourier decomposition represents a data sequence as a linear combination of a set of sine and cosine basis functions. Although the data and Fourier sequences are each discrete, the basis functions are continuous and infinite in duration. It is thus possible to reconstruct the signal for any time within the range of the original sequence:

$$y(t) = \sum_{k=1}^{Nspec} A_k \sin(2\pi v_k t + \theta_k)$$

The Fourier basis functions can be treated as complex exponentials, zero phase sine and cosine pairs, phase bearing cosines, or phase bearing sines. In the above equation, 'A' is the amplitude reported, 'nu' is the frequency, and 'theta' is the phase. The signal at any time't' can be reconstructed by summing the value of all 'Nspec' sinusoids in the spectrum evaluated at that time. The amplitudes are adjusted so that all power is represented in the positive frequencies. This option enables the following window_

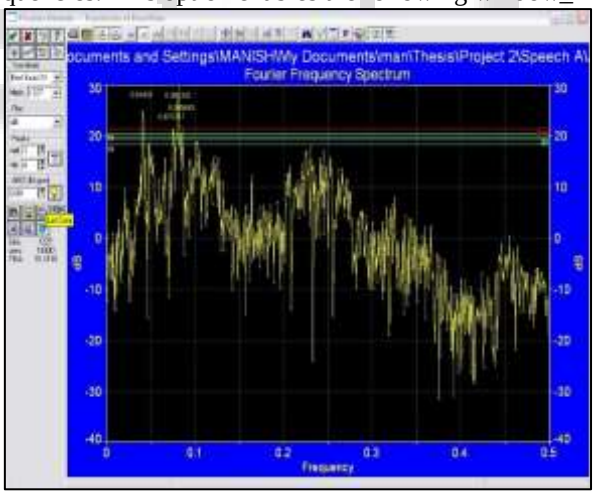

Fig. 1:

It allows creating a list of data which contains hundreds of rows of nine spectral features in term of numeric values. We have to follow the a procedure to get such data list

1) Transform: The Best Exact N composite algorithm is the default and will use for our analysis.
2) Nmin: The initial Nmin value will be the data size. This value can very, for each signal that can be automatically select. To zero pad, enter any value greater than the data size. We may also select from a power of 2
3) Plot: The frequency domain information can be plotted in a variety of formats. We draw Amplitude Plot. It drawn between the amplitude and frequency

All rights reserved by *www.ijsrd.com*

4) Peaks: The spectral peaks are identified by a local maxima detection algorithm. Both the amplitude and the frequency locations of the detected peaks are based upon a cubic spline bin interpolation procedure.The sig item sets the target number of peaks (signal components) to detect. Up to 50 peaks can be detected. Peaks are ranked by interpolated amplitude. The wid item sets the bin width tolerance for defining a peak. A peak must exist across this number of FFT bins to be counted. The default is a single bin.
5) List: This offers an extended FFT data summary. The FFT channel number, frequency, and magnitude are always listed. The use the selection of the following:

– Add Frequency
– Add Magnitude
– Add Real, Imag
– Add Amplitude
– Add Wavelength
– Add Phase (Sine-based)
– Add dB Normalized
– Add Power Spectral Density, Mean Squared Amplitude

The amplitude and phase of each component in the FFT is derived from sine-based conversion. Each of the components in the FFT can be reconstructed using: Y=Amplitude*sin (2*PI*Frequency*X+Phase).

The following type of sample data obtained as results after synthesis procedure for each sample.

Chnl Frequency Magnitude Real Imag Amplitude Wavelength Phase dB Norm PSD MSA
0 0.00000000 4.38281250 -4.3828125 0.00000000 0.00367686 0.00000000 4.71238898 -21.892529 1.3519e-05
1 0.00083893 0.40858057 -0.1926980 0.36028541 0.00068554 1192.00000 3.63272855 -42.502032 2.3498e-07
2 0.00167785 0.29242501 0.10259350 -0.2738375 0.00049065 596.000000 0.35846473 -45.407296 1.2037e-07
3 0.00251678 0.81491939 -0.7211255 -0.3795676 0.00136731 397.333333 5.19689718 -36.505293 9.3477e-07
4 0.00335570 1.19838834 0.36287457 -1.1421281 0.00201072 298.000000 0.30763147 -33.155635 2.0215e-06
5 0.00419463 0.47013252 0.25331958 -0.3960477 0.00078881 238.400000 0.56904275 -41.283181 3.1111e-07
6 0.00503356 0.20293792 -0.0344754 -0.1999881 0.00034050 198.666667 6.11247574 -48.580322 5.797e-08
7 0.00587248 1.25075039 1.14592781 -0.5012245 0.00209857 170.285714 1.15847301 -32.784174 2.202e-06
8 0.00671141 0.59582962 -0.2363085 0.54696549 0.00099971 149.000000 3.54940722 -39.225145 4.9971e-07
9 0.00755034 0.76829973 -0.6258962 -0.4455765 0.00128909 132.444444 5.33105806 -37.016973 8.3088e-07
10 0.00838926 0.34760154 0.23184458 -0.2589883 0.00058322 119.200000 0.73015316 -43.905953 1.7008e-07

The above data is synthesized output from speech sample associated with 'HAPPY' emotional state. Here we are illustrating small part of output actual output contained about 600 rows.

Similar synthesis had been performed for all 100 speech samples. Averaging performed for compacting the demonstration.

IV. Simulation of ANN Model for Recognition of Emotional States

Following is the processed data, converted and transposed in to matrix format. Here columns representing Emotional States and rows representing synthesized parameters. This matrix was further used as input for the training and simulation of 'Emotional State Recognition'.

```
>> b = [p]
b =
    0.4788    0.2499    0.4401    0.2498    0.2498
    2.3454    2.6936    3.6908    2.8968    2.2186
    0.0035   -0.0004   -0.0149    0.0092    0.0026
   -0.0614   -0.0195    0.0039   -0.0014   -0.0085
    0.0057    0.0048    0.0068    0.0081    0.0074
   11.8695   13.2341   13.7731   12.9858   12.6602
    3.0798    3.1055    3.1798    2.9877    3.1373
  -27.7375  -30.2547  -27.4979  -27.8173  -29.0379
    0.0006    0.0012    0.0001    0.0001    0.0001
```

The target (Output) was set to the numeric values as_
[1 2 3 4 5]
, according to the emotion, it indicates_
'1' for Disgust, '2' for Sad, '3' for Happy, '4' for Anger, and '5' for Fear.

Entire training and simulation was performed with MATLAB's neural network toolbox. Multiple models were tested based on following parameters of neural network design –

1) Type of ANN
2) Number of Layers in ANN
3) Number of neurons in each layer
4) Appropriate Transfer and output Function
5) Adoption of suitable training / learning algorithm
6) Number of training iterations
7) Setting up learning rate &
8) Defining permitted deviation from expected output.

The above mentioned training parameters plays an important role in entire process and various combinations depends upon type and complexity of problem. The one found suitable in accordance with our problem is given below_

```
load ((‘input.mat’),’p’)
load ((‘target.mat’),’t’);
net =
newff(minmax(p),[10,6,1],{'tansig','tansig','purelin'},'traingda')
net.trainParam.show = 1000;
net.trainParam.lr = 0.0000005;
net.trainParam.epochs = 250000;
net.trainParam.goal = 1e-10;
```

All rights reserved by *www.ijsrd.com*

```
[net,tr]=train(net,p,t);
save (('Emotion2.mat'),'net');
%a = sim(net,p)
```

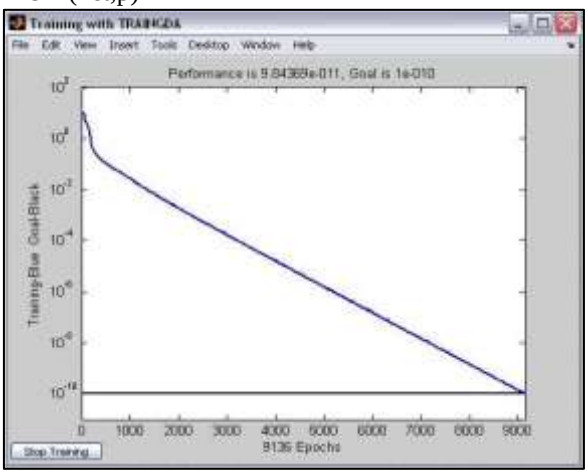

Fig. 2:

The performance graph shown above was found as the result of the training process. This network has the test input as 'p' with the same target 't', indicating the numbers in sequence in order to emotional states. Used network is feed forward back propagation network with gradient decent adaptation training function. It is a multilayer forward network using extend gradient-descent based delta learning rule, commonly known as backpropagation (of errors) rule. Backpropagation provides a computationally efficient method for updating the weight in feed forward network, with differentiable activation function units, to learn training set of input / output examples. Being a gradient-descent method it minimize the total squared error of the output computed by the net. Many of the other networks had trained with different training function which was fast in training and took less number of epochs also, but they exhibits a lot spikes in performance graph i.e. training with trainrp function gives the following performance graph.

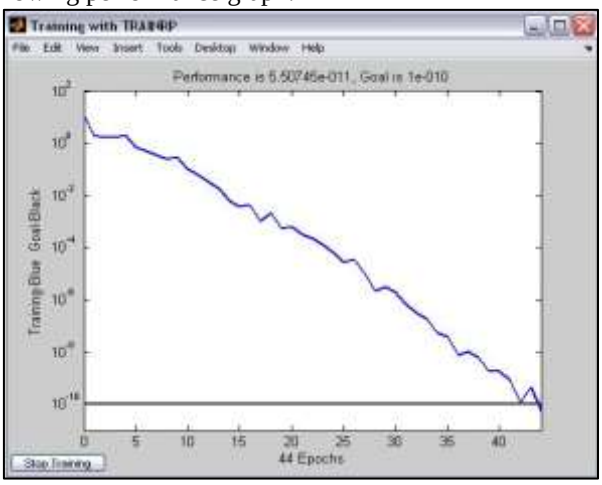

Fig. 3:

In comparison of this graph previous graph have a very smooth line. This means that the bias values and the weights are set in a proper sequential increment manner. The reason is clear from the test input data where, variations and difference between two consecutive data is very slight, so we preferred gradient decent adaptation. We will see the impact of this scenario, directly on the output in term of figures in result area. In addition we gave relatively small learning rate and goal to get better results. Finally we simulate the trained network to test the results.

```
>> clear
>> load Emotion2.mat
>> p = [0.4401
3.6908
-0.0149
0.0039
0.0068
13.7731
3.1798
-27.4979
0.0001
]
p = 0.4401
3.6908
-0.0149
0.0039
0.0068
13.7731
3.1798
-27.4979
0.0001
>> a = sim(net,p)
a = 3.0000
```

The test data within the variable 'p' is refer to 'HAPPY' emotional state of a speaker 'M' and the desired output for this state is '3'. The actual output within the variable 'a' is same as the expected target.

V. RESULTS

1) The database of 100 speech sample had been collected, which contains five emotional states involved in human speech.
2) A total of nine features were extracted form each samples, which contains thousands of rows of data in numeric values.
3) Modeling of neural network and its simulation has successfully done and emotion recognition rate of approximately 70% was obtained against five basic emotional states.
4) We found the impact of two different training functions on emotion recognition rate. Resilient backpropagation (trainrp) given 53% while gradient descent with adaptive learning (traingda) given 70%. Results clearly shows that Gradient Descent algorithm provides better learning environments for this kind of problem.

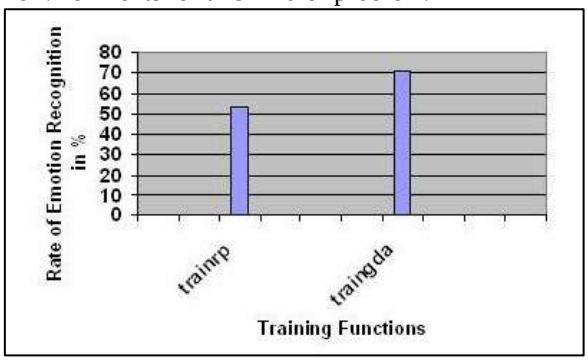

Fig. 4:

All rights reserved by *www.ijsrd.com*

5) Breakdown and equilibrium observed in learning rate, that is training the trained network repetitively can lower the training iterations required but after a breakeven point it may observe highs & lows in next few training cycles and finally settled down to a stable level where putting the ANN on more training cycle does require approximately same number of iterations. This optimum level of training is recommended to achieve.

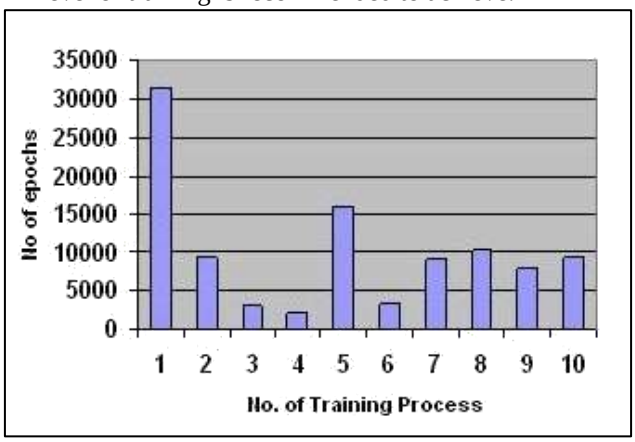

Fig. 5:

The results may vary and differ with other experiments, depends upon the data size, synthesis method and the ANN modeling & simulation technique.

VI. CONCLUSION

With the objective of testing applied methodology towards emotional states recognition through speech synthesis, we carried an experiment with quite different approach. We collected and processed samples in a more practical way to achieve better results even with just 3-4 word sentence in reference to possibility of real time implementation.

A model which should be compact and less complex can be integrated with new age smart devices. Study shows that carefully designed model can learn easily and quickly.

Future aspects depends on the further research, that will be necessary to evaluate the emotion recognition algorithm by collecting a speech database uttered by hundreds or more then this count of speakers and carrying out recognition experiments. In addition to that, the database should collected from all age group, various languages and ethnicity and accents.

Furthermore, it would be evaluated that, which and how much features having the significant values for training procedures or what would be the better combinations of such features.

And finally, it must need to be evaluated that if neuro-fuzzy system may have the better ability to solve the same problem when we consider the huge and complex domains of sample data.

REFERENCES

[1] S. McGilloway, R. Cowie, and E. D. Cowie: Prosodic Signsof Emotion in Speech: Preliminary Results from a New Technique for Automatic Statistical Analysis. ICPhS, Vol. I, p. 250(1995).
[2] T. Shimizu et al: Spontaneous Dialogue Speech Recognition Using Cross-Word Context Constrained Word Graph. Proceedings ofICASSP'96, Vol. 1, pp. 145-148 (1996).
[3] Ryohei Nakatsu, Joy Nicholson and Naoko Tosa: Emotion Recognition and Its Application to Computer Agents with Spontaneous Interactive Capabilities. '98, pp. 228-232 (1998).
[4] Introduction to Neural Networks using Matlab 6.0, by_ S N Sivanandam, S Sumathi, S N Deepa.(Tata McGraw-Hill Pub.)
[5] Artificial Neural Networks: By_ B Yegnanarayana.(PHI Pub.)
[6] Neural Networks: J A Freeman, D M Skapura.(AWL Pub.)
[7] AutoSignal™ v1.6 Manual, ©copyright 1999-2002 AISN Software Incl.
[8] MATLAB® 7.0.1 User Guide, ©copyright Math Works Corporation.

Websites:

[9] Mathworks.com
[10] SYSTATE.com

All rights reserved by *www.ijsrd.com*