ROBUST STRESS SPEECH RECOGNITION USING MFCC AND NEURAL NETWORK

Nitin kumar, Sharmelee Thangjam,Harvinder Kaur
[1]Student,[2]Assistant Professor,[3]Assistant professor
[1]UIET,PU, Chandigarh, India.

Abstract : A number of methods have been developed in the field of human-computer interaction. The aim of this research is to observe and analyze the impact of emotions on the performance of individuals while performing their duties. In this article, the detection accuracy of the designed stress speech recognition system is measured for the students of collage during the exam time. To extract features, Mel Frequency Cepstral Coefficient is used along with grasshopper and Artificial Neural Network (ANN) as an optimization and classification technique respectively. The speech sample of students is collected manually and dataset of 30 students is formed. From the test results examined in MATLAB, the emotions are categorized into three types, high, medium and low. The detection accuracy of about 93.93 % is obtained.

***Index Terms* -** Stress, Speech, MFCC, Grasshopper, ANN.

I. INTRODUCTION

A number of methods have been developed in the field of human-computer interaction. Recently, the popular theme is "emotional intelligence”. In this research, the main goal is to examine and investigate the impact of emotions on the performance of individuals while performing their duties [1]. In military and civilian applications, a speaker is under stress and must be assessed whether the multilingual communications and security systems are becoming more and more important [2]. Emergency call centres and police units all over the world are bombarded with different types of calls, only part of them is of great importance. Then it is of particular interest to improve the effectiveness of decisions and to identify stressful words to save their lives[3]. The feelings of living things basically represent the energy level of an individual. In the modern era, as the competition among students goes on increasing, therefore, the level of stress also being increased [4]. Due to the change in the lifestyle as well as in education, the students face the problem and become a challenge in the competitive environment. Multiple factors have been considered to analyze stress[5]. For the last ten years, researchers working in the field of signal processing and have presented various approaches to emotion detection as well as classification and also provide different methods for stress detection along with classification problems[6]. A few researchers have worked on facial expression (image processing) and a few have been used brain signal by using the concept of signal processing [7]. Sound or speech in the field of communication is a good way to convey emotions. These output signals are related to various parameters and features. Therefore, the signal formulation using digital signal processing (DSP) can be feasible for stress detection and its classification [8].

The working principle of recognition of emotions basically depends on the acoustic difference measured with respect to an appropriate uttering. Emotions and speeches are directly related to each other. The speech is an efficient tool that consists of essential information related to the human mental state along with the physical condition. The amplitude of speech signal comprises of a number of features depending upon the mood. This article contributed to analyze the stress level of the students during the exam time [9].

RELATED WORK

Casale et al. (10, 2007) offered a novel feature vector that enabled a better classification of emotional examined during stressful states. A feature vector component has been derived from genetic algorithms based on a subset selection procedure. A good distinction has been achieved between neutral, furious, loud and Lombard states for the Speech under Simulated & Actual Stress (SUSAS) database databases.

Vignolo et al. (11, 2016) proposed a methodology to study the presentation of data by optimizing a filter bank to improve the results of the highlighted words. An evolutionary algorithm has been developed to choose a filter bank, which contributed to obtained high accuracy. To provide shaping to the filter bank spline function has been used that helps in the reduction of different parameters. The accuracy using Filter bank technique has been increased.

Besbes et al. (12, 2016) have used Support Vector Machine (SVM) with various kernel functions to identify utterances of speech under stress. The authors have worked to classified four different stress categories namely; neutral, Angry, Lombard and Loud.

Deb et al. (13, 2016) have employed Harmonic peak to Energy Ratio (HPER). This technique measure speech features using Fourier Spectra. HPER scheme measures the harmonic level of breathiness speech level, which is different for different stress levels. Also, for classification binary classifier SVM has been used.

Gulhane et al. (14, 2019) presented the MFCC approach along with SVM for the detection and classification of stress among students during exam time. The voice signal of 50 students has been recorded (male & female both). For depressive and aggressive stress, the accuracy of up to 90 % has been observed.

Ahmad et al. (15, 2018) have examined the stress level during the emergency call period by utilizing SVM along with a deep learning approach. The effect along with background noise has also been considered.

Dubey et al. (16, 2018) have utilized Fuzzy based approach and perform an experiment in MATLAB simulator. The input in terms of speech frequency and pitch has been provided and the voice has been classified as the voice of men, women and children.

In the existing work, some of the researchers utilized the fuzzy-based algorithmic approach for recognizing the emotions whereas a few researchers have utilized SVM as a classification algorithm to identify the status of speech. But, the efficiency of the existing speech recognition model is not obtained as per the researchers need. Therefore, to enhance the efficiency of the work, multiclass classifier named as an artificial neural network with optimization technique grasshopper is used. The extracted features are optimized using a grass hopper algorithm that reduces the memory space as well as the processing time by selecting the appropriate features of the speech signal.

PROPOSED WORK

In this section, the process is explained, which is used for the identification and classification of stress level among the students of Punjab University during the exam time. A total of 30 numbers of samples have been recorded using a microphone and save in .wave format. The proposed model used different phases as illustrated in figure 1.

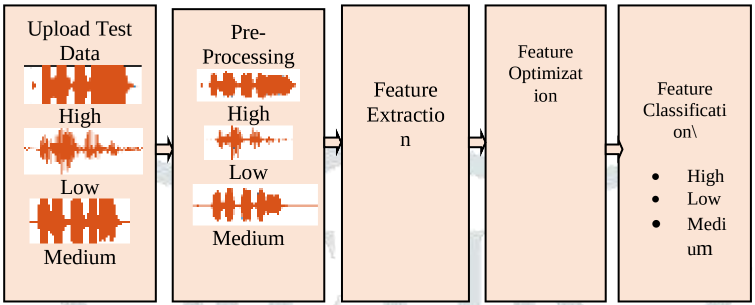

Figure 1 Block Diagram of Propose Work

pre-processing

As the speech signals have been recorded through the microphone in the lab, therefore the chances of noise or hiss produced by electronic components are more. To obtain a high quality signal, pre-processing of data is necessary. Various schemes are considered to decrease the effect of interferences due to the power line.

Pre-processing involves smoothing of signal, DWT and de-noising, in this proposed work. The steps are described below.

i. Smoothing procedure

Generally smoothing is used to reduce noise inside the signal then create a signal having value of pixel is low.

i. De-noising

We use a thresholding method that is applicable to the ECG signal in the de-noising of signal, phase. The process of thresholding is carried out using the "WTHRESH" function that gives soft/hard thresholding of the input signal.

i. DWT (Discrete wavelet transform)

Two kinds of filters such as low-pass filter and high-pass filter were primarily used in the DWT method. It is a wavelet transformation that samples the wavelets. It consists of signal frequency as well the position of the signal with respect to time. DWT mainly comprises of $\boldsymbol{N}$ steps. Initially, DWT generates two coefficient namely approximation coefficient and detail coefficient.

Feature Extraction

MFCC was used to extract attributes from speech signals. MFCC is focused on the recognized human ears differences with crucial bandwidth having frequencies below 1000 Hz. The MFCC algorithm's primary objective is to duplicate the behavior of the ears.

Algorithm: MFCC as a Feature Extraction

```
Input: Speech signal with different stress
Output: MFCCs Feature Set

1. To extract features, speech signals have to load
2. Parameters that are defined - Sample frequency = 16 kHz
                                 - Length of Frame = 25 ms
3. Total number of frames = $\frac{(Frame\ Length\ X\ Sample\ frequency)}{1000}$
4. If Total number of frames is approximately equal to Even number
       Total number of frames is equal to Pad (Zero)
```

```
Else
    Total number of frame’s = Total number of frame’s
Stop
5. Use the specified given equation to apply the Discrete Fourier Transform on all frame
$Signal_i(k) = \sum_{n=1}^{m} \quad signal_i(n)h(n)e^{-\frac{j2\pi kn}{m}} \qquad 1 \le k \le K$
6. Calculate the estimation of the power spectrum periodogram with the equation provided.
$Power_i(k) = \frac{1}{m}|Signal_i(k)|^2$
7. The filter bank Mel-spaced applied
8. Energies of filter bank= Vector’s Energy
9. Vector’s Log Energy = Filter bank energies (log)
10. Coefficients of Cepstral = Log Energy Vector (DCT)
11. Mel Frequency Cepstral Coefficients (MFCCs) = Smaller of 12-13 from 26 coefficients of cepstral
12. Return: MFCCs like a set of feature
13. End
```

Feature Optimization

The features of test speech signals after extraction has been optimized or selected by using a swarm intelligence algorithm, which is developed by Seyedali Mirjalili. This algorithm copies the behavior of grasshopper insect and their social interaction. Grasshoppers are disparaging insects as this damage crop in high quantity. This insect has two lifecycles named as Nymph and adult. In Nymph phase, the insect have not winged and hence move slowly on earth surface to eat vegetables and find its route [17]. In the adult phase, wings appear on the insect and cover larger search space. This algorithm has been used to select only those features of speech.

The process of grasshopper algorithm used in the work is written below.

Algorithm: Grasshopper Optimization

```
Input: MFCC Feature □ MFCC feature of Speech
Output: OMFCC Feature □ Optimized MFCC Feature

Initialize GOA parameters    – Iterations (T)
                               – Number of Population (P)
                               – Lower Bound (LB)
                               – Upper Bound (UB)
                               – Fitness function
                               – Number of Selection (N)
Calculate T = Size (MFCC Feature)
Fitness function: $f(fit) = \{1, If\ feature\ is\ optimial\ 0, otherwise$
For □ T
    $fs = \sum_{i=1}^{P} \quad f(i)$

    $ft = \frac{\sum_{i=1}^{P} \quad f(i)}{Length\ of\ feature}$

    $f(fit)$ = fitness function which define by above given equation

    $OMFCC = GOA(P, Iterations, LB, UB, N, f(fit))$

End

Return; OMFCC as a set of Optimized MFCC Feature
End
```

D. Feature Classification

These optimized features are used as input data to train the system and the trained data of speech is stored into the test database. Therefore, during testing the speech test features are compared with the test database and classified as Low, Medium and High. ANN is used as a classification scheme, which is one of the artificial intelligent approaches used to classify speech signal [18]. The algorithm is as follows.

Algorithm: ANN

```
Input: Student speech data as Training (T), Target as a centroid (Gp) and Neurons (N)
Output: Classified Type of speech (Low, Medium and High)

Initialize ANN with parameters
```

```
 – Number of Iteration (I)
 – Performance metrics:Gradient, Mutation (Mu), MSE, and check Validation
 – Algorithm for Training: Levenberg Marquardt (Trainlm)
 –Distribution of Data: Randomly
On each set of T

Target = Speech types as a Training data

Stop

Initialization of ANN with trained and then group

Net = Newff ($T, Gp, N$)

Set the necessary training variables and train the model

Net is assigned by: Train (Net, Training data , Group)

Return; Classified Type of speech (Low, Medium and High)

End
```

RESULT AND DISCUSSION

The basic terminology that has been taken into consideration is recognition accuracy that measured the performance of the designed stress detection signal. The analysis is conducted in different ways. All installation and practice have been implemented on the Matlab tool. Training and testing process is used to analyze the performance of the system. Different parameters that have been determined are listed in table 1.

Number of Speech Signal	Precision	Recall	F-measure	Accuracy	Error rate
1	0.986	0.977	0.981	87.42	12.57
2	0.986	0.977	0.982	96.74	3.25
3	0.987	0.987	0.985	89.27	8.76
4	0.986	0.974	0.980	98.19	1.8
5	0.982	0.972	0.981	94.89	2.99
6	0.986	0.970	0.978	94.73	5.26
7	0.987	0.985	0.972	96.28	10.58

Table I Computed result

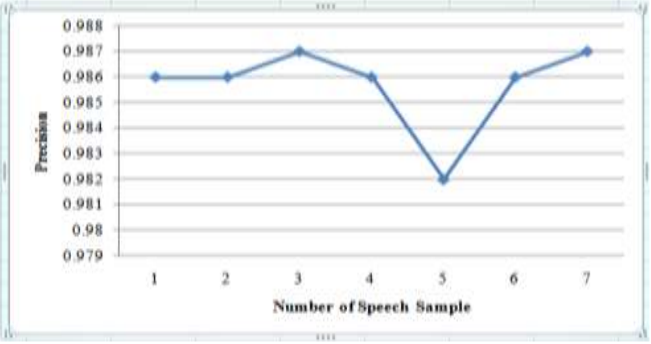

Figure 2: Precision

The precision parameters are used to determine the actual speech samples selected from the true as well as false positive samples. At 3rd iteration, the precision value is high, which represents that the detection of samples with respect to true as well as falsely positive sample is high. The average value of precision determine for the research work is 0.986. The formula is written below.

$$\text{Precision} = \frac{\text{Truely detected } stress\ Speech\ samples}{Total\ Predicted\ positive\ stress\ speech\ sample}$$

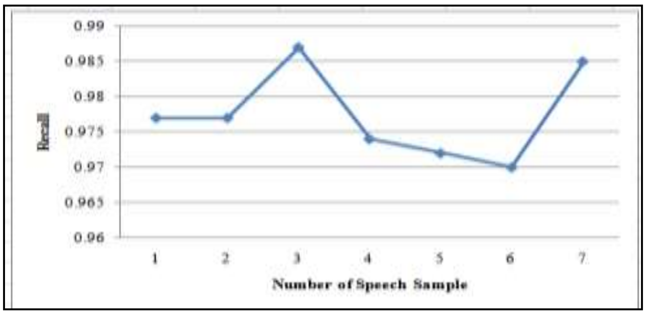

Figure 3 Recall

The recall parameter is determined by using the formula written below.

$$\text{Recall} = \frac{\text{True Positive stress speech sample}}{\text{Total Actual Positive stress speech sample}}$$

The average recall obtained for the proposed work is determined as 0.977.

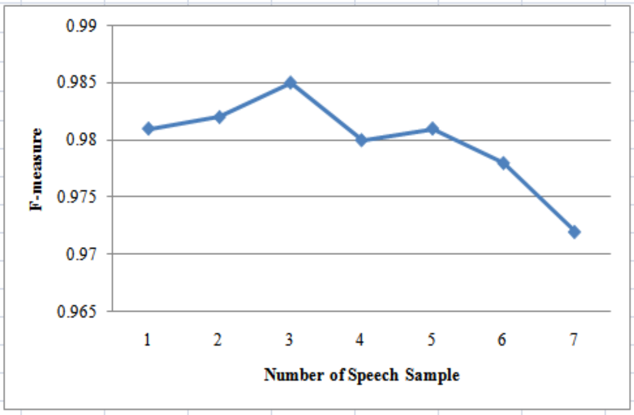

Figure 4: F-measure

To determine the balance between precision and recall the F-measure parameter is used. The average value measured for the proposed work is 0.979.

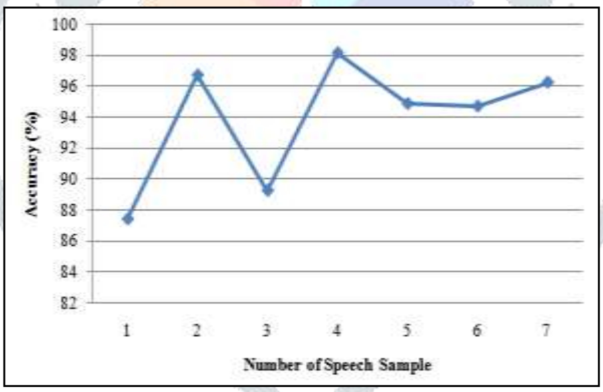

Figure 5: Accuracy

The detection accuracy has been used to classify the three kinds of speech signal named as high, low and medium. The average detection accuracy determined for the proposed work is about 93.93 % and highest at sample 4 (98.19%).

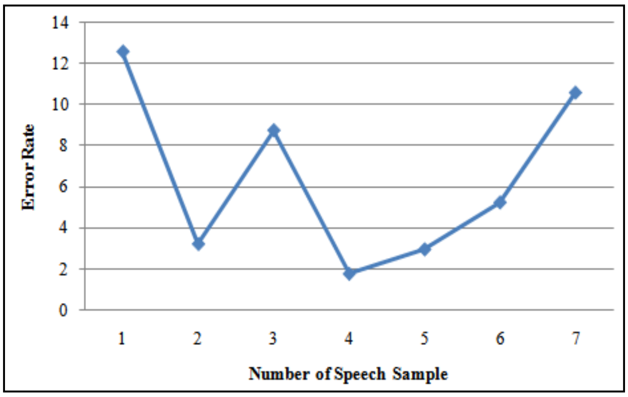

Figure 6: Error Rate

The error rate parameter is used to determine the error occurred during the detection and classification process. The average error rate measured for the proposed work is 6.45.

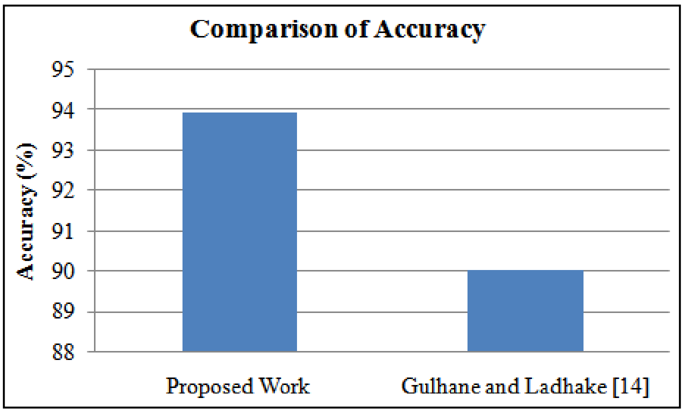

Figure 7: Comparison of Accuracy

The comparison of accuracy with respect to the traditional work performed by Gulhane and Ladhake [14] has been represented in figure 7. The accuracy has been improved, because of the better selection of speech feature using multiclass classification algorithm as compared to SVM as a single class classification algorithm. The enhancement in the work of about 4.37 % has been obtained.

CONCLUSION

In this research, stress in speech signal detection system has been presented. Stress is the reaction that represents the mental state of human. The increase in stress level increase the heart rate, therefore, it is essential for the society to determine a solution to measure the instant stress level. To enhance the detection rate, MFCC along with grasshopper and ANN have been used. Results demonstrated the feasibility of the presented system with minimum error rate of 6.45 for seven different types of stress speech samples. The detection accuracy up to 93.93 % has been obtained. At last, the comparison between existing works has been provided to demonstrate the efficiency of the proposed work. The detection accuracy has been increased by 4.37%.

REFERENCES

1. Hasrul, M. N., Hariharan, M., & Yaacob, S. (2012, February). Human Affective (Emotion) behaviour analysis using speech signals: A review. In *Biomedical Engineering (ICoBE), 2012 International Conference on* (pp. 217-222). IEEE.
2. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, *71*, 10-49.
3. Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. *International journal of speech technology*, *15*(2), 99-117.
4. Ramakrishnan, S. (2012). Recognition of emotion from the speech: A review. In *Speech Enhancement, Modeling and Recognition-Algorithms and Applications*. InTech.
5. Jayanna, H. S., & Prasanna, S. M. (2009). Analysis, feature extraction, modeling and testing techniques for speaker recognition. *IETE Technical Review*, *26*(3), 181-190.
6. El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572-587.
7. Hansen, J. H., & Patil, S. (2007). Speech under stress: Analysis, modeling and recognition. In *Speaker classification I*(pp. 108-137). Springer, Berlin, Heidelberg.
8. Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003, December). Detection of stress and emotion in speech using traditional and FFT based log energy features. In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on* (Vol. 3, pp. 1619-1623). IEEE.

9. Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, *21*(1), 93-120.

10. Casale, S., Russo, A., & Serrano, S. (2007). Multistyle classification of speech under stress using feature subset selection based on genetic algorithms. *Speech Communication*, *49*(10-11), 801-810.

11. Vignolo, L. D., Prasanna, S. M., Dandapat, S., Rufiner, H. L., & Milone, D. H. (2016). Feature optimisation for stress recognition in speech. *Pattern Recognition Letters*, *84*, 1-7.

12. Besbes, S., & Lachiri, Z. (2016, March). Multi-class SVM for stressed speech recognition. In *Advanced Technologies for Signal and Image Processing (ATSIP), 2016 2nd International Conference on* (pp. 782-787). IEEE.

13. Deb, S., & Dandapat, S. (2016). Classification of speech under stress using harmonic peak to energy ratio. *Computers & Electrical Engineering*, *55*, 12-23.

14. Gulhane, Y., & Ladhake, S. A. (2019). Stress Analysis Using Speech Signal. In *International Conference on Innovative Computing and Communications* (pp. 31-40). Springer, Singapore.

15. Ahmad, J., Sajjad, M., Rho, S., Kwon, S. I., Lee, M. Y., & Baik, S. W. (2018). Determining speaker attributes from stress-affected speech in emergency situations with hybrid SVM-DNN architecture. *Multimedia Tools and Applications*, *77*(4), 4883-4907.

16. Dubey, S., Kumar, H. A., Abhilash, R., & Chinnaiah, M. C. (2018). Fuzzy Logic Based Speech Recognition and Gender Classification. In *Microelectronics, Electromagnetics and Telecommunications* (pp. 495-503). Springer, Singapore.

17. Saremi, S., Mirjalili, S., & Lewis, A. (2017). Grasshopper optimisation algorithm: theory and application. *Advances in Engineering Software*, *105*, 30-47.

18. Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis*, *22*(5), 717-727.