International Journal of Engineering and Techniques - Volume 4, Issue 4, July - Aug 2018

RESEARCH ARTICLE OPEN ACCESS

Performance Evaluation of Detection Process of Emotions from Speech Signal According to Various Parameters

Komal Rajvanshi[1], Dr. Ajay Khunteta[2]

[1](PG Scholar, Department of Computer Science, Poornima College of Engineering, Jaipur, Rajasthan)
[2] (Associate Professor, Department of Computer Science, Poornima College of Engineering, Jaipur, Rajasthan)

Abstract:

Within this paper the execution of voice signal detection is described. Several features like timbral, pitch and rhythm are explored according to capability of features for identifying various audio types. Allocation of main characteristics and also the common methods utilized for classification are described. In this paper we have implemented algorithm on some database and detected human emotions from audio signal contained in that database. We have implemented this approach using neural networks and observed the results with the help of different performance and error graphs.

Keywords — **Affective computing, Emotions, Signal acquisition, Feature extraction, Neural network**.

I. INTRODUCTION

Speech recognition procedure is usually performed by the Speech Recognition System. In the speech recognition procedure, speech input signal is executed into identification of speech in the form of text. Speech Recognition System supports the technology to bound humans and computers more strongly. There is a common fact that one should know in order to apply or develop a Speech Recognition System.

Speech generation has usually been represented as a linear convolution between source and filter. Speech signals result from air pressure fluctuations generated by the vocal system. The lungs give the power (air) to the system, and the vocal folds located in the larynx produce the fundamental sound of speech, with fundamental frequency F0. The filtering step is finally executed by the vocal tract, which attenuates or increases certain frequencies through resonance e□ects. The equivalent spectrum of a speech signal can therefore be treated as the product of an excitation spectrum and a vocal tract spectrum.

Before any audio signal can be categorized under a particular class, the features in that audio signal have to be extracted. These features will finalize the class of the signal. Feature extraction inherits the analysis of the input of the audio signal. The feature extraction methods can be classified as temporal analysis and spectral analysis method. Temporal analysis utilizes the waveform of the audio signal itself for analysis. Spectral analysis utilizes spectral demonstration of the audio signal for analysis. Each audio feature is extracted by breaking the input signal into a succession of analysis windows or frames, each of approximate 10-40-ms length, and computing one feature value for each of the windows. One method is to take the values of all features for a particular analysis window to generate the feature vector for the classification decision, so that class assignments can be achieved almost in real time, thus realizing a real-time classifier.

User input speech is known as utterances, in other words when user speaks something it is known as utterances.

Single word exhibits multiple meanings and multiple identifications. It is completely based on pronunciation. A single word is spoken in different means in accordance to country, age etc.

It is the performance calculation tool. It is computed by various means but in this case, if speaker utters “NO”, then Speech Recognition System must detect it as word “NO”. If it is performed precisely then accuracy of speech recognition system is efficiently very good or else.

II. DATABASE

A data base is the gathering of data. In our proposed research work we have utilized voice samples for the database. In that database we identify characteristics of the speech signals and then we stock them into the database. The main question arises that how we are going to collect hundreds of files in the database. The method would be as follows. Initially we would obtain the properties of the speech. All those properties which are needed would be calculated and then it would be stocked into an array. The array would move on as the files move. We would obtain the features and would take the average in the end and then collect them into the database for each class of the voice which we have considered i.e. HAPPY, SAD, AND NEUTRAL.

III. PROPOSED WORK

The mechanism which has been adopted in this research is represented by block diagram shown in Fig. 1.

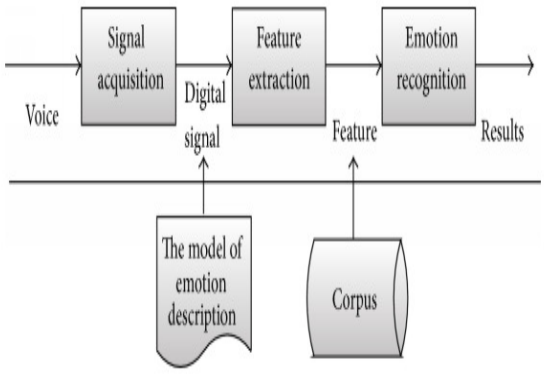

Fig. 1 Block diagram of proposed method

Functioning of each block is defined as follows:

Signal Acquisition: It is the method of sampling signals that compute real world physical situations and transforming the resulting samples into digital numeric signals that can be measured by a computer. It typically transforms analog signals into digital signals for processing. The constituents of signal acquisition systems exhibit:

i. Sensors, to transform physical parameters to electrical signals.
ii. Signal conditioning component, to transform sensor signals into a form that can be transformed to digital values.
iii. Analog-to-digital converters, to transform conditioned sensor waveforms to digital values.

It is utilized in commercial and industrial electronics field and also called as data logger. It generates digital output and provides it to feature extraction block mixing it with model of emotion.

Feature Extraction: Typical emotional feature extraction method focuses on the analysis of the emotional properties in the speech from speech time construction, amplitude construction, and frequency construction.

Some parameters have been measured in this simulation which are as follows:

Epochs: Epochs is the number of repetitions which the neural network performs. For example suppose that there are three repetitions out of 500 which have successfully detected the category of the speech file. If the network system runs 3 iterations, it does not detect that the best result would be obtained at third iteration via the network takes an average of the three repetitions and represents the best results out of time.

Time: It finds out the total time consumed in the detection.

Performance Measure: It shows the elapsed performance obtained in the network architecture which can be plotted via the SOM plots of the neural network.

Validation Checks: It represents the number of validations which the network can implement to the testing method.

IV. EXPERIMENTAL ANALYSIS

In this work we have investigated the performance of neural network model on the features extracted from the audio files.

Basically three categories of audio files are incorporated in this research work, related to sad, happiness and neutral moods of person. Simulation results are represented below:

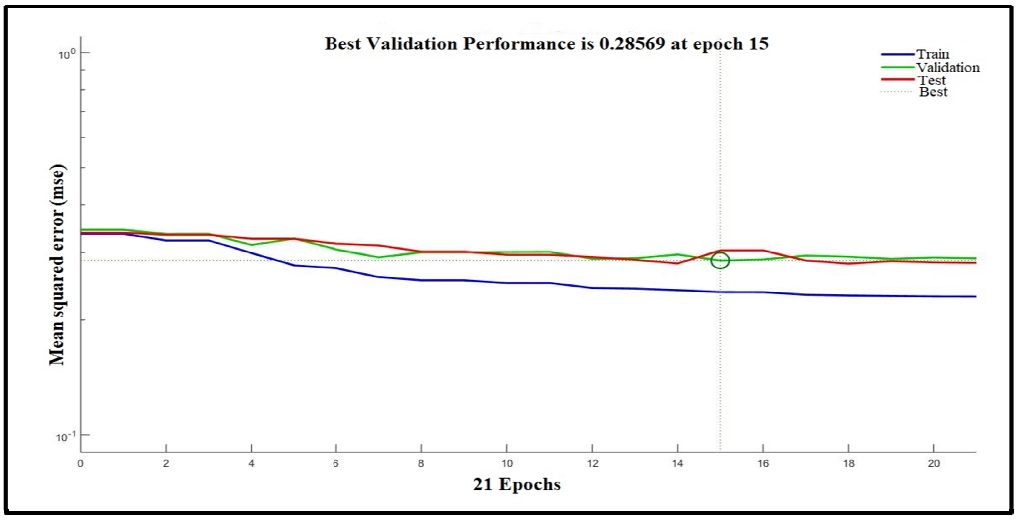

Fig. 2 Neural network performance

In Fig. 2 graph exhibits three lines for three different stages of training, validation, and test. Training procedure on the training vectors continue to start until the network gets to the point that the training decreases the error of network on the validation vectors which would result in neglecting the over-fitting of the data sets. As it is represented in the figure, the best validation performance is occurred at epoch 15, and after 6 error iterations, the process is stopped at epoch 21.

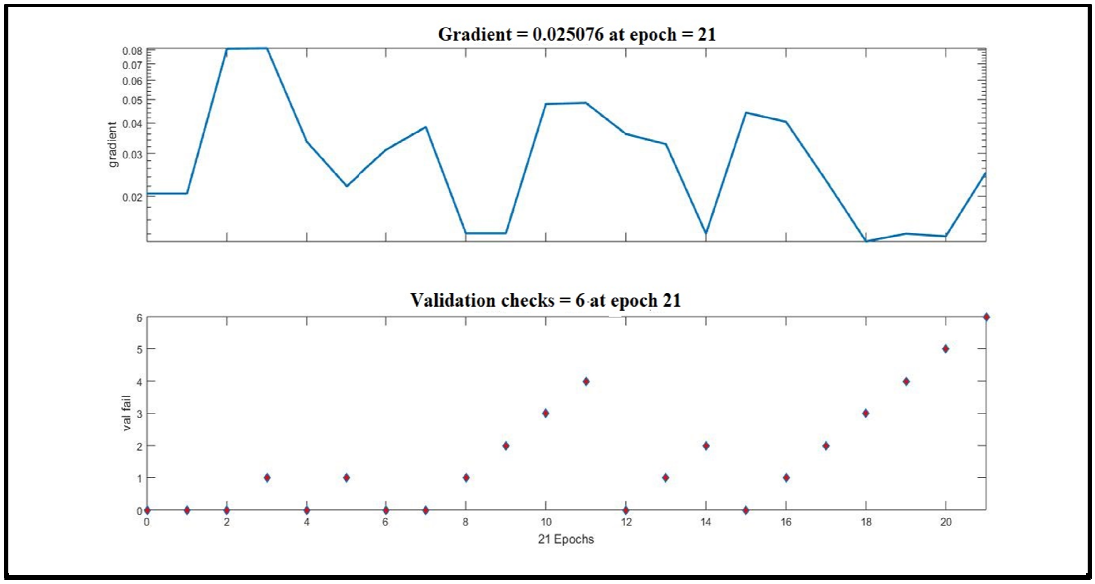

Fig. 3 Neural network training state

From Fig. 3, it is concluded that the errors are repeated 6 times after epoch 15 and the simulation test is stopped at epoch 21. This error iterates beginning at epoch 15 shown over-fitting of the data. Hence, the epoch 14 is considered as the base and its weights are selected as the final weights. Additionally, the validation check is equal to 6, due to the fact that the errors are repeated 6 times before stopping the procedure.

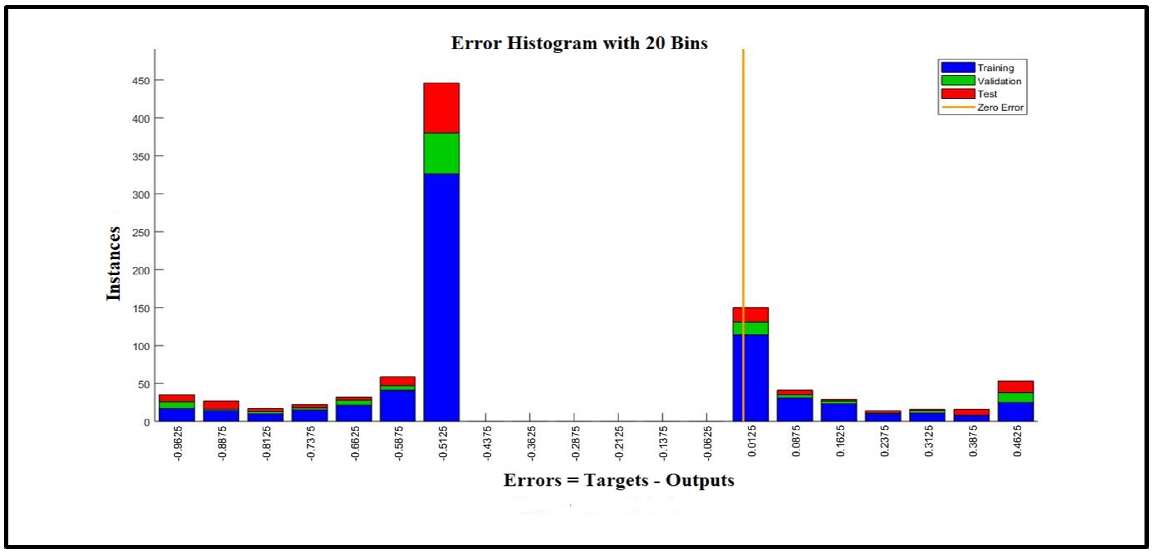

Fig. 4 Error histogram

Fig. 4 represents the error histogram diagram with 20 bins for the three stages of training, validation, and test in artificial neural network system. As it is represented in the Figure, the zero error is demonstrated with a yellow line in the middle part with 13 instances in the training set.

Bins are represented as number of vertical bars demonstrated on the graph. The aggregate error from neural network ranges from -0.9625 (leftmost bin) to 0.4625 (rightmost bin). This range of error is divided into 20 smaller bins, so each bin has a width of

$$\{0.4625 - (-0.9625)\}/20 = 0.07125$$

Each vertical bar demonstrates the number of samples from your database, which exists in a particular bin.

The final confusion matrix depicting the best performance shown by the proposed model at layer size of 75 neurons is depicted in below Fig. 5.

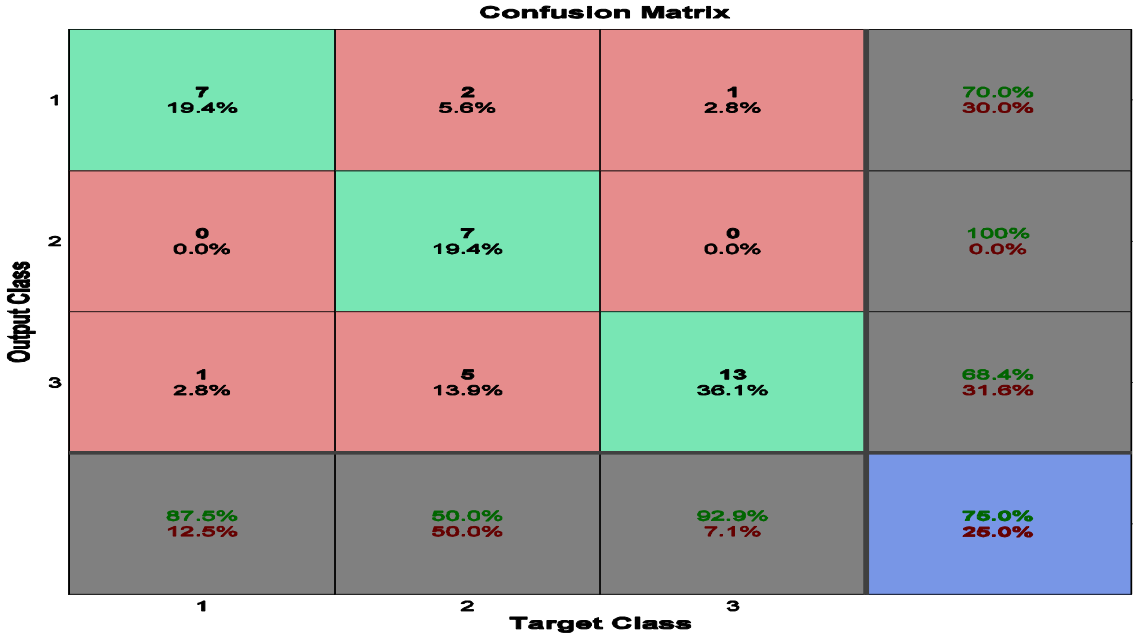

Fig. 5 Confusion matrix

V. CONCLUSION

In fig. 5 the diagonal entries shows correct number of classification and off diagonal entries represents the misclassified entries. The right-most columns and lower most row represents the true positive rate, false positive rate and correctly classified accuracies for three classes of emotions. We can conclude that the maximum achieved accuracy is 75 percents at 75 numbers of neurons layer size.

REFERENCES

1. *V.V. Nanavare and S.K. Jagtap, "Recognition of Human Emotions from Speech Processing", ScienceDirect, 2015.*

2. *Assel Davletcharova, Sherin Sugathan, Bibia Abraham and Alex Pappachen James, "Detection and Analysis of Emotion from Speech Signals", ScienceDirect, 2015.*

3. *Ji Zhengbiao, Zhou Feng and Zhu Ming, "An Algorithm Study for Speech Emotion Recognition Based Speech Feature Analysis", International Journal of Multimedia and Ubiquitous Engineering, 2015.*

4. *S. Lugović, I. Dunđer and M. Horvat, "Techniques and Applications of Emotion Recognition in Speech", MIPRO, 2016.*

5. *Margarita Kotti and Yannis Stylianou, "Effective Emotion Recognition in Movie Audio Tracks", IEEE, 2017.*

6. *Murray, I. R., & Arnott, J. L., "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", Journal of the Acoustical Society of America, 1993.*

7. *Ayadi, E., Moataz, Kamel, M. S., & Karray, F., "Survey on speech emotion recognition features, classification schemes and databases", Pattern Recognition, 2011.*

8. *Ververidis, D., & Koropoulos, C., "Emotional speech recognition: Resources, features, and methods", Speech Communication, 2006.*